Psychological Methods
2009, Vol. 14, No. 3, 275–299

1082-989X/09/$12.00 DOI: 10.1037/a0015825

Estimation of IRT Graded Response Models: Limited Versus Full Information Methods

Carlos G. Forero and Alberto Maydeu-Olivares
University of Barcelona

The performance of parameter estimates and standard errors in estimating F. Samejima's graded response model was examined across 324 conditions. Full information maximum likelihood (FIML) was compared with a 3-stage estimator for categorical item factor analysis (CIFA) when the unweighted least squares method was used in CIFA's third stage. CIFA is much faster in estimating multidimensional models, particularly with correlated dimensions. Overall, CIFA yields slightly more accurate parameter estimates, and FIML yields slightly more accurate standard errors. Yet, across most conditions, differences between methods are negligible. FIML is the best election in small sample sizes (200 observations). CIFA is the best election in larger samples (on computational grounds). Both methods failed in a number of conditions, most of which involved 200 observations, few indicators per dimension, highly skewed items, or low factor loadings. These conditions are to be avoided in applications.

Keywords: limited information estimation, two-parameter logistic model, ordinal factor analysis

Supplemental materials: http://dx.doi.org/10.1037/a0015825.supp

The use of rating scales for measuring psychological constructs is an integral part of behavioral sciences measurement, particularly in assessing personality and attitudinal constructs. Increasingly, applied researchers use more sophisticated techniques for modeling rating scales, such as item response theory (IRT) models, instead of more classical procedures, such as factor analysis. The factor analysis model and IRT models are members of the broader class of latent trait models (Bartholomew & Knott, 1999). The factor analysis model is a linear model originally proposed for continuous data. In contrast, IRT models are nonlinear latent trait models for categorical data. Thus, in principle, IRT models are better suited than factor analysis for modeling the categorical ordered data arising from the application of rating scales (Bartholomew & Knott, 1999; Maydeu-Olivares, 2005b; McDonald, 1999).

Carlos G. Forero and Alberto Maydeu-Olivares, Department of Personality, Evaluation and Psychological Treatment, Faculty of Psychology, University of Barcelona, Barcelona, Spain.

This study was partially supported by Spanish Ministry of Education Grant SEJ2006-08204/PSIC (Albert Maydeu-Olivares, principal investigator) and by a 2006 dissertation support award from the Society of Multivariate Experimental Psychology.

Correspondence concerning this article should be addressed to Carlos G. Forero, Faculty of Psychology, University of Barcelona, Paseo Vall d'Hebrón 171, Barcelona 08035, Spain. E-mail: carlos.garcia.forero@ub.edu

There are many IRT models that can be applied to rating data (for an overview of models, see van der Linden and Hambleton, 1997). Possibly, the most widely used IRT model for rating data is Samejima's (1969) graded response model (GRM). Also, there are several estimation procedures that can be used to estimate IRT models. A thorough description of IRT estimation methods is given in Baker and Kim (2004; see also Bolt, 2005). The current standard estimation method in IRT is full information maximum likelihood (FIML) via the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). It is termed *full information* because all the information contained in the response patterns is used to estimate the model parameters. FIML is asymptotically efficient, in the sense that in infinite samples, no other estimator yields parameter estimates with smaller variances. Yet, FIML estimation may be computationally demanding, and, as a result of the computational requirements involved, issues such as goodness of fit and standard errors have been off the IRT agenda until recent times.

IRT estimation methods such as FIML were developed to model data arising from educational applications (see Lord, 1952; Lord & Novick, 1968). In typical educational applications large sample sizes are often available, tests consist of a large number of indicators, and interest lies in modeling unidimensional constructs. However, in applications of rating scales we often encounter situations that are far from those typically encountered in educational settings. First,

multidimensional constructs are more often of interest. Second, there is significant demand from practitioners for short assessment tools that gather the maximum amount of information in the minimum possible time. These very short questionnaires are, for instance, frequently encountered in behavioral research within medical settings. To complete the picture, we increasingly find applications that focus on very specific populations; as a result, only small samples are available for analysis. How suitable is FIML in these situations? The optimal properties of FIML are asymptotic and need not hold in finite samples. Yet, it is precisely the finite sample behavior of the estimator that is of interest in applications. Also, what are the limits of the "good" behavior of FIML? Is FIML the best option for multidimensional models, small numbers of items, and small sample sizes?

An alternative perspective for the estimation of latent trait models with ordinal indicators arose from within the factor analysis tradition. More specifically, when the observed responses are assumed to arise from a standard factor analysis model whose responses are categorized according to a set of thresholds, a model formally equivalent to a variant of Samejima's (1969) graded response IRT model is obtained (Takane & de Leeuw, 1987). This variant of Samejima's model is also known as the normal ogive model (McDonald, 1997). Within a factor analysis tradition, estimation of this model proceeds differently. For brevity, we will refer to these estimation procedures as categorical item factor analysis (CIFA). In CIFA, parameters are generally estimated in several stages through the use of polychoric correlations. Also, only univariate and bivariate information is used for parameter estimation. Accordingly, CIFA methods estimated in stages have been called *limited information* methods.

Both FIML and CIFA estimation procedures yield parameter estimates with good statistical properties. Their estimates are consistent and are asymptotically normally distributed. However, from a statistical viewpoint FIML estimation is preferable in principle to CIFA estimation, as the former yields parameter estimates with smaller variance. This result is, however, asymptotic and need not hold in finite samples. On the other hand, CIFA estimators have some clear advantages over FIML estimation:

1. They are computationally much faster than FIML.

2. Models with many latent traits and correlated latent traits pose no particular computational difficulty to CIFA, in contrast to FIML.

3. As CIFA belongs to the broad family of structural equation models (SEM), very complex models involving exogenous variables and categorical, continuous, and censored dependent variables can be estimated with ease within a comprehensive measurement model (Muthén, 1983, 1984).[1]

Some simulation studies have addressed the performance of FIML in estimating IRT models in finite samples (e.g., Boulet, 1996; Finger, 2001; Gosz & Walker, 2002; Knol & Berger, 1991; Reise & Yu, 1990; Reiser & VanderBerg, 1994; Stone, 1992; Tate, 2003; Tuerlinckx & De Boeck, 2001). However, due to the computational burden of FIML, only a few of these studies involved more than 100 replications per condition. More important, taken together they covered only a small subset of the situations of interest in applications. For instance, the behavior of FIML parameter estimates in multidimensional models for rating data has never been investigated with at least 100 replications per condition. Additional research is needed on the behavior of FIML standard errors. Due to the computational ease of CIFA, more simulation studies have assessed the performance of CIFA methods (e.g., DiStefano, 2002; Dolan, 1994; Flora & Curran, 2004; Kaplan, 1991; Muthén & Kaplan, 1985, 1992; Oranje, 2003; Parry & McArdle, 1991; Potthast, 1993; Rigdon & Ferguson, 1991), and, as a result, we have a more comprehensive view of the empirical behavior of CIFA estimation methods. Furthermore, even though several studies have pitted FIML against CIFA estimation methods (Boulet, 1996; Finger, 2001; Gosz & Walker, 2002; Knol & Berger, 1991; Reiser & VanderBerg, 1994; Stone, 1992; Tate, 2003), in each case only a few conditions were considered, the number of replications was clearly insufficient, or some aspects (e.g., the comparison of standard errors) were not investigated. Taken together, these studies provide us with a fragmentary view of the empirical performance of FIML versus CIFA.

To fill this gap, we performed an extensive simulation study that compared the empirical performance of these methods in a wide range of settings. We experimentally manipulated different conditions of sample size, number of items, number of response categories, number of latent traits, item discrimination, and item skewness to create a full array of possible conditions that may be found in empirical applications of IRT methods. For each condition, we investigated the performance of FIML and CIFA parameter estimates as well as the performance of their standard errors. In so doing, we aimed to provide guidelines for applied researchers on the boundaries of good performance of IRT estimation methods for rating data and recommendations on the most advisable method of estimation under each research setting investigated.

The remainder of this article is organized as follows. In the next section, we describe Samejima's GRM, the IRT

[1] Recently, some commercial software, such as GLLAMM (Rabe-Hesketh, Pickles, & Skrondal, 2001) or Mplus versions from 4.1 onward (Muthén & Muthén, 2006), has implemented the possibility of estimating IRT models with exogenous variables by FIML.

model employed in our study. Next, we describe how FIML and CIFA methods proceed in estimating this IRT model. We follow by reviewing the results of previous studies that investigated the performance of these estimation methods. Then, we describe the simulation study performed and report its results. The article concludes with a summary report, a set of guidelines for applied researchers, and a discussion of further research topics.

The Normal GRM

IRT models are a family of latent variable models for categorical indicators. Consider modeling the responses to the n items of a questionnaire. Each of the items is to be rated using one of m response alternatives. Thus, for ease of exposition we assume that the number of response alternatives is the same for all items. Items will be labeled as y_i, $i = 1, \ldots, n$. Response categories will be labeled $k = 0, \ldots, m - 1$.

IRT models are intended to provide the probability of each of the m^n possible response patterns that may be observed. In IRT models, this probability depends on (a) the conditional probability of endorsing a response category, given the latent traits, and (b) the distribution of the latent traits (see the Appendix for further details). Often, the latent traits are assumed to be normally distributed, and this assumption will be used in this paper.

The IRT model that is most familiar to applied researchers is likely the two-parameter logistic model (2PL; Birnbaum, 1968). In this model, the conditional probability of endorsing an item is

$$\Pr(y_i = 1|\eta) = \frac{1}{1 + \exp[-a_i(\eta - b_i)]}, \tag{1}$$

where $\Pr(y_i = 0|\eta) = 1 - \Pr(y_i = 1|\eta)$. In this equation, a_i is the discrimination parameter, b_i is the item difficulty, and η denotes the latent trait. In the IRT literature, θ is used instead of η to denote the latent trait. Here, we use η for consistency with the notation used in the SEM literature. This model can be written in terms of the logistic distribution function $\Psi(\bullet)$ as

$$\Pr(y_i = 1|\eta) = \Psi(-a_i(\eta - b_i)). \tag{2}$$

Although the Lord and Novick (1968) parameterization used in Equation 1 is most popular in IRT applications, it cannot be extended to models that depend on more than one latent trait. For multidimensional models (i.e., models with $p > 1$ latent traits), the parameterization that is most widely used is

$$\Pr(y_i = 1|\eta) = \Psi(\alpha_i + \beta_i\eta), \tag{3}$$

where α_i is the intercept parameter and β_i is the slope parameter. This parameterization readily extends to multidimensional models, such as

$$\Pr(y_i = 1|\boldsymbol{\eta}) = \Psi(\alpha_i + \boldsymbol{\beta}_i'\boldsymbol{\eta}), \tag{4}$$

where $\boldsymbol{\beta}_i$ and $\boldsymbol{\eta}$ are now $p \times 1$ vectors. The relationship between Lord and Novick's parameterization and the intercept/slope parameterization used in Equation 3 is given by

$$\alpha_i = a_i b_i \quad \beta_i = a_i. \tag{5}$$

The 2PL is suitable for modeling rating items with two response alternatives. Arguably, the most widespread model for rating data is Samejima's (1969) GRM.[2] The GRM is obtained using

$$\Pr(y_i = k_i|\boldsymbol{\eta}) = \begin{cases} 1 - \Psi(\alpha_{i,k} + \boldsymbol{\beta}_i'\boldsymbol{\eta}) & \text{if } k_i = 0 \\ \Psi(\alpha_{i,k} + \boldsymbol{\beta}_i'\boldsymbol{\eta}) - \Psi(\alpha_{i,k+1} + \boldsymbol{\beta}_i'\boldsymbol{\eta}) & \text{if } 0 < k_i < m - 1. \\ \Psi(\alpha_{i,k+1} + \boldsymbol{\beta}_i'\boldsymbol{\eta}) & \text{if } k_i = m - 1 \end{cases} \tag{6}$$

When the number of response alternatives is two, this model reduces to the 2PL.

A normal distribution function $\Phi(\bullet)$ may be used instead of the logistic distribution function $\Psi(\bullet)$ in Samejima's model. When a logistic function is used, the model is called logistic GRM, and when a normal function is used, it is called normal ogive GRM. In the special case where the items are dichotomous, it is referred to as the normal ogive model instead of the 2PL.

CIFA Formulation of the Normal Ogive GRM

The normal ogive GRM can be alternatively derived from a factor analytic framework. Within this framework, it is assumed that a latent response variable y_i^* underlies each observed categorical response variable y_i. The latent responses y_i^* are related to the latent traits $\boldsymbol{\eta}$ via a standard factor analytic model,

$$y_i^* = \boldsymbol{\beta}_i'\boldsymbol{\eta} + \varepsilon_i, \tag{7}$$

where $\boldsymbol{\beta}_i'$ is a $1 \times p$ vector of factor loadings and ε_i is a measurement error. The latent traits and measurement errors are assumed to be normally distributed, so that the latent response variables are normally distributed.

In turn, the latent response variables are related to the observed categorical responses via a threshold relation,

[2] There is also some evidence suggesting that it may be the best fitting parametric IRT model for rating data (Maydeu-Olivares, 2005a).

$$y_i = k \text{ if } \alpha_{i,k} < y_i^* < \alpha_{i,k+1}, \quad (8)$$

where $\alpha_{i0} = -\infty$ and $\alpha_{i,m-1} = +\infty$. That is, under this model, a respondent chooses a response alternative based on her location on the response variable y_i^* relative to a set of $m - 1$ item threshold parameters, $\alpha_{i,k}$. Response alternative k will be endorsed when the respondent's latent response value y_i^* lies between thresholds $\alpha_{i,k}$ and $\alpha_{i,k+1}$.

Thus, the ordinal factor analysis model defined by Equations 7 and 8 is simply a standard factor analysis model to which a threshold process has been added to take into account the ordinal nature of the observed data. However, unlike in the factor analysis model for continuous responses, in the ordinal factor analysis model the variances of the measurement errors $\boldsymbol{\varepsilon}$ are not identified. These variances can be identified using one of two types of constraints and will result in two equivalent parameterizations of the model. See the Appendix for further details.

One way to identify the model is by setting the variances of $\boldsymbol{\varepsilon}$ to 1 or to some other constant. If they are fixed to 1, the parameters being estimated are the $\alpha_{i,k}$ and $\boldsymbol{\beta}_i$ parameters of the normal ogive version of the GRM given in Equation 6. Another way is to constrain the variances of $\boldsymbol{\varepsilon}$ to be equal to $1 - \boldsymbol{\beta}_i' \boldsymbol{\Psi} \boldsymbol{\beta}_i$, where $\boldsymbol{\Psi}$ denotes the $p \times p$ matrix of correlations among the latent traits. In this case, the latent responses y_i^* are standardized (their variance is 1) and the parameters being estimated are the standardized thresholds $\tau_{i,k}$ and the standardized factor loadings $\boldsymbol{\lambda}_i$. The relationship between the standardized and unstandardized parameters is

$$\tau_{i,k} = \frac{-\alpha_{i,k}}{\sqrt{1 + \boldsymbol{\beta}_i' \boldsymbol{\Psi} \boldsymbol{\beta}_i}}, \quad \boldsymbol{\lambda}_i = \frac{\boldsymbol{\beta}_i}{\sqrt{1 + \boldsymbol{\beta}_i' \boldsymbol{\Psi} \boldsymbol{\beta}_i}}. \quad (9)$$

Notice that the $\alpha_{i,k}$ and $\boldsymbol{\beta}_i$ parameters receive two names. In IRT terminology they are referred to as intercepts and slopes. In SEM terminology they are unstandardized thresholds and unstandardized factor loadings. We use both terms interchangeably here, and we use the same notation in Equations 7 and 8 to emphasize that they are the same parameters (when the normal ogive version of the GRM model is used).

In summary, Samejima's graded model has two variants that depend on which link function (logistic and normal ogive) is used to relate the parameters to the conditional probability of observing a response category. In the special case of dichotomous items, the logistic form of the model reduces to the 2PL, whereas the normal ogive form reduces to the normal ogive model. Regardless of the number of response alternatives and link function, three parameterizations can be used. The first one, a_i and b_i, is the most widely used for unidimensional IRT models. However, it cannot be used for multidimensional models. For this reason, it will not be considered here. The second parameterization, $\boldsymbol{\tau}_{i,k}$ and $\boldsymbol{\lambda}_i$, is the one used in SEM programs for CIFA. We will refer to this choice as standardized parameterization, because it results in factor loadings that are bounded between -1 and 1. The third parameterization, $\alpha_{i,k}$ and $\boldsymbol{\beta}_i$, will be referred to as unstandardized parameterization. In this case, the $\boldsymbol{\beta}_i$ parameters are unbounded. Note that in the IRT literature the $\{\alpha_{i,k}, \boldsymbol{\beta}_i\}$ parameterization is often called intercept/slope parameterization and the $\{\tau_{i,k}, \boldsymbol{\lambda}_i\}$ parameterization is called threshold/factor loading parameterization.

IRT Estimation Methods

FIML via the EM Algorithm

Possibly, the most widespread full information estimation method in IRT modeling is marginal maximum likelihood via the EM algorithm (Bock & Aitkin, 1981; Bock et al., 1988).[3] In this method, the probability of each observed pattern of responses is estimated at each iteration of the estimation process, and each of these pattern probabilities involves an integral in p dimensions (the number of latent traits). The integrals do not yield close-form expressions, and, as a result, they must be approximated numerically, usually by means of Gauss–Hermite quadrature (Davis & Rabinovitz, 1975, ch. 2). In this procedure, Q points are defined along each dimension, which produces a grid with Q^p points in the p-dimensional latent trait space. Then the integral is approximated by a weighted sum of the function evaluated at each point of this grid. As a result, computational requirements increase exponentially with the number of latent trait dimensions.

Maximum likelihood estimates of model parameters are obtained iteratively by means of the EM algorithm (Dempster, Laird, & Rubin, 1977) in two steps. During the E step, a provisional set of item parameters is regarded as the true item parameters, and the proportion of examinees choosing a certain category is estimated given these parameters and the response patterns. In the M step, the proportion of responses obtained in the E step is regarded as the true probability, and item parameters are estimated. These resulting item parameters are regarded as true parameters in the next E step. This process is repeated until a certain convergence criterion is reached.

FIML requires heavy computations, particularly in multidimensional models and especially if the latent traits are correlated. Also, the behavior of FIML depends on how the numerical integration is performed to obtain the probability of the response patterns. The more quadrature points per dimension the better, but the exponential rate of growth of the total number of points makes estimation unfeasible if there are more than a few latent traits. To reduce the amount

[3] What we refer to as FIML is often denoted as marginal maximum likelihood (MML) in the IRT literature. Some authors may refer to it as ML.

of computation in multidimensional IRT models, one uses only a few points per dimension. However, this strategy has the drawback of producing inaccurate parameter estimates, especially in long questionnaires (Meng & Schilling, 1996; Schilling & Bock, 2005). Estimation of the standard errors, which requires the inversion of the matrix of second derivatives at the end of the last M step, adds additional computational burden.

CIFA Estimation

CIFA procedures are specifically designed for the normal ogive form of the GRM. They use only low-order associations among the observed variables, and they are performed in several stages for improved computational efficiency. The most widely used estimation procedures in CIFA consist of three stages. In the first stage, the thresholds $\boldsymbol{\tau}$ are estimated for each variable separately using maximum likelihood. Thus, only univariate information is used in the first stage. In the second stage, polychoric correlations are estimated. A polychoric correlation is the correlation between two latent response variables y_i^*. Each polychoric correlation is estimated separately, with the thresholds estimated in the first stage and by maximum likelihood. Only bivariate information is used in the second stage. In the third stage, all estimated thresholds and polychoric correlations are gathered into a vector $\hat{\boldsymbol{\kappa}}$ and the model parameters are estimated by minimizing the function[4]

$$F = (\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\theta}))' \hat{\mathbf{W}} (\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\theta})). \quad (10)$$

In Equation 10, $\boldsymbol{\kappa}(\boldsymbol{\theta})$ denotes the restrictions imposed by the model on the thresholds and polychoric correlations and $\boldsymbol{\theta}$ denotes the model parameters. These can be the unstandardized parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ or the standardized parameters $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$, depending on which parameterization is used. If the latent traits are correlated, $\boldsymbol{\theta}$ also includes the parameters of their correlation matrix $\boldsymbol{\Psi}$.

Different weight matrices $\hat{\mathbf{W}}$ can be used in Equation 10. Let $\boldsymbol{\Gamma}$ be the asymptotic covariance matrix of the thresholds and polychoric correlations estimated in the first two stages. Some popular choices of $\hat{\mathbf{W}}$ are (a) $\hat{\mathbf{W}} = \hat{\boldsymbol{\Gamma}}^{-1}$ (weighted least squares [WLS]; Muthén, 1978, 1984); (b) $\hat{\mathbf{W}} = (diag(\hat{\boldsymbol{\Gamma}}))^{-1/2}$ (diagonally weighted least squares [DWLS]; Muthén, du Toit, & Spisic, 1997); and (c) $\hat{\mathbf{W}} = \mathbf{I}$ (unweighted least squares [ULS]; Muthén, 1993).

Consistent and asymptotically normal parameter estimates as well as standard errors can be obtained for all three estimation methods. Asymptotically, the WLS parameter estimates have smallest variance among the class of estimators obtained by minimizing Equation 10. Thus, WLS is asymptotically efficient within this class of estimators. However, in practice, the behavior of WLS is very poor unless the sample size to model size ratio is very large. DWLS and ULS yield much better parameter estimates in small samples (Dolan, 1994; Flora & Curran, 2004; Muthén, 1993), and some evidence suggests that there is little difference between DWLS and ULS when modeling categorical data (Maydeu-Olivares, 2001).

In general, CIFA estimation of Samejima's GRM is computationally much more efficient than FIML estimation. Integration is performed for each item separately (to estimate the thresholds) or for pairs of items separately (to estimate the polychoric correlations). As a result, regardless of the number of latent traits, only univariate and bivariate integrals are involved. However, the computational efficiency of CIFA estimation methods is achieved at the expense of disregarding three-way and higher order associations among the items.[5] As the studies described in the next section show, it is not yet clear if, through disregard of three-way and higher order information, CIFA estimation methods result, in finite samples, in worse parameter estimates and standard errors than do FIML methods.

Previous Research on the Empirical Behavior of FIML and CIFA Estimators

Table 1 lists all simulation studies that have investigated the performance of either FIML or CIFA methods in estimating IRT models for rating data.[6] We have also included the major studies that have compared the behavior of FIML and CIFA methods in applications. Finally, we have included the major review articles (e.g., McDonald & Mok, 1995; Mislevy, 1986).

As it can be seen in Table 1, a number of studies have compared the behavior of FIML with that of one or more CIFA methods in applications (e.g., Bolt, 2005; Janssen

[4] When the standardized parameters are estimated, and if no restrictions are imposed among the thresholds $\boldsymbol{\tau}$, the third stage estimation can be performed by minimizing a function of the polychoric correlations alone.

[5] Three-way and four-way associations among the items are used to estimate the asymptotic covariance matrix $\boldsymbol{\Gamma}$ of the estimated thresholds and polychoric correlations. The estimated matrix $\boldsymbol{\Gamma}$ is used for parameter estimation in CIFA–WLS. In CIFA–ULS estimation, this matrix is used solely to compute the standard errors, not to estimate parameters.

[6] There is some confusion in the literature about labeling of CIFA estimators. For instance, the WLSM and WLSMV CIFA estimators available in Mplus (Muthén & Muthén, 2004) are in fact the same estimator, a DWLS CIFA estimator. WLSM and WLSMV differ solely in the choice of goodness-of-fit test. The CIFA procedures (WLS, DWLS, and ULS) implemented in Mplus are equivalent to the CIFA procedures implemented in Lisrel except for the formula used to compute the asymptotic covariance matrix of the sample thresholds and polychoric correlations. The formulas used in Lisrel and Mplus to compute this matrix are asymptotically equivalent but may yield very slight differences in finite samples.

Table 1
Major Studies on Factors Affecting Estimation Method Performance of Latent Trait Models for Categorical Variables (in Descending Chronological Order)

Paper	Type of study[a]	No. replications	Estimator				
			FIML	CIFA–WLS	CIFA–DWLS	CIFA–ULS	CIFA–NOHARM[a]
Muthén & Kaplan (1985)	S	1,000		√			
Mislevy (1986)	R		√	√		√	
Reise & Yu (1990)	S	1	√				
Baker (1991)	S	100	√				
Parry & McArdle (1991)	S	1		√		√	√
Knol & Berger (1991)	S	10	√				√
Kaplan (1991)	S	100		√			
Rigdon & Ferguson (1991)	S	300		√	√	√	
Stone (1992)	S	100	√				
Muthén (1993)	S	500				√	
Potthast (1993)	S	100		√			
Reiser & VanderBerg (1994)	S	500	√	√			
Dolan (1994)	S	100		√			
McDonald & Mok (1995)	R	—	√	√			√
Boulet (1996)	S	100	√				√
Janssen & De Boeck (1999)	E	—	√				√
Schumaker & Beyerlein (2000)	E	—					√
Tuerlinckx & De Boeck (2001)	S	50	√				
Finger (2001)	S	5	√			√	√
Gosz & Walker (2002)	S	100	√				√
DiStefano (2002)	S	100		√			
Oranje (2003)	S	1,000		√	√		
Tate (2003)	E/S	1	√	√	√		√
Flora & Curran (2004)	S	500		√	√		
Bolt (2005)	R	—	√	√		√	√
Beauducel & Herzberg (2006)	S	500			√		
Wirth & Edwards (2007)	R	—	√	√	√		
This study	S	1,000	√			√	

Note. FIML = full information maximum likelihood; CIFA = categorical item factor analysis; WLS = weighted least squares; DWLS = diagonally weighted least squares; ULS = unweighted least squares; NOHARM = program that estimates the normal ogive form of Samejima's model for dichotomous variables in two stages with ULS; S = simulation study assessing the performance of at least one method via Monte Carlo simulation; R = review or analytical study about the performance of at least one method; E = empirical study comparing two estimation methods with real data; U = unidimensional; M = multidimensional; P = polytomous; D = dichotomous.
[a] CIFA–NOHARM differs from CIFA–ULS in that estimation proceeds in two stages in CIFA–NOHARM and in three stages in CIFA–ULS.

& De Boeck, 1999; Oranje, 2003; Schumaker & Beyerlein, 2000). In general, these studies have found small differences among the estimators compared. However, simulation studies in which the true model is known are needed for verification of the theoretically superior performance of FIML over CIFA methods. Only Reiser and VandenBerg (1994), Boulet (1996), and Gosz and Walker (2002) have compared these methods using at least 100 replications per condition. In addition, Reiser and VanderBerg used WLS, the CIFA method that yields poorer results in finite samples.[7] Boulet and Gosz and Walker performed CIFA using NOHARM (Fraser & McDonald, 1988), a program that estimates the normal ogive form of Samejima's model for dichotomous variables in two stages with ULS.[8]

Reiser and VanderBerg (1994) compared the performance of FIML and CIFA–WLS in models with 4 to 10 variables and a single latent trait. Sample size was 500 observations in all conditions, and parameter values were the same for all condi-

[7] Reiser and VanderBerg (1994) estimated the model in a single stage, as did Christoffersson (1975), rather than in three stages.

[8] First, each threshold is estimated separately, as in other CIFA procedures. In a second stage, the remaining parameters of the model are estimated with ULS using bivariate information and holding the first-stage estimates fixed. CIFA–NOHARM is closely related to CIFA–ULS, in which estimation proceeds in three stages, with tetrachoric/polychoric correlations being computed in an intermediate stage. In practice, NOHARM and CIFA–ULS yield very similar results (Maydeu-Olivares, 2001).

Table 1 (*continued*)

Study variables						Assessed outcomes		
No. dimensions	No. categories	Sample size	Model size	Item skewness	Item slope	Parameter estimates	Standard errors	Convergence
U	P	√		√		√	√	
U/M	D	√	√			√	√	
U	P	√			√	√		
U	D	√	√			√		
U	D	√		√	√	√		√
U/M	D	√	√			√		
U/M	P	√		√		√	√	
M	P	√		√	√	√	√	
U	D	√	√			√		
U	D/P	√	√				√	
M	P	√		√		√	√	
U	D		√			√	√	
U	D/P	√		√		√	√	
U/M	—	√				√		
U	D	√	√			√		
M	D					√		
U	D					√		
U	D		√		√	√		
U/M	D	√	√			√		
M	D					√		
M	P	√		√	√	√	√	
U/M	D/P	√	√			√		√
U/M	D			√		√		
U/M	D/P	√	√	√		√	√	√
M	D					√	√	
U/D	D/P	√	√			√	√	
U/D	D/P	√	√	√	√	√		
U/M	D/P	√	√	√	√	√	√	√

tions. Reiser and VanderBerg concluded that the use of high-order marginals gives FIML slight advantages in holding down parameter bias of the estimator in finite samples. Boulet (1996) compared ULS and FIML using a unidimensional 2PL with 15 to 60 indicators, 250 to 1,000 observations, and latent traits differing in skewness. Boulet concluded that ULS and FIML showed similar trends in terms of relative bias and that ULS recovered item parameters more accurately when the latent trait was normal. Gosz and Walker (2002) compared FIML and CIFA in fitting data generated from a 2PL. Using 2,500 observations and 40 items, they concluded that parameter estimates were more accurate when ULS was used.

Simulation results comparing the performance of different CIFA estimators have clearly revealed that WLS is the worst estimator in small samples (e.g., Muthén, 1993), in terms of both parameter estimates and standard errors. The weight matrix in WLS estimation depends on four-way sample moments, which are very unstable unless sample size is very large relative to model size (Muthén & Kaplan, 1992). In contrast, DWLS and ULS show a much better performance in small samples. Thus, Flora and Curran (2004) found that DWLS yields good results with sample sizes of 100 observations. Research on ULS performance suggests good parameter recovery, with small amounts of negative bias decreasing with increasing sample sizes (Finger, 2001; Gosz & Walker, 2002; Knol & Berger, 1991; Parry & McArdle, 1991; Reise & Yu, 1990; Tate, 2003).

For its part, FIML has exhibited good parameter recovery with small-to-moderate bias that diminishes as sample size and number of indicators per factor increase (Baker, 1987;

Oranje, 2003; Reise & Yu, 1990; Tate, 2003). Parameter bias has been found to be worse when items are extreme in either factor loading or skewness (Drasgow, 1989). Standard error estimation of loadings has been found to be problematic with sample sizes of about 200 observations (Drasgow, 1989; Reiser & VanderBerg, 1994).

In closing, existing literature suggests that FIML and CIFA methods perform similarly in terms of parameter estimation. WLS yields good parameter estimates provided that sample size is large enough to meet its asymptotical properties, but DWLS and ULS have shown better performance in small samples. Concerning standard errors, the literature suggests that FIML and WLS yield poorer results, in particular when indicators have extreme item parameters. However, all in all, the previous studies provide us with a fragmentary view of the empirical performance of FIML and CIFA methods in estimating Samejima's GRM. Various studies have investigated the effect of different factors that may influence the performance of the estimators, but no study has investigated the effect of all possible factors simultaneously. This approach is needed and could reveal the existence of possible interactions among the factors. In addition, there are a number of aspects that has never been investigated, such as the behavior of FIML standard errors for multidimensional models.

A Monte Carlo Investigation of the Performance of FIML and CIFA–ULS Parameter Estimates and Standard Errors for the Graded Model

We performed a simulation study to compare the performance of FIML and CIFA–ULS in estimating Samejima's GRM under varied conditions of dimensionality, factor loading, sample size, number of items per factor, number of response alternatives per item, and item skewness. All simulations were performed with Mplus Version 3.13 (Muthén & Muthén, 2004).[9] Default convergence criteria were used for both methods. For FIML, Gauss–Hermite integration with 64 points was used for unidimensional models and Gauss–Hermite integration with 8 points per dimension (for a total of 512 points) was used for the three-dimensional models. We verified the results for both FIML and ULS against our own code in several conditions and obtained comparable results. ULS was the CIFA method of choice, because previous research has revealed that it performs much better than WLS in small samples.

We investigated 324 conditions per estimation method and obtained 1,000 replications for each condition. The 324 conditions were obtained using a factorial design by crossing.

1. Three sample sizes (200, 500, and 2,000 respondents).
2. Two levels of latent trait dimensionality (one and three latent traits).
3. Three test lengths (9, 21, and 42 items).
4. Three levels of factor loadings λ (or, alternatively, β parameters): low ($\lambda = .4$, $\beta = 0.74$),[10] medium ($\lambda = .60$, $\beta = 1.27$), and high ($\lambda = .8$, $\beta = 2.26$). Factor loadings were set equal across items in generating the data (not in estimating the model) to facilitate the reporting of the findings.
5. Six item types (three types consist of items with two categories, and three types consist of items with five categories) that varied in skewness and/or kurtosis.

The sample sizes were chosen to be small to large in typical applications. Also, small-to-medium test lengths were chosen because prior results have suggested that the performance of parameter estimates and standard errors improves with increasing test length. Finally, we include items with typical low (.4) to large (.8) factor loadings.

The item types used in the study are depicted in Figure 1. These item types were chosen to be typical of a variety of applications. Items of Types I to III consist of only two categories. Type III items have the highest item skewness and kurtosis. The threshold was chosen such that only 10% of respondents endorse the items. Type II items are endorsed by 15% of respondents, resulting in smaller values of skewness and kurtosis. Items of Types II and III are typical of applications in which items are seldom endorsed. On the other hand, Type I items are endorsed by 40% of respondents. These items have low skewness, and their kurtosis is smaller than that of a standard normal distribution.[11] Items of Types IV through VI consist of five categories. The skewness and kurtosis of Type IV items closely match those of a standard normal distribution. Type IV items are also symmetric (skewness = 0); however, the kurtosis is higher than that of a standard normal distribution. These items can be found in applications in which the middle category reflects an

[9] Similar results would be expected when using any other software program for SEM that implements CIFA estimation.

[10] Throughout the article, α and β parameter values are given in the logistic scale implied by Equation 5.

[11] The skewness and kurtosis of a standard normal distribution are 0 and 3, respectively. We subtracted 3 from the kurtosis values, so that 0 indicated no excess kurtosis, a positive value indicated excess kurtosis greater than that of a normal distribution, and a negative value indicated excess kurtosis less than that of a normal distribution.

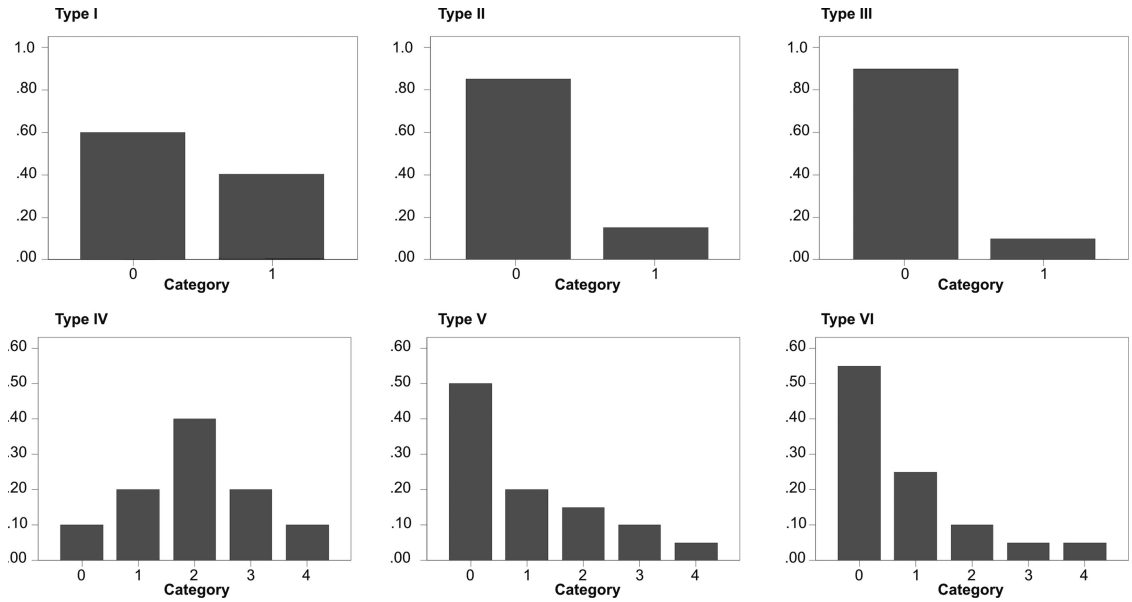

Figure 1. Bar graphs of the different types of items employed in the simulation study.

undecided position and a large number of respondents choose this middle category. Finally, Type V and Type VI items show a substantial amount of skewness and kurtosis. For these items, the probability of endorsing each category decreases as the category label increases.

In the case of three-dimensional models, the factors were set up to be orthogonal, with one third of the items serving as indicators for each dimension. This setup resulted in six conditions of number of indicators per factor (3, 7, 9, 14, 21, and 42 items).

Choice of Link Function and Parameterization

A comparison of FIML and CIFA is difficult, because unstandardized parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) and the logistic GRM are most often used in FIML, whereas standardized parameters ($\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$) are used and only the normal GRM can be estimated with the sequential CIFA procedures. Regarding the choice of model, we chose to use the logistic GRM for FIML to increase face validity. Thus, for FIML, data were generated and estimated with the logistic GRM. For CIFA, data were generated and estimated with the normal GRM. In both cases, we used the true parameter population values as starting values to maximize the probability of convergence. Regarding the choice of parameterization, each estimation method has a "natural" parameterization (i.e., unstandardized parameters for FIML and standardized parameters for CIFA). For both estimation methods, use of the alternative parameterization leads to complex model constraints that may hinder the convergence of the estimation process and, as a result, may affect the accuracy of parameter estimates and standard errors.[12] To thoroughly investigate the effect of the choice of parameterization, we performed CIFA estimation twice, in one case minimizing with respect to the unstandardized parameters and in the second case minimizing with respect to the standardized parameters.[13]

Parameter estimates and standard errors obtained using the normal and logistic variants of the GRM model can be put in the same metric by using the scaling constant $D = 1.702$. Haley (1952) showed that the use of this constant puts the estimates on the same scale (within .01 units), and the constant D has been used in many published descriptions of the 2PL to put the logistic parameters of this model in the scale of the normal ogive parameters. That is, often

$$\Pr(y_i = 1|\eta) = \frac{1}{1 + \exp[-Da_i(\eta - b_i)]}$$

[12] If the unstandardized parameters are used in CIFA, the polychoric correlations implied by the model include products of inverses of square roots of functions of the model parameters (see the Appendix). Alternatively, if the standardized parameters are used in FIML, the conditional probabilities in Equation 6 include inverses of square roots of functions of the model parameters.

[13] Selection of parameterization (standardized or unstandardized) is performed in Mplus by choosing "delta" or "theta" parameterization, respectively. We used the same starting seed in the simulations to ensure that the data sets being analyzed were identical.

is used instead of Equation 1. Here, FIML is treated as the benchmark method. Consequently, CIFA parameter estimates and their standard errors are transformed to the logistic metric using the constant *D*. See the Appendix for details.

In summary, in this simulation study we investigated the performance of three procedures.

1. FIML: FIML estimation of the logistic GRM model using unstandardized parameters.
2. Unstandardized ULS: CIFA–ULS estimation of the normal ogive GRM model using unstandardized parameters.
3. Standardized ULS: CIFA–ULS estimation of the normal ogive GRM model using standardized parameters.

Parameter estimates and standard errors for CIFA were transformed into the unstandardized logistic metric used in FIML. All results are provided via this parameterization and link function to ensure direct comparability of the estimates.

The following outcomes were investigated: (a) proportion of proper solutions per condition, (b) relative bias and root-mean-square error of approximation (RMSE) of parameter estimates, (c) relative bias of standard errors, and (d) coverage rates.

The reader must be aware that these three methods constitute only two estimators, namely FIML and CIFA–ULS. We stress that although it may appear that three different estimators were used, procedures 2 and 3 above are in fact one and the same. The results obtained should not be affected by the parameterization used for minimization. However, in very rare and extreme cases, always less than 1/100 of the replications and sometimes only 1/1,000 of the replications, different estimates are obtained for purely numerical reasons. When this occurs, substantial differences between both sets of estimates could appear, and aggregating results over replications within an experimental condition exaggerates differences that are in fact due to a handful of replications. We present both sets of results to illustrate how the choice of parameterization used for the minimization may indeed yield different results, if only in extremely rare applications.

Further details on Samejima's model and the parameterizations employed in this paper are given in the Appendix. In addition, we provide as supplementary materials that can be downloaded a data file taken from Maydeu-Olivares, Rodríguez-Fornells, Gómez-Benito, and D'Zurilla (2000) and detailed Mplus input files that can be used to estimate the GRM with CIFA–ULS and FIML.

Results

Convergence Rates

We define convergence rate as the percentage of replications per condition that converged with the Mplus default values for each method, excluding improper solutions. A solution was deemed improper when at least one estimated λ parameter was larger than .999 in absolute value (or equivalently when $|\hat{\beta}| > |22.35|$). As in Flora and Curran (2004), nonconvergent and improper solutions were considered invalid observations and were removed from analysis.[14]

Across the 324 conditions investigated, the average convergence rate was 98.3% for FIML, 96.7% for unstandardized ULS, and 96.4% for standardized ULS. Thus, on average, convergence rates obtained by the three methods were satisfactory (and somewhat better for FIML). However, convergence rates differed depending on the number of indicators per dimension, item skewness, and sample size.

For all methods, convergence rates were substantially worse when only three indicators per dimension were present. Again, in this situation, convergence rates were better for FIML: Average convergence was 90.6% for FIML, 87.3% for unstandardized ULS, and 85.4% for standardized ULS. When the number of indicators per dimension was seven or more, convergence rates were similar on average across methods (99%). Also, convergence rates for all methods worsened as skewness increased. Again, convergence rates were somewhat better for FIML. Thus, when item skewness was greater than or equal to 1.5, average convergence was 98.2% for FIML, 96.6% for unstandardized ULS, and 96.4% for standardized ULS. When item skewness was less than 1.5, convergence performance was similar on average across methods (99%). Finally, convergence improved as sample size increased.

Table 2 summarizes convergence rates by estimation method across number of indicators per factor (3 or $\geq$7),[15] item skewness ($<$1.5 or $\geq$1.5), and sample size (200, 500, or 2,000 observations). As it can be seen in Table 2, sample size is the main threat for an accurate estimation. Neverthe-

[14] An improper or nonconvergent solution is of no use to the applied researcher. Thus, removal of these cases took us one step closer to applied research settings, as we increased the "external validity" of the results (for further implications of this strategy, see Chen, Bollen, Paxton, Curran, & Kirby, 2001). Nonetheless, we conducted additional analyses that included the improper solutions. Though these analyses resulted in changes in the outcome statistics, the conclusions did not change qualitatively. Evidence suggests that improper solutions mainly affect the problematic cases.

[15] Recall that we do not consider models with 4 to 6 indicators per dimension.

Table 2
Percentage of Valid Replications Across Conditions

		Indicators per dimension			
		3		≥7	
		Valid replications		Valid replications	
N and method	Skewness	Minimum	*M*	Minimum	*M*
200					
FIML	<1.5	28.7	87.2	98.9	99.9
	>1.5	41.0	78.8	79.8	99.1
Unstandardized ULS	<1.5	42.9	86.0	98.8	99.9
	>1.5	13.7	59.5	15.4	92.8
Standardized ULS	<1.5	23.7	77.8	80.6	97.7
	>1.5	4.00	46.9	46.7	93.4
500					
FIML	<1.5	57.3	93.3	100.0	100.0
	>1.5	48.9	85.3	95.8	99.9
Unstandardized ULS	<1.5	80.0	97.1	100.0	100.0
	>1.5	36.7	83.3	71.7	98.7
Standardized ULS	<1.5	65.3	94.7	100.0	100.0
	>1.5	16.4	75.9	95.7	99.9
2,000					
FIML	<1.5	98.3	99.8	100.0	100.0
	>1.5	95.2	99.2	100.0	100.0
Unstandardized ULS	<1.5	99.7	99.9	100.0	100.0
	>1.5	88.0	98.3	100.0	100.0
Standardized ULS	<1.5	99.7	99.9	100.0	100.0
	>1.5	76.5	96.7	100.0	100.0

Note. Valid replications are defined as the number of converging replications with proper solutions ($\lambda > 0.999$ for standardized ULS or $\beta > 22.35$ for unstandardized ULS and FIML). FIML = full information maximum likelihood; ULS = unweighted least squares.

less, even at the lowest sample size considered (200 observations), convergence rates are on average 99% when the number of indicators per dimension is at least 7. When sample size is at least 500 observations and there are at least 7 indicators per dimension, convergence rates are at least 95% for standardized ULS and FIML and 72% for unstandardized ULS. In contrast, average convergence rates are unacceptable (i.e., less than 80% on average) for all methods when the number of indicators per dimension is 3, skewness is greater than 1.5, and sample size is 200 observations.

Relative Bias and RMSE of Parameter Estimates

To compare the performance of the different methods considered, we computed the relative bias of parameter estimates as a percentage using $((\bar{\theta} - \theta)/\theta)*100$, where $\bar{\theta}$ is the average parameter estimate across valid replications and θ denotes the true parameter value. We considered that values of relative bias of less than 10% are acceptable, values from 10% to 20% indicate substantial bias, and values larger than 20% indicate an unacceptable degree of bias. The percentage of conditions falling within each bias category is displayed in Table 3.

In regard to the intercept parameters (α), Table 3 shows that the relative bias for most conditions was less than 10%, regardless of the estimation method. However, the performance of FIML was slightly superior. Several trends were readily apparent. First, relative bias is generally very small and seldom negative. Second, on average, relative bias is smaller for FIML. Third, on average, relative bias increases as item skewness increases, provided the number of indicators per dimension is three and item skewness is greater than 1.5. This effect is more pronounced for standardized ULS. Fourth, the variability of the relative bias increases with increasing item skewness.

Bias decreased with increasing sample size. Indeed, almost all conditions in which relative bias was larger than 10% consisted of 200 observations. Relative bias was also affected by the size of the factor loadings, but the trend was different for each method. In limited information methods, higher bias was associated with low slopes ($\beta = 0.74$, $\lambda = 0.40$), whereas for FIML higher bias was associated with high slopes ($\beta = 2.27$, $\lambda = 0.80$).

RMSE was computed to assess the combined effect of parameter bias and parameter variance. As criterion, RMSE has no accepted cutoff value with which to decide whether

Table 3
Percentage of Conditions in Each Method Showing 10%, 10%–20%, 20%–100%, and More Than 100% Relative Bias

Estimate and method	Relative bias			
	<10%	10%–20%	20%–100%	>100%
Intercept				
FIML	97.5	1.5	0.9	0.0
Unstandardized ULS	95.4	1.9	2.8	0.0
Standardized ULS	94.4	2.5	3.1	0.0
Slope				
FIML	95.7	3.1	1.2	0.0
Unstandardized ULS	93.8	1.9	3.7	0.6
Standardized ULS	97.8	2.2	0.0	0.0
Intercept *SE*				
FIML	93.8	1.9	1.9	2.5
Unstandardized ULS	92.0	1.5	0.0	6.5
Standardized ULS	92.9	0.9	2.2	4.0
Slope *SE*				
FIML	93.5	1.2	3.1	2.2
Unstandardized ULS	92.0	0.0	1.5	6.5
Standardized ULS	93.2	0.6	2.2	4.0

Note. Comparison was performed in α, β logistic parameterization. FIML = full information maximum likelihood; ULS = unweighted least squares; *SE* = standard error.

an estimate is acceptable or not,[16] but it is useful when comparing the quality of two estimators, as it represents a trade-off between bias and variability. Nevertheless, the pattern of results with RMSE was almost identical to the pattern found with relative bias. This is due to the fact that estimation performance in the present study was mainly driven by the bias of parameter estimates, rather than their variance.

This comparison is shown in Table 4, which displays the average values for relative bias and RMSE by number of observations, indicators per factor (with three, seven, and more than seven indicators per factor), item skewness (<1.5 and ≥1.5), and true β parameter. As the table shows, it was mainly conditions with three indicators per factor and $n = 200$ that showed substantial amounts of positive bias. With this sample size and 3 indicators per factor, FIML was more accurate than limited information methods. This difference disappeared when the skewness was high and the slope was at least 1.27, a setting at which standardized ULS slightly outperforms FIML in terms of RMSE. Differences in estimation precision disappeared as the number of indicators per factor increased. Limited information methods were marginally more precise in terms of relative bias when 7 or more indicators were used per factor, and it was so with even the smallest sample size ($n = 200$) and particularly when true slopes were low ($\beta = 0.74$, $\lambda = 0.40$).

With regard to intercept RMSEs, limited information and FIML yielded comparable values in almost every condition with more than three indicators per factor and 200 observations. Standardized ULS was the method with the smallest RMSEs. There were certain conditions in which this was not the case. FIML outperformed limited information methods in the case of three indicators per factor, $n = 200$, and skewness of less than 1.5 for all slope values and when skewness was greater than 1.5 and slopes were less than or equal to 1.27. There were other conditions for which FIML showed smaller RMSEs than did standardized ULS, such as in the case of three indicators per factor, low skewness (skewness <1.5) and low slope ($\beta = 0.74$), no matter the sample size.

With regard to the slope parameters (β), Table 3 shows that in most conditions acceptable levels of bias were obtained regardless of the estimation method. Much as with the intercept parameters, increasing skewness increased estimation bias, whereas increasing sample size, the number of indicators per factor, and the number of categories per item improved the accuracy of the estimates. Also, when all other factors were held constant, a higher slope improved estimation.

The sign of the bias for slopes was most often positive, although in some conditions both methods underestimated the true parameters. We see from Table 3 that, on average, relative bias was larger for slopes than it was for intercepts. Again, bias increased in models with only three indicators per factor. Also, bias increased with increasing skewness, particularly when the number of indicators was only three and skewness was larger than 1.5. Almost all conditions in which more than 10% bias was obtained had just three indicators per factor.

Table 5 displays the average values for β relative bias and RMSE by number of observations, indicators per factor (with three, seven, and more than seven indicators per factor), item skewness (<.5 and ≥1.5), and true β parameter. As this table shows, estimation inaccuracies were more serious in the case of unstandardized ULS, which was more affected by these factors. Also, for all methods, parameter estimation improved with increasing true slope. Estimation inaccuracies appeared in

[16] Any choice of statistic for comparing the estimators has its pros and cons. RMSE is a common criterion when comparing estimators. As RMSE considers bias and parameter variability, it is very appealing when there is a trade-off between the two. However, the RMSE metric is dependent on the scale of the data, so it is not possible to provide an accepted cutoff criterion to point out when estimators fails. Another disadvantage of RMSE is that it is not possible to compute a RMSE for standard errors.

On the other hand, relative bias may be used for identifying unacceptable conditions through a cutoff criterion. Nevertheless, this might lead to misleading results whenever a parameter estimate is very close to zero, as relative bias would be artificially inflated by the denominator, even for very small departures from the true parameter.

There are other indices that could have been used, such as absolute bias or mean square error, but we decided that using RMSE and relative bias as indices to compare the estimators performance would simplify comparisons for the reader. Comparability has been maintained throughout the paper because we computed all statistics and indices using the logit intercept/slope (α, β) parameterization familiar to most IRT users. This was achieved in standardized ULS by using Equation A. 13 (see Appendix).

Table 4
Average Percentage of Relative Bias and Average RMSE (in Parentheses) of α Parameter Estimates for Each Method by Number of Observations, Indicators per Factor, Item Skewness, and True β Parameter

Observations and indicators per factor	Skewness	Method								
		FIML (β)			Unstandardized ULS (β)			Standardized ULS (β)		
		0.74	1.27	2.27	0.74	1.27	2.27	0.74	1.27	2.27
200										
3	<1.5	06 (0.37)	05 (0.36)	03 (0.43)	**18 (1.10)**	05 (0.43)	04 (0.44)	**−24 (0.97)**	**44 (0.64)**	06 (0.74)
	>1.5	**12 (0.74)**	09 (1.11)	**16 (2.66)**	**62 (4.06)**	**28 (2.56)**	**21 (2.42)**	**16 (1.04)**	**11 (0.72)**	**12 (0.98)**
7	<1.5	02 (0.23)	02 (0.27)	02 (0.36)	02 (0.22)	02 (0.25)	02 (0.35)	**51 (0.65)**	02 (0.25)	02 (0.35)
	>1.5	09 (0.86)	05 (0.60)	06 (0.89)	**25 (2.53)**	04 (0.59)	06 (0.89)	04 (0.35)	06 (0.64)	05 (0.67)
>7	<1.5	02 (0.22)	02 (0.25)	02 (0.34)	01 (0.21)	01 (0.24)	02 (0.33)	01 (0.21)	01 (0.24)	02 (0.33)
	>1.5	04 (0.48)	04 (0.43)	05 (0.72)	05 (0.74)	02 (0.41)	04 (0.71)	03 (0.44)	03 (0.39)	04 (0.63)
500										
3	<1.5	02 (0.19)	02 (0.19)	01 (0.25)	05 (0.47)	02 (0.18)	02 (0.25)	**−40 (1.39)**	02 (0.18)	02 (0.25)
	>1.5	**12 (1.44)**	**15 (2.00)**	**12 (2.23)**	**30 (2.57)**	09 (0.99)	06 (1.01)	07 (0.45)	08 (0.71)	07 (0.83)
7	<1.5	01 (0.14)	01 (0.16)	01 (0.22)	01 (0.13)	01 (0.15)	01 (0.21)	01 (0.13)	01 (0.15)	01 (0.21)
	>1.5	03 (0.53)	02 (0.28)	02 (0.45)	02 (0.33)	01 (0.26)	02 (0.43)	02 (0.33)	01 (0.26)	02 (0.43)
>7	<1.5	01 (0.14)	01 (0.16)	01 (0.21)	00 (0.13)	00 (0.15)	01 (0.20)	00 (0.13)	00 (0.15)	01 (0.20)
	>1.5	01 (0.20)	01 (0.24)	02 (0.39)	01 (0.22)	01 (0.22)	01 (0.38)	01 (0.18)	01 (0.22)	01 (0.38)
2,000										
3	<1.5	01 (0.08)	00 (0.09)	00 (0.12)	01 (0.08)	00 (0.09)	00 (0.12)	01 (0.08)	00 (0.09)	00 (0.12)
	>1.5	02 (0.23)	01 (0.32)	02 (0.65)	05 (0.88)	01 (0.23)	01 (0.30)	05 (0.75)	01 (0.23)	01 (0.30)
7	<1.5	00 (0.07)	00 (0.08)	00 (0.11)	00 (0.07)	00 (0.08)	00 (0.10)	00 (0.07)	00 (0.08)	00 (0.10)
	>1.5	00 (0.11)	00 (0.13)	01 (0.20)	00 (0.10)	00 (0.12)	00 (0.20)	00 (0.10)	00 (0.12)	00 (0.20)
>7	<1.5	00 (0.07)	00 (0.08)	00 (0.10)	00 (0.06)	00 (0.07)	00 (0.10)	00 (0.06)	00 (0.07)	00 (0.10)
	>1.5	00 (0.09)	00 (0.12)	01 (0.18)	00 (0.08)	00 (0.11)	00 (0.18)	00 (0.08)	00 (0.11)	00 (0.18)

Note. Comparison was performed in α, β logistic parameterization. The true values β = 0.74, 1.27, 2.27 are equivalent to λ = .4, .6, .8. Conditions with more than 10% bias are in boldface. RMSE = root-mean-square error of approximation; FIML = full information maximum likelihood; ULS = unweighted least squares.

similar conditions: small sample sizes ($n = 200$) that were used to estimate a few, highly skewed indicators. In these conditions, standardized ULS and FIML maintained the amount of bias within a more restricted range than did unstandardized ULS, which was the worst method. All in all, the amount of bias for standardized ULS was acceptable except with $n = 200$ and three indicators per factor. The same results were found for FIML, except that in this case, and if the slope was 1.27 or more and skewness was low, FIML yielded accurate estimates. No method performed accurately with three indicators per factor and skewness of over 1.5 until 2,000 observations were used, although standardized ULS performed slightly better in such a setting than did the other two methods with $n = 500$.

The behavior of RMSEs for slope parameters was similar to that of the intercepts. FIML was superior in conditions with more three indicators per factor and $n = 200$, but standardized ULS was the method that otherwise displayed, in general, smaller RMSEs. Again, there were exceptions to this pattern. FIML showed smaller RMSEs than did standardized ULS in the case of three indicators per factor, low skewness (<1.5), and low slope (β = 0.74), regardless of sample size.

As far as slope and intercept parameter estimates are concerned, there are only small differences in RMSE between standardized ULS and FIML, and standardized ULS is in general slightly superior to FIML. FIML is superior to standardized ULS in terms of RMSE only when the number of indicators per factor is three and sample size is 200. Also, with three indicators per dimension and low slopes (β = 0.74), FIML is sometimes superior to standardized ULS.

Relative Bias of Standard Errors

We computed the relative bias of the standard errors using $((\overline{SE}_\theta - sd_\theta)/sd_\theta)*100$, where $\overline{SE}_\theta$ is the average standard error of a parameter estimate across valid replications and sd_θ denotes the standard deviation of the parameter estimates across valid replications. Notice that it is not possible to compute a RMSE for standard errors, which are an essential part of this study. To facilitate interpretation and give further insight on relative bias interpretation, we provide standard deviations of parameter estimates as a reference for comparisons between the magnitudes of relative biases.

As shown in Table 3, acceptable levels of bias were obtained in most conditions regardless of the estimation method (over 90% of the conditions). However, for all three methods the performance of the parameter estimates was better than the performance of the standard errors. Also, the percentage of

Table 5
Average Percentage of Relative Bias and Average RMSE (in Parentheses) of β Parameter Estimates for Each Method by Number of Observations, Indicators per Factor, Item Skewness, and True β Parameter

Observations and indicators per factor	Skewness	Method								
		FIML (β)			Unstandardized ULS (β)			Standardized ULS (β)		
		0.74	1.27	2.27	0.74	1.27	2.27	0.74	1.27	2.27
200										
3	<1.5	**17 (0.59)**	08 (0.51)	04 (0.52)	**51 (1.73)**	09 (0.66)	05 (0.61)	**21 (0.74)**	**14 (0.71)**	**10 (0.71)**
	>1.5	**16 (0.91)**	**12 (0.94)**	**15 (1.96)**	**120 (3.74)**	**44 (2.32)**	**26 (2.16)**	**41 (1.17)**	**19 (1.07)**	**14 (1.34)**
7	<1.5	04 (0.29)	02 (0.27)	02 (0.37)	03 (0.28)	02 (0.26)	02 (0.35)	04 (0.36)	02 (0.26)	02 (0.35)
	>1.5	09 (0.85)	07 (0.59)	05 (0.77)	**10 (2.22)**	05 (0.60)	06 (0.81)	07 (0.51)	08 (0.68)	06 (0.67)
>7	<1.5	02 (0.21)	02 (0.23)	02 (0.33)	02 (0.20)	01 (0.22)	02 (0.31)	02 (0.20)	01 (0.22)	02 (0.31)
	>1.5	05 (0.52)	05 (0.42)	06 (0.63)	−03 (0.77)	02 (0.43)	04 (0.65)	02 (0.47)	02 (0.43)	04 (0.60)
500										
3	<1.5	07 (0.33)	02 (0.27)	01 (0.30)	**15 (0.81)**	03 (0.27)	02 (0.31)	05 (0.54)	03 (0.27)	02 (0.31)
	>1.5	**26 (1.14)**	**21 (1.46)**	**11 (1.59)**	**80 (2.45)**	**15 (0.97)**	08 (0.94)	**22 (0.79)**	**19 (0.78)**	**11 (0.83)**
7	<1.5	01 (0.17)	01 (0.16)	01 (0.22)	01 (0.16)	01 (0.16)	01 (0.21)	01 (0.16)	01 (0.16)	01 (0.21)
	>1.5	05 (0.50)	02 (0.30)	01 (0.40)	03 (0.40)	02 (0.30)	02 (0.41)	03 (0.40)	02 (0.30)	02 (0.41)
>7	<1.5	01 (0.13)	01 (0.14)	01 (0.20)	01 (0.13)	01 (0.14)	01 (0.19)	01 (0.13)	01 (0.14)	01 (0.19)
	>1.5	02 (0.25)	02 (0.24)	02 (0.35)	−01 (0.27)	01 (0.24)	02 (0.36)	−01 (0.23)	01 (0.24)	02 (0.36)
2,000										
3	<1.5	01 (0.15)	00 (0.12)	00 (0.14)	02 (0.15)	01 (0.12)	00 (0.14)	02 (0.15)	01 (0.12)	00 (0.14)
	>1.5	06 (0.35)	02 (0.31)	00 (0.49)	**14 (0.85)**	02 (0.27)	02 (0.29)	08 (0.72)	02 (0.27)	02 (0.29)
7	<1.5	00 (0.08)	00 (0.08)	00 (0.11)	00 (0.08)	00 (0.08)	00 (0.10)	00 (0.08)	00 (0.08)	00 (0.10)
	>.5	01 (0.16)	00 (0.14)	−01 (0.19)	00 (0.15)	00 (0.14)	01 (0.19)	00 (0.15)	00 (0.14)	01 (0.19)
>7	<1.5	00 (0.06)	00 (0.07)	00 (0.10)	00 (0.06)	00 (0.07)	00 (0.09)	00 (0.06)	00 (0.07)	00 (0.09)
	>.5	00 (0.12)	00 (0.12)	00 (0.17)	00 (0.11)	00 (0.12)	00 (0.17)	00 (0.11)	00 (0.12)	00 (0.17)

Note. Comparison was performed in α, β logistic parameterization. The true values β = 0.74, 1.27, 2.27 are equivalent to λ = .4, .6, .8. Conditions with more than 10% bias are shown in boldface. RMSE = root-mean-square error of approximation; FIML = full information maximum likelihood; ULS = unweighted least squares.

conditions for which acceptable levels of bias were obtained is similar across methods for the standard errors of the intercepts. For the slopes, a similar number of conditions showed acceptable bias for FIML and standardized ULS; for unstandardized ULS, bias was somewhat worse. Remarkably, more than 100% bias was obtained for all methods in a number of conditions. Extreme bias is most likely for unstandardized ULS, followed by standardized ULS and then FIML.

Tables 6 and 7 show the average bias for intercept and slope standard errors by method, skewness level, model size, and true parameter value. They also provide standard deviations of parameter estimates. Overall, the behavior of unstandardized ULS and FIML standard errors was similar. Thus, we found that (a) most often, standard errors for intercepts are overestimated and those for slopes are underestimated; (b) increasing skewness increased the variability of the bias; (c) increasing skewness increased the bias for models with only three indicators per dimension; and (d) in some conditions with three indicators per factor, the bias of the standard errors was positive and unacceptably high. The main difference between the performance of these two methods is that when the standard errors were unacceptable, the magnitude of the bias was much larger for unstandardized ULS than for FIML. The performance of standardized ULS fell between that of these two methods.

As shown in Tables 6 and 7, all estimation methods yielded unacceptable standard errors for both the slope and the intercept parameters whenever the number of indicators per dimension was three, item skewness was large (≥1.5), and sample size was no more than 500 observations. This pattern also appeared when skewness was less than 1.5 and sample size was 200 observations, although FIML provided accurate standard errors when the slope parameter was high enough (β ≥ 1.27). ULS methods yielded unacceptable standard errors for both intercepts and slopes when the number of items per dimension was only three and sample size was 200, especially with low slopes. This circumstance changed when observations were increased to 500: All methods yielded good standard errors, provided that the item skewness was low (<1.5). Even so, this increase in sample size is not enough in the case of highly skewed items (skewness ≥ 1.5). FIML and unstandardized ULS (but not standardized ULS) also yielded unacceptable standard errors for slope parameters when there were seven indicators per dimension, item skewness was large, and slopes were small (β = 0.74).

Table 6
Average Percentage of Relative Bias of the Standard Errors and Average Standard Deviations (in Parentheses) for α Parameter for Each Method by Number of Observations, Indicators per Factor, Item Skewness, and True β Parameter

Observations and indicators per factor	Skewness	Method								
		FIML (β)			Unstandardized ULS (β)			Standardized ULS (β)		
		0.74	1.27	2.27	0.74	1.27	2.27	0.74	1.27	2.27
200										
3	<1.5	**18 (0.38)**	−01 (0.37)	−01 (0.45)	**1490 (1.20)**	**23 (0.45)**	−01 (0.45)	**222 (1.06)**	**109 (0.45)**	**205 (0.45)**
	>1.5	**174 (0.68)**	**238 (1.09)**	**576 (2.60)**	**5535 (3.82)**	**1245 (2.46)**	**152 (2.32)**	**762 (3.82)**	**122 (2.46)**	**33 (2.32)**
7	<1.5	−02 (0.24)	−02 (0.27)	−01 (0.37)	−02 (0.22)	−01 (0.25)	−02 (0.35)	04 (0.24)	−01 (0.25)	−02 (0.35)
	>1.5	**74 (0.84)**	**13 (0.59)**	**−13 (0.87)**	**601 (2.46)**	−03 (0.58)	**−11 (0.87)**	04 (2.46)	**15 (0.58)**	**−10 (0.87)**
>7	<1.5	−01 (0.23)	00 (0.26)	00 (0.35)	−01 (0.21)	−01 (0.24)	−00 (0.33)	−01 (0.21)	−01 (0.24)	−01 (0.33)
	>1.5	**−11 (0.46)**	−06 (0.42)	−09 (0.70)	**33 (0.73)**	−07 (0.40)	−09 (0.69)	**−20 (0.78)**	−07 (0.40)	−08 (0.68)
500										
3	<1.5	08 (0.20)	−01 (0.20)	01 (0.26)	**425 (0.51)**	−01 (0.19)	00 (0.25)	−01 (0.44)	−01 (0.19)	00 (0.25)
	>1.5	**252 (1.42)**	**445 (1.96)**	**110 (2.19)**	**3392 (2.48)**	**106 (0.96)**	**00 (0.99)**	**380 (2.48)**	**233 (0.96)**	**28 (0.99)**
7	<1.5	−01 (0.14)	−01 (0.17)	−01 (0.23)	−01 (0.14)	−01 (0.16)	−01 (0.22)	−01 (0.14)	−01 (0.16)	−01 (0.22)
	>1.5	**38 (0.52)**	−03 (0.28)	−04 (0.44)	−01 (0.33)	−03 (0.26)	−04 (0.43)	−01 (0.33)	−03 (0.26)	−04 (0.43)
>7	<1.5	00 (0.14)	−01 (0.16)	−01 (0.22)	00 (0.13)	00 (0.15)	−01 (0.21)	00 (0.13)	00 (0.15)	−01 (0.21)
	>1.5	−04 (0.20)	−02 (0.24)	−03 (0.38)	03 (0.22)	−02 (0.22)	−03 (0.38)	−02 (0.22)	−02 (0.22)	−03 (0.38)
2,000										
3	<1.5	01 (0.08)	01 (0.09)	01 (0.12)	−01 (0.08)	−01 (0.09)	−01 (0.12)	−01 (0.08)	−01 (0.09)	−01 (0.12)
	>1.5	03 (0.23)	**−16 (0.32)**	**−22 (0.65)**	**496 (0.87)**	−06 (0.23)	−03 (0.29)	**10 (0.87)**	−06 (0.23)	−03 (0.29)
7	<1.5	00 (0.07)	00 (0.08)	01 (0.11)	00 (0.07)	−01 (0.08)	−01 (0.11)	00 (0.07)	−01 (0.08)	−01 (0.11)
	>1.5	−01 (0.11)	00 (0.13)	00 (0.20)	−02 (0.10)	−01 (0.12)	−01 (0.20)	−02 (0.10)	−01 (0.12)	−01 (0.20)
>7	<1.5	00 (0.07)	00 (0.08)	00 (0.11)	00 (0.06)	00 (0.07)	00 (0.10)	00 (0.06)	00 (0.07)	00 (0.10)
	>1.5	00 (0.09)	−01 (0.12)	−01 (0.18)	−01 (0.08)	00 (0.11)	−01 (0.18)	−01 (0.08)	00 (0.11)	−01 (0.18)

Note. Comparison was performed in α, β logistic parameterization. The true values β = 0.74, 1.27, 2.27 are equivalent to λ = .4, .6, .8. Conditions with more than 10% bias are shown in boldface. FIML = full information maximum likelihood; ULS = unweighted least squares.

Even if few in number, standard error inaccuracies were quite dramatic for ULS methods—especially for unstandardized ULS, for which almost every positively biased condition showed much more than 100% relative bias—and they were found across all skewness levels. In contrast, the FIML standard errors with unacceptable bias were confined to the more extreme item skewness levels.

Parameter Coverage

Figure 2 shows the coverage of 95% confidence intervals for parameter estimates for all 324 conditions investigated. Coverage was adequate (between 92.5% and 97.5% for 95% confidence intervals) for most conditions across methods. For α parameters, coverage was similar across methods, except for three conditions for which FIML yielded somewhat unacceptable coverages. These conditions involved models with only three extremely skewed indicators per dimension and medium-to-high slopes (β ≥ 1.27) and were estimated with 500 or fewer observations. For β parameters, FIML resulted in more accurate coverages than did ULS. The latter yielded somewhat inaccurate coverages, regardless of the number of indicators per dimension, when sample size was small (200 observations) and items were skewed. It is interesting to compare the coverage rates for standardized and unstandardized ULS. As shown in Figure 2, in general, coverage rates for unstandardized ULS were more accurate and less affected by item skewness than were coverage rates for standardized ULS. However, in those conditions for which coverage rates were unacceptable, they were far more unacceptable for unstandardized than standardized ULS. Across methods, slope coverage was acceptable as long as sample size was larger than 200 observations. Models with few indicators per factor were prone to yield inflated coverage values.

Discussion

Our purpose in this simulation study was to investigate the limits of the good performance of FIML in estimating IRT models by manipulating a comprehensive set of factors that could affect its performance. Due to the computational demands for this estimation method, previous research on the finite sample behavior of this asymptotically optimal estimator was rather fragmentary; only a few conditions were investigated, and most often the number of replications was insufficient. Two issues that had scarcely been addressed in the literature and that have been investigated in this study are the behavior of FIML standard errors and the behavior of FIML in multidimensional models.

Table 7
Average Percentage of Relative Bias of the Standard Errors and Average Standard Deviations (in Parentheses) for β Parameter for Each Method by Number of Observations, Indicators per Factor, Item Skewness, and True β Parameter

Observations and indicators per factor	Skewness	Method								
		FIML (β)			Unstandardized ULS (β)			Standardized ULS (β)		
		0.74	1.27	2.27	0.74	1.27	2.27	0.74	1.27	2.27
200										
3	<1.5	**33 (0.58)**	−02 (0.50)	−01 (0.51)	**1530 (1.69)**	**24 (0.65)**	−04 (0.60)	**243 (0.72)**	**110 (0.70)**	**189 (0.69)**
	>1.5	**125 (0.90)**	**211 (0.92)**	**568 (1.92)**	**5025 (3.61)**	**1145 (2.25)**	**144 (2.08)**	**735 (1.13)**	**120 (1.04)**	**30 (1.30)**
7	<1.5	−06 (0.28)	−03 (0.27)	−02 (0.36)	−09 (0.28)	−05 (0.26)	−04 (0.35)	−01 (0.36)	−05 (0.26)	−04 (0.35)
	>1.5	**48 (0.85)**	**11 (0.58)**	**−11 (0.76)**	**518 (2.21)**	−04 (0.60)	**−10 (0.80)**	−04 (0.51)	**10 (0.67)**	−09 (0.66)
>7	<1.5	−03 (0.21)	−02 (0.23)	−02 (0.32)	−06 (0.20)	−04 (0.22)	−04 (0.31)	−06 (0.20)	−04 (0.22)	−04 (0.31)
	>1.5	**−12 (0.52)**	−06 (0.42)	−07 (0.62)	**19 (0.77)**	−09 (0.43)	−08 (0.64)	**−24 (0.47)**	−09 (0.43)	−07 (0.59)
500										
3	<1.5	**16 (0.33)**	−02 (0.27)	00 (0.30)	**419 (0.80)**	−03 (0.27)	−02 (0.31)	−02 (0.53)	−03 (0.27)	−02 (0.31)
	>1.5	**239 (1.12)**	**408 (1.43)**	**109 (1.56)**	**3097 (2.38)**	**97 (0.95)**	01 (0.92)	**356 (0.77)**	**210 (0.76)**	**27 (0.81)**
7	<1.5	−02 (0.16)	−02 (0.16)	−02 (0.22)	−04 (0.16)	−02 (0.16)	−02 (0.21)	−04 (0.16)	−02 (0.16)	−02 (0.21)
	>1.5	**33 (0.50)**	−03 (0.30)	−03 (0.40)	−02 (0.40)	−03 (0.29)	−03 (0.41)	−02 (0.40)	−03 (0.29)	−03 (0.41)
>7	<1.5	−01 (0.13)	−01 (0.14)	−01 (0.20)	−02 (0.13)	−02 (0.13)	−02 (0.19)	−02 (0.13)	−02 (0.13)	−02 (0.19)
	>1.5	−04 (0.25)	−02 (0.24)	−02 (0.34)	−01 (0.26)	−03 (0.24)	−03 (0.36)	−06 (0.23)	−03 (0.24)	−03 (0.36)
2,000										
3	<1.5	−01 (0.15)	−01 (0.12)	00 (0.14)	−02 (0.15)	−01 (0.12)	00 (0.14)	−02 (0.15)	−01 (0.12)	00 (0.14)
	>1.5	06 (0.35)	**−12 (0.31)**	**−18 (0.49)**	**450 (0.84)**	−03 (0.27)	−02 (0.29)	**13 (0.70)**	−03 (0.27)	−02 (0.29)
7	<1.5	00 (0.08)	−01 (0.08)	00 (0.11)	−01 (0.08)	00 (0.08)	00 (0.10)	−01 (0.08)	00 (0.08)	00 (0.10)
	>1.5	−02 (0.16)	−01 (0.14)	00 (0.19)	−02 (0.15)	00 (0.14)	−01 (0.19)	−02 (0.15)	00 (0.14)	−01 (0.19)
>7	<1.5	00 (0.06)	00 (0.07)	00 (0.10)	−01 (0.06)	00 (0.07)	00 (0.09)	−01 (0.06)	00 (0.07)	00 (0.09)
	>1.5	−01 (0.11)	00 (0.12)	−01 (0.16)	−01 (0.11)	−01 (0.12)	−00 (0.17)	−01 (0.11)	−01 (0.12)	−01 (0.17)

Note. Comparison was performed in α, β logistic parameterization. The true values β = 0.74, 1.27, 2.27 are equivalent to λ = 4, .6, .8. Conditions with more than 10% bias are shown in boldface. FIML = full information maximum likelihood; ULS = unweighted least squares.

Also of interest was comparison of the behavior of FIML with that of a CIFA estimator based on polychorics, as the latter involves less computation. We performed CIFA–ULS estimation using two different parameterizations, unstandardized parameters and standardized parameters, to assess their effect on IRT estimation.

What Are the Limits of the Good Performance of FIML in Estimating IRT Models?

On the whole, the performance of FIML under the conditions investigated was excellent. In only 36 of the 324 conditions investigated was parameter or standard error bias larger than 10% (our cutoff criterion for "good" performance).

FIML failed in conditions involving the combination of (a) three latent traits, (b) a small number of indicators per dimension, (c) binary items, (d) low item slopes, and (e) high skewness. As more of these factors were involved, the higher the likelihood that FIML would fail to yield adequate parameter estimates and/or standard errors. Thus, of the failed conditions, 77% involved three dimensions, 61% involved three indicators per dimension, 86% involved binary items, 55% involved true item slopes of β = 0.74 (or equivalently λ = .4), and 75% involved items with skewness ≥1.5. For instance, FIML failed in all conditions involving three latent traits, each with three indicators, when sample size was 200 observations and the items were dichotomous (i.e., regardless of item skewness and item slope). It also failed under the above conditions when sample size was 500 if the true item slopes were β = 0.74.

Of the 36 conditions for which FIML failed, 22 involved models with three uncorrelated latent traits, each with 3 indicators. This is an unrealistic setting in applications, as the latent traits are generally correlated when so few indicators are used. However, it is interesting that FIML failed to estimate the three-dimensional model with 3 dichotomous indicators per dimension, even when sample size was 2,000, when the items had the highest skewness considered (2.67) and the smallest slope (β = 0.74 or, equivalently, λ = .4). Of the conditions for which FIML failed that did not include 3 indicators per dimension, five involved three latent traits with 7 indicators each when the items were dichotomous, sample size was 200, item skewness was large (≥1.5), and item slopes were not large (β ≤ 1.27 or, equivalently, λ ≤ .6). Two more conditions involved three latent traits with a sample size of 500, 7 and 14 indicators per dimension, highest item skewness (2.67), and lowest item slopes (β = 0.74 or, equivalently, λ = .4). Finally, FIML also failed in seven of the nine conditions that involved one latent trait, 9 dichotomous indicators, and 200 observations: those in which item skewness was ≥1.5.

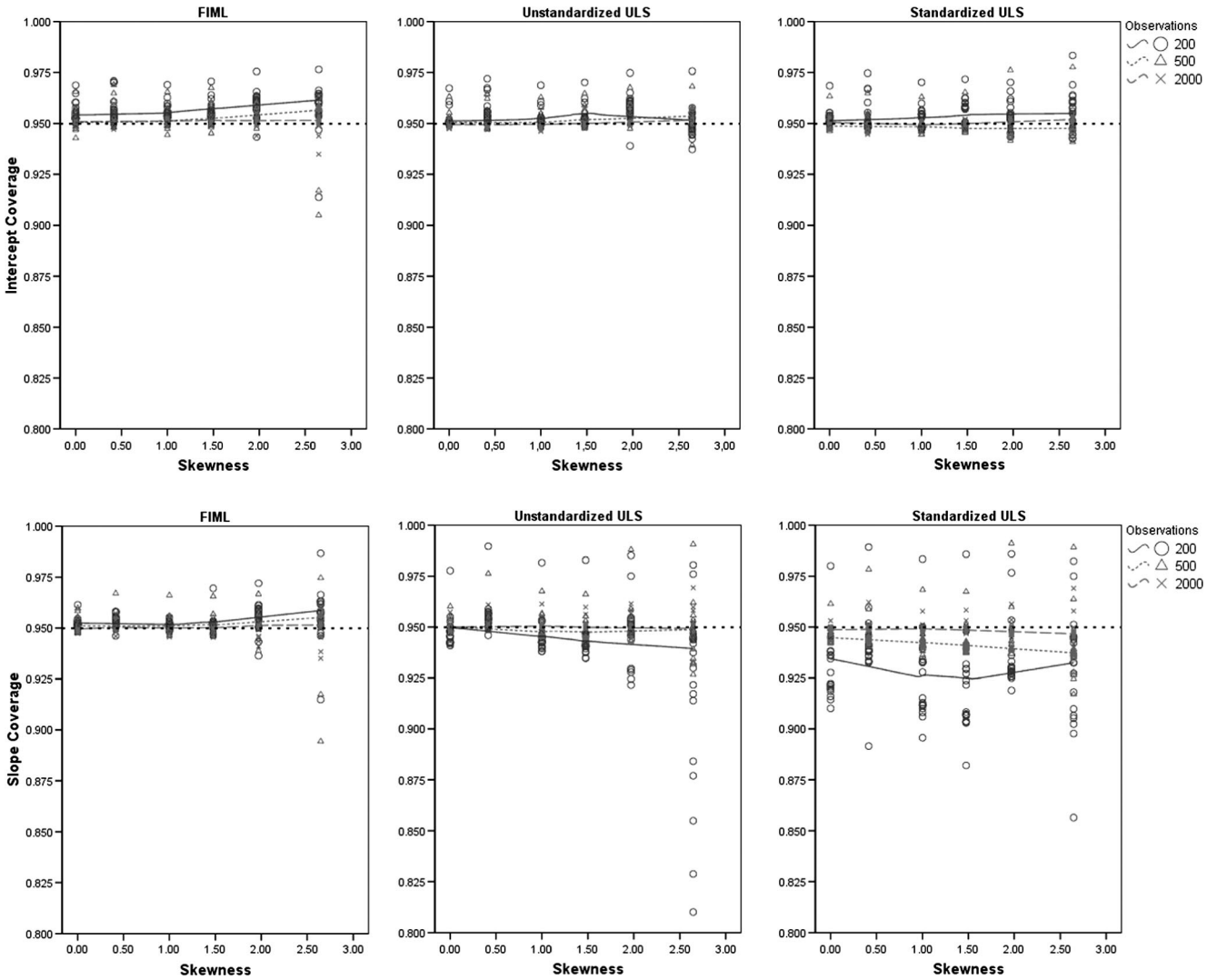

Figure 2. Proportion of times (coverage) that 95% confidence intervals for parameter estimates included the population parameter. Coverage should be close to a nominal rate (95%). A nonparametric procedure has been used to model the relationship between coverage and item skewness by sample size. Except when sample size was 200 observations and item skewness was extreme (>2.5), coverage was adequate for all methods. ULS = unweighted least squares; FIML = full maximum likelihood.

What Are the Limits of the Good Performance of CIFA–ULS? Effect of the Choice of Parameterization

Overall, the performance of ULS proved to be as good as that of FIML. Both standardized and unstandardized CIFA–ULS failed for 27 conditions for which FIML failed. Eleven of these shared failed conditions involved the combination of three factors, low factor loading ($\beta = 0.74$), and a sample size of 200 observations, whereas in the remaining 16 conditions at least two of these settings were present. Unstandardized ULS failed for 13 additional conditions (for a total of 40) for which FIML succeeded. Eight of them involved three latent traits with 3, 7, and 14 indicators per factor and low-to-medium slopes ($\beta = 0.74$ and $\beta = 1.27$). The remaining 5 conditions were unidimensional and involved a combination of high skewness (>1.5), a sample size of 200 observations, and models with 21 and 42 indicator per factor. Standardized ULS failed for just 35 conditions, 1 less than FIML. In addition to its failure in the conditions for which all three methods failed (described above), ULS failed in estimating six conditions with three latent traits and highly skewed dichotomous indicators (skewness > 1.5) with 3 to 14 indicators per dimension. In this setting the method fails even with 2,000 observations if the item slope is low ($\beta = 0.74$). Standardized ULS failed for two additional conditions with 5-point indicators that involved 3

indicators per dimension in three-dimensional models and 200 or 500 observations. It is interesting that both standardized and unstandardized ULS failed in estimating the hardest model (9 dichotomous indicators, three latent traits, lowest slope) for sample size 2,000 not only when skewness was highest (2.67)—as did FIML—but also when the item skewness was the second largest (1.96).

Across the 324 conditions of the study, the average performance of CIFA was somewhat better when standardized parameters were used for the estimation. However, this was due to differential performance in just a handful of replicates within each condition (always less than 1 in a 100, often 1 in 1,000). In the vast majority of cases, minimizing with respect to standardized or unstandardized yielded identical results. The effect of parameterization on results at the aggregated level was marked in difficult estimation settings (i.e., low sample size, few indicators per factor, and highly skewed items). Otherwise, the effect of parameterization on averaged results was almost negligible. The main effect on aggregated results of using the unstandardized parameterization for CIFA–ULS was to increase the variability of the parameter estimates. Tables 6 and 7 provide the standard deviation of intercept and slope parameter estimates by method, model size, number of observations, and indicators per factor, skewness level, and true parameter value. As these tables show, the standard deviation of the parameter estimates was reduced by half in certain instances related with the aforementioned estimation conditions; sometimes, the reduction was threefold when standardized parameters were used. Partly as a result of this, the bias of unstandardized ULS was larger than that of standardized ULS for conditions in which the standardized parameterization yielded unacceptable results. Also as a result of this, coverage for the slope parameters was, in general, better when estimation was based on unstandardized rather than standardized parameters.

However, in practice, applied researchers should expect the same results regardless of which parameterization is used for estimation. In cases of nonconvergence, or when an improper solution is found when unstandardized parameters are used, the results obtained here suggest that a proper solution might be found minimizing with respect to standardized parameters.[17]

Which Method to Use?

A striking result of our study is that ULS and FIML are not that different with regard to the accuracy of the parameter estimates and standard errors. Nevertheless, to provide insight on which method is more accurate, we have computed the percentage of conditions in which standardized ULS is more accurate than FIML across four criteria (intercept relative bias, slope relative bias, intercept standard error relative bias, and slope standard error relative bias). These data are summarized in Table 8.

Clearly, standardized ULS outperforms FIML in parameter accuracy (particularly in estimating the intercepts), whereas the standard errors for the slope parameters are more accurate for FIML. Similar categories of successful performance are obtained when using RMSE, although it would be more complicated to propose a cutoff performance. Nevertheless, when using minimum RMSE criteria one would still favor standardized ULS as the method of choice.

Table 8
Summary Table Comparing FIML and Standardized ULS Performance

	Standardized ULS		
	Succeeds		
Criterion	Performance better than FIML	Performance worse than FIML	Fails
α			
FIML succeeds	67.0%	19.4%	2.5%
FIML fails	1.9%	0.9%	8.3%
β			
FIML succeeds	51.2%	35.2%	2.5%
FIML fails	1.6%	1.2%	8.3%
α *SE*			
FIML succeeds	42.0%	44.4%	2.5%
FIML fails	2.8%	0.0%	8.3%
β *SE*			
FIML succeeds	21.9%	64.5%	2.5%
FIML fails	2.5%	0.3%	8.3%

Note. Successful conditions are defined as those in which the relative biases of α estimates, α standard errors, β estimates, and β standard errors are smaller than 10% (in absolute value). Failed conditions are defined as those in which at least one of these four criteria is not met. For conditions where at least one method succeeds, the table provides the percentage of all conditions where standardized ULS outperforms FIML in each of the four criteria: α relative bias, β relative bias, α standard error relative bias, and β standard error relative bias. Notice that the entries in each section sum to 100%. For the 324 conditions investigated, FIML failed but standardized ULS succeeded in 2.8% of conditions. FIML = full information maximum likelihood; ULS = unweighted least squares; *SE* = standard error.

Because the observed differences between standardized and unstandardized ULS are caused by differential performance on a handful of replicates, the percentages for those conditions in which unstandardized ULS is more accurate than FIML are given in Table 9. The results shown in this table are very similar to those shown in Table 8. Regardless of which parameterization is used for the ULS estimator, estimates are more accurate for ULS, and FIML is more accurate in its standard errors.

Which is the most advisable method then, FIML or CIFA–ULS? In general, there is not much to choose from in terms of parameter estimation accuracy and standard error accuracy between CIFA–ULS and FIML. The latter clearly exhibits better performance than the former only in models that involve three indicators per dimension, that are estimated with

[17] Mplus can perform the minimization with respect to unstandardized or standardized parameters. When unstandardized parameters are used, standardized parameters are available as an option. When standardized parameters are used, unstandardized parameters can be obtained with program statements.

Table 9
Summary Table Comparing FIML and Unstandardized ULS Performance

Criterion	Unstandardized ULS: Succeeds: Performance better than FIML	Unstandardized ULS: Succeeds: Performance worse than FIML	Unstandardized ULS: Fails
α			
FIML succeeds	71.6%	13.6%	3.7%
FIML fails	2.2	0.3%	8.6%
β			
FIML succeeds	55.2%	29.9%	3.7%
FIML fails	1.5%	0.9%	8.6%
α *SE*			
FIML succeeds	41.0%	44.1%	3.7%
FIML fails	2.5%	0.0%	8.6%
β *SE*			
FIML succeeds	21.9%	63.3%	3.7%
FIML fails	2.2%	0.3%	8.6%

Note. Successful conditions are defined as those in which the relative biases of α estimates, α standard errors, β estimates, and β standard errors are smaller than 10% (in absolute value). Failed conditions are defined as those in which at least one of these four criteria is not met. For conditions where at least one method succeeds, the table provides the percentage of all conditions where unstandardized ULS outperforms FIML in each of the four criteria: α relative bias, β relative bias, α standard error relative bias, and β standard error relative bias. Notice that the entries in each section sum to 100%. For the 324 conditions investigated, FIML failed but unstandardized ULS succeeded in 2.5% of conditions. FIML = full information maximum likelihood; ULS = unweighted least squares; *SE* = standard error.

only 200 observations, and that exhibit low item skewness ($\leq$1.5). Models that involve three indicators per dimension are estimated with only 200 observations and have high item skewness that is not well estimated by either method. In all other conditions, the behavior of these estimators is similar, with CIFA–ULS marginally outperforming FIML in terms of parameter estimation and FIML marginally outperforming CIFA–ULS in terms of standard errors.

However, in terms of estimation speed, CIFA–ULS has a clear advantage over FIML, particularly for multidimensional models. On a 3-Ghz machine with 2 Gb of RAM memory, CIFA–ULS took, averaging across conditions involving one latent trait, 7 s to perform one replication, regardless of whether standardized or unstandardized parameters were used. At most, it took 41 s. In contrast, FIML took an average of 10 s and a maximum of 84 s. However, for three-dimensional models, FIML took an average of 167 s and a maximum of 1,800 s (i.e., 30 min). In contrast, computing time for CIFA–ULS is unaffected by the number of dimensions involved. The computational difference is important in applied settings where researchers typically fit a set of models, especially if the number of variables is large and as the number of latent dimensions increases. Future improvements in computer power will make FIML more attractive computationally, but we conjecture that the computational advantage of CIFA–ULS will remain.

An additional benefit of CIFA is that more complex models can be estimated with ease. Large models with correlated latent traits or covariates can be readily estimated with the standard software that implements CIFA. These more complex models can also be estimated with FIML, but computational difficulties have prevented the implementation of FIML for estimating general SEM with mixed measurement models until recent times. On the other hand, an obvious drawback of the sequential CIFA methods is that only one IRT model, namely the normal ogive version of Samejima's model, can be estimated.

Conclusions

Due to the computational burden involved in FIML estimation of IRT models, previous research offered a fragmentary view of the finite sample performance of this estimator. In particular, the behavior of FIML estimates in multidimensional models and the behavior of FIML standard errors had scarcely been investigated. In this study we have examined the performance of FIML parameter estimates and standard errors in 324 different conditions involving unidimensional as well as multidimensional models. Also, the performance of FIML has been pitted against that of CIFA–ULS, an estimator that is asymptotically inferior to but presents clear computational advantages over FIML.

One of the most relevant results for applied researchers refers to convergence rates. In applications, a model is useless if its estimation does not converge. All methods showed similar convergence trends, and their probability of converging is not high enough in dichotomous models with few indicators and small samples. Nonetheless, convergence was near 100% for all conditions in which sample size was larger than 500 observations.

Another relevant result for applied researchers is the existence of some conditions for which no method yields adequate results. Whenever possible, these conditions are to be avoided in applications. In particular, IRT estimation fails in estimating models involving (a) a small number of indicators per dimension, (b) binary items, (c) low item slopes (around $\beta = 0.74$ or, equivalently, $\lambda = .4$), (d) high item skewness ($\geq$1.5), and (e) small sample size (around 200 observations). The more of these factors are involved, the higher the likelihood that IRT estimation will fail to yield adequate parameter estimates and/or standard errors.

Another relevant result in applied research is the performance of FIML in relation to CIFA–ULS. FIML is the best election in harsh conditions. As a result, (a) when both estimators fail badly, CIFA–ULS fails more markedly than FIML, and (b) FIML is the best election in small sample sizes (200 observations). In all remaining conditions, the performance of these estimators is comparable, with CIFA–ULS yielding slightly more accurate parameter estimates and FIML yielding slightly more accurate standard errors. As a result, it can be

argued that in nonharsh conditions (e.g., when sample size is at least 500 observations), the least computationally expensive estimator (CIFA–ULS) is preferable.

Also, although only of theoretical interest, the results of this study clearly show that the asymptotic advantage of FIML over CIFA (smaller variability of parameter estimates) is not realized in finite samples. Indeed, the standard deviation of the estimates does not differ appreciably when FIML or CIFA is employed. Furthermore, when CIFA estimation is performed with standardized parameters, the variability of parameter estimates is considerably smaller than is that for FIML.

A final result concerns the behavior of standard errors. Parameter estimates and standard errors are biased in similar conditions. However, the amount of bias in estimating standard errors was much larger than that in estimating parameters for both FMIL and CIFA. The amount of bias increased with increasing skewness, but problems generally appeared only when the number of indicators per dimension was very low.

As with any other simulation study, our study was limited by the specification of the conditions employed. Thus, in multidimensional IRT models, only models with uncorrelated traits were considered. We conducted additional simulations to investigate the performance of the methods when the latent traits were correlated. That is, we replicated some of the conditions involving Type I and II items, seven indicators per dimension, and 500 observations. The effects of low and moderate (.2 and .6) correlations between the dimensions were investigated. For CIFA–ULS, the accuracies of slope parameters and standard errors were unaffected by the magnitude of the correlations among the latent traits. In contrast, FIML slope estimates and standard errors worsened as the magnitude of the correlations decreased. Furthermore, FIML computing time increased roughly by 125% when latent traits were correlated, whereas CIFA–ULS computing time was practically unchanged. The results obtained for CIFA–ULS are consistent with previous results (e.g., Flora & Curran, 2004).

Also, larger models could have been considered, but the present results clearly reveal that IRT estimation is more problematic as the number of indicators per dimension becomes smaller. Nevertheless, minimum sample size for IRT estimation in very large models (e.g., 100 variables) is an interesting topic for future research. Future research should also investigate the effects of performing FIML with standardized parameters instead of unstandardized parameters. The results obtained in this study with CIFA suggest that, in some rare applications, minimizing with respect to standardized parameters in FIML may lead to more accurate parameter estimates. Another topic for future research is the behavior of FIML when the model is misspecified. Again, prior studies (e.g., Flora & Curran, 2004; Maydeu-Olivares, 2006) suggest that the performance of CIFA parameter estimates and standard errors is somewhat robust to mild model misspecification, and we expect similar results for FIML.

Another limitation, by design, of the present study is that in all instances we conducted simulations using complete data to avoid missingness issues. No conclusions as to the performance of these methods in the presence of missing data can be drawn, and these results might not generalize to cases in which missing data are present. It is well established how to conduct FIML estimation in the presence of data missing at random, but we are unaware of any statistical theory for CIFA estimation when data are missing at random. Thus, in the presence of missing data, researchers using CIFA must resort to listwise or pairwise deletion of missing data, which may lead to inconsistent parameter estimates.

Finally, goodness-of-fit testing has not been considered in this study. Until recently, goodness-of-fit testing was not feasible due to the sparse data conditions encountered in IRT applications. Recent limited information statistics proposed by Maydeu-Olivares and Joe (2005, 2006) have effectively solved this problem, and now goodness-of-fit testing can reliably be performed for both FIML and CIFA (the same sample sizes are needed for accurate model estimation).[18]

The results of this study reveal that there is room for improving IRT estimation accuracy. Future research should investigate if new estimators are able to yield adequate results in the conditions identified in this study for which both FIML and CIFA–ULS fail. Two candidate estimators that should be pitted against FIML are the bivariate composite likelihood (BCL; Maydeu-Olivares & Joe, 2006; Zhao & Joe, 2005)[19] and Markov chain Monte Carlo methods (see Wirth & Edwards, 2007). The BCL estimator is slightly more computationally involved than CIFA but much less involved than FIML. Monte Carlo methods are generally much more involved computationally than FIML. The BCL estimator is especially attractive because it has good computational features and only slightly less asymptotical efficiency than does FIML. Moreover, it is capable of easily handling complex SEM models, with observed variables of mixed measurement types (i.e., continuous, categorical). Also, unlike CIFA, other IRT models, such as Bock's (1972), can be estimated with BCL.

[18] Goodness-of-fit statistics routinely printed by SEM programs implementing CIFA do not assess how well the model fits the data. Rather, they assess how well the model reproduces the sample statistics used in the third estimation stage (thresholds and tetrachoric/polychoric correlations). These goodness-of-fit measures can be misleading when the assumption of discretized multivariate normality of the data does not hold (Muthén, 1993). This assumption can be assessed by using triplets of dichotomous variables (Muthén & Hofacker, 1988) or pairs of polytomous variables. However, it is not clear what to conclude if the assumption is blatantly violated for some pairs of variables but not for others. Recently, Maydeu-Olivares (2006) has provided a test of discretized multivariate normality, and a direct test of model fit has been proposed by Maydeu-Olivares and Joe (2005, 2006).

[19] Jöreskog and Moustaki (2001) refer to this estimator as "underlying bivariate normal" estimator, which is an appropriate term within the context of the model they investigated. However, because the estimator is completely general, we feel the term *bivariate composite likelihood* is more appropriate.

However, this outlook for future work should not distract from the main findings of this study, namely, that FIML yields adequate parameter estimates and standard errors for IRT models with samples of size 500. With samples of this size, CIFA provides an attractive alternative that researchers may wish to consider, particularly if their application involves a complex model. Adequate parameter estimates can be obtained in some conditions with as few as 200 observations, in which case FIML is likely to behave better than CIFA.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11,* 111–141.

Baker, F. B. (1991). Comparison of minimum logit chi-square and Bayesian item parameter estimation. *British Journal of Mathematical and Statistical Psychology, 44,* 299–313.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis.* London: Arnold.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186–203.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Ed.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 32,* 771–775.

Bolt, D. M. (2005). Limited and full-information IRT estimation. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Mahwah, NJ: Erlbaum.

Boulet, J. R. (1996). *The effect of nonnormal ability distribution in IRT parameter estimation using full-information methods.* Unpublished doctoral dissertation, University of Ottawa, Ottawa, Ontario, Canada.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Mehods and Research, 29,* 468–508.

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

Davis, P. J., & Rabinovitz, P. (1975). *Methods of numerical integration.* New York: Academic Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

Dolan, C. V. (1994). 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, 47,* 309–326.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77–90.

Finger, M. S. (2001). *A comparison of full-information and unweighted least-squares limited-information methods used with the 2-parameter normal ogive model.* Unpublished doctoral dissertation, University of Minnesota, Twin Cities Campus.

Flora, D., & Curran, P. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267–269.

Gosz, J. K., & Walker, C. M. (2002, April). *An empirical comparison of multidimensional item response data using TESTFACT and NOHARM.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error.* (Tech. Rep. No.15, Office of Naval Research Contract No. 25140, NR-342–022). Stanford, CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Janssen, R., & De Boeck, P. (1999). Confirmatory analysis of componential test structures using multidimensional item response theory. *Multivariate Behavioral Research, 34,* 245–268.

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

Kaplan, D. (1991). The behaviour of three weighted least squares estimators for structured means analysis with non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 44,* 333–346.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457–477.

Lord, F. (1952). A theory of test scores. *Psychometrika Monograph* (Whole No. 7), *17.*

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

Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics, 26,* 51–71.

Maydeu-Olivares, A. (2005a). Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40,* 261–279.

Maydeu-Olivares, A. (2005b). Linear IRT, non-linear IRT, and factor analysis: A unified framework. In A. Maydeu-Olivares &

J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift to Roderick P. McDonald* (pp. 73–100). Mahwah, NJ: Erlbaum.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika, 71,* 57–77.

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

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

Maydeu-Olivares, A., Rodríguez-Fornells, A., Gómez-Benito, J., & D'Zurilla, T. J. (2000). Psychometric properties of the Spanish adaptation of the Social Problem-Solving Inventory—Revised (SPSI–R). *Personality and Individual Differences, 29,* 699–708.

McDonald, R. P. (1997). Normal ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 257–269). New York: Springer.

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

McDonald, R. P., & Mok, M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 54,* 483–495.

Meng, X. L., & Schilling, S. (1996). Fitting full-information factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association, 91,* 1254–1267.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3–31.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551–560.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22,* 43–65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript.

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika, 53,* 563–578.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45,* 19–30.

Muthén, B., & Muthén, B. (2004). MPLUS (Version 3.13). Los Angeles: Author.

Muthén, B., & Muthén, B. (2006). MPLUS (Version 4.1). Los Angeles: Author.

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

Parry, C. D. H., & McArdle, J. J. (1991). An applied comparison of methods for least squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15,* 35–46.

Potthast, M. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology, 46,* 273–286.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). *GLLAMM manual* (Tech. Rep. No. 2001801). London: Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational and Behavioral Statistics, 27,* 133–144.

Reiser, M., & VanderBerg, M. (1994). Validity of chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology, 47,* 85–107.

Rigdon, E. E., & Ferguson, C. E., (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, 28,* 491–497.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34*(Suppl. 4).

Schilling, S. G., & Bock, R. D. (2005). High dimensional ML item factor analysis by adaptive quadrature. *Psychometrika, 70,* 533–555.

Schumaker, R. E., & Beyerlein, S. T. (2000). Confirmatory factor analysis with different correlation types and estimation methods. *Structural Equation Modeling, 7,* 629–636.

Stone, C. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16,* 1–16.

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

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27,* 159–203.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interaction on the estimated discrimination parameters in item response theory. *Psychological Methods, 6,* 181–195.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

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

Zhao, Y., & Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics, 33,* 335–356.

Appendix

Technical Details

Unstandardized Versus Standardized Parameters

IRT models express the probability of each of the m^n possible response patterns as a function of p latent traits via the equation

$$\Pr(y_1 = k_1, \ldots, y_n = k_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n} \mathrm{P}(y_i = k_i|\boldsymbol{\eta})\varphi_p(\boldsymbol{\eta})d\boldsymbol{\eta}. \quad \text{(A.1)}$$

Specific IRT models are obtained by choosing a specific expression for (a) the conditional probabilities, $\mathrm{P}(y_i = k_i|\boldsymbol{\eta})$, and (b) the density of the latent traits, $\varphi_p(\boldsymbol{\eta})$. In this paper we have assumed that $\varphi_p(\boldsymbol{\eta}) = \phi_n(\boldsymbol{\eta}; \mathbf{0}, \boldsymbol{\Psi})$, (i.e., that $\boldsymbol{\eta}$ is normally distributed with mean zero and correlation matrix $\boldsymbol{\Psi}$). Also, we have considered two variants (logistic and normal ogive) of an IRT model, Samejima's GRM (see Equation 6).

The normal ogive GRM can alternatively be derived from a factor analytic perspective. As in Equation (7), assume that $\boldsymbol{y}^* = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$, where

$$\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega} \end{pmatrix}\right), \quad \text{(A.2)}$$

and $\boldsymbol{\Omega}$ is a diagonal matrix. The latent response variables $\boldsymbol{y}^*$ are related to the observed categorical responses $\boldsymbol{y}$ via the threshold relationship of Equation 8. From Equations 7, 8, and A.3, it follows that

$$\Pr(y_1 = k_1, \cdots, y_n = k_n) = \int_{\mathbf{R}} \cdots \int \phi_n\left(\boldsymbol{y}^*; \boldsymbol{\mu}_{y^*}, \sum_{y^*}\right)d\boldsymbol{y}^*, \quad \text{(A.3)}$$

where $\phi_n(\bullet)$ denotes an n-variate normal density with mean $\boldsymbol{\mu}_{y^*} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{y^*} = \boldsymbol{B}\boldsymbol{\Psi}\boldsymbol{B}' + \Omega$, and $\mathbf{R}$ is the n-dimensional rectangular region formed by the product of intervals

$$R_i = \begin{cases} (\alpha_{i,m-1}, \infty) & \text{if } y_i = m - 1 \\ \vdots & \\ (\alpha_{i,1}, \alpha_{i,2}) & \text{if } y_i = 1 \\ (-\infty, \alpha_{i,1}) & \text{if } y_i = 0 \end{cases}. \quad \text{(A.4)}$$

Because $\boldsymbol{y}^*$ is not observed, its variance is not identified, which in turn implies that the variances of the random errors ε are not identified. The easiest way to achieve identification is to set the error variances equal to 1 (i.e., $\boldsymbol{\Omega} = \mathbf{I}$. In this case, the ordinal factor analysis parameters and the normal ogive version of the GRM—expressed as a special case of Equation A.1—are equal. These equations are simply two alternative ways of describing the same model, as a function of a set of latent traits (Equation A.1) or a function of a set of underlying response variables (Equation A.4). See Takane and de Leeuw (1987) for further details.

Now, consider standardizing $\boldsymbol{y}^*$ using

$$\mathbf{z}^* = \boldsymbol{\Delta}(\boldsymbol{y}^* - \boldsymbol{\mu}_{y^*}), \qquad \boldsymbol{\Delta} = \left(\mathrm{Diag}\left(\sum_{y^*}\right)\right)^{\frac{1}{2}}, \quad \text{(A.5)}$$

so that the standardized latent response variables have mean $\boldsymbol{\mu}_{z^*} = 0$ and correlation structure

$$\boldsymbol{P}_{z^*} = \boldsymbol{\Delta}\sum_{y^*}\boldsymbol{\Delta} = \boldsymbol{\Delta}(\mathbf{B}\boldsymbol{\Psi}\mathbf{B}' + \boldsymbol{\Omega})\boldsymbol{\Delta}. \quad \text{(A.6)}$$

These are polychoric correlations, as $\boldsymbol{z}^*$ is a vector of categorized normal variables.

The probabilities (Equation A.4) are unchanged when standardizing $\boldsymbol{y}^*$, in which case we write

$$\Pr(y_1 = k_1, \ldots, y_n = k_n) = \int_{\breve{\mathbf{R}}} \cdots \int \phi_n(\mathbf{z}^*; \mathbf{0}, \mathbf{P}_{z^*})d\mathbf{z}^* \quad \text{(A.7)}$$

where now $\breve{\mathbf{R}}$ is a rectangular region formed by the product of intervals

(*Appendix continues*)

$$\check{R}_i = \begin{cases} (\tau_{i,m-1}, \infty) & \text{if } y_i = m-1 \\ \vdots & \\ (\tau_{i,1}, \tau_{i,2}) & \text{if } y_i = 1 \\ (-\infty, \tau_{i,1}) & \text{if } y_i = 0 \end{cases}. \quad \text{(A.8)}$$

To see this, let μ_i^* be an element of $\boldsymbol{\mu}_{y^*}$ and let $\sigma_i^{2^*}$ be a diagonal element of $\boldsymbol{\Sigma}_{y^*}$. Then, at the threshold $y_i^* = \alpha_{i,k}$, z_i^* takes the value $\tau_{i,k} = \frac{\alpha_{i,k} - \mu_i^*}{\sqrt{\sigma_i^{2^*}}} = \delta_i \alpha_{i,k}$ where δ_i is a diagonal element of $\boldsymbol{\Delta}$. In matrix form,

$$\boldsymbol{\tau}_k = \boldsymbol{\Delta}\alpha_k. \quad \text{(A.9)}$$

The correlation structure (Equation A.6) can be reparameterized using

$$\boldsymbol{\Lambda} = \boldsymbol{\Delta}\boldsymbol{B}, \quad \text{(A.10)}$$

$$\boldsymbol{\Theta} = \boldsymbol{\Delta}\boldsymbol{\Omega}\boldsymbol{\Delta} = \boldsymbol{\Theta}^{1/2}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1/2}\boldsymbol{\Theta}^{1/2}. \quad \text{(A.11)}$$

Equations A.10 and A.11 imply that

$$\boldsymbol{P}_{z^*} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}. \quad \text{(A.12)}$$

In summary, two parameterizations can be used for the normal ogive version of Samejima's model. One parameterization uses the parameters $\boldsymbol{\alpha}$, $\boldsymbol{B}$, and $\boldsymbol{\Omega}$; the other parameterization uses the parameters $\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Theta}$. The latter arise from standardization of the former.

The diagonal elements of $\boldsymbol{\Theta}$ in the standardized parameterization are not free parameters, as it can be verified that when $\boldsymbol{\Omega}$ is a diagonal matrix,

$$\boldsymbol{\Theta} = \boldsymbol{\Delta}\boldsymbol{\Omega}\boldsymbol{\Delta} = \mathrm{I} - \mathrm{Diag}(\boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'). \quad \text{(A.13)}$$

Also, from Equation A.11,

$$\boldsymbol{\Delta}^{-1} = \boldsymbol{\Omega}^{1/2}\boldsymbol{\Theta}^{-1/2}, \quad \text{(A.14)}$$

and the inverse relationship between both parameterizations is

$$\boldsymbol{\alpha}_k = \boldsymbol{\Delta}^{-1}\boldsymbol{\tau}_k = \boldsymbol{\Theta}^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\tau}_k, \; \boldsymbol{B} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Theta}^{-1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\Lambda}. \quad \text{(A.15)}$$

The scalar counterparts of Equations A.9 and A.10 is

$$\tau_{i,k} = \frac{\alpha_{i,k}}{\sqrt{\omega_i + \boldsymbol{\beta}_i'\boldsymbol{\Psi}\boldsymbol{\beta}_i}}, \; \lambda_i = \frac{\beta_i}{\sqrt{\omega_i + \boldsymbol{\beta}_i'\boldsymbol{\Psi}\boldsymbol{\beta}_i}}, \quad \text{(A.16)}$$

where ω_i is the ith diagonal element of $\boldsymbol{\Omega}$. Equation 9 gives the special case of Equation A.16 when the model is identified using $\boldsymbol{\Omega} = \mathbf{I}$, except for a sign change for the thresholds.[20] Also, using Equation A.13, the scalar counterpart of Equation A.15 is

$$\alpha_{i,k} = \frac{\sqrt{\omega_i} \times \tau_{i,k}}{\sqrt{1 - \boldsymbol{\lambda}_i'\boldsymbol{\Psi}\boldsymbol{\lambda}_i}}, \; \beta_i = \frac{\sqrt{\omega_i} \times \boldsymbol{\lambda}_i}{\sqrt{1 - \boldsymbol{\lambda}_i'\boldsymbol{\Psi}\boldsymbol{\lambda}_i}}. \quad \text{(A.17)}$$

For $\boldsymbol{\Omega}$ symmetric (not necessarily diagonal), Equation A.14 does not hold. And neither does Equation A.13, although it is true that the diagonal elements of $\boldsymbol{\Theta}$ are still of the form $1 - \boldsymbol{\lambda}_i'\boldsymbol{\Psi}\boldsymbol{\lambda}_i$.

Transformation From Normal Ogive Scale to Logistic Scale

Unstandardized parameter estimates obtained using a logistic link function can be put in the metric of a normal ogive link by multiplication by the scaling constant $D = 1.702$. Conversely, unstandardized parameter estimates obtained using a normal link function can be put in a logistic metric by division by the scaling constant D. In the formulas above, unstandardized parameters are in a normal ogive metric when $\boldsymbol{\Omega} = \mathbf{I}$. They are approximately on a logistic matrix when $\boldsymbol{\Omega} = D^2\mathbf{I}$. In this study, when CIFA estimation was performed by minimizing use of unstandardized parameters, we used $\boldsymbol{\Omega} = D^2\mathbf{I}$. As a result, unstandardized parameter estimates and standard errors were on a logistic metric. When CIFA estimation was performed with respect to standardized parameters, these were transformed to unstandardized parameters on a logistic metric using Equation A.17. Standard errors were obtained using the delta method (see Agresti, 1990). Consult the supplementary materials to this article for further details on implementation.

Correlation Structure Implied by Standardized and Unstandardized Parameters

Consider a test consisting of n items that is assumed to depend on a single dimension. Setting for identification the variance of the latent trait to 1 and $\boldsymbol{\Omega} = \mathbf{I}$, the correlation structure implied by the model is

$$\mathbf{P}_{z^*} = \begin{pmatrix} 1 & & & \\ \lambda_2\lambda_1 & 1 & & \\ \vdots & \ddots & 1 & \\ \lambda_n\lambda_1 & \cdots & \lambda_n\lambda_{n-1} & 1 \end{pmatrix}, \quad \text{(A.18)}$$

[20] As shown in this Appendix and as implemented in Mplus, unstandardized and standardized thresholds are of the same sign. However, in the literature unstandardized and standardized thresholds are always depicted as being of opposite signs, and we have used this convention in the body of the paper (see Equation 9). The rationale for using different signs for both thresholds is presented, for instance, in Takane and de Leeuw (1987).

when standardized parameters are used (see Equation A.12) and

$$\mathbf{P}_{z^*} = \begin{pmatrix} 1 & & \\ \frac{\beta_2\beta_1}{\sqrt{\beta_2^2+1}\sqrt{\beta_1^2+1}} & 1 & \\ \vdots & \ddots & 1 \\ \frac{\beta_n\beta_1}{\sqrt{\beta_n^2+1}\sqrt{\beta_1^2+1}} & \cdots & \frac{\beta_n\beta_{n-1}}{\sqrt{\beta_n^2+1}\sqrt{\beta_{n-1}^2+1}} \quad 1 \end{pmatrix}, \tag{A.19}$$

when unstandardized parameters are used (see Equation A.6). We attribute the poorer performance of unstandardized ULS to the complex nonlinear constraints (Equation A.19) imposed by the unstandardized parameterization on the threshold and correlation structure.

Received May 21, 2007
Revision received December 16, 2008
Accepted January 14, 2009 ■