MULTIVARIATE BEHAVIORAL RESEARCH, *40*(2), 261–279


# Further Empirical Results on Parametric Versus Non-Parametric IRT Modeling of Likert-Type Personality Data

Albert Maydeu-Olivares
*Faculty of Psychology. University of Barcelona*
*Marketing Dept. Instituto de Empresa*

Chernyshenko, Stark, Chan, Drasgow, and Williams (2001) investigated the fit of Samejima's logistic graded model and Levine's non-parametric MFS model to the scales of two personality questionnaires and found that the graded model did not fit well. We attribute the poor fit of the graded model to small amounts of multidimensionality present in their data. To verify this conjecture, we compare the fit of these models to the Social Problem Solving Inventory-Revised, whose scales were designed to be unidimensional. A calibration and a cross-validation sample of new observations were used. We also included the following parametric models in the comparison: Bock's nominal model, Masters' partial credit model, and Thissen and Steinberg's extension of the latter. All models were estimated using full information maximum likelihood. We also included in the comparison a normal ogive model version of Samejima's model estimated using limited information estimation.

We found that for all scales Samejima's model outperformed all other parametric IRT models in both samples, regardless of the estimation method employed. The non-parametric model outperformed all parametric models in the calibration sample. However, the graded model outperformed MFS in the cross-validation sample in some of the scales.

We advocate employing the graded model estimated using limited information methods in modeling Likert-type data, as these methods are more versatile than full information methods to capture the multidimensionality that is generally present in personality data.

---

The author is indebted to the editor and two anonymous reviewers for helpful comments that greatly improved this article. This research has been supported by the Dept. of Universities, Research and Information Society (DURSI) of the Catalan Government, and by grants BSO2000-0661 and BSO2003-08507 of the Spanish Ministry of Science and Technology.

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

This article is dedicated to the memory of Michael V. Levine, an inspiring figure for many of us.

Modeling the responses to personality questionnaires using latent variable models is at the heart of personality research. Modeling is most often performed using linear latent variable models (i.e., factor analysis). However, when the items of personality questionnaires are categorical then the factor model is known upfront to be an incorrect model. This is because the relationship between a categorical variable and a continuous latent trait can never be exactly linear (McDonald, 1999). When modeling categorical items, a non-linear latent variable model is in principle more appropriate. These models are generally referred to as IRT (item response theory) models. For an overview of item response theory models see van der Linden and Hambleton (1997).

An important difference between linear (factor analysis) and non-linear latent variable modeling (IRT) is that within a linear framework the aim is to model the correlations (or covariances) among the observed categorical variables, whereas within an IRT framework the aim is to reproduce the data patterns. In statistical terms, in factor analysis the aim is to model a function of the first and second order joint moments of the data whereas in IRT the aim is to model a function of all joint moments of the data. Thus, IRT involves greater computational efforts than factor analysis. It also involves greater modeling efforts. This is because in factor analysis one simply needs to specify the number of latent traits and their mean and variance. In contrast, in IRT modeling one needs to specify (a) the number of latent traits, (b) a category response function (CRF), a function for the conditional probability of endorsing a response category given the latent traits, and (c) a function for the density of the latent traits. Broadly speaking IRT models can be classified as parametric, non-parametric or semi-parametric. When the CRFs and the density depend on parameters that can be interpreted substantively the IRT model is parametric. When the CRFs and the density depend on parameters which can not be interpreted substantively, the IRT model is said to be non-parametric. Finally, when either the CRFs or the density but not both depend on parameters which can be interpreted substantively, the IRT model may be denoted as semiparametric.

Since the objective in IRT modeling is to reproduce the observed proportions of data patterns, estimation generally is performed from these pattern proportions (full information estimation). Yet, full information estimation of multidimensional IRT models is generally very cumbersome computationally. As a result, most IRT research has focused on unidimensional models.

## UNIDIMENSIONAL IRT MODELS SUITABLE FOR MODELING LIKERT-TYPE DATA

Consider a personality questionnaire consisting of $n$ statements, $y_i$, $i = 1, ..., n$. We assume that each of these statements is to be rated using one of $m$ categories, $k_i = 0$,

..., $m-1$. The number of possible patterns that can be observed is therefore $m^n$. For each of these patterns of responses all unidimensional IRT models assume that

$$\Pr\left[\bigcap_{i=1}^{n}(y_i = k_i)\right] = \int_{-\infty}^{\infty}\prod_{i=1}^{n}\Pr(y_i = k_i \mid \theta) f(\theta)\, d\theta, \tag{1}$$

where $f(\theta)$ denotes the density of the latent trait, and $\Pr(y_i = k_i|\theta)$ is denoted a *category response function* (CRF). As pointed out in the introduction, the model shown by Equation 1 can be either parametric, non-parametric, or semi-parametric. Thissen and Steinberg (1986) pointed out the existence of two large classes of parametric IRT models for Likert-type data, difference models and divide-by-total models. The only difference model proposed to date is Samejima's (1969) *graded response model*. In this model, $f(\theta)$ is a standard normal density function and

$$\Pr(y_i = k_i \mid \theta) = \begin{cases} 1 - G(\alpha_{i,1} + \beta_i\theta) & \text{if} \quad k_i = 0 \\ G(\alpha_{i,k} + \beta_i\theta) - G(\alpha_{i,k+1} + \beta_i\theta) & \text{if} \quad 0 < k_i < m-1, \\ G(\alpha_{i,m-1} + \beta_i\theta) & \text{if} \quad k_i = m-1 \end{cases} \tag{2}$$

where $G(\alpha_{i,k} + \beta_i\theta)$ equals either the standard normal distribution function

$$\Phi(\alpha_{i,k} + \beta_i\theta) = \int_{-\infty}^{\alpha_{i,k}+\beta_i\theta} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx, \tag{3}$$

or the standard logistic distribution function

$$\Psi(\alpha_{i,k} + \beta_i\theta) = \frac{1}{1 - \exp\left[-(\alpha_{i,k} + \beta_i\theta)\right]}. \tag{4}$$

Thus, in this model, for each item there is one slope parameter $\beta_i$ and $m-1$ intercept parameters $\alpha_{i,k}$.

Among the divide-by-total models we can distinguish between models with or without lower asymptote (or "guessing") parameters. Reise and Waller (1990) have argued that lower asymptote parameters are not needed to model personality data and Chernyshenko et al. (2001) have provided empirical support to this argument. Thus, here we shall focus on the divide-by-total models without lower as-

ymptote parameters. In these models $f(\theta)$ is also a standard normal density function. Furthermore, these models have a CRF of the form

$$\Pr\left(y_i = k_i \mid \theta\right) = \frac{\exp\left(c_{i,k} + a_{i,k}\theta\right)}{\sum_{k=0}^{m-1} \exp\left(c_{i,k} + a_{i,k}\theta\right)}. \quad (5)$$

Three *nested* divide-by-total models will be considered in this article: (a) Bock's (1972) nominal model, (b) Thissen and Steinberg's (1986) ordinal model, and (c) Masters's (1982) partial credit model. Bock's *nominal model* is obtained by introducing just the minimal constraints in the parameters of Equation 5 needed to identify the model. For each item only $m - 1$ slope parameters $a_{i,k}$ and $m - 1$ intercept parameters $c_{i,k}$ are identified. These are the parameters of Bock's model. In contrast, for each item Thissen and Steinberg's *ordinal model* has one slope parameter and $m - 1$ intercept parameters. This model is equivalent to the *generalized partial credit model* proposed by Muraki (1992). Finally, for each item Masters's *partial credit model* has $m - 1$ intercept parameters and it is assumed that all items have the same slope parameter. Hence, of these three models, Bock's is the most general and Masters's the most restricted.

Several non-parametric IRT models have also been proposed. Consider for instance Levine's (1984) MFS (*multilinear formula score*) model. In this model one simply takes $J$ orthonormal basis, $h_j(\theta)$, for the linear span of the posterior latent trait density and the CRF is simply a linear function of these basis functions

$$\Pr\left(y_i = k_i \mid \theta\right) = \sum_{j=1}^{J} \alpha_{i,k,j} h_j(\theta). \quad (6)$$

Thus, whereas the previous parametric models use functions with a relatively small number of parameters that combine non-linearly, Levine's model uses a large number of parameters that combine linearly yielding CRFs with arbitrary shapes (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Linear inequality constraints among these parameters are generally used to reduce the size of parameter space and obtain CRFs that are smooth. Further inequality constraints among the model parameters can be introduced so that the resulting CRFs are monotone or concave over a specified range of the latent trait continuum. Thus, the parameters of this model, $\boldsymbol{\alpha}$, do not have any substantive interpretation.

## CHOOSING AN IRT MODEL TO FIT PERSONALITY DATA

There have been few applications of full information IRT modeling to personality data (for a review see Chernyshenko et al., 2001). One obvious reason for this is that IRT involves greater computational and modeling efforts. Yet another

reason may be that in personality substantive models are often multidimensional yet full information estimation of multidimensional IRT models is still problematic. Nonetheless, many personality questionnaires consist of several scales each of which is often assumed to tap a unidimensional construct. In these cases, although fitting a multidimensional IRT model to the entire questionnaire can be difficult, one can fit a unidimensional IRT model to each of the scales separately. Then the question arises: If one is willing to leave aside the linear latent variable model in favor of a non-linear model, which non-linear model should be used? There is of course no general answer to this question, which has to be addressed on an application by application basis. Thus, applications where competing IRT models are applied to the same data are needed to reach an agreement on which IRT model is generally more appropriate for modeling the Likert-type variables that are commonly found in personality applications.

To our knowledge, the only study that has fitted competing unidimensional IRT models to personality data is Chernyshenko et al. (2001). Using a cross-validation sample, they compared the fit of Samejima's (1969) logistic graded model and Levine's (1984) non-parametric MFS model to each of the scales of the 16PF Questionnaire (Conn & Rieke, 1994) and to each of the scales of Goldberg's (1998) public domain measure of the Big Five personality factor markers. The 16PF items consist of three *unordered* categories (No, Don't Know, Yes), whereas Goldberg's Big Five items consist of five ordered categories. Chernyshenko et al. also compared the fit of several IRT models after dichotomizing the data of the 16PF and of the Big Five questionnaire. In this case, the models compared were the two and three parameter logistic models (2PL and 3PL), and Levine's MFS model.

Their conclusions regarding the dichotomized version of the 16PF were (a) some but not all scales were well fitted by the 2PL model, (b) across scales there were no substantial fit differences between the 2PL and 3PL models so there is little need for a lower asymptote parameter, and (c) the non-parametric dichotomous MFS model provided a good fit to all scales and in all cases outperformed the logistic models. For the polytomous analyses of the 16PF their conclusions were similar: (a) across scales Samejima's (1969) graded logistic model provided an unacceptable poor fit, and (b) the polytomous MFS provided a good fit to all scales. Regarding the Big Five questionnaire, Chernyshenko et al. (2001) concluded that both in the dichotomous and polytomous case (a) no scale was well fitted by any parametric model, and (b) MFS provided a good fit to all scales with CRFs differing markedly across items. In summary, Chernyshenko et al. concluded that personality data can be adequately modeled by a non-parametric IRT model but that the fit of "traditional" parametric IRT models was "a matter of concern."

That a non-parametric model can adequately reproduce personality data is clearly good news to personality researchers. However, although a good-fitting non-parametric model is helpful in terms of modeling and prediction, it need not be helpful to understand the response process. The parameters of non-parametric models are not amenable to substantive interpretation. Only the resulting CRFs are

amenable to substantive interpretation. But substantive interpretation of non-parametric CRFs may prove difficult in applications. For instance, Chernyshenko et al. (2001) found that the CRFs for some options and items had a V or M shape. Why? What do they mean?

As for the parametric models, the poor fit of the 2PL and 3PL models may be caused by recoding items with an odd number of categories into just two. Also, the poor fit of the logistic graded model to the 16PF may be caused by using a model for ordered data to fit unordered data. Yet, the poor fit of the logistic graded model to the Big Five scales is indeed a cause for concern. Why did this model failed to fit in this case? Chernyshenko et al. (2001) offered two possible explanations: (a) poor fit is caused by small multidimensional components in these scales, and (b) models based on a cumulative response process such as the 2PL, 3PL and the graded model are inappropriate for modeling personality processes.

Both conjectures are appealing but they take us on different directions. Their first conjecture suggests that parametric models based on a cumulative response process such as Samejima's may be appropriate to model personality data but that they failed to provide a good fit in their application because unidimensional models were used while multidimensional models were needed. Their second conjecture suggests that parametric models based on an ideal point response process (Coombs, 1964) may be more appropriate for personality data (see Roberts, Donoghue, & Laughlin, 2000). In these models the probability of a person endorsing an item depends on the position of the person relative to the position of the item on the latent trait continuum. The closer they are, the most likely it is that the respondent endorses the item. Thus, in these models, CRFs are non-monotonic.

In our opinion, ideal point models (or unfolding models as they are also referred to) may be more appropriate than cumulative response models for some attitude data. But for personality data where often respondents are asked the degree with which a description applies to them, or how often they perform certain behaviors, cumulative response models (such as Samejima's, 1969) should be more appropriate. We believe that Chernyshenko et al.'s (2001) poor fit results with Samejima's model may be caused by multidimensionality in their data (their first conjecture).

In this study we explore this conjecture by fitting several competing models to a set of personality scales using a methodology similar to that employed by Chernyshenko et al. (2001). However, the personality scales used in our study were designed to be unidimensional. We conjecture that in this situation, the fit of Samejima's (1969) logistic graded model may be comparable to the fit of the MFS model. Also, we include in this study Samejima's normal graded model (also known as normal ogive model). If our conjecture is correct, then a good fit to personality scales such as those from the Big Five questionnaire fitted by Chernyshenko et al.'s should be obtained with multidimensional parametric IRT models. A multidimensional normal ogive version of Samejima's graded model

can be estimated using the limited information methods implemented in structural equation modeling software such as Lisrel (Jöreskog & Sörbom, 2001) or Mplus (L. Muthén & Muthén, 2001).

To sum up, in the present study we further explore empirically whether "standard" unidimensional parametric IRT models are able to adequately fit personality data. To do so, we use the same goodness of fit procedures employed by Chernyshenko et al. (2001) and the same polytomous models (MFS and Samejima's graded logistic model). However, there are three key differences between our study and that of Chernyshenko et al.: (a) we fit IRT models to personality scales that were designed to be unidimensional, (b) we also include in our study the normal ogive graded model estimated by limited information methods (i.e., from the univariate and bivariate margins of the contingency table), and (c) we also investigate the fit of divide-by-total models.

## METHOD

### Measures

We shall fit the scales of the SPSI-R (D'Zurilla, Nezu, & Maydeu-Olivares, 2002) whose items are Likert-type with five ordered categories. This questionnaire consist of 52 items at measuring the process by which people attempt to resolve problems they experience in everyday living. More specifically, the SPSI-R measures two constructive or adaptive problem-solving dimensions—positive problem orientation (PPO) and rational problem solving (RPS)—and three dysfunctional dimensions—negative problem orientation (NPO), impulsivity/carelessness style (ICS), and avoidance style (AS). The number of items composing the PPO, NPO, RPS, ICS and AS scales are 5, 10, 20, 10 and 7, respectively.

### Participants

We used two samples of undergraduate students from a large public east coast university to whom the SPSI-R was administered as part of a course requirement. The first sample consists of 1,053 respondents to whom the questionnaire was administered between 1987 and 1990. The second sample consists of 953 respondents to whom the questionnaire was administered between 1991 and 1992. The item parameters were estimated using the first sample. The sample of new observations was then used to cross-validate the results using the parameters estimated in the calibration sample.

It is important to point out that in this study we use a different type of cross-validation study than Chernyshenko et al. (2001). These authors started with one set of data and randomly halved it to obtain their calibration and

cross-validation samples. In contrast, here we use a sample of new observations from the same population. Because in the present study respondents are not randomly assigned to one sample or another, we should expect larger fit discrepancies between calibration and cross-validation samples, and poorer goodness of fit indices in our cross-validation sample than in Chernyshenko et al.

## IRT Models

The following unidimensional models will be compared: (a) Samejima's (1969) logistic graded model, (b) Samejima's normal ogive graded model, (c) Bock's (1972) nominal model, (d) Thissen and Steinberg's (1986) ordinal model, (e) Masters's (1982) partial credit model, and (f) Levine's (1984) MFS model.

The first five models are parametric and the last one is non-parametric. We shall use Levine's non-parametric MFS model as a benchmark for the parametric models as Chernyshenko et al. (2001) found this model to satisfactorily fit all the scales they considered. Also, the parametric models considered here were chosen to cover the two most popular classes within these models. Masters's model was selected because it is a Rasch-type model and we were interested in assessing the suitability of Rasch models to fit actual personality data. The ordinal model was chosen because it has the same number of parameters as Samejima's (1969) and we were interested in comparing the two classes of models, divide-by-total and difference, using models with the same number of parameters. Finally, Bock's model was selected to assess if improved fit to polytomous ordered data is obtained by employing a model that does not assume that the categories are ordered.

## Estimation of the IRT Models

Samejima's (1969) logistic graded model and all divide-by-total models were estimated using Multilog (Thissen, 1991). This program employs full information marginal maximum likelihood (MML) estimation via the EM algorithm as described by Bock and Aitkin (1981). MFS was estimated using Forscore (Williams & Levine, 1993) which also employs full information MML estimation. In all cases, default values were used.

To explore the effects of using limited information vs. full information to estimate polytomous IRT models, the graded model with normal CRFs was estimated using Lisrel (Jöreskog & Sörbom, 2001) with unweighted least squares (ULS) in the third stage of the estimation procedure. This is because B. Muthén (1993) has shown that employing ULS in the third stage of the estimation procedure leads to considerably less biased estimates than using weighted least squares (WLS).

## Assessing the Goodness of Fit

Comparing the fit of alternative IRT models is not a trivial task. The asymptotic approximations to the distribution of the statistics used for assessing the goodness of fit in categorical data analysis (i.e., Pearson's $X^2$ statistic and the likelihood ratio test $G^2$) are not appropriate in the very sparse contingency tables usually encountered in personality applications (see Agresti, 1990; Reiser & VandenBerg, 1994).

Furthermore, overall tests may fail to reveal much about the nature of the departure of an IRT model. Thus, there is a growing trend advocating the use of limited information residuals to assess the fit of IRT models (Bartholomew, 1998; Drasgow et al., 1995; Maydeu-Olivares, 2001; Reiser, 1996). Here, we shall assess how well each of the models reproduces the first, second, and third order marginals of the observed contingency tables computing Pearson's $X^2$ test for each item separately, for each pair of items, as well as for each triplet of items as in Drasgow et al. (1995)—see also Chernyshenko et al. (2001). Since the joint distribution of these limited information statistics is presently unknown, following Drasgow et al., in this article we shall simply report the mean and standard deviation of the $X^2$ to degrees of freedom ratio for item singles, pairs, and triples for each model and for each scale. Using these statistics, we shall simply discuss the rank ordering of the competing models, seeking a consistency of the rank ordering of the models across the personality scales. Since some of the models being compared have more parameters than others we shall use a sample of new observations from the population in addition to the calibration sample, as one should expect models with more parameters to better reproduce the calibration sample (Drasgow, Levine, Williams, McLaughlin, & Candell, 1989).

## Expected Results

MFS non parametric model has many more parameters than any parametric model. Hence, we expect this model to reproduce the calibration sample much better than any parametric model. Furthermore, Chernyshenko et al.'s (2001) study suggests that this model will provide the best fit also to our sample of new observations. However, we conjecture that the graded model's inability to reproduce the data in Chernyshenko et al.'s study is due to small multidimensional components in the questionnaires they modeled. We have chosen a questionnaire whose scales were purposely constructed to be unidimensional and thus we expect some parametric model to be able to adequately fit these scales in the sample of new observations.

Among the parametric models, we expect the following results.

1. A very similar fit in both samples of the graded and ordinal models as they have the same number of parameters. In fact, using artificial data Maydeu-Olivares, Drasgow, and Mead (1995) showed that although the

ordinal model and the graded model are not equivalent models, they yield virtually indistinguishable predictions.
2. A negligible fit improvement in the sample of new observations when the nominal model is used rather than the ordinal model (as the data is ordinal).
3. The ordinal model fitting better than Masters in both samples, as the latter assumes that all items are equally discriminating and the SPSI-R items' were not chosen to be equally discriminating.
4. As suggested by numerical studies in the cognitive domain (e.g., Bock & Aitkin, 1981), a similar fit of the graded model when estimated using limited and full information procedures.

## RESULTS

A reviewer pointed out the need to include some index of the dimensionality of the scales analyzed. In particular, it was suggested to report the first few eigenvalues of the reduced (i.e., squared multiple correlations in the diagonal) product-moment correlation matrix. In Table 1 we display the first three eigenvalues of these matrices, which suggest that all scales (in both samples) are highly unidimensional.

In Tables 2 and 3 we provide for each scale the means and standard deviations of the $X^2/df$ ratio for item singles, pairs, and triples for each model. In Table 2 we provide the results for the calibration sample, and in Table 3 for the cross-validation sample. In these tables we have highlighted using bold face the parametric model with the smallest mean for each scale at the item, pairs, and triplets level. Also, in these tables we have underlined the results for the non-parametric model when it outperforms the best parametric model.

It should be noted that because of the nature of the sequential estimation approach employed in Lisrel, it must reproduce almost perfectly the first order margins of the contingency table in the calibration sample. Therefore limited and full

TABLE 1
First Three Eigenvalues of the Reduced
Product Moment Correlation Matrix

| | *PPO* | *NPO* | *RPS* | *ICS* | *AS* |
|---|---|---|---|---|---|
| Calibration sample | 2.04 | 5.04 | 7.90 | 3.46 | 3.71 |
| | 0.21 | 0.38 | 0.63 | 0.60 | 0.20 |
| | 0.09 | 0.26 | 0.29 | 0.23 | 0.15 |
| Sample of new observations | 2.21 | 5.25 | 9.07 | 4.30 | 3.67 |
| | 0.17 | 0.42 | 0.48 | 0.67 | 0.23 |
| | 0.09 | 0.17 | 0.35 | 0.21 | 0.07 |

*Note.* Squared multiple correlations on the diagonal.

TABLE 2
Means and Standard Deviations of $X^2/df$ Ratio in the Calibration Sample

| | *Models* | | | | | |
|---|---|---|---|---|---|---|
| *Scale (# Items)* | *Graded-N* | *Graded-L* | *Masters* | *Ordinal* | *Bock* | *MFS* |
| PPO (5) | | | | | | |
| Items | **0.004** | 0.026 | 0.008 | 0.011 | 0.008 | 0.108 |
| | (0.001) | (0.017) | (0.005) | (0.004) | (0.006) | (0.107) |
| Pairs | 2.052 | **1.638** | 2.317 | 2.268 | 2.103 | 0.485 |
| | (0.640) | (0.490) | (0.790) | (0.738) | (0.791) | (0.180) |
| Triples | 1.590 | **1.313** | 1.869 | 1.744 | 1.505 | 0.594 |
| | (0.190) | (0.137) | (0.407) | (0.214) | (0.190) | (0.132) |
| NPO (10) | | | | | | |
| Items | **0.004** | 1.454 | 0.386 | 0.541 | 0.475 | 0.309 |
| | (0.001) | (0.483) | (0.040) | (0.151) | (0.159) | (0.118) |
| Pairs | 1.888 | **2.066** | 2.429 | 2.174 | 2.081 | 0.893 |
| | (0.843) | (0.880) | (1.385) | (1.030) | (1.028) | (0.597) |
| Triples | 1.576 | **1.643** | 2.026 | 1.824 | 1.763 | 1.042 |
| | (0.455) | (0.453) | (0.843) | (0.543) | (0.564) | (0.351) |
| RPS (20) | | | | | | |
| Items | **0.005** | 1.744 | 0.673 | 1.087 | 1.353 | 1.237 |
| | (0.002) | (0.749) | (0.049) | (0.506) | (0.584) | (0.426) |
| Pairs | 3.006 | **3.565** | 4.318 | 3.977 | 4.128 | 1.185 |
| | (1.098) | (1.146) | (2.042) | (1.365) | (1.442) | (0.441) |
| Triples | 2.198 | **2.121** | 3.033 | 2.658 | 2.534 | 1.230 |
| | (0.585) | (0.551) | (1.312) | (0.774) | (0.724) | (0.298) |
| ICS (10) | | | | | | |
| Items | **0.004** | 0.036 | 0.050 | 0.033 | 0.021 | 0.036 |
| | (0.001) | (0.021) | (0.009) | (0.016) | (0.008) | (0.072) |
| Pairs | 2.276 | **1.950** | 2.596 | 5.346 | 2.236 | 0.831 |
| | (1.228) | (1.227) | (1.182) | (3.630) | (1.454) | (0.827) |
| Triples | 1.701 | **1.505** | 2.021 | 1.961 | 1.584 | 0.985 |
| | (0.609) | (0.569) | (0.646) | (0.633) | (0.588) | (0.478) |
| AS (7) | | | | | | |
| Items | **0.004** | 1.283 | 0.326 | 0.390 | 0.637 | 0.460 |
| | (0.002) | (0.583) | (0.030) | (0.173) | (0.341) | (0.097) |
| Pairs | 2.483 | **2.462** | 3.941 | 2.986 | 2.788 | 0.765 |
| | (0.578) | (0.671) | (1.496) | (0.739) | (0.747) | (0.350) |
| Triples | **1.827** | 1.832 | 3.039 | 2.391 | 2.182 | 0.914 |
| | (0.327) | (0.384) | (1.301) | (0.468) | (0.456) | (0.203) |

*Note.* $N$ = 1,053; standard deviations in parentheses. The lowest mean in each row among the parametric models is indicated in bold. We have underlined the results for the non-parametric model when this outperforms the best parametric model. Graded-N = Samejima's (1969) normal ogive graded model; Graded-L = Samejima's logistic graded model; Masters = Masters's (1982) partial credit model; ordinal = Thissen and Steinberg's (1986) ordinal model-equivalent to Muraki's (1992) generalized partial credit model; Bock = Bock's (1972) norminal model; MFS = Levine's (1984) MFS model; PPO = Positive Problem Orientation; NPO = Negative Problem Orientation; RPS = Rational Problem Solving; ICS = Impulsive/Careless problem solving; AS = Avoidant problem solving.

TABLE 3
Means and Standard Deviations of $X^2/df$ Ratio
in the Sample of New Observations

| | *Models* | | | | | |
|---|---|---|---|---|---|---|
| *Scale (# Items)* | *Graded-N* | *Graded-L* | *Masters* | *Ordinal* | *Bock* | *MFS* |
| PPO (5) | | | | | | |
| Items | 6.761 | 6.871 | 6.840 | 6.931 | **6.538** | 6.897 |
| | (2.332) | (2.512) | (2.332) | (2.429) | (2.195) | (2.520) |
| Pairs | 4.587 | **4.164** | 5.205 | 5.033 | 4.362 | 3.742 |
| | (0.701) | (0.681) | (1.020) | (0.907) | (0.634) | (0.693) |
| Triples | 3.493 | **3.207** | 4.151 | 3.894 | 3.304 | 2.960 |
| | (0.451) | (0.391) | (0.894) | (0.614) | (0.363) | (0.490) |
| NPO (10) | | | | | | |
| Items | 25.085 | **23.181** | 25.533 | 24.237 | 24.879 | 26.489 |
| | (15.808) | (15.346) | (17.006) | (16.215) | (16.205) | (16.348) |
| Pairs | 10.193 | **10.046** | 11.198 | 10.756 | 11.168 | 10.382 |
| | (3.854) | (4.155) | (4.245) | (4.174) | (4.335) | (4.207) |
| Triples | 5.860 | **5.440** | 6.594 | 6.203 | 6.309 | 5.724 |
| | (1.620) | (1.422) | (1.703) | (1.572) | (1.458) | (2.010) |
| RPS (20) | | | | | | |
| Items | 11.422 | **10.615** | 15.332 | 17.410 | 18.458 | 15.293 |
| | (5.121) | (5.051) | (6.735) | (7.873) | (8.136) | (7.310) |
| Pairs | **6.704** | 7.106 | 10.397 | 11.031 | 10.511 | 6.784 |
| | (1.958) | (2.237) | (3.828) | (3.427) | (3.048) | (1.853) |
| Triples | **4.389** | 4.860 | 6.887 | 6.545 | 6.431 | 4.502 |
| | (1.074) | (1.477) | (3.358) | (2.879) | (2.340) | (1.101) |
| ICS (10) | | | | | | |
| Items | 24.384 | **23.882** | 25.360 | 24.890 | 25.039 | 25.053 |
| | (16.045) | (15.527) | (16.848) | (16.431) | (16.384) | (15.818) |
| Pairs | 12.002 | **11.307** | 12.364 | 12.240 | 12.456 | 12.012 |
| | (4.849) | (4.441) | (4.897) | (4.762) | (4.903) | (5.412) |
| Triples | 7.143 | **6.694** | 7.602 | 7.554 | 6.860 | 6.116 |
| | (2.039) | (2.003) | (2.083) | (2.091) | (1.943) | (2.093) |
| AS (7) | | | | | | |
| Items | 15.841 | **12.818** | 17.572 | 16.190 | 14.322 | 14.701 |
| | (4.392) | (3.139) | (4.682) | (4.448) | (4.109) | (4.206) |
| Pairs | 6.996 | **6.008** | 8.575 | 7.787 | 7.394 | 5.691 |
| | (1.013) | (0.952) | (2.247) | (1.211) | (1.165) | (0.683) |
| Triples | 4.885 | **4.219** | 6.624 | 5.641 | 5.167 | 3.426 |
| | (0.568) | (0.519) | (1.923) | (0.784) | (0.787) | (0.497) |

*Note.* $N = 943$; standard deviations in parentheses. The lowest mean in each row among the parametric models is indicated in bold. We have underlined the results for the non-parametric model when this outperforms the best parametric model. Graded-N = Samejima's (1969) normal ogive graded model; Graded-L = Samejima's logistic graded model; Masters = Masters's (1982) partial credit model; ordinal = Thissen and Steinberg's (1986) ordinal model-equivalent to Muraki's (1992) generalized partial credit model; Bock = Bock's (1972) norminal model; MFS = Levine's (1984) MFS model; PPO = Positive Problem Orientation; NPO = Negative Problem Orientation; RPS = Rational Problem Solving; ICS = Impulsive/Careless problem solving; AS = Avoidant problem solving.

information methods should not be compared on how they reproduce the first order margins in a calibration sample.

### Comparison of Parametric Models Estimated Using Multilog (Logistic Graded, Partial Credit Model, Ordinal, and Nominal)

In the *calibration* sample we observe two clear results.

1. The graded model consistently outperforms all divide-by-total models (i.e., it has the smallest mean of the $X^2/df$ ratio for item pairs and triplets for all five scales). This result was unexpected as Maydeu-Olivares et al. (1995) had found in a simulation study that the graded and ordinal models provided virtually undistinguishable predictions.
2. Among the divide-by-total models, the nominal model consistently outperforms the ordinal model, which in turn consistently outperforms the partial credit model. These results were somewhat expected as they simply reflect the number of parameters in each model. The more parameters in a model, the better the model is able to capture the idiosyncrasies of the calibrating sample (even after controlling for the different number of degrees of freedom).

In the *cross-validation* sample, we still observe the first of the trends observed in the calibration sample, namely, the graded logistic model consistently outperforms all divide-by-total models. As for the second of these trends, we find that the nominal model generally—but not always—outperforms the ordinal model. However, we consistently find that Masters's partial credit model is the worst fitting model.

### Comparison of Limited Versus Full Information Estimation of Samejima's (1969) Graded Model

As the graded model consistently outperforms all divide-by-total models in both the calibration and the cross-validation sample, it is of interest to compare the performance of limited information estimation procedures (which use normal CRFs) with the full information procedure (which uses logistic CRFs) for this model. Note that in this study we are not able to disentangle the effects of limited versus full information estimation from the effects of using normal versus logistic CRFs. However, since the normal and logistic CRFs are well known to be very close to each other, it is reasonable to attribute much of the fit difference to the estimation method employed.

As can be seen in Tables 2 and 3, in both samples there are not large differences between estimation methods, although overall the full information estimation procedure tends to yield smaller goodness of fit indices. In any case, we see in these tables that the normal graded model estimated using limited information methods

consistently outperforms all divide-by-total models in both the calibration and cross-validation sample. Thus, regardless of how the graded model is estimated, it consistently outperforms all divide-by-total models.

## Comparison of Parametric Versus Non-Parametric Models

The non-parametric model considered, MFS, outperforms all parametric models in the calibration sample at the bivariate and trivariate levels for all the SPSI-R scales. This was expected, as this model has considerable more parameters than any parametric model (13 parameters per item although subject to inequality constraints to yield smooth CRFs). However, the results of Chernyshenko et al.'s (2001) study would suggest that MFS would substantially outperform all parametric models also in the cross-validation study. This is not the case in our study, as it only outperforms the best parametric model in the shortest scales (PPO and AS). For the longer scales (NPO, ICS, and RPS) the graded model generally outperforms MFS (although the differences in fit do not seem substantial).

It is interesting to display graphically the CRFs and ability density for the MFS model and compare them with those of Samejima's (1969) logistic graded model (the best performing parametric model). In Figures 1 and 2 we show the CRFs for the second item of the PPO scale as estimated with these two models. In these figures, using the cross-validation sample, empirical proportions represented by a * for 25 equally spaced points on the PPO latent trait continuum are drawn along with their estimated 95% confidence intervals. These confidence intervals were computed as in Drasgow et al. (1995). Whenever a small number of people in an interval of the latent trait continuum chose a particular option, the confidence interval around the empirical proportions was not drawn to indicate that that particular empirical proportion (and its confidence interval) may be very poorly estimated.

In comparing the curves shown in these figures where the confidence intervals are drawn, we see that the MFS curves closely resemble those of the graded model, except for the highest category, which in the graded model must be monotonically increasing—see Equation 2, whereas the MFS CRF increases up to $\theta = 1.3$ and then decreases. These figures also suggest that the graded model reproduces more closely the first order margins of the PPO scale than MFS. This is reflected in Table 3, where the graded model has a smaller mean $X^2/df$ ratio than MFS at the item level. Yet, as we pointed out, the MFS model reproduces slightly more accurately the bivariate and trivariate margins of the PPO cross-validation data than the graded model.

In Figure 3, we plot the PPO latent trait density estimated using Forscore along with the standard normal density used in estimating the graded model. As can be seen in this figure, the density of the PPO latent trait estimated by MFS has heavier tails than those of a standard normal density; also, it is somewhat bimodal, with its

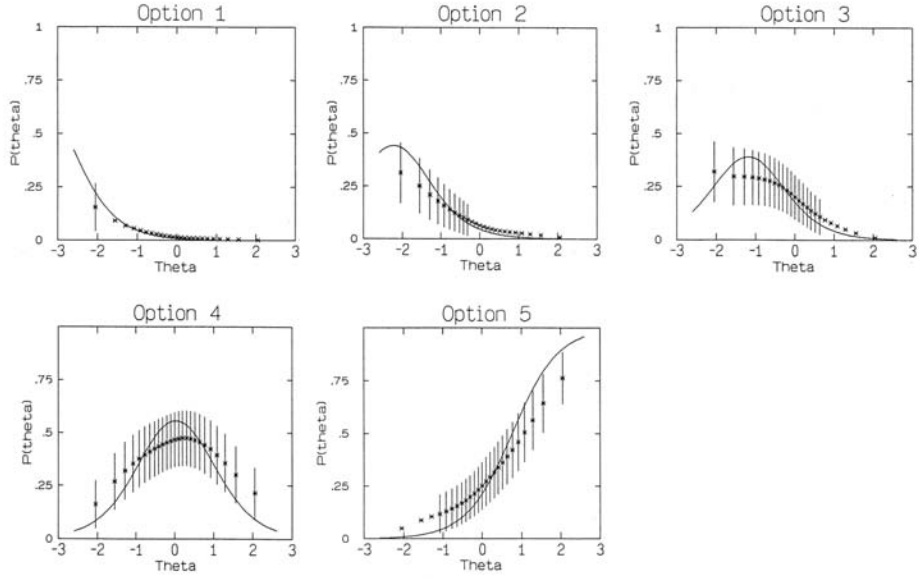


FIGURE 1 Fitplots of the CRFs of the logistic graded model for PPO's item 2 in the cross-validation sample.

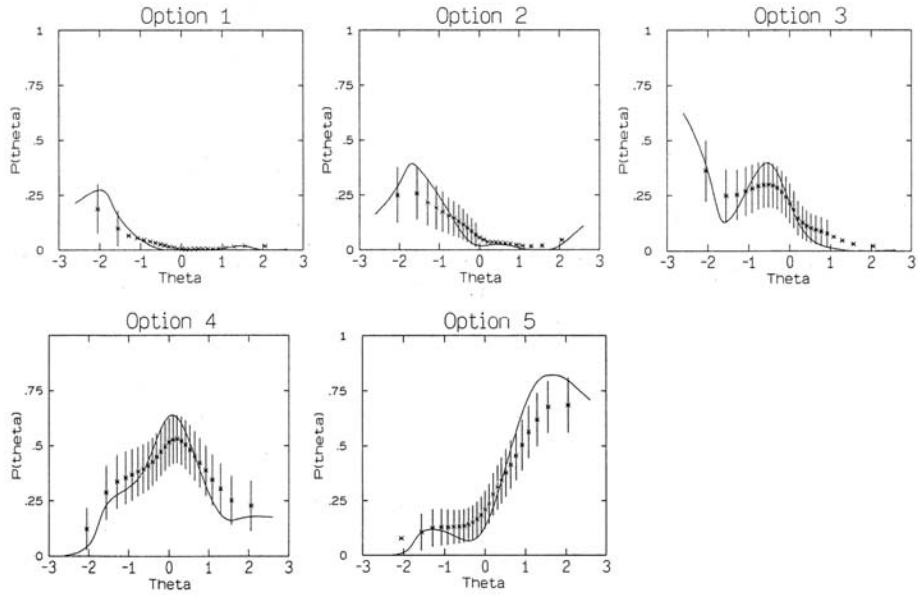


FIGURE 2 Fitplots of the CRFs of the MFS non-parametric model for PPO's item 2 in the cross-validation sample.

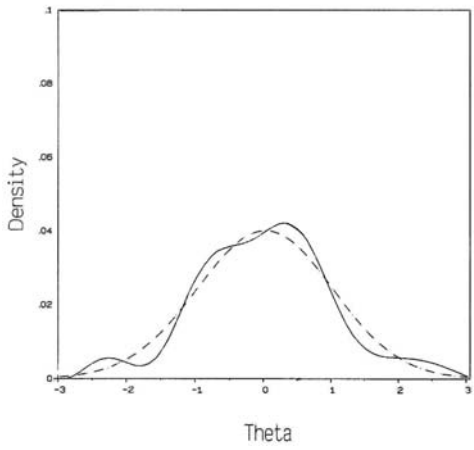


FIGURE 3 Plot of the density of the PPO latent trait estimated non-parametrically using the MFS model. The dashed line represents a standard normal density, the solid line the latent trait density estimated by MFS.

modes located at $\theta = -1$ and $\theta = 0.5$. However, taking into account the sample size employed, the MFS density closely resembles the standard normal density assumed by the graded model.

It is interesting to point out that Figures 1 and 2 suggest—in concordance with Chernyshenko et al.'s (2001) study—that no lower asymptote parameters are needed to model personality data.

## DISCUSSION AND CONCLUSIONS

We believe that several conclusions may be safely drawn from the present research regarding the use of unidimensional IRT models in personality research.

1. Among parametric models, the graded model consistently outperforms any divide-by-total model in both the calibration and the cross-validation sample. This is true even when the graded model is estimated by limited information procedures.

2. The logistic graded model estimated by full information procedures consistently outperforms the normal graded model estimated by limited information procedures. However, the differences are generally small. Furthermore, we are not able to disentangle the effects of the response function from the effects of the estimation procedure in this study.

3. Among the divide-by-total models, the more parameters a model has, the better it will fit a calibration sample. Thus, the nominal model outperforms the ordinal model, and this in turn outperforms the partial credit model. When a cross-validation sample is employed the nominal model not always outperforms the ordinal model, although the former has more parameters. Thus, we find some evidence for our hypothesis that when fitting ordinal data, a model for nominal data does not necessarily yield a better fit. Furthermore, we found that the Rasch-type partial credit model yields the poorest fit to these data. This is not surprising, as the SPSI-R items were not chosen to be equally discriminating.

4. The MFS non-parametric model consistently outperforms all parametric model in the calibration sample. This is as it should be as it has a considerably larger number of parameters. However, it needs not necessarily fit better than the graded model in a cross-validation sample. The latter is in open conflict with the results of Chernyshenko et al.'s (2001) study. We conjecture that the different results concerning the appropriateness of the unidimensional graded model to fit personality scales are the result of different amounts of multidimensionality present in their data and ours. For scales that are substantially unidimensional—like those considered here—the unidimensional graded model may be an appropriate model. For scales with moderate amounts of multidimensionality, one must resort to a non-parametric unidimensional model which essentially projects the multidimensional space into a unidimensional one (Levine, 1994), or to a multidimensional graded model, which can be effortlessly estimated using limited information methods.

We find MFS's ability to reproduce our cross-validation data remarkable and extremely encouraging. MFS is a model with a large number of parameters per item which captures every idiosyncrasy of the calibrating sample—as reflected in Table 2. Yet, it succeeded in the stiff challenge posed by the design of the cross-validation study. Our cross-validation sample was obtained by gathering new observations from the population rather than by taking a random subset of the available data. Also, this study and Chernyshenko et al.'s (2001) modeling of the Big Five items suggest than smaller sample sizes than previously thought are needed for successful MFS estimation. Levine (1994, personal communication) suggested that sample sizes of 3000 or even 5000 observations were needed for successful MFS estimation. Our results suggest that sample sizes as small as 1000 observations suffice for MFS to successfully fit cross-validation samples.

A serious limitation of the present study is that since the sampling distribution of the goodness of fit statistics used is not known, we could only report consistent trends across scales in terms of fit of competing IRT models. However, we are unable to establish if any of the models actually fits the SPSI-R data.

Despite the limitations of the goodness of fit assessment employed in this research, we believe that the following recommendations can be made to applied researchers interested in fitting an IRT model to their personality data.

1. If it is of interest to fit a unidimensional model, the best choice seems to be a graded model, regardless of whether it is estimated using limited or full information procedures.
2. If a unidimensional graded model does not yield a satisfactory fit to the data, and a unidimensional model is of substantive interest, then it may be worth considering employing a non-parametric unidimensional model, such as Levine's (1984) MFS.
3. If multidimensionality is suspected and it is of substantive interest to model it, then a multidimensional normal graded model is called for. This model can be readily estimated using limited information methods with such popular software as Lisrel (Jöreskog & Sörbom, 2001) or Mplus (L. Muthén & Muthén, 2001).

Further research is clearly needed along these lines. To date, competing IRT models have only been applied to only a handful of personality questionnaires. More studies are needed that consider alternative questionnaires. Finally, both our study and Chernyshenko et al.'s (2001) used calibration sample sizes of 1000 observations. These are rather large samples for personality studies. Do IRT models succeed in modeling personality data with smaller calibration samples? If so, how small? Unfortunately, to answer these questions new goodness of fit statistics with known sampling distributions and accurate Type I errors in sparse tables are needed. Perhaps this is the most critical issue to foster IRT modeling in Personality research.

## REFERENCES

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

Bartholomew, D. J. (1998). Scaling unobservable constructs in the social sciences. *Applied Statistics*, *47*, 1–13.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nomial 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.

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*, 563–562.

Conn, S., & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual*. Champaign, IL: IPAT.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polychotomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, *13*, 285–299.

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.

Goldberg, L. R. (1998). *International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences*. Available at http://ipip.ori.org.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8. User's reference guide*. Chicago: Scientific Software.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84–4). Champaign: University of Illinois, Model Based Measurement Laboratory.

Levine, M. V. (1994, June). *Every data set that is well fit by a two-dimensional IRT model can be equally well fit by a lower dimensional model*. Paper presented at the 59th Annual Meeting of the Psychometric Society, Champaign, IL.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Maydeu-Olivares, A. (2001). Multidimensional IRT modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, *26*, 49–69.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1995). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, *18*, 245–256.

McDonald, R. P. (1999). *Test theory. A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Muthén, B. (1993). Goodness of fit with categorical and other non normal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Muthén, L., & Muthén, B. (2001). *MPLUS user's guide*. Los Angeles: Author.

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

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3–32.

Samejima, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *17*.

Thissen, D. (1991). *MULTILOG 6: Multiple, categorical item analysis and test scoring using Item Response Theory*. Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern Item Response Theory*. New York: Springer-Verlag.

Williams, B., & Levine, M. V. (1993). *FORSCORE: A computer program for nonparametric item response theory*. Champaign: University of Illinois, Model Based Measurement Laboratory.

*Accepted August 2004*