Structural Equation Modeling, 18:333–356, 2011

ISSN: 1070-5511 print/1532-8007 online
DOI: 10.1080/10705511.2011.581993

Psychology Press
Taylor & Francis Group

Comparing the Fit of Item Response Theory and Factor Analysis Models

Alberto Maydeu-Olivares
Faculty of Psychology, University of Barcelona

Li Cai
Graduate School of Education and Information Studies, UCLA

Adolfo Hernández
Department of Statistics and Operations II, Universidad Complutense

Linear factor analysis (FA) models can be reliably tested using test statistics based on residual covariances. We show that the same statistics can be used to reliably test the fit of item response theory (IRT) models for ordinal data (under some conditions). Hence, the fit of an FA model and of an IRT model to the same data set can now be compared. When applied to a binary data set, our experience suggests that IRT and FA models yield similar fits. However, when the data are polytomous ordinal, IRT models yield a better fit because they involve a higher number of parameters. But when fit is assessed using the root mean square error of approximation (RMSEA), similar fits are obtained again. We explain why. These test statistics have little power to distinguish between FA and IRT models; they are unable to detect that linear FA is misspecified when applied to ordinal data generated under an IRT model.

Keywords: approximate fit, attitudes, categorical data analysis, goodness-of-fit, ordinal factor analysis, personality, Samejima's graded model, structural equation modeling

The common factor model is used to relate linearly a set of observed variables to a smaller set of unobserved continuous variables, the common factors or latent traits. When the model holds, the latent traits are then used to explain the dependencies between the variables observed. However, the factor model is most often applied to response variables that are discrete, such as binary variables, or ratings scored using successive integers. Indeed, when applied to discrete data we know a priori that the factor model is always misspecified to some extent, because the predicted responses under this model can never be exact integers (McDonald, 1999).

Models do not have to be correct to be useful and over the years the factor model has repeatedly been used to model ratings yielding psychologically meaningful results (Waller,

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

Tellegen, McDonald, & Likken, 1996). However, it is preferable to use models that might be correct in the population rather than a model that is known up front to be false. As a result, when modeling rating data, in principle practitioners should use latent trait models that take into account the discrete ordinal nature of rating responses, rather than the factor analysis (FA) model. Latent trait models suitable for rating data, known as item response theory (IRT) models, have existed for 30 years now. A special IRT model for rating data is the ordinal factor analysis model,[1] which relates a FA model to the observed ordinal data via a threshold relationship and estimation proceeds through the use of tetrachoric or polychoric correlations. For an overview of IRT models and methods, see Hulin, Drasgow, and Parsons (1983), Hambleton and Swaminathan (1985), Hambleton, Swaminathan, and Rogers (1991), van der Linden and Hambleton (1997), and Embretson and Reise (2000). For the relationship between IRT and ordinal FA see, for instance, Bartholomew and Knott (1999) or Bolt (2005).

Yet, if IRT models can be correct when applied to discrete ordinal data, whereas the FA model cannot, IRT models should fit the ordinal data better than FA models in applications. Also, if they yield a different fit, they might lead to different substantive conclusions. Do they? Do we need to rewrite some of our theories developed using the common factor model applied to ordinal data because a better fitting IRT model suggests a different substantive theory? These are undoubtedly interesting questions. However, to answer them we need a test statistic that can be used to pit the fit of an FA model to the fit of an IRT model to the same data set. In other words, we need a common measure of fit that can be used with both types of models.

In the next subsections we discuss how such fit comparison can be performed, but before we do that we first review how goodness-of-fit is assessed for IRT models and then for FA models. We conclude the introduction with a discussion of which IRT model should be compared in terms of data fit to the factor model, as there are many IRT models that could be used.

GOODNESS OF FIT ASSESSMENT IN IRT

In IRT one models the probabilities of all the response patterns that can be obtained given the variables being modeled. The standard test statistics for discrete data, Pearson's χ^2 statistic and the likelihood ratio statistic G^2, can be used to test the null hypothesis that the IRT model for the pattern probabilities is correctly specified. When the model holds, the distribution of both statistics is well approximated by a chi-square distribution in large samples. Yet, at least since Cochran (1952) it has been known that when some pattern probabilities are small the chi-square approximation to the sampling distribution of χ^2 and G^2 yields incorrect p values. Furthermore, if the number of possible response patterns is large, then the probabilities for some patterns must necessarily be small as the sum of all pattern probabilities must be equal to one (Bartholomew & Tzamourani, 1999). In other words, if the number of response patterns is large, the chi-square approximation to χ^2 and G^2 cannot be used to test the overall fit of the model, regardless of sample size. This means that unless the questionnaire being modeled consists of only a few questions, to be rated using just a few categories, the χ^2 and G^2 statistics are useless. As Thissen and Steinberg (1997) put it, "when the number of categories is five or

[1]LISREL (Jöreskog & Sörbom, 2008), M*plus* (L. Muthén & Muthén, 1998–2010), or EQS (Bentler, 2004) estimate this model when a factor model is fitted to variables declared to be categorical.

more, the approximation becomes invalid for any model as soon as the number of variables is greater than six with any conceivable sample size" (p. 61). Of course, having a small sample size only makes matters worse. In contrast, there are a number of reliable procedures for goodness-of-fit assessment of the factor model.

GOODNESS-OF-FIT ASSESSMENT FOR FACTOR MODELS

When the data are normally distributed, the goodness of fit of a factor model can be assessed by using the minimum of the discrepancy function used for estimation (rescaled by sample size) for many discrepancy functions. This is the usual chi-square test used in structural equation modeling (SEM). However, when the observed data are responses to questionnaire items, the normal distribution is likely to provide a poor approximation to the distribution of the item scores, and the p values obtained by the usual chi-square test might be unreliable.

Two methods have been proposed to obtain p values that are robust to nonnormality. One method consists of rescaling the usual chi-square test statistic as suggested by Satorra and Bentler (1994). The rescaled test statistics have been found to yield accurate p values when the data are not normally distributed. Another method is to use the test statistic based on residual covariances proposed by Browne (1984) under asymptotically distribution-free assumptions. In principle, this approach is preferable to the Satorra–Bentler approach as it yields a statistic that is asymptotically distributed as a chi-square. However, previous research suggests that Browne's statistic, T_B, overrejects the model unless the sample size is very large and the number of variables being modeled is small (e.g., Curran, West, & Finch, 1996; Muthén & Kaplan, 1992). More recently, Yuan and Bentler (1997, 1999) proposed two alternative test statistics based on residual covariances that might yield better results in applications. The first of these statistics, T_{YB}, is a correction of T_B and follows the same asymptotic distribution. The second statistic, T_F, has an asymptotic F distribution.

HOW TO COMPARE THE FIT OF FA AND IRT MODELS

The fit of FA and IRT models can only be compared using univariate and bivariate information (e.g., covariances) because under the usual factor model assumptions (see next section) only univariate and bivariate moments are specified. In other words, unless additional assumptions are made, one cannot compute higher moments (trivariate, four-way moments, etc.) using the common factor model. Therefore one cannot compute the probability of a response pattern in this model as one does in an IRT model.[2]

However, Maydeu-Olivares (2005b) pointed out that under some conditions the test statistic based on residual covariances proposed by Browne (1982, 1984) for the FA model could be used to assess the goodness-of-fit of IRT models. In other words, these test statistics provide a common measure of goodness-of-fit for FA and IRT models. The purpose of this article is to explore this suggestion in detail. By means of a simulation study, we investigate whether

[2]A linear model with additional assumptions that enables the computation of pattern probabilities was introduced in Maydeu-Olivares (2005b), who referred to it as a linear IRT model to distinguish it from the common factor model.

these test statistics based on residual covariances yield accurate p values when used to assess the goodness-of-fit of IRT models. If so, unlike the χ^2 and G^2 statistics generally used in IRT, they can be reliably used to assess the model–data misfit of these models. Then, we use these statistics to gauge the fit of the IRT models against that of a factor model.

However, before proceeding we must address the issue of which IRT model should be compared with the factor model. Currently, if applied researchers want to replace the common FA model by an IRT model it is not clear which IRT model they should use.

CHOICE OF IRT MODEL AND MODEL–DATA FIT ISSUES

For rating data it has been argued (Reise & Waller, 1990) that there is no guessing or any similar phenomenon that requires lower asymptote parameters in IRT models. Chernyshenko, Stark, Chan, Drasgow, and Williams (2001) provided empirical support to this argument. Even restricting ourselves to IRT models without guessing parameters, the list of possible IRT models that can be used to model rating responses is extensive (see van der Linden & Hambleton, 1997). In a recent study, Maydeu-Olivares (2005a) heuristically compared the fit of some well-known parametric IRT models to five personality scales. The models considered were Samejima's (1969) graded model, Bock's (1972) nominal model, Masters's (1982) partial credit model, and Thissen and Steinberg's (1986) ordinal model.[3] The author concluded that Samejima's model provided the best fit to each of the five scales considered. Consequently, in this article we focus on Samejima's IRT model, which is formally equivalent to the ordinal FA model (see Takane & de Leeuw, 1987) implemented in some popular SEM programs.

The remainder of this article is organized as follows. First, we briefly describe the factor model. Because IRT modeling focused on unidimensional models, this article often focuses on unidimensional models. Next, we briefly describe Samejima's IRT model. Then we describe the test statistics based on residual covariances under consideration. The following section addresses via a simulation study the question of whether these statistics yield accurate p values when applied to IRT models. Next, we gauge the performance of the IRT model against a factor model to fit the items of two well-known questionnaires. The last section explores the power of the statistics to distinguish between these models. The final section provides a discussion of the main findings and concluding remarks.

THE FACTOR ANALYSIS MODEL

Consider a test composed of p rating items Y_i, $i = 1, \ldots p$, each consisting of m categories. We assume these categories have been scored using successive integers: $k = 0, \ldots, m-1$. A one-factor model for these variables is

$$\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\eta + \varepsilon. \tag{1}$$

In this equation, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of intercepts, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of factor loadings and η denotes the latent variable (the common factor). Also, ε is a $p \times 1$ vector of measurement

[3]This model is formally equivalent to Muraki's (1992) generalized partial credit model.

errors. A factor model assumes that (a) the mean of the latent trait is zero, (b) the variance of the latent trait is one, (c) the mean of the random errors is zero, (d) the latent traits and random errors are uncorrelated, and (e) the random errors are mutually uncorrelated, so that their covariance matrix, $\mathbf{\Omega}$, is a diagonal matrix with elements ω_i.

The assumptions of this model imply that the variance of an item is of the form

$$\sigma_i^2 = \beta_i^2 + \omega_i, \tag{2}$$

whereas the covariance between two items is of the form

$$\sigma_{ij} = \beta_i \beta_j. \tag{3}$$

Letting $\boldsymbol{\theta} = (\beta_i, \cdots, \beta_p, \omega_i, \cdots, \omega_p)'$ be the parameters of a one-factor model involved in Equations 2 and 3, and letting $\boldsymbol{\sigma}$ be the t-dimensional vector of variances and nonduplicated covariances, $t = \frac{p(p+1)}{2}$, these restrictions can be compactly written as $\boldsymbol{\sigma}(\boldsymbol{\theta})$.

Estimation and Testing

Consider a sample of size N from the population of interest, and let $\mathbf{s}$ be the sample counterpart of $\boldsymbol{\sigma}$ (i.e., the sample variances and nonduplicated sample covariances). The factor model can be estimated by minimizing one of several discrepancy functions between $\mathbf{s}$ and $\boldsymbol{\sigma}$ (see Browne & Arminger, 1995). Here, we shall use the maximum likelihood discrepancy function for normally distributed data. When the data are not normally distributed it yields consistent estimates but standard errors and goodness-of-fit tests are in general invalid (e.g., Satorra & Bentler, 1994).

For testing the model when the observed data are not multivariate normal, Browne (1982, 1984) proposed a test statistic, T_B, based on the residual covariances $\mathbf{s} - \boldsymbol{\sigma}$. This statistic can be written as

$$T_B = N(\mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}}))'\hat{\mathbf{U}}(\mathbf{s} - \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})), \qquad \mathbf{U} = \boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\boldsymbol{\Gamma}^{-1}. \tag{4}$$

Here, $\boldsymbol{\Gamma}$ denotes the asymptotic covariance matrix of $\sqrt{N}\mathbf{s}$, and $\boldsymbol{\Delta}$ is a matrix of derivatives, $\boldsymbol{\Delta} = \frac{\partial \boldsymbol{\sigma}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$. To compute $\mathbf{U}$ we evaluate $\boldsymbol{\Delta}$ at the parameter estimates and we use sample moments (up to the fourth order) to estimate $\boldsymbol{\Gamma}$ (see the Appendix for details). Let q be the number of estimated parameters and let $t = \frac{p(p+1)}{2}$ be the number of population covariances in $\boldsymbol{\sigma}$. When the model is identified, T_B is asymptotically distributed as a chi-square variable with degrees of freedom $t - q$. Unfortunately, it has repeatedly been found in simulation studies that T_B overrejects the model unless the model is small and the sample size is large. More recently, Yuan and Bentler (1997) proposed using a combination of sample and expected moments to estimate $\boldsymbol{\Gamma}$ (see Appendix), which leads to a modification of T_B (with its same asymptotic distribution) that has a better small sample performance than T_B. This is

$$T_{YB} = \frac{N \times T_B}{N + T_B} = \frac{T_B}{1 + T_B/N}. \tag{5}$$

Even more recently Yuan and Bentler (1999) proposed another modification of T_B. This is

$$T_F = \frac{N - df}{N \times df} T_B, \tag{6}$$

where $df = t - q$, the degrees of freedom of the model. Unlike T_B and T_{YB}, T_F is asymptotically distributed as an F distribution with $(df, N - df)$ degrees of freedom.

Some Remarks

In the FA model, as we have described it here, no assumptions are made on the distribution of the latent trait η nor the random errors **e**. As a result, no assumptions are made on the distribution of the observed variables **y**. Only moment assumptions are made. Under these assumptions, the factor model implies the structure given in Equations 2 and 3 for the population variances and covariances. In FA applications only the bivariate associations present in the data are modeled: either the central moments (variances and covariances) or the standardized central moments (correlations). In contrast, in IRT, the categorical nature of the item responses is taken into account. In IRT applications, the probability of observing each response pattern is modeled. With p items each consisting of m categories there are m^p possible response patterns. We now consider how each of these response patterns is modeled using Samejima's (1969) graded response model.

SAMEJIMA'S GRADED MODEL

Again, consider a test composed of p rating items Y_i, $i = 1, \dots, p$, each consisting of m categories scored using successive integers $k = 0, \dots, m - 1$. For example, consider a three-item test where each item consists of five options. One possible response pattern is $(Y_1 = 1) \cap (Y_2 = 3) \cap (Y_3 = 2)$. In general, a response pattern for a p-item test can be written as $\bigcap_{i=1}^{p}(Y_i = k_i)$. IRT models are a class of models for multivariate categorical data. In these models, the probability of any of the possible response patterns can be written as

$$\Pr\left[\bigcap_{i=1}^{p}(Y_i = k_i)\right] = \int_{-\infty}^{\infty} \prod_{i=1}^{p} \Pr(Y_i = k_i|\eta) f(\eta) d\eta. \tag{7}$$

This equation holds for *any* IRT model that assumes a single underlying latent trait (Bartholomew & Knott, 1999). In Equation 7, $f(\eta)$ denotes the distribution (density) of the latent trait. Here, we assume throughout that the latent trait is normally distributed, so that $f(\eta)$ is a standard normal density function. Also, $\Pr(Y_i = k_i|\eta)$ is the conditional probability that item i takes the value k_i given the latent trait, the option response function.[4]

A popular member of the IRT class of models suitable for modeling binary responses is the two-parameter logistic model (Lord & Novick, 1968). In this model, the conditional probability of endorsing an item given the latent trait is given by

$$\Pr(Y_i = 1|\eta) = \Psi(\alpha_{i,1} + \beta_i \eta) = \frac{1}{1 + \exp[-(\alpha_{i,1} + \beta_i \eta)]}, \tag{8}$$

The conditional probability of not endorsing the item is $\Pr(Y_i = 0|\eta) = 1 - \Psi(\alpha_{i,1} + \beta_i \eta)$.

[4]This function is also called category response function. When the items are binary, the conditional probability of endorsing an item, $\Pr(Y_i = 1|\eta)$, is called item response function.

Samejima's graded response model is an extension of the two-parameter logistic model suitable for modeling ordered responses like those obtained from rating items. More specifically, a two-parameter logistic model is used to model the probability of endorsing categories 1 or higher, 2 or higher, and so on. As a result, in this model the option response function is specified as

$$\Pr(Y_i = k|\eta) = \begin{cases} 1 - \Psi(\alpha_{i,1} + \beta_i \eta) & \text{if } k = 0 \\ \Psi(\alpha_{i,k} + \beta_i \eta) - \Psi(\alpha_{i,k+1} + \beta_i \eta) & \text{if } 0 < k < m - 1, \\ \Psi(\alpha_{i,m-1} + \beta_i \eta) & \text{if } k = m - 1 \end{cases} \tag{9}$$

where $\Psi(\alpha_{i,k} + \beta_i \eta)$ equals the standard logistic distribution function evaluated at $\alpha_{i,k} + \beta_i \eta$,

$$\Psi_{i,k} = \Psi(\alpha_{i,k} + \beta_i \eta) = \frac{1}{1 + \exp[-(\alpha_{i,k} + \beta_i \eta)]}. \tag{10}$$

Note that this model reduces to the two-parameter logistic model when the number of categories per item, *m,* is two.

In the IRT literature, Equation 10 is often written as

$$\Psi_{i,k} = \frac{1}{1 + \exp[-a_i(\eta - b_{i,k})]}. \tag{11}$$

This is, for instance, the parameterization used in MULTILOG (Thissen, Chen, & Bock, 2003). The two parameterizations are equivalent, with the equivalence being $\alpha_{i,k} = -a_i b_{i,k}$ and $\beta_i = a_i$.

Note that in Samejima's model m parameters are estimated for each item ($m - 1$ intercepts $\alpha_{i,h}$ and 1 slope β_i). Thus, the total number of parameters to be estimated in a p item test is $q = pm$.

Estimation and Testing

In IRT modeling the observed ratings are treated as multinomial variables. IRT models are generally estimated by maximizing the multinomial likelihood function as described in Bock and Aitkin (1981). This is the estimation procedure used in this article.

As for the goodness-of-fit assessment of these models, in this article we investigate the feasibility of the suggestion given in Maydeu-Olivares (2005b) to use Browne's T_B statistic, and we also empirically compare the performance of this statistic against the T_{YB} and T_F statistics. When these statistics are used, the null hypothesis is $H_0 : \boldsymbol{\sigma} = \boldsymbol{\sigma}(\boldsymbol{\theta})$ for some parameter vector $\boldsymbol{\theta}$ to be estimated from the data, against the alternative $H_1 : \boldsymbol{\sigma} \neq \boldsymbol{\sigma}(\boldsymbol{\theta})$. In Samejima's graded logistic model $\boldsymbol{\theta} = (\alpha_{1,1}, \cdots, \alpha_{1,m-1}, \cdots, \alpha_{p,1}, \cdots, \alpha_{p,m-1}, \beta_1, \cdots, \beta_p)'$.

Now, to compute these goodness-of-fit statistics, the population variances and covariances implied by Samejima's graded logistic model are needed. We show in the Appendix that, under

Samejima's model, the variance of the scores for a single item is

$$\sigma_i^2 = Var[Y_i] = \int_{-\infty}^{+\infty} \left\{ \sum_{k=1}^{m-1} (2k-1)\Psi_{i,k} \right\} f(\eta) d\eta - \mu_i^2, \tag{12}$$

whereas the covariance between the scores for two items is

$$\sigma_{ij} = Cov[Y_i Y_j] = \left(\int_{-\infty}^{+\infty} \left[\sum_{l=1}^{m-1} \Psi_{j,l} \right] \left[\sum_{k=1}^{m-1} \Psi_{i,k} \right] f(\eta) d\eta \right) - \mu_i \mu_j, \tag{13}$$

where

$$\mu_i = E[Y_i] = \int_{-\infty}^{+\infty} \left[\sum_{k=1}^{m-1} \Psi_{i,k} \right] f(\eta) d\eta \tag{14}$$

is the mean of the scores for an item under the model. In the Appendix we also provide details for computing the T_B, T_{YB}, and T_F statistics under Samejima's logistic model.

Now, to use the T_B, T_{YB}, and T_F statistics to test the goodness-of-fit of tests of an IRT model, the model must be identified from the covariance matrix.[5] In addition, the number of degrees of freedom needs to be greater than zero. For Samejima's model, these conditions are satisfied provided that the number of items is equal or larger than twice the number of categories per item. That is, these statistics can be used to test Samejima's model provided that $p \geq 2m$. They cannot be used to test the goodness-of-fit of tests composed of only a few items and a large number of categories per item. For instance, these statistics cannot be used to assess the goodness-of-fit of Samejima's model to the Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). This test consists of $p = 5$ items each with $m = 7$ categories. With this number of items there are only $t = \frac{p(p+1)}{2} = 15$ distinct variances and covariances, whereas the number of estimated parameters is $q = pm = 35$. The degrees of freedom are negative. Only the fit of more constrained IRT models (with fewer than 15 parameters) can be tested using covariances. An alternative is to collapse categories prior to the analysis to reduce the number of estimated parameters.

Assessing the Error of Approximation

So far, our presentation has focused on testing whether a factor model or an IRT model hold exactly in the population of interest. However, no model can be expected to hold exactly in applications, but only approximately. Browne and Cudeck (1993) introduced a framework for assessing the error of approximation of a model. More specifically, they proposed using Steiger and Lind's (1980) root mean square error of approximation (RMSEA) to assess how well a

[5]When an IRT model is identified from the covariance matrix then $\boldsymbol{\Delta}$ is of full rank. Also, the model can be estimated using only the information contained in the sample covariance matrix.

model reproduces the observed covariances relative to its degrees of freedom. The general expression for the RMSEA is

$$\mathrm{RMSEA} = \sqrt{\frac{T - df}{N \times df}}, \tag{15}$$

where T denotes in our case any test statistic based on residual covariances. When T is asymptotically chi-square distributed (as the T_B and T_{YB} statistics discussed here), Browne and Cudeck showed that the RMSEA is asymptotically distributed as a noncentral chi-square under a sequence of local alternatives assumptions (for details, see Browne & Cudeck, 1993). Thus, we can use an RMSEA computed on T_B or T_{YB} and its confidence interval to gauge the error of approximation of a factor model and also of an IRT model (under the same conditions specified previously) to a given data set.

Can the p values obtained using the T_B, T_{YB}, and T_F statistics be trusted when applied to test Samejima's model (and its special case, the two-parameter logistic model)? To address this question, the next section reports the results of a simulation study. Also, to benchmark the performance of the statistics in assessing IRT models, we also report the results of a simulation study to assess the fit of the common factor model.

SIMULATION STUDY FOR CORRECTLY SPECIFIED MODELS

Factor Analysis Model

Four sample size conditions and two model size conditions were investigated. A total of 1,000 replications were used in each cell of the 4×2 factorial design. The sample sizes were 500, 1,000, 2,000, and 5,000 observations. The two different model sizes were 10 and 20 variables. The sample sizes were chosen to be medium to very large for typical applications, whereas the model sizes were chosen to be small to medium, also for typical applications. The factor loadings used in the 10-item condition were $\boldsymbol{\beta}' = (.4, .5, .6, .7, .8, .8, .7, .6, .5, .4)$, and the variances of the measurement errors were $\boldsymbol{\omega}' = (.6, .7, .8, .9, 1, .6, .7, .8, .9, 1)$; zero intercepts $\boldsymbol{\alpha}$ were used. These parameters were chosen to be typical in personality and attitudinal research applications. For the 20-item condition these parameters were simply duplicated.

Data were generated using a multivariate normal distribution. Therefore, continuous data were used in the simulations for the factor model. This is because we are only interested in the statistics' rejection rates when the factor model holds to benchmark the performance of the statistics when Samejima's model holds. The estimation was performed by maximum likelihood using a version of Rubin and Thayer's (1982) EM algorithm. All replications converged for all conditions. No Heywood cases were encountered.

Samejima's Model

Here, four sample size conditions, two model size conditions, and two different conditions for the number of categories were investigated. A total of 1,000 replications were used in each cell of the $4 \times 2 \times 2$ factorial design. The sample sizes were 500, 1,000, 2,000, and

5,000 observations, the two different model sizes were 10 and 20 variables, and the number of categories were two and five. Thus, we investigated the performance of these statistics for data conforming to Samejima's model with five categories per item, and to the two-parameter logistic model because Samejima's model reduces to the two-parameter logistic model when all variables are binary.

The slopes were obtained by transforming the factor loadings to a logistic scale. In the 10-item condition we used $\boldsymbol{\beta}' = (.74, .98, 1.28, 1.66, 2.27, 2.27, 1.66, 1.28, .98, .74)$. For the binary case, the intercepts were $\boldsymbol{\alpha}' = (.97, 0, \ldots .97, 0)$. For the five-category condition the intercepts were $\boldsymbol{\alpha}' = (.97, 0, \ldots, .97, 0)$. For the 20-item condition these parameters were simply duplicated.

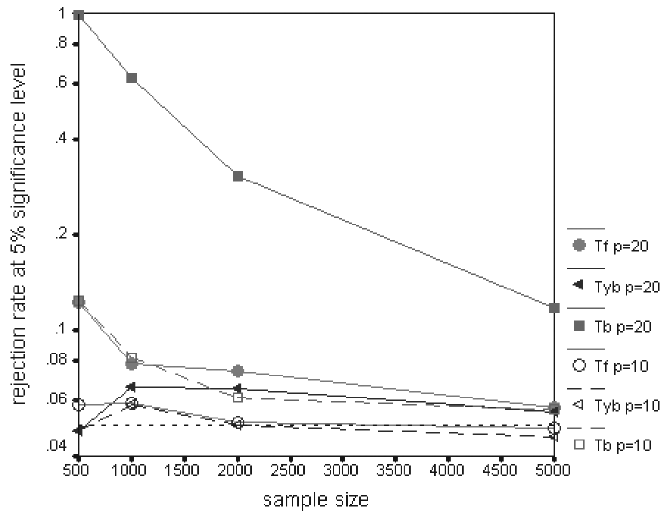
In all cases the estimation was performed using (marginal) maximum likelihood estimation via an EM algorithm (see Bock & Aitkin, 1981). The parameter estimation subroutines were written in GAUSS (Aptech Systems, 2003) and produced identical results as MULTILOG (Thissen et al., 2003) in trial runs. To ensure the numerical accuracy of the estimates, 97 quadrature points, equally spaced between -6 and 6 were used in the numerical integration of response pattern probabilities. In all conditions all 1,000 replications converged.

Simulation Results

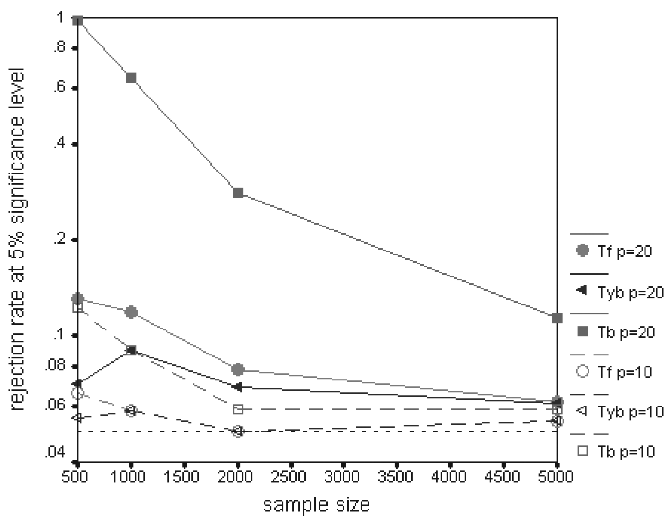
Figure 1a depicts graphically the behavior of the three statistics considered (T_B, T_{YB}, and T_F) when used to assess the fit of the factor model. Figures 1b and 1c depict the behavior of these statistics when used to assess the fit of Samejima's graded model to two-category and five-category data. In these figures, the empirical rejection rates at 5% significance level are plotted as a function of sample size (500, 1,000, 2,000, and 5,000) and model size (10 and 20 variables). These figures show that the behavior of the statistics is very similar when it is used to assess the fit of the factor model and the IRT model. They also show that among the three statistics, and consistent with previous findings, T_B requires very large sample sizes to yield reliable p values, particularly in large models. Samples larger than 5,000 observations are needed when the number of observed variables is 20. With this model size, the statistic almost always rejects the model when 500 observations are used. Fortunately, much smaller sample sizes are needed for the T_{YB} and T_F statistics to yield adequate p values. Our simulations reveal that when using T_{YB} only 500 observations are needed to obtain accurate p values in the critical region (significance level between .01 and .10) with a 20-item test. Above that region this statistic slightly underestimates the p value unless the sample size increases considerably. As shown in Figure 1, the behavior of T_F is somewhat worse than that of T_{YB}, particularly at the smallest sample size and largest model size considered. Thus, the T_{YB} statistic appears preferable to the T_F statistic.

In summary, our simulation studies support Maydeu-Olivares's (2005b) proposal. Provided the degrees of freedom are positive, test statistics based on residual covariances can be used to reliably assess the goodness-of-fit of IRT models. In fact, the behavior of these statistics when assessing the fit of Samejima's model is very similar to their behavior when assessing the fit of a factor model.

However, do these statistics have the power to distinguish both models? This is an important question, because if the statistics lack power to distinguish both models it will be unlikely that different substantive conclusions can be reached. We address this question in the next section.

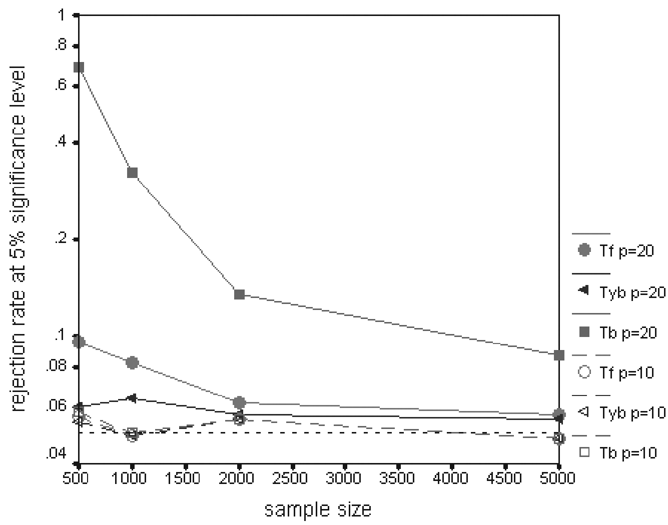


(a)



(b)

FIGURE 1 Empirical rejection rates at the 5% significance level of three statistics (T_B, T_{YB}, and T_F) to assess the goodness-of-fit of (a) a one-factor model, (b) Samejima's graded logistic model to two-category items (two-parameter logistic model), and (c) Samejima's model to five-category items. In all cases, the fitted model is correct. Rejection rates should be close to the nominal rate (5%), drawn as a thin line parallel to the x-axis. Rejection rates are plotted as a function of sample size (500, 1,000, 2,000, and 5,000) and model size (10 and 20 variables). The y-axis is on an exponential scale. The behavior of the statistics is very similar across models. T_B requires very large sample sizes to yield reliable p values. Only 500 observations are needed for T_{YB} to obtain accurate p values. The behavior of TF is somewhat worse than that of T_{YB}. (*continued*)



(c)

FIGURE 1 (*Continued*).

POWER OF THE TEST STATISTICS

Ordinal data were generated using Samejima's model and the same specifications as before but a common factor model was fitted. In this case the fitted model is misspecified (data are not continuous and the relationship between the observed variables and the latent traits is not strictly linear). Thus, we would like rejection rates to be as high as possible. Figure 2 depicts the difference of rejection rates at the 5% significance level when the fitted model is incorrect (a one-factor model is fitted) and when the fitted model is correct (Samejima's graded model). Only the results for T_{YB} are shown in Figure 2. Very similar results are obtained for the other two statistics. As Figure 2 shows, this adjusted rejection rate (adjusted empirical power rate) is very low, even for sample sizes of 5,000 observations, below 10% in all cases but one. In general, then, power increases only very slowly with increasing sample size, and with increasing model size (number of items). Finally, not surprisingly, the power is slightly higher when the number of response alternatives is two than when the number of response alternatives is five. Figure 2 contains an outlier condition. When the number of items is 20 and the number of response alternatives is two, and sample size is 5,000, the adjusted rejection rate jumps to 22%. However, even for sample sizes of 2,000 observations, the adjusted rejection rate is below 8%.

These results suggest that goodness-of-fit statistics based on residual covariances have very little power to reject the FA model when it is incorrectly applied to discrete rating data. As a result, we believe that in applications where an IRT model and a FA model of the same dimensionality are applied to the same data, the fit of the IRT model as assessed by these test statistics will be only marginally better than the fit of the factor model, simply because these statistics have low power to distinguish between both types of models.

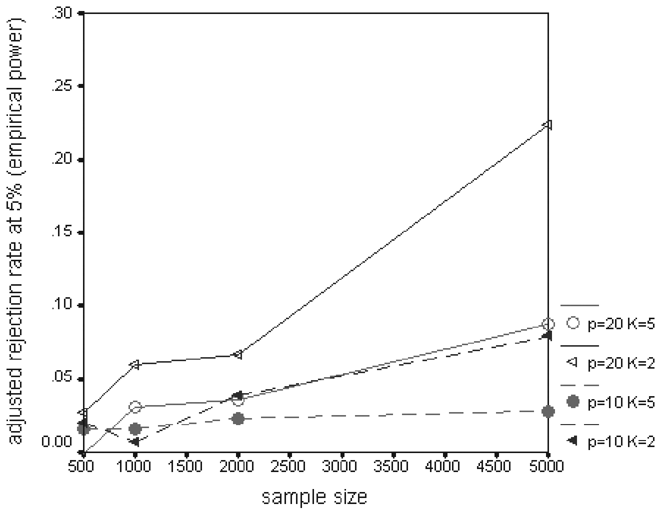


FIGURE 2 Adjusted empirical rejection rates at the 5% significance level (adjusted empirical power) of T_{YB} when fitting a one-factor model to data generated according to Samejima's graded model. The empirical rejection rate when the model is correct has been subtracted from the empirical rejection rate when the model is incorrect. This difference has been plotted as a function of sample size (500, 1,000, 2,000, and 5,000), number of alternatives per item (two and five), and number of items (10 and 20). The difference in rejection rates should be as large as possible.

In the next section we provide two applications in which we compare the fit of Samejima's model against the fit of a factor model. The first study involves binary items; the second five-category rating items.

APPLICATIONS

Binary Data: Result for the EPQ–R

In this application, we examine the fit of the two-parameter logistic model and of a one-factor model to each of the three scales of Eysenck's Personality Questionnaire–Revised (EPQ–R; Eysenck, Eysenck, & Barrett, 1985), separately for male and female respondents. Sample size was 610 for males and 824 for females. The EPQ–R is a widely used personality inventory that consists of three scales, each one measuring a broad personality trait: Extraversion (E; 23 items), Neuroticism or Emotionality (N; 24 items), and Psychoticism or Tough-Mindness (P; 32 items). All the items are dichotomous. Here we analyze the UK normative data for this questionnaire, kindly provided by Paul Barrett and Sybil Eysenck. After reverse-coding the negatively worded items, we obtained the goodness-of-fit statistics shown in Table 1. For completeness, the results for all three statistics are shown. However, practitioners need only compute the T_{YB} statistic, as it provides the most accurate results.

TABLE 1
Goodness-of-Fit Tests for the Eysenck's Personality Questionnaire–Revised Data

	Two-Parameter Logistic Model				*Factor Analysis Model*			
Scale	*Stat*	*Value*	*df*	*p Value*	*Stat*	*Value*	*df*	*p Value*
Male Sample								
E	T_B	904.47	230	0	T_B	1,290.7	230	0
	T_{YB}	363.59	230	0	T_{YB}	413.31	230	0
	T_F	2.45	230, 380	0	T_F	3.50	230, 380	0
	T_B	1,070.7	252	0	T_B	1,157.8	252	0
N	T_{YB}	387.79	252	0	T_{YB}	398.65	252	0
	T_F	2.50	252, 358	0	T_F	2.70	252, 358	0
	T_B	3,235.7	464	0	T_B	3,561.2	464	0
P	T_{YB}	511.83	464	.062	T_{YB}	519.34	464	.038
	T_F	1.67	464, 146	0	T_F	1.84	464, 146	0
Female Sample								
E	T_B	1,193.6	230	0	T_B	1,612.7	230	0
	T_{YB}	486.78	230	0	T_{YB}	544.48	230	0
	T_F	3.75	230, 594	0	T_F	5.06	230, 594	0
	T_B	1,149.3	252	0	T_B	1,475.0	252	0
N	T_{YB}	479.24	252	0	T_{YB}	527.85	252	0
	T_F	3.17	252, 572	0	T_F	4.07	252, 572	0
	T_B	1,689.3	464	0	T_B	2,014.9	464	0
P	T_{YB}	552.94	464	.003	T_{YB}	583.82	464	0
	T_F	1.59	464, 360	0	T_F	1.90	464, 360	0

Note. $N = 610$ for males and 824 for females; N = Neuroticism or Emotionality (24 items); E = Extraversion (23 items); P = Psychoticism or Tough-Mindedness (32 items); T_B = Browne's residual test statistic; T_{YB} = Yuan and Bentler chi-square residual test statistic; T_F = Yuan and Bentler's F test statistic.

Inspecting the results for the statistic of choice, T_{YB}, we see in Table 1 that both the factor model and the two-parameter logistic model fail to reproduce these scales. The only exception is the Psychoticism scale in the male sample. Using a 5% significance level, the hypothesis that the two-parameter logistic model reproduces this scale's covariances cannot be rejected (just barely). On the other hand, the hypothesis that a one-factor model reproduces these covariances is barely rejected.

The results presented in Table 1 also illustrate how the difference between the T_{YB} and T_B statistics can be substantial in applications. Even though the statistics follow the same asymptotic distribution, estimated T_B values can be more than three times larger than estimated T_{YB} values. Furthermore, there are some discrepancies in the p values obtained when using the T_{YB} and T_F statistics.

Finally, the results presented in Table 1 reveal that, as expected given the power results, there are only modest improvements in fit (as assessed by the T_{YB} statistic) when the two-parameter logistic model is applied to these data instead of the factor model. Fit improvement (as assessed by the T_{YB} statistic) ranges from a minimum of 1% for the Psychoticism scale in the male sample to a maximum of 12% for the Extroversion scale, also in the male sample. The modest improvements in fit obtained when using the two-parameter logistic model instead of the factor model are perhaps best seen when inspecting Table 2, which provides the RMSEA estimates based on T_{YB}. As can be seen in Table 2, even though the two-parameter logistic

TABLE 2
RMSEA Confidence Intervals Based on T_{YB} for the Eysenck's Personality Questionnaire–Revised Data

		Two-Parameter Logistic Model		*Factor Analysis Model*	
Sample	*Scale*	*RMSEA*	*90% CI*	*RMSEA*	*90% CI*
Males	E	0.031	(.025; .037)	0.036	(.031; .042)
	N	0.030	(.024; .035)	0.031	(.025; .037)
	P	0.013	(.0; .019)	0.014	(.004; .020)
Females	E	0.037	(.032; .041)	0.041	(.036; .045)
	N	0.033	(.029; .038)	0.036	(.032; .041)
	P	0.015	(.009; .020)	0.018	(.013; .022)

Note. RMSEA = root mean square error of approximation; E = Extraversion (23 items); N = Neuroticism or Emotionality (24 items); P = Psychoticism or Tough-Mindedness (32 items).

model provides a better fit in all cases, the 90% confidence interval for the two-parameter logistic RMSEA includes in all cases the RMSEA estimate obtained when fitting the factor model. Interestingly, we also see in Table 2 that even though a test of exact fit almost invariably rejects both the common factor model and the two-parameter logistic model fitted to the EPQ–R scales, these models cannot be rejected if a test of close fit (RMSEA $\leq$ 0.05) is used.

Polytomous Ordinal Data: The Social Problem Inventory–Revised

In this application, we examine the fit of Samejima's logistic model and of a one-factor model to each of the scales of the Social Problem Inventory–Revised (SPSI–R; D'Zurilla, Nezu, & Maydeu-Olivares, 2002), separately for male and female respondents. This inventory consists of 52 rating items, each consisting of five categories. The SPSI–R aims at measuring the process by which people attempt to resolve everyday problems. It consists of five scales: Positive Problem Orientation (PPO), Rational Problem Solving (RPS), Negative Problem Orientation (NPO), Impulsivity/Carelessness Style (ICS), and Avoidance Style (AS). The number of items making up the PPO, NPO, RPS, ICS, and AS scales is 5, 10, 20, 10, and 7, respectively. Each scale was explicitly designed to be unidimensional (D'Zurilla et al., 2002). For this analysis we analyze the normative U.S. young adult sample. Sample size is 492 for males and 551 for females.

The fit of Samejima's model to the PPO and AS scales cannot be assessed with these statistics based on residual covariances. This is because with five categories per item the scales must consist of at least 10 items but they consist of only 5 and 7 items, respectively. Therefore, the results are presented only for the ICS, NPO, and RPS scales in Table 3. As can be seen in Table 3, a one-factor model does not provide a good fit to either the ICS or the NPO scale, nor does a one-factor model fit the RPS scale, although the fit is somewhat better than for the other scales. In contrast, Samejima's graded model marginally fits both the ICS and NPO

TABLE 3
Goodness-of-Fit Tests for the Social Problem Inventory–Revised Data

	Graded Logistic Model				Factor Analysis Model			
Scale	*Stat*	*Value*	*df*	*p Value*	*Stat*	*Value*	*df*	*p Value*
Male Sample								
ICS	T_B	11.62	5	.040	T_B	128.95	35	0
	T_{YB}	11.35	5	.045	T_{YB}	100.83	35	0
	T_F	2.31	5; 487	.044	T_F	3.376	35; 457	0
NPO	T_B	9.47	5	.092	T_B	100.32	35	0
	T_{YB}	9.29	5	.098	T_{YB}	83.27	35	0
	T_F	1.88	5; 487	.097	T_F	2.668	35; 457	0
RPS	T_B	231.86	110	0	T_B	409.32	170	0
	T_{YB}	157.39	110	.002	T_{YB}	223.16	170	.004
	T_F	1.64	110; 382	0	T_F	1.581	170; 322	0
Female Sample								
ICS	T_B	23.96	5	0	T_B	138.45	35	0
	T_{YB}	22.96	5	0	T_{YB}	110.56	35	0
	T_F	4.76	5; 546	0	T_F	3.71	35; 516	0
NPO	T_B	37.93	5	0	T_B	129.61	35	0
	T_{YB}	35.48	5	0	T_{YB}	104.86	35	0
	T_F	7.53	5; 546	0	T_F	3.47	35; 516	0
RPS	T_B	233.12	110	0	T_B	442.88	170	0
	T_{YB}	163.64	110	0	T_{YB}	245.13	170	.006
	T_F	1.70	110; 441	0	T_F	1.81	170; 381	0

Note. $N = 492$ for males and 551 for females; ICS = Impulsivity/Carelessness scale (10 items); NPO = Negative Problem Orientation scale (10 items); RPS = Rational Problem Solving scale (20 items); T_B = Browne's residual test statistic; T_{YB} = Yuan and Bentler chi-square residual test statistic; T_F = Yuan and Bentler's F test statistic.

scales, but only in the male sample. In the female sample, the model fits all three scales quite poorly. The fit of the IRT model to the RPS in the male sample is poor.

As this application illustrates, in some instances Samejima's graded model might yield a substantially better exact fit to the observed covariances than the factor model when the number of categories is larger than two. However, it does so at the expense of more parameters. A factor model uses two parameters for each item regardless of its number of categories. In contrast, Samejima's model uses m parameters to fit an m-categories item. As a result, when fitting items consisting of more than two categories, the number of degrees of freedom available for testing is larger for the factor model than for Samejima's model (see Table 3).

Because of the different number of parameters involved when $m > 2$, a goodness-of-fit comparison between Samejima's and the factor model adjusting for degrees of freedom is of particular interest. In Table 4 we provide the RMSEA point estimates for this application along with 90% confidence intervals. The confidence intervals were computed as in Browne and Cudeck (1993). As can be seen in Table 4, the RMSEA's point estimates for the factor model are smaller than for Samejima's model in four of the six cases considered (in all cases for the female sample, and for the RPS in the male sample). Furthermore, the RMSEA estimate for Samejima's model is outside the 90% confidence intervals for the factor model's RMSEA for the ICS and NPO scales in the female sample. On the other hand, when Samejima's model fits better than the factor model, in no instance is the factor model

TABLE 4
RMSEA Confidence Intervals Based on T_{YB} for the Social Problem Inventory–Revised Data

		Graded Logistic Model		*Factor Analysis Model*	
Sample	*Scale*	*RMSEA*	*90% CI*	*RMSEA*	*90% CI*
Males	ICS	0.051	(.007; .091)	0.062	(.048; .076)
	NPO	0.042	(0.; .083)	0.053	(.039; .068)
	RPS	0.030	(.018; .040)	0.025	(.015; .034)
Females	ICS	0.081	(.046; .117)	0.063	(.05; .076)
	NPO	0.105	(.074; .139)	0.060	(.047; .074)
	RPS	0.030	(.020; .039)	0.028	(.02; .036)

Note. RMSEA = root mean square error of approximation; ICS = Impulsivity/Carelessness scale (10 items); NPO = Negative Problem Orientation scale (10 items); RPS = Rational Problem Solving scale (20 items).

RMSEA outside the 90% confidence interval for Samejima's RMSEA. In terms of error of approximation to the covariances, the factor model fits these data as well as if not better than Samejima's model.

DISCUSSION

The simulation results presented here suggest that test statistics based on residual covariances can be fruitfully used to assess the goodness-of-fit of IRT models to rating items. Their empirical behavior under correct model specification is not substantially different when used to assess the fit of Samejima's model (i.e., an ordinal FA model) or a factor model. Of the three statistics investigated, the one proposed by Yuan and Bentler (1997), T_{YB}, yielded the most accurate p values and it is therefore the one recommended for applications. For both the FA model and Samejima's model, 500 observations seem to suffice to provide adequate p values in the critical region (rejection rates between 1%–10%). Above this region the asymptotic p values for this statistic are a little conservative.

The use of the statistic T_{YB} for IRT research has two clear limitations. The first limitation is that the IRT model must be identified from the covariances to use statistics based on residual covariances such as T_{YB}. Some IRT models cannot be tested using this statistic when the questionnaire consists of only a few items each with a large number of categories. For instance, for Samejima's model the number of items must be twice the number of categories to use these statistics. The second limitation is that test statistics based on covariances are meaningless for assessing the fit of IRT models to polytomous unordered data.

Clearly, more research is needed before any definitive conclusions can be reached on the usefulness of this approach for goodness-of-fit assessment of IRT models to rating data. First, more simulations under correct model specification are needed. Although we presented only one set of simulations here, we run additional simulations with greater variation in the spacing of the intercept values obtaining similar results to those presented here. Still, it is questionable to present results from a few points in the parameter space and draw generalizable conclusions from them. Thus, additional independent simulations are needed. Second, simulations under correct model specification for alternative IRT models, such as Masters's (1982) partial credit model or Muraki's (1992) generalized partial credit model are needed. Third, an extensive set of simulations under different conditions of model misspecification is needed. Do these statistics have the power to detect different violations of the model assumptions, such as the presence of lower asymptotes, the presence of multidimensionality, and so on? Additional work is needed to answer these questions.

Here, we have provided expressions for the derivative matrices involved for unidimensional graded response models. The extension of these results to multidimensional models is straightforward and will be presented in a separate report. For ordinal FA models (i.e., an FA model linked to the observed ratings via a threshold process) estimated via tetrachoric/polychoric correlations, test statistics based on residual covariances provide an alternative to the current two-stage testing procedures. Currently, the structural restrictions on the polychoric correlations are tested via the regular chi-square test (see Muthén, 1993), and the assumption of discretized bivariate normality is assessed piecewise for pairs of variables (see Maydeu-Olivares, García-Forero, Gallardo-Pujol, & Renom, 2009). In contrast, tests based on residual covariances provide a test of both assumptions simultaneously.

More generally, test statistics based on residual covariances are an alternative to recently developed limited information tests for IRT models (e.g., Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares, 2006; Maydeu-Olivares & Joe, 2005, 2006; Reiser, 1996, 2008). Further research should compare the empirical performance of test statistics based on residual covariances and of these limited information tests.

CONCLUSION

Our study was motivated by the investigation of whether tests based on residual covariances could distinguish between an IRT model and a factor model. This is one particular violation of the IRT model assumptions. Our results suggest that these test statistics have little power to distinguish between data generated using a factor model and data generated using an IRT model. In fact, FA models have been found in numerous applications to provide a good fit to rating data even though FA models are misspecified when applied to rating data. Our results suggest that this is due to the lack of power of these statistics to detect this particular model misspecification.

We have also addressed the issue of whether IRT models provide a better fit in rating data applications than FA models. Our experiences with this test suggest that when the IRT model and the FA model have the same number of parameters per item (e.g., in the case of the two-parameter logistic model), the fit of the two models is likely to be rather similar. This was illustrated with the EPQ–R data. On the other hand, when the IRT model has more parameters per item than the factor model (e.g., in the case of Samejima's model applied to items with more than 2 categories), the statistics might suggest that the IRT model fits better than the common factor model. However, this difference in fit does not take into account the different number of parameters used. When fit is adjusted for model parsimony (e.g., using the RMSEA), our experience suggests that a similar fit is likely to be obtained when using the IRT and factor model. This was illustrated with the SPSI–R data.

Given that statistics based on residual covariances appear to be unable to distinguish the factor model and IRT models, is there any other way we could use to distinguish them? We do not believe so. FA models can only be tested using covariances; they cannot be tested using higher order moments unless the factor model is extended to higher order moments. Hence, to be able to compare the fit of an IRT model and a factor model using the same statistic, only statistics based on covariances can be used.

In closing, we feel that much more work is needed on the issue of comparing the fit of FA and IRT models before any definitive conclusions can be reached, and we have outlined some lines for future work we feel are worth pursuing.

ACKNOWLEDGMENTS

This research was supported by the Department of Universities, Research and Information Society (DURSI) of the Catalan Government, and by grant BS02003–08507 of the Spanish Ministry of Science and Technology.

REFERENCES

Aptech Systems, Inc. (2003). GAUSS (version 6.0.8) [Computer program]. Maple Valley, WA: Author.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London, UK: Arnold.

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

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

Bentler, P. M. (2004). EQS 6 [Computer program]. Encino, CA: Multivariate Software.

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.

Bolt, D. (2005). Limited vs. full information estimation in IRT modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics: A festschirft to Roderick P. McDonald* (pp. 27–72). Mahwah, NJ: Lawrence Erlbaum Associates.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, UK: Cambridge University Press.

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

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean and covariance structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185–249). New York, NY: Plenum.

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

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

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

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics, 23,* 315–345.
Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29.
Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49,* 71–75.
D'Zurilla, T. J., Nezu, A. M., & Maydeu-Olivares, A. (2002). *Manual of the Social Problem-Solving Inventory–Revised.* North Tonawanda, NY: Multi-Health Systems.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.
Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. T. (1985). A revised version of the Psychoticism scale. *Personality and Individual Differences, 6,* 21–29.
Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer Academic.
Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.
Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory.* Homewood, IL: Dow Jones-Irwin.
Jöreskog, K. G., & Sörbom, D. (2008). LISREL 8 [Computer program]. Chicago, IL: Scientific Software.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.
Maydeu-Olivares, A. (2005a). Further empirical results on parametric vs. non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research, 40,* 275–293.
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: Lawrence Erlbaum Associates.
Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika, 71,* 57–77.
Maydeu-Olivares, A., García-Forero, C., Gallardo-Pujol, D., & Renom, J. (2009). Testing categorized bivariate normality with two-stage polychoric correlation estimates. *Methodology, 5,* 131–136.
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.
McDonald, R. P. (1999). *Test theory: A unified approach.* Mahwah, NJ: Lawrence Erlbaum Associates.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Muthén, B. O. (1993). Goodness of fit with categorical and other non normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–234). Newbury Park, CA: Sage.
Muthén, B. O., & 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, L., & Muthén, B. O. (1998–2010). M*plus* 6 [Computer program]. Los Angeles, CA: Muthén & Muthén.
Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14,* 45–58.
Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika, 61,* 509–528.
Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology, 61,* 331–360.
Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47,* 69–76.
Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* No. 17.
Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications to developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
Steiger, J. H., & Lind, J. M. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.
Thissen, D., Chen, W. H., & Bock, D. (2003). MULTILOG 7 [Computer program]. Chicago, IL: SSI International.
Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567–577.
Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–66). New York, NY: Springer Verlag.
van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York, NY: Springer Verlag.
Waller, N. G., Tellegen, A., McDonald, R. P., & Likken, D. T. (1996). Exploring non-linear models in personality research: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64,* 545–576.
Yuan, K.-H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92,* 767–774.
Yuan, K.-H., & Bentler, P. M. (1999). *F* tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 24,* 225–243.

APPENDIX

Model-Implied Variances and Covariances Under Samejima's Logistic Model

Because rating items are categorical, we use the definition of population variances and covariances for categorical variables scored as $k = 0, \ldots, m-1$. These are

$$\sigma_{ii} = Var[Y_i] = \left(\sum_{k=0}^{m-1} k^2 \Pr(Y_i = k) \right) - \mu_i^2, \tag{16}$$

$$\sigma_{ij} = Cov[Y_i Y_j] = \left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} kl \Pr[(Y_i = k) \cap (Y_j = l)] \right) - \mu_i \mu_j, \tag{17}$$

where

$$\mu_i = E[Y_i] = \sum_{k=0}^{m-1} k \Pr(Y_i = k). \tag{18}$$

Thus, computing the model-implied variances and covariances under Samejima's graded logistic model requires computing the univariate and bivariate marginal probabilities under this model. These are obtained as a special case of Equation 7:

$$\Pr(Y_i = k) = \int_{-\infty}^{+\infty} \Pr(Y_i = k|\eta) f(\eta) d\eta, \tag{19}$$

and

$$\Pr[(Y_i = k) \cap (Y_j = l)] = \int_{-\infty}^{+\infty} \Pr(Y_i = k|\eta) \Pr(Y_j = l|\eta) f(\eta) d\eta. \tag{20}$$

Using Equations 16 to 20 and Equations 9 and 10 we obtain the desired results:

$$\begin{aligned}
\mu_i &= E[Y_i] = \sum_{k=0}^{m-1} k \Pr(Y_i = k) = \\
&= \sum_{k=1}^{m-2} k \int_{-\infty}^{+\infty} (\Psi_{i,k} - \Psi_{i,k+1}) f(\eta) d\eta + (m-1) \int_{-\infty}^{+\infty} \Psi_{i,m-1} f(\eta) d\eta = \\
&= \int_{-\infty}^{+\infty} [(\Psi_{i,1} - \Psi_{i,2}) + 2(\Psi_{i,2} - \Psi_{i,3}) + \ldots + (m-1)\Psi_{i,m-1}] f(\eta) d\eta = \\
&= \sum_{k=1}^{m-1} \int_{-\infty}^{+\infty} \Psi_{i,k} f(\eta) d\eta = \int_{-\infty}^{+\infty} \left[\sum_{k=1}^{m-1} \Psi_{i,k} \right] f(\eta) d\eta
\end{aligned} \tag{21}$$

$$\begin{aligned}
\sigma_i^2 &= Var[Y_i] = \left(\sum_{k=0}^{m-1} k^2 \Pr(Y_i = k) \right) - \mu_i^2 = \\
&= \int_{-\infty}^{+\infty} [(\Psi_{i,1} - \Psi_{i,2}) + 2^2(\Psi_{i,2} - \Psi_{i,3}) + \ldots + (m-1)^2 \Psi_{i,m-1}] f(\eta) d\eta - \mu_i^2 = \\
&= \int_{-\infty}^{+\infty} \left\{ \sum_{k=1}^{m-1} [k^2 - (k-1)^2] \Psi_{i,k} \right\} f(\eta) d\eta - \mu_i^2 = \\
&= \int_{-\infty}^{+\infty} \left\{ \sum_{k=1}^{m-1} (2k-1) \Psi_{i,k} \right\} f(\eta) d\eta - \mu_i^2
\end{aligned} \tag{22}$$

$$\begin{aligned}
\sigma_{ij} &= Cov[Y_i Y_j] = \left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} kl \Pr[(Y_i = k) \cap (Y_j = l)] \right) - \mu_i \mu_j = \\
&= \left(\sum_{k=1}^{m-1} \sum_{l=1}^{m-1} kl \int_{-\infty}^{+\infty} [\Pr(Y_i = k|\eta) \Pr(Y_j = l|\eta)] f(\eta) d\eta \right) - \mu_i \mu_j = \\
&= \left(\sum_{k=1}^{m-1} k \int_{-\infty}^{+\infty} \Pr(Y_i = k|\eta) \left[\sum_{l=1}^{m-1} l \Pr(Y_j = l|\eta) \right] f(\eta) d\eta \right) - \mu_i \mu_j = \\
&= \left(\sum_{k=1}^{m-1} k \int_{-\infty}^{+\infty} \Pr(Y_i = k|\eta) \left[\sum_{l=1}^{m-1} \Psi_{j,l} \right] f(\eta) d\eta \right) - \mu_i \mu_j = \\
&= \left(\int_{-\infty}^{+\infty} \left[\sum_{l=1}^{m-1} \Psi_{j,l} \right] \left[\sum_{k=1}^{m-1} k \Pr(Y_i = k|\eta) \right] f(\eta) d\eta \right) - \mu_i \mu_j = \\
&= \left(\int_{-\infty}^{+\infty} \left[\sum_{l=1}^{m-1} \Psi_{j,l} \right] \left[\sum_{k=1}^{m-1} \Psi_{i,k} \right] f(\eta) d\eta \right) - \mu_i \mu_j
\end{aligned} \tag{23}$$

Computational Details

For computing T_B, T_{YB}, and T_F, one needs to compute the matrix of derivatives $\boldsymbol{\Delta}$, and to compute $\boldsymbol{\Gamma}$. For a given data set $\boldsymbol{\Gamma}$ needs to be computed only once, whereas $\boldsymbol{\Delta}$ depends on the model being tested.

$\boldsymbol{\Gamma}$ can be consistently estimated using sample moments as follows (e.g., Satorra & Bentler, 1994): Let $\overline{\mathbf{y}}$ be the $n \times 1$ vector of sample means, and let $\mathbf{y}_j$ be a $n \times 1$ vector of observed responses for subject j. Also, let $\mathbf{d}_j = vecs((\mathbf{y}_j - \overline{\mathbf{y}})(\mathbf{y}_j - \overline{\mathbf{y}})')$ where vecs($\bullet$) denotes an operator that stacks the elements on or below the diagonal of a matrix onto a column vector. Then,

$$\hat{\boldsymbol{\Gamma}} = \frac{1}{N}\sum_{j=1}^{N}(\mathbf{d}_j - \mathbf{s})(\mathbf{d}_j - \mathbf{s})'. \tag{24}$$

Alternatively, $\boldsymbol{\Gamma}$ can be consistently estimated using a combination of sample and expected moments as follows:

$$\hat{\boldsymbol{\Gamma}}^* = \frac{1}{N}\sum_{j=1}^{N}(\mathbf{d}_j - \boldsymbol{\sigma}(\boldsymbol{\theta}))(\mathbf{d}_j - \boldsymbol{\sigma}(\boldsymbol{\theta}))' \tag{25}$$

where $\boldsymbol{\sigma}(\boldsymbol{\theta})$ denote the model implied variances and covariances, given by Equations 2 and 3 for the common factor model, and by Equations 12 and 13 for Samejima's model. The use of Equation 25 instead of Equation 24 in the quadratic form of Equation 4 leads to Equation 5, with Equation 24 in its quadratic form (see Yuan & Bentler, 1997).

Turning now to $\boldsymbol{\Delta}$, for the factor model the derivatives involved are

$$\frac{\partial \sigma_{ij}}{\partial \beta_r} = \begin{cases} 2\beta_i & \text{when } r = i = j \\ \beta_j & \text{when } r = i \\ \beta_i & \text{when } r = j \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad \frac{\partial \sigma_{ij}}{\partial \omega_r} = \begin{cases} 1 & \text{when } r = i = j \\ 0 & \text{otherwise} \end{cases}. \tag{26}$$

The derivatives involved in $\boldsymbol{\Delta}$ under Samejima's logistic graded model are as follows. For variances we have

$$\frac{\partial \sigma_i^2}{\partial \alpha_{j,k}} = \begin{cases} 0 & \text{when } j \neq i \\ (2k - 1 - 2\mu_i)\displaystyle\int_{-\infty}^{+\infty} [1 - \Psi_{i,k}]\Psi_{i,k} f(\eta) d\eta & \text{when } j = i \end{cases}, \tag{27}$$

$$\frac{\partial \sigma_i^2}{\partial \beta_j} = \begin{cases} 0 & \text{when } j \neq i \\ \displaystyle\sum_{k=1}^{m-1}\left((2k - 1 - 2\mu_i)\int_{-\infty}^{+\infty} \eta[1 - \Psi_{i,k}]\Psi_{i,k} f(\eta) d\eta\right) & \text{when } j = i \end{cases}. \tag{28}$$

For covariances we have

$$\frac{\partial\sigma_{ij}}{\partial\alpha_{r,q}} = \begin{cases} 0 & \text{when } r \neq i, j \\ \int_{-\infty}^{+\infty} \left[\left(\sum_{l=1}^{m-1} \Psi_{j,l} \right) - \mu_j \right] [1 - \Psi_{i,q}] \Psi_{i,q} f(\eta) d\eta & \text{when } r = i \\ \int_{-\infty}^{+\infty} \left[\left(\sum_{k=1}^{m-1} \Psi_{i,k} \right) - \mu_i \right] [1 - \Psi_{j,q}] \Psi_{j,q} f(\eta) d\eta & \text{when } r = j \end{cases}, \tag{29}$$

$$\frac{\partial\sigma_{ij}}{\partial\beta_r} = \begin{cases} 0 & \text{when } r \neq i, j \\ \int_{-\infty}^{+\infty} \left[\left(\sum_{l=1}^{m-1} \Psi_{j,l} \right) - \mu_j \right] \left[\sum_{k=1}^{m-1} ([1 - \Psi_{i,k}] \Psi_{i,k}) \right] \eta f(\eta) d\eta & \text{when } r = i \\ \int_{-\infty}^{+\infty} \left[\left(\sum_{k=1}^{m-1} \Psi_{i,k} \right) - \mu_i \right] \left[\sum_{l=1}^{m-1} ([1 - \Psi_{j,l}] \Psi_{j,l}) \right] \eta f(\eta) d\eta & \text{when } r = j \end{cases}. \tag{30}$$

In closing this section of computing notes, an alternative expression for $\mathbf{U}$ in Equation 4 is the following. Letting $\boldsymbol{\Delta}_c$ be an orthogonal complement to $\boldsymbol{\Delta}$ such that $\boldsymbol{\Delta}_c'\boldsymbol{\Delta} = 0$.

$$\mathbf{U} = \boldsymbol{\Delta}_c(\boldsymbol{\Delta}_c'\boldsymbol{\Gamma}\boldsymbol{\Delta}_c)^{-1}\boldsymbol{\Delta}_c'. \tag{31}$$

The use of this equation avoids having to invert the large matrix $\boldsymbol{\Gamma}$.