Multivariate Behavioral Research, *36* (4), 523-562

Fitting Item Response Theory Models to Two Personality Inventories: Issues and Insights

Oleksandr S. Chernyshenko, Stephen Stark, Kim-Yin Chan,
Fritz Drasgow and Bruce Williams
University of Illinois at Urbana-Champaign

The present study compared the fit of several IRT models to two personality assessment instruments. Data from 13,059 individuals responding to the US-English version of the Fifth Edition of the Sixteen Personality Factor Questionnaire (16PF) and 1,770 individuals responding to Goldberg's 50 item Big Five Personality measure were analyzed. Various issues pertaining to the fit of the IRT models to personality data were considered. We examined two of the most popular parametric models designed for dichotomously scored items (i.e., the two- and three-parameter logistic models) and a parametric model for polytomous items (Samejima's graded response model). Also examined were Levine's nonparametric maximum likelihood formula scoring models for dichotomous and polytomous data, which were previously found to provide good fits to several cognitive ability tests (Drasgow, Levine, Tsien, Williams, & Mead, 1995). The two- and three-parameter logistic models fit some scales reasonably well but not others; the graded response model generally did not fit well. The nonparametric formula scoring models provided the best fit of the models considered. Several implications of these findings for personality measurement and personnel selection were described.

Personality assessment is currently enjoying a rebirth in Industrial and Organizational psychology (Hough & Schneider, 1996). Since the mid-1980s, several meta-analytic studies have demonstrated the usefulness of personality variables for predicting important work outcomes (e.g., Barrick & Mount, 1991, 1993; Ones, Viswesvaran, & Schmidt, 1993), while others have suggested that personality variables have less adverse impact against minorities than measures of cognitive ability (e.g., Feingold, 1994; Hough, 1996; Ones et al., 1993; Sackett, Burris, & Callahan, 1989). Because personality does not correlate strongly with intelligence, it is hoped that

We greatly appreciate the help of Michael Levine on providing his insightful comments on drafts of this article have substantially improved its quality. This research was supported in part by an NSF grant to Michael Levine, grant # 9515038.

Please address correspondence regarding this manuscript to Oleksandr S. Chernyshenko, Department of Psychology, University of Illinois at Urbana-Champaign, E. Daniel St., Champaign, IL, 61820. Electronic mail may be sent to ochernys@s.psych.uiuc.edu.

personality measures can provide incremental validity for predicting job-related criteria. This has important implications for the practice of personnel selection (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Tett, Jackson, & Rothstein, 1991).

The increased use of personality constructs in selection, training and promotion inevitably brings attention to the quality and fairness of personality testing. Adequately addressing these issues requires sophisticated mathematical methods. Traditional classical test theory approaches that evaluated psychological measures at the level of total scores have been complemented by more recent item response theory (IRT) approaches that focus on item level data. For example, Waller, Tellegen, McDonald and Lykken (1996) used IRT for the design and development of personality scales. Others have used IRT to detect items that are biased against gender and ethnic groups (e.g., Gratias & Harvey, 1998; Jennings & Schmitt, 1998) and to address the issue of faking on personality tests (e.g., Flanagan, Raju, & Haygood, 1998; Zickar & Robie, 1998).

Item response theory relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of choosing each of the response categories. This probabilistic relationship is mathematically defined by the item response function (IRF), which is a nonlinear regression of the probability of choosing to an item response category on a latent trait, θ. There are several families of item response functions that can be used to model unidimensional or multidimensional data having dichotomous or polytomous response formats.

IRT methods allow personality researchers to improve test construction and evaluate the quality of individual items. At a broader level, IRT may even help us understand how people respond to personality items. Some of the advantages of IRT over classical methods include: (a) item parameters are not subpopulation dependent; (b) the person parameter is not specific to the set of items forming the test; and (c) measurement precision is not assumed to be constant; instead IRT methods allow researchers to calculate conditional standard errors of measurement (see Rouse, Finger, & Butcher, 1999, for a recent review). Another advantage of IRT is that it can be used to develop computerized adaptive assessment instruments.

Despite the appeal of IRT, we argue that researchers and practitioners need to pay more attention to the fundamental issue of model-data fit when using IRT models to describe personality data. Without evidence of model fit, IRT results may be suspect. Throughout this study we will examine how commonly used IRT models (dichotomous and polytomous) fit data from two widely used personality inventories. Our purpose is three-fold: (a) raise awareness of model-data fit issues in the personality domain; (b) highlight

new indices of model-data fit; and (c) stimulate research on applications of IRT to noncognitive data.

There is no a priori justification why any specific IRT model should describe data adequately. More general models that have less restrictive assumptions will generally fit better but they require larger samples to estimate their increasing numbers of parameters and quickly become impractical for test developers. Regardless of mathematical complexity, the use of any model must eventually be justified on the basis of empirical results, and not only on a priori grounds (Lord, 1980). One can use a model with confidence only after repeatedly verifying that model predictions correspond to the observed data. For example, extensive research has been conducted to assess the fit of IRT models to multiple-choice cognitive ability tests. Overall, good correspondence has been established between the three-parameter logistic (3PL) model and cognitive ability data. Similar conclusions have yet to be reached for personality data.

Intuitively, one expects that unidimensional IRT models should fit personality data well, provided that a single factor model adequately describes the data. This is a reasonable assumption because the majority of personality scales have been developed using the factor analytic method of scale construction (Hogan, 1991). However, IRT makes some assumptions that are stronger than those made by the common factor model. As noted by McDonald (1999, p. 255), the common factor model assumes weak local independence, which means that the covariance of all *pairs* of items is zero for respondents with a fixed latent trait level; IRT on the other hand, assumes strong local independence meaning that the probability of a response pattern factorizes; that is, it can be written as the product of the probabilities of item responses for a given subpopulation of respondents. Consequently, it is possible to have violations of strong local independence, but no violations of weak local independence. Thus, poorly fitting unidimensional IRT models may be observed even though factor analytic methods suggest that a single common factor underlies the responses. In our opinion, only empirical studies can sufficiently address the issue of model-data fit in personality.

Current Research on Model-Data Fit in Personality

To date, few attempts have been made to determine which IRT models (if any) can fit personality data adequately. One influential study that specifically addressed this issue was Reise and Waller's (1990) paper on the fit of the one-parameter logistic (1PL) and two-parameter-logistic (2PL) models to the dichotomously scored Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982). They utilized several methods for

assessing the suitability of personality data for IRT analyses, including the assessment of unidimensionality and model-data fit. Their results showed that the 2PL model provided an adequate fit to the MPQ data. Consequently, they went on to advocate the use of IRT methods for the assessment of normal personality.

Given these findings, personality researchers have apparently assumed that IRT models, commonly applied to cognitive data, are also appropriate for personality data. Most researchers have chosen one of the IRT models (e.g., the 2PL or 3PL) and estimated item parameters with conveniently available computer programs such as BILOG (Mislevy & Bock, 1991). For example, Ellis, Becker and Kimmel (1993) used the 3PL model to evaluate the measurement equivalence of the Trier Personality Inventory (Becker, 1989), and Waller et al. (1996) used the 2PL model in their analysis of the MPQ. Rouse, Finger and Butcher (1999) also used the 2PL model to evaluate scales of the Personality Psychopathology Five (Psy-5, Harkness & McNulty, 1994), while Cooke and Michie (1997) fit the 2PL to data from the Hare Psychopathy Checklist-Revised (Hare, 1991). Schmit and Ryan (1997) used the graded response model to address the specificity of item content in the NEO-PI Conscientiousness scale. These researchers checked whether their data were unidimensional, but did not report how well their model fit the observed data.

Assessing the fit of IRT models is difficult. As Reise, Widaman and Pugh (1993) note, there are many useful fit statistics for factor analysis and structural equation modeling, but few parallel measures for IRT. Clearly, developing measures of fit for IRT is an important area for future research by psychometricians.

In sum, these studies show that IRT can improve the assessment of personality, and, consequently, aid personnel selection in many ways. Nevertheless, these findings are meaningful only to the extent that the IRT models adequately fit personality data. The research on fit is limited to one study that used only two IRT models and one personality questionnaire and, thus, many fundamental questions remain unanswered. For example, we do not know whether the 3PL model is more appropriate than the 2PL model for personality data: Reise and Waller (1990) did not attempt to fit the 3PL model even though they noted that several items had "lower asymptotes" corresponding to the *c*-parameter in the 3PL model. Reise and Waller argued that guessing was not expected on a personality test such as the MPQ. Subsequently, other researchers have suggested that guessing on cognitive ability items may be analogous to faking on personality items (see Rouse et al., 1999). Thus, examination of the fit of the 3PL model to personality data is needed. It is also unclear how well IRT models can fit

polytomous responses that are typical of many personality inventories. Reise and Waller (1990) only reported model-data fit for a dichotomously scored inventory (i.e., the MPQ). Hence, it is time to take a hard look at the fit of IRT models to personality data by evaluating a variety of IRT models and personality item formats.

Here, it is necessary to acknowledge the trade off between searching for models that adequately describe item responses and rejecting items that do not fit a chosen model. Although increasing the complexity of models will improve fit, it will also increase the sample size needed for IRT analyses, as well as possibly hinder applications of IRT in practical settings. On the other hand, outright rejection of items that do not fit a particular model may result in elimination of classes of items from personality instruments [see Roberts, Laughlin & Wedell (1999) for a specific example]. We believe that it is important to have psychometric models that are general enough to describe what are considered to be psychometrically "good" personality items. In this manuscript we worked with existing personality inventories. They contain items that were selected from pools of hundreds of items using careful screening processes. Thus, we believe that these items can be considered "good" and our task was to find a model general enough to describe them. If we could not assume that the items were of "good" quality, then the issue of determining the generality required of an IRT model would be more complex because it becomes difficult to separate problems due to "bad" items from problems due to inadequate IRT models.

Models

In the present article, a *series* of unidimensional IRT models of increasing complexity was applied to data from two personality inventories to examine the degree of generality needed to fit personality items adequately. Models examined included those used for dichotomously scored items (i.e., the 2PL model and 3PL model) and a model for polytomous items (i.e., Samejima's, 1969, graded response [SGR] model). The 2PL, 3PL and SGR models are nested parametric models in the sense that the 2PL model can be obtained from the two more complex models by setting some parameters to zero. Also examined were Levine's maximum likelihood formula scoring (MFS) models for dichotomous and polytomous data (Levine, 1984), which have provided a good fit for several cognitive ability tests (Drasgow et al., 1995). The polytomous MFS model can be considered the most general of the models studied because it does not require dichotomizing item responses, its item response functions do not have a specific parametric form, and, hence the item response functions may

assume a wide variety of shapes. It is particularly useful when parametric models do not fit the data and a researcher wants to discover the shape of item response functions (Levine, 1984).

Dichotomous Models

The Two-Parameter Logistic Model. The 2PL model has been used extensively with personality data because of its simplicity and some evidence of model-data fit (Cooke & Miche, 1997; Reise & Waller, 1990; Waller et al., 1996). It is a model for dichotomously scored responses and has item response functions of the following form

$$P(u_i = 1|\theta = t) = \frac{1}{1+\exp[-1.7a_i(t-b_i)]}, \quad (1)$$

where a_i is the discrimination parameter for item i(i = 1, …, n), b_i is the extremity parameter for item i, u_i is the response of the person with trait level θ to item i, and 1.7 is a scaling constant.

According to the 2PL model, very low θ individuals have almost no chance of making a positive response to items with large, positive extremity parameters. This model seems appropriate for modeling responses to items where "guessing" or acquiescent responding is unlikely. Hulin, Drasgow and Parsons (1983) have pointed out that the 2PL model may be appropriate for attitude items that intermix positive and negative stems to minimize or eliminate acquiescent response sets.

The Three-Parameter Logistic Model. According to this model, the probability of selecting the correct or positive response on item i is written as

$$P(u_i = 1|\theta = t) = c_i + \frac{1-c_i}{1+\exp[-1.7a_i(t-b_i)]}, \quad (2)$$

where a_i is the item discrimination parameter, b_i is the item extremity parameter, and c_i is the lower asymptote of the item response function and corresponds to the probability of a correct or positive response among respondents with low trait levels. The 3PL model might be appropriate when individuals with low trait levels can occasionally respond correctly to difficult items (e.g. cognitive ability tests with multiple-choice format). The 3PL model may also be useful if acquiescent responding or faking is expected. The 3PL model has received less attention in the personality literature than

the 2PL model. The only example we could find was the Ellis et al. (1993) paper that used this model to evaluate the measurement equivalence of English and German versions of the Trier Personality Inventory.

Levine's Maximum Likelihood Formula Scoring (MFS) Model for Dichotomous Responses. All previously discussed models could be defined by the item/option response functions based on a relatively small number of parameters. The MFS model has similar features, but its item response function is represented by a linear combination of a finite set of orthogonal functions, such as orthogonal polynomials and trigonometric functions. Because the MFS model uses linear combinations of these functions to define the IRFs, the IRFs are able to assume a wide variety of shapes. The basic formula for the dichotomous MFS model is:

$$P(u_i = 1 | \theta = t) = \sum_j \alpha_{ij} h_j(t), \tag{3}$$

where h_j is an orthogonal function and α_{ij} is the weight given to the j^{th} function in defining the option response function for item *i*. The summation index *j* indexes the orthogonal functions needed to account for data adequately. Williams (1986) found that no more than *eight* orthogonal functions were needed to fit dichotomously scored ability test data.

If every item except *i* has been modeled, then the conditional likelihood of a positive response to item *i* can be written in the following linear form

$$P(u_i = 1, \text{ pattern } \mathbf{v}^* \text{ on the remaining } n\text{-}1 \text{ items} | \theta = t) = \sum_j \alpha_{ij} h_j(t) l(\mathbf{v}^*, t), \tag{4}$$

where **v*** is a vector containing the item responses (without item *i*) and $l(\mathbf{v}^*, t)$ denotes the likelihood of **v*** at *t*. Thus, the marginal likelihood of the *n*-item response pattern is

$$\sum_j \alpha_{ij} \int h_j(t) l(\mathbf{v}^*, t) f(t)\, dt, \tag{5}$$

where *f* is the density of θ, so that $l(\mathbf{v}^*, t) f(t)$ is proportional to the posterior θ density given $n - 1$ item responses.

Because the total number of response patterns, although very large, is finite, the set of posterior densities can be written as a linear combination of a finite number of functions. The MFS model makes a simplifying assumption that the vector spaces of functions obtained as linear

combinations of posterior densities of the *n* - 1 item patterns are nearly the same no matter what item is excluded. This implies that the IRF for the item being studied can be closely approximated as a linear combination of the posterior densities computed using the remaining items.

The MFS model selects a set of *J* orthogonal functions that can approximate the entire array of posterior densities with the smallest mean squared error. The number of orthogonal functions depends on the complexity of the data.

The MFS model is implemented in the computer program FORSCORE (Williams & Levine, 1993), which uses a constrained optimization process similar to typical IRT programs for parametric models (such as BILOG; Mislevy & Bock, 1991) where the parameters to be estimated are restricted. With FORSCORE, a researcher must translate qualitative assumptions about the shape of response functions into linear inequalities (usually first and second order derivatives of the orthogonal functions) that must be satisfied during the optimization process. For example, the assumption of a nondecreasing item response function at *t* translates into the following inequality

$$\frac{d}{dt} P(u_i = 1 | \theta = t) = \sum_j \alpha_{ij} \frac{d}{dt} h_j(t) \geq 0. \tag{6}$$

Although estimating item response functions for the MFS model requires larger sample sizes and more parameters than logistic models, the MFS model is especially useful when the shapes of item response functions are unknown. With any type of data, the MFS model is able to generate a best fitting item response function — one that does not need to be logistic or monotonic in form. To do this, a researcher imposes no constraints on the shape. This flexibility allows a researcher to determine the "true" form of the item response functions in any data (although, as noted below, there are significant complications).

Polytomous Models

The models discussed above are appropriate only for dichotomously scored data. If items have more than two response options (polytomous), then artificial dichotomization of responses is usually performed with one or more options designated as the "correct" or "positive" response and all remaining options recoded as negative responses. If dichotomous IRT models are found not to fit polytomous data, the misfit could be an artifact

of dichotomization (Jansen & Roskam, 1986). Because most personality measures use a polytomous item response format, the use of dichotomous models may be jeopardized by the dichotomization and polytomous models may be preferred.

There are numerous polytomous IRT models (see van der Linden & Hambleton, 1997). Given space limitations, we decided to apply models that seemed most appropriate for the response format of personality items. In most personality scales, response options are ordered according to the level of agreement with a particular statement. Two polytomous models suitable for ordered response categories were selected for the present study: Samejima's graded response (SGR) model (Samejima, 1969) and the polytomous MFS model (Levine & Willams, 1993). Other well-known polytomous IRT models, such as Bock's (1972) nominal model, Samejima's (1979) multiple-choice model and Thissen and Steinberg's (1984) multiple-choice model, were not included in the present study because they were designed originally to model cognitive ability tests with multiple-choice formats and they assume no ordering of response options.

Samejima's Graded Response (SGR) Model. For dichotomously scored items, it is common practice to discuss only the item response function for the correct response to an item although a response function also exists for the negative category. In polytomous IRT terminology, these functions are called option response functions because they relate a person's probability of endorsing a particular response option to the trait level. According to the SGR model, the probability of selecting option k on item i is

$$(7)\quad P(v_i = k \mid \theta = t) = \frac{1}{1+\exp[-1.7a_i(t-b_{i,k})]} - \frac{1}{1+\exp[-1.7a_i(t-b_{i,k+1})]},$$

where v_i denotes the person's response to the polytomously scored item i; k is the particular option selected by the respondent ($k = 1, \ldots, s_i$, where s_i refers to the number of options for item i); a_i is the item discrimination parameter, which is assumed to be the same for each option within a particular item; b is the extremity parameter that varies from option to option given the constraints $b_{k-1} < b_k < b_{k+1}$, and b_{Si+1} is taken as $+\infty$.

Note that the 2PL model is a special case of the SGR model when the number of response options is equal to two. A detailed discussion about fitting the SGR model to Likert-type data can be found in Muraki (1990). In the personality domain, the SGR model has been used to address the specificity of item content in the NEO-PI Conscientiousness scale (Schmit & Ryan, 1997) and to model faking on the Assessment of Biographical and

Life Events (ABLE; Peterson, Hough, Dunnette, Rosse, Toquam, & Wing, 1990) personality inventory. Zickar and Drasgow (1996) presented some evidence that the SGR model did not fit the ABLE data well, but did not explore that issue in detail.

Polytomous MFS Model for Ordered Responses. The extension of the dichotomous MFS model to the polytomous case is straightforward. In the polytomous MFS model, the probability of selecting option *k* on item *i* is written

$$P(v_i = k|\theta = t) = \sum_j \alpha_{ijk} h_j(t), \tag{8}$$

where v_i denotes a person's response to the polytomously scored item; *k* is the particular option selected by the respondent, α_{ijk} is the weight for item *i*, function *j* and option *k* to be estimated by marginal maximum likelihood estimation. The description of the constraints that were imposed on ordered response categories for the polytomous MFS model is given in the Methods section.

Assessing Model-Data Fit

The model-data fit issue can be addressed in two ways. First, the data must conform to model assumptions about dimensionality. Second, predictions based on the estimated model should be examined in cross-validation samples. This can be done using a variety of statistical tests of goodness of fit and graphical methods.

Checking Model Assumptions. Most IRT models make the basic assumption of unidimensionality and hence local independence. Unidimensionality asserts that the response probability is a function of a single latent characteristic θ of an individual.

Despite the importance of the unidimensionality assumption, there is little agreement on an adequate test of unidimensionality. Hattie (1984, 1985) empirically assessed over 30 indices of unidimensionality and found problems with nearly all of them. Drasgow and Lissak (1983) developed a procedure called modified parallel analysis (MPA) that circumvented the majority of the problems associated with traditional approaches. MPA is a combination of IRT and factor analysis of tetrachoric correlations. The procedure extends Humphreys and Montanelli's method of parallel analysis (1975), which compares the eigenvalues from a synthetically created data set to those estimated from real data. In several Monte Carlo studies, Drasgow and Lissak (1983) showed that MPA was effective in determining whether an item pool was sufficiently unidimensional for the application of IRT.

Another frequently used method for assessing unidimensionality is Stout's nonparametric DIMTEST (Stout, 1987). It is a conditional covariance based hypothesis testing procedure that assesses whether two subtests are dimensionally distinct. Unfortunately, the utility of the DIMTEST procedure for personality measurement is limited due to the short length of personality scales. Simulation studies (personal communication with William Stout) indicated that at least 20 items are needed to obtain accurate results for the DIMTEST procedure, however, the majority of personality scales rarely exceed 15 items.

The confirmatory factor analysis (CFA) approach that tests a one-factor model is also appropriate for assessing scale unidimensionality. It is easier to implement than MPA and has many well-established goodness of fit indices. On the other hand, we are unaware of any simulation studies that used fit statistics from CFA to determine whether an item pool was sufficiently unidimensional for IRT analyses. Hence it is difficult to judge the appropriateness of using a unidimensional IRT model based on CFA goodness of fit indices.

In this article, we chose to use MPA and CFA approaches to assess the unidimensionality of personality scales. Stout's DIMTEST procedure was not utilized because the longest scale we examined had only 14 items.

Checking Model-Data Fit. Drasgow et al. (1995) advocated a combination of complementary graphical and statistical methods to evaluate the adequacy of model predictions. In our study, both graphical fit plots and chi-square goodness of fit tests for single items, pairs, and triples were used to investigate the fit of IRT models to personality data.

Graphical fit plots are one of the most widely used methods for examining model-data fit. The idea is to plot item/option response functions, estimated from a calibration sample, as well as the empirical proportions of positive responses obtained from a cross-validation sample. In the simplest version, a fit plot is constructed by dividing the θ continuum into, say, 25 strata. Then θ is estimated for each examinee, and the total number of examinees in each θ stratum is counted. An empirical proportion is computed as the number of examinees who selected the positive option divided by the total number of examinees in the stratum. Samejima's (1983) simple sum procedure provides an example of such a traditional fit plot. The simple sum estimate of a point $P_i(t)$ on an item response function is computed as

$$\hat{P}_i(t) = \frac{\sum_{(A:A \epsilon S^+)} P\left(\theta = t | \hat{\tau} = \hat{\tau}_A\right)}{\sum_{A} P\left(\theta = t | \hat{\tau} = \hat{\tau}_A\right)} \tag{9}$$

where the summation in the denominator is over all examinees in the sample, the summation in the numerator is over only the examinees who correctly answered the item *i*, $\hat{\tau}$ is a θ estimator computed from responses to all items except the target item, and $\hat{\tau}_A$ is the value of $\hat{\tau}$ computed from examinee A's data.

The problem with this straightforward approach is that the θ estimate ($\hat{\tau}$) for the individual is hardly ever equal to the true θ due to estimation error. Error in $\hat{\tau}$ can change the fit plot such that even with a very large sample and perfectly estimated response functions, the sample fit plot may differ substantially and systematically from the true response function. This problem is especially pronounced for short tests where θ estimates have larger error.

Levine and Williams (1991, 1993) found an elegant solution to this problem. By extending Samejima's simple sum procedure, they showed that appropriately constructed fit plots can have the same shape as the true item/option response function even for a biased estimator of θ with a substantial sampling error. In Samejima's model, the simple sum estimate of a point on the item/option response function, $\hat{P}_i(t)$, is computed with the $\hat{\tau}$ estimate. Levine and Williams proposed replacing this estimate with the vector-valued statistic that is simply the respondent's response pattern **u*** to a given set of items (including the target item). Levine and Willams's empirical estimate of an item response function can be written as

$$\hat{P}_i(u_i = 1|\theta = t) = \frac{N^+}{N} \frac{\sum_{(A:A \epsilon S^+)} P\left(\theta = t | u = u_A^*\right) / N^+}{\sum_{A} P\left(\theta = t | u = u_A^*\right) / N} \tag{10}$$

where N^+ is the number of respondents answering the target item positively; N is the total number of respondents; $\mathbf{u}^*_A$ is the dichotomously scored response pattern for respondent A; and S^+ is the set of respondents answering the target item positively. The resulting fit plot is proportional to the ratio of two averaged posterior densities: The numerator is the density for respondents who answered the target item positively at a particular level of θ and the denominator is the density of all respondents at that level. Levine mathematically showed that $\hat{P}_i(t)$ approaches a point on the true item response function as the sample increases. Levine and Willams's proof of the asymptotic properties of Equation 10 requires that all items, including the one being studied, be incorporated in $\mathbf{u}^*$. It is important to note that the empirical response functions must be recalculated for each new model even though the data are identical. Thus, they can assume different shapes.

Statistical tests of goodness of fit (i.e., χ^2 fit statistics) are probably the most widely used in model-data fit assessment. Unfortunately, they are often viewed as inconclusive evidence of adequate model-data fit because of their sensitivity to sample size and their insensitivity to certain forms of model-data misfit. As with the fit plots, we used an improved method of computing this statistic: The *adjusted* chi-square to degrees of freedom ratio.

The ordinary χ^2 for item *i* is computed from the expected and observed frequencies,

$$\chi_i^2 = \sum_{k=1}^{s} \frac{\left[O_i(k) - E_i(k)\right]^2}{E_i(k)}, \tag{11}$$

where *s* is the number of keyed options, $O_i(k)$ is the observed frequency of endorsing option *k*, and $E_i(k)$ is the expected frequency of option *k* under the specific IRT model. The expected frequency of respondents selecting each option is computed using

$$E_i(k) = N \int P(v_i = k | \theta = t) f(t) dt, \tag{12}$$

where $f(\bullet)$ is the θ density, usually taken to be the standard normal because item/option response functions are scaled in reference to this distribution. In our study, the above integral was evaluated by numerical quadrature using 61 grid points on the interval (-3, +3). To bypass the sensitivity to sample size and to allow comparisons between different samples and tests, the χ^2 was first adjusted to the magnitude that would be expected in a sample of 3,000. Then, the ratio of chi-square to the degrees of freedom was computed. A ratio of more than 3.0 for any given item was viewed as indicative of model-data misfit.

Van der Wollenberg (1982) showed that χ^2 statistics for single items are in many instances insensitive to unidimensionality violations; instead, for the Rasch model, χ^2 statistics for item singles are mainly sensitive to whether the items vary in discrimination. In addition, the χ^2 statistic for a single item, as implemented here, is insensitive to certain types of misfit. Figure 1 shows an example of this problem where the empirical IRF consistently lies above the estimated IRF at low trait levels and below it at high trait levels. Although it is visually clear that the data do not fit the IRT model, the χ^2 for an individual item will be close to zero, because it is a marginal statistic and the estimated IRF is integrated with a normal theta density; consequently, many different integrands can integrate to the same constant (i.e., the observed marginal number endorsing the item). To avoid these problems, the χ^2 statistic should

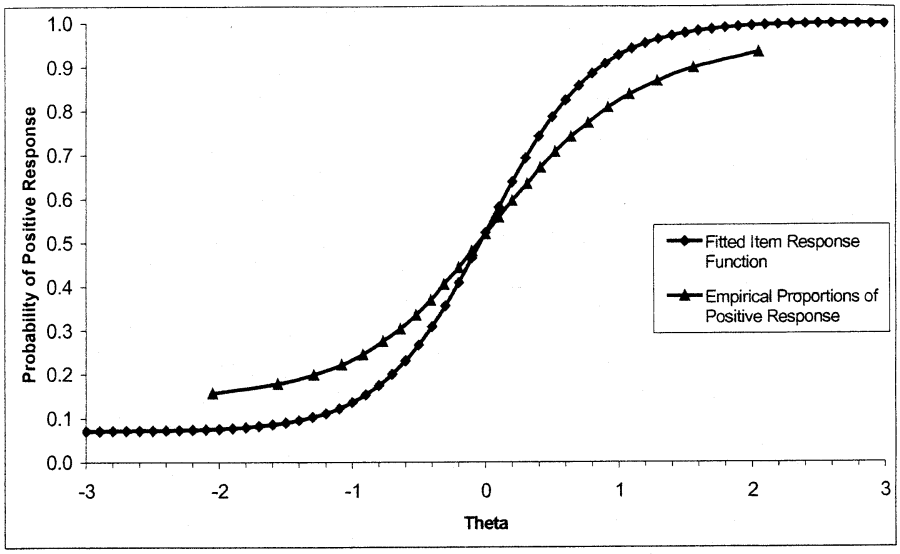

Figure 1
Example of the Model-Data Misfit to Which χ^2 Statistic for Single Item is Insensitive

be computed for pairs and triples of items. Pairs and triples of items with similar misfits will have large χ^2 statistics.

Our use of χ^2 statistics for pairs and triples of items is similar in spirit to statistics developed by van den Wollenberg (1982) and Glas (1988). Using the Rasch model, these researchers showed that statistics computed from pairs of items are sensitive to violations of local independence. Specifically, number correct is a minimal sufficient statistic for the latent trait θ in the Rasch model and consequently van den Wollenberg and Glas computed indices of pairwise dependence conditional on number right. These indices are large when local independence fails to hold.

The χ^2 statistics for item pairs and triples provide strong tests of model-data fit. The goal of any mathematical modeling is to represent data in a simpler form. In IRT, we want to reconstruct the *patterns* of responses using the chosen model. Traditionally, we assess the fit of models one item at a time, which implies that if the individual items are fit well, the entire pattern of responses will fit satisfactorily also due to local independence; however, the assumption of local independence is never tested directly. The χ^2 statistics for item pairs and triples provide an explicit means for examining how successful a model is in predicting patterns of responses. They are sensitive to unidimensionality violations because the expected frequencies are computed under a unidimensional model.

The expected frequency for a pair of items in the $(k, k')^{th}$ cell of the two-way table for items i and i' is computed as follows:

$$E_{i,i'}(k,k') = N \int P(v_i = k|\theta = t) P(v_{i'} = k'|\theta = t) f(t) dt, \quad (13)$$

and, the observed frequencies are counted in each cell. Some cells are combined so that the expected frequencies exceed 5. The usual χ^2 for a two-way table is then calculated. A similar procedure is carried out with item triples. Algebraically, if model-data misfit occurs for an item pair or triple at the same trait level, the χ^2 will increase even for the kinds of misfit presented in Figure 1.

Previous research with cognitive ability data found the *adjusted* chi-square to degrees of freedom ratio statistic for item singles, pairs and triples to be very useful for comparing several competing IRT models (Drasgow et al., 1995). The best fitting models had small (below 3) adjusted chi square to degrees of freedom ratios for item singles as well as small ratios for pairs and triples.

In sum, if the ratio of chi-square to the degrees of freedom exceeds 3.0 for item singles, pairs, or triples, one can infer that the parametric form of the item/option response function is violated or that the data are multidimensional. Both of these outcomes indicate that the chosen IRT model does not fit the data.

Method

Data

The five models described previously were fitted to data from two personality inventories. These included the Fifth Edition of the Sixteen Personality Factor Questionnaire (16PF, Conn & Rieke, 1994) and Goldberg's 50-item measure of the Big Five Factor markers (Goldberg, 1997, 1998). Both the 16PF and Goldberg's Big Five are among the most widely used personality inventories in research and practice today.

16PF. The first data set consisted of 13,059 individuals responding to the US-English version of the Fifth Edition of the 16PF. The data were test protocols sent to the Institute for Personality and Ability Testing (IPAT) by its customers for the purpose of creating computer-generated interpretive reports in 1995 and 1996 for the purposes of research, counseling/ development, and selection. 170 items from 16 noncognitive scales were analyzed. The 16 scales were Warmth (11 items), Emotional Stability (10 items), Dominance (10 items), Liveliness (10 items), Rule-Consciousness (11 items), Social Boldness (10 items), Sensitivity (11 items), Vigilance (10

items), Abstractedness (11 items), Privateness (10 items), Apprehension (10 items), Openness to Change (14 items), Self-Reliance (10 items), Perfectionism (10 items), Tension (10 items), and Impression Management (12 items).

The raw polytomous data (three response options) were prepared for analysis by reverse scoring items that were negatively worded and, when appropriate, dichotomizing. The responses were scored so that the "high" option ("c") indicated high standing on the trait continuum. For the dichotomous models, the middle option ("b") was coded as "high" and assigned a score of one. Past research (and our experience as well) indicated that combining the middle response option with either the "low" or the "high" response options has little effect on the resulting item parameters.

The IRT calibration sample consisted of 6,530 respondents. This sample was formed by taking every second respondent from the total data set, beginning with the first respondent. Thus, only respondents with odd count numbers from the original data set were used for IRT calibration of the items. The remaining 6,529 cases formed the "empirical sample" that was used to cross-validate the estimated IRT models. Table 1 presents the summary statistics for the 16PF scales.

The Big Five Personality Factor Scales. The second data set consisted of 1,594 Singapore military recruits and 274 Junior College students (total: 1,768) who had responded to Goldberg's (1997, 1998) public domain, 50-item measure of the Big Five personality factor markers. The data were collected as part of a dissertation research effort by Chan (1999), in which several self-report measures were administered in English, the official language of Singapore. The age of the respondents ranged from 17 to 24. All had at least a high school education conducted primarily in English in Singapore. The military sample was entirely male, whereas 60% of the student sample was female. Although the soldiers were asked to indicate their identification numbers due to the longitudinal nature of Chan's research, there was no reason to suspect that they may have faked their responses because they were not job applicants — military service is compulsory in Singapore. The students answered the questionnaire anonymously.

The measure consisted of five 10-item subscales measuring Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellectance or Openness to Experience. Respondents were asked how well each of the 50-items described them on a five-point scale as follows: (1) Very inaccurate, (2) Moderately inaccurate, (3) Neither inaccurate nor accurate, (4) Moderately accurate, (5) Very accurate. Table 1 presents the summary statistics for the Big Five scales, and the valid cases used for the

Table 1
Summary Statistics for Scales and Data

Scale	Number of Items	Number of Valid Cases			Mean	SD	Alpha
		Total	Calibration	Validation			
			16 PF				
Warmth	11	12936	6462	6474	14.72	4.53	0.68
Emotional Stability	10	12985	6492	6493	14.55	5.05	0.81
Dominance	10	12957	6483	6474	13.70	4.18	0.67
Liveliness	10	12933	6477	6456	11.50	4.77	0.71
Rule-Consciousness	11	12922	6457	6465	14.90	4.86	0.74
Social Boldness	10	12955	6468	6487	11.81	6.20	0.86
Sensitivity	11	12911	6455	6456	11.59	5.83	0.79
Vigilance	10	12970	6495	6475	10.19	4.65	0.74
Abstractedness	11	12906	6465	6441	7.12	4.98	0.74
Privateness	10	12973	6486	6487	10.62	5.14	0.76
Apprehension	10	12939	6470	6469	10.91	5.51	0.78
Openness to Change	14	12894	6447	6447	17.77	5.48	0.69
Self-Reliance	10	12996	6501	6495	7.88	5.24	0.78
Perfectionism	10	12953	6482	6471	12.07	4.94	0.73
Tension	10	12988	6497	6491	9.89	5.24	0.77
Impression Management	12	12979	6494	6485	11.42	5.12	0.67
			Goldberg's 50-item Measure of Big Five Personality Factor Markers				
Emotional Stability	10	1848	922	926	30.70	7.36	0.84
Extraversion	10	1857	930	927	30.87	6.67	0.80
Intellect	10	1848	928	920	33.86	5.84	0.77
Agreeableness	10	1848	922	926	37.63	4.92	0.68
Conscientiousness	10	1857	928	929	34.22	6.00	0.76

purposes of IRT calibration and cross-validation. Note that cases with missing data were dropped from the analyses for each of the five scales. The calibration and cross-validation samples were formed using a random split of the combined data set. For the purpose of fitting the 2PL and 3PL models, the data were dichotomized after negatively worded items were reverse scored, such that the first three options representing a low level of a trait were scored as 0, and the last two options indicating a high level of a trait were scored as 1.

Analyses

Modified Parallel Analysis (MPA) of Scale Unidimensionality. To test the assumption of scale unidimensionality, MPA was performed for each scale (Drasgow & Lissak, 1983). First, inter-item tetrachoric correlations were computed using PRELIS (Jöreskog & Sörbom, 1989). A principal axis factoring (PAF) was then conducted to extract the common factors. Scree plots of eigenvalues were then constructed (Hambleton, Swaminathan, & Rogers, 1991).

Next, BILOG (Mislevy & Bock, 1991) was used to estimate item parameters for the 3PL model for each of the 16PF and Big Five scales. These item parameters were used to create synthetic data sets that were truly unidimensional and contained the same number of simulated examinees as our original samples. As before, inter-item tetrachoric correlations for the simulated data of each scale were computed and factor analyzed using PAF to obtain the eigenvalues. The eigenvalues from the synthetic data sets were superimposed on the scree plots of the real data for each scale, and the differences in second eigenvalues were examined.

Assessing Unidimensionality with Confirmatory Factor Analysis (CFA). The LISREL-8 program (Jöreskog & Sörbom, 1993) was used to fit a one factor model to the polychoric correlation matrix for the items from each personality scale. Each factor loading in the factor loading matrix ($\mathbf{\Lambda}$) was set to be a free parameter. The factor variance was fixed at 1.0 and maximum likelihood parameter estimation was used.

IRT Calibration of Items. The parameters of the 2PL and 3PL models were estimated using BILOG (Mislevy & Bock, 1991). Thissen's (1991) MULTILOG program was used for the SGR model. Williams and Levine's (1993) FORSCORE program was used for both MFS models. All programs utilized marginal maximum likelihood estimation to obtain parameters. Both data sets were analyzed using the same convergence criteria for all models. With all parametric models, the lower asymptote parameter was constrained to be between 0 and 1, the item discrimination parameter was constrained to be positive, and the item extremity parameter was constrained to lie between –3 and + 3 to avoid implausible values.

With the MFS models, several qualitative assumptions about the shapes of item response functions were imposed to constrain optimization. FORSCORE can accommodate a variety of constraints including monotonicity, concavity and smoothness; they can be assumed globally or just over an interval, for some or all items. In the present article, only

smoothness constraints were imposed for the dichotomous MFS model. In the case of the polytomous MFS analysis of the 16PF data, we placed inverted-U constraints on the middle option and monotonicity constraints on other options. The polytomous MFS model was not used for the Big Five items because the sample size was inadequate for estimated increased number of parameters.

Model-Data Fit. After calibrating a test, the fit of each model was evaluated using a cross-validation sample. Fit plots for each item were constructed using Williams's (1999) EMPOCC program. As explained previously, rather than assigning each examinee to a cell and incrementing the examinee count in that cell, as in the usual fit plot histogram, the examinees' posterior densities were used to distribute the counts over the θ continuum represented by the 25 θ strata. Chi-square statistics were computed for single items and all possible pairs and triples of items within each personality scale and then adjusted to a sample size of 3,000. The ratios of χ^2 statistics to their degrees of freedom were then computed. To summarize the large number of adjusted ratios, we sorted them into six intervals: very small (<1), small (≥ 1 and <2), medium (≥ 2 and <3), moderately large (≥ 3 and <4), large (≥ 4 and <5), and very large (≥ 5). Means and standard deviations of adjusted χ^2/df ratios were also computed for each scale.

Results

16PF

Unidimensionality. Figure 2 presents four typical eigenvalue plots obtained by modified parallel analyses of the 16PF scales (space limitations prohibit an extended presentation of the results for each scale). As recommended by Drasgow and Lissak (1983), a visual comparison of the second eigenvalues was performed. None of the plots exceeded the recommended criteria for data with little or no guessing (These criteria are presented graphically in Drasgow and Lissak.). Therefore, based on the MPA method for dimensionality assessment, we were satisfied that the noncognitive scales of the 16PF were sufficiently unidimensional for IRT analysis.

The values of four commonly utilized goodness of fit indices that indicate how well a one- factor model fit each personality scale are reported in Table 2. These include the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980), Goodness of Fit Index (GFI; Tanaka & Huba, 1984), Normed Fit Index (NFI; Bentler & Bonett, 1980) and Comparative Fit Index

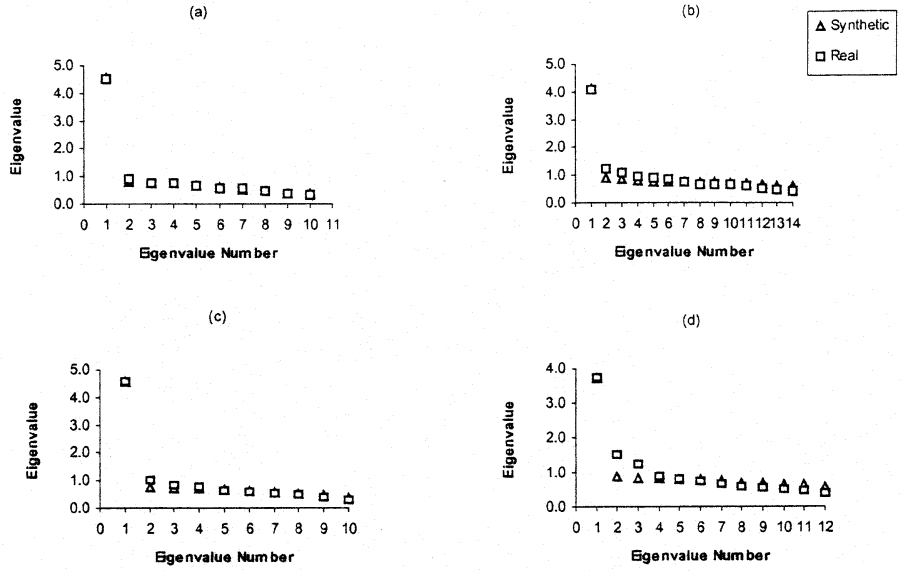

Figure 2
Modified Parallel Analysis Eigenvalue Plots Obtained from Item Responses from Synthetic and Real Data Sets for (a) the Emotional Stability, (b) Openness to Change, (c) Social Boldness, and (d) Dominance Scales of the Fifth Edition of the 16PF

(CFI; Bentler, 1990). As we already indicated, there is currently no rule available that allows us to determine sufficient unidimensionality specifically for IRT analysis using these indices. The RMSEA values were .08 or less for almost all 16PF scales indicating an adequate fit of the one-factor model (Browne & Cudeck, 1993). The GFI indices also supported that conclusion (values of .95 and above indicate good fit). The NFI and CFI indices for some scales were somewhat lower than the recommended .95 level (Hu & Bentler, 1999).

Model-Data Fit. Table 3 contains a summary of the adjusted χ^2/df ratios for five of the 16PF personality scales. Relatively small χ^2/df statistics for single items were obtained for all models and scales (i.e., the average adjusted χ^2/df for single items ranged between 0.87 and 3.0 for all models). These results are comparable to those obtained in previous investigations of cognitive ability data (see Drasgow et al., 1995) in that they indicate good correspondence between the estimated model and the observed data. This situation changed, however, when item pairs and triples were taken into account. The SGR model had noticeably larger χ^2/df ratios

Table 2
The CFA Fit Indices of One-factor Model for the 16PF and "Big Five" Scales

Scale	RMSEA	GFI	NFI	CFI
16 PF				
Warmth	.07	.95	.80	.80
Emotional Stability	.05	.98	.95	.95
Dominance	.05	.98	.89	.89
Liveliness	.07	.97	.87	.87
Rule-Consciousness	.07	.96	.86	.87
Social Boldness	.07	.96	.94	.94
Sensitivity	.08	.95	.87	.87
Vigilance	.05	.98	.94	.95
Abstractedness	.07	.95	.85	.85
Privateness	.08	.95	.88	.88
Apprehension	.07	.97	.90	.90
Openness to Change	.05	.97	.82	.83
Self-Reliance	.06	.97	.93	.93
Perfectionism	.08	.96	.85	.86
Tension	.07	.97	.90	.91
Impression Management	.07	.95	.75	.75
"Big Five"				
Emotional Stability	.10	.92	.91	.91
Extraversion	.07	.96	.94	.95
Intellect	.10	.92	.89	.90
Agreeableness	.06	.96	.95	.96
Conscientiousness	.06	.97	.95	.96

than any other model. For example, 90 out of 91 possible ratios for item pairs and 358 out of 364 possible ratios for item triples fell in the "very large misfit" interval (≥ 5) for the Openness to Change scale. The other 16PF scales, also exhibited χ^2/df ratios of similar magnitude for items pairs and triples. These findings indicate that the SGR model did not fit the 16PF data adequately.

The χ^2/df ratios for item pairs and triples for the dichotomous models (i.e. 2PL, 3PL and MFS) and polytomous MFS model varied in magnitude

Table 3
Frequency, Means and *SD* of χ^2/df Ratios for Five 16PF Scales

Model	Item	Frequency Distribution of Adjusted χ^2 to *df* Ratio						Mean	SD
		<1	1 - <2	2 - <3	3 - <4	4 - <5	>=5		
				16PF Liveliness Scale					
2PL									
	Singles	6	2	2	0	0	0	1.21	0.82
	Doubles	6	9	17	6	2	5	3.26	4.40
	Triples	1	16	36	32	17	18	4.08	3.56
3PL									
	Singles	6	2	2	0	0	0	1.20	0.76
	Doubles	6	9	18	3	6	3	3.19	4.44
	Triples	1	18	42	31	14	14	3.93	3.6
MFS dichotomous									
	Singles	6	2	2	0	0	0	1.13	0.75
	Doubles	8	22	10	4	0	1	1.89	1.70
	Triples	6	64	32	9	1	8	2.19	1.35
SGR									
	Singles	5	3	2	0	0	0	1.15	0.67
	Doubles	0	0	0	0	2	43	9.82	3.89
	Triples	0	0	0	0	2	118	8.53	2.86
MFS polytomous									
	Singles	6	1	1	2	0	0	1.61	1.15
	Doubles	4	17	17	3	3	1	2.38	1.89
	Triples	0	50	57	5	0	8	2.37	1.08
				16PF Sensitivity Scale					
2PL									
	Singles	8	2	0	1	0	0	0.98	0.83
	Doubles	12	12	7	4	6	14	4.05	4.21
	Triples	8	20	30	26	16	65	5.45	4.16
3PL									
	Singles	8	2	1	0	0	0	0.87	0.60
	Doubles	13	12	6	5	4	15	3.89	4.13
	Triples	5	27	29	27	16	61	5.23	4.02
MFS dichotomous									
	Singles	2	1	5	0	1	2	2.91	1.74
	Doubles	1	19	20	10	3	2	2.61	1.23
	Triples	0	54	81	19	9	2	2.42	0.87
SGR									
	Singles	8	2	1	0	0	0	0.99	0.60
	Doubles	0	1	0	6	4	44	7.76	3.46
	Triples	0	0	2	7	20	136	7.12	2.58

Table 3 (cont.)

Model	Item	Frequency Distribution of Adjusted χ^2 to *df* Ratio						Mean	SD
		<1	1 - <2	2 - <3	3 - <4	4 - <5	>=5		
MFS polytomous									
	Singles	4	4	2	0	1	0	1.55	1.09
	Doubles	3	22	12	5	8	5	2.68	1.39
	Triples	0	52	66	34	11	2	2.58	0.93
		16PF Openness to Change Scale							
2PL									
	Singles	7	3	2	1	0	1	1.72	1.31
	Doubles	7	39	23	7	4	11	2.59	1.81
	Triples	0	139	91	59	43	32	2.83	1.36
3PL									
	Singles	7	3	2	1	0	1	1.66	1.29
	Doubles	10	43	17	8	3	10	2.56	1.94
	Triples	5	134	101	47	47	30	2.79	1.41
MFS dichotomous									
	Singles	7	3	1	2	1	0	1.64	1.21
	Doubles	16	42	19	5	2	7	2.10	1.51
	Triples	20	186	84	42	25	7	2.17	1.09
SGR									
	Singles	4	6	3	0	0	1	1.69	1.31
	Doubles	0	0	0	0	1	90	9.67	2.71
	Triples	0	0	0	0	6	358	8.32	2.41
MFS polytomous									
	Singles	6	5	2	0	0	1	1.49	1.16
	Doubles	1	35	37	11	4	3	2.40	0.98
	Triples	0	138	195	30	1	0	2.23	0.52
		16PF Tension Scale							
2PL									
	Singles	5	3	0	0	0	2	2.08	2.42
	Doubles	8	9	12	6	4	6	3.04	2.92
	Triples	1	33	39	21	7	19	3.41	2.22
3PL									
	Singles	7	1	0	0	0	2	1.97	2.34
	Doubles	12	9	10	7	2	5	2.84	2.75
	Triples	0	32	37	26	6	19	3.37	2.03
MFS dichotomous									
	Singles	2	4	1	0	1	2	3.00	2.85
	Doubles	5	16	9	6	6	3	2.72	2.09
	Triples	6	40	49	17	0	8	2.51	1.47

Table 3 (cont.)

Model	Item	Frequency Distribution of Adjusted χ^2 to *df* Ratio						Mean	SD
		<1	1 - <2	2 - <3	3 - <4	4 - <5	>=5		
SGR									
	Singles	2	3	3	1	0	1	2.09	1.32
	Doubles	0	0	0	0	1	44	9.76	3.77
	Triples	0	0	0	1	6	113	7.87	2.44
MFS polytomous									
	Singles	2	4	3	1	0	0	2.02	0.86
	Doubles	0	23	14	5	1	2	2.33	1.32
	Triples	0	80	22	14	4	0	2.06	0.75
		16PF Impression Management Scale							
2PL									
	Singles	8	1	1	1	0	1	1.56	1.59
	Doubles	7	14	19	3	8	15	4.72	5.67
	Triples	1	30	46	37	16	90	5.81	4.50
3PL									
	Singles	8	1	1	1	0	1	1.53	1.46
	Doubles	5	17	16	5	9	14	4.68	5.67
	Triples	1	31	48	35	15	90	5.76	4.45
MFS dichotomous									
	Singles	3	5	2	2	0	0	1.80	0.84
	Doubles	4	31	19	6	4	2	2.49	2.46
	Triples	1	94	78	26	9	12	2.71	2.04
SGR									
	Singles	6	3	1	0	2	0	1.56	1.31
	Doubles	1	5	4	4	7	45	6.81	3.72
	Triples	0	0	7	20	33	160	6.25	2.02
MFS polytomous									
	Singles	7	4	0	0	0	1	1.40	1.33
	Doubles	2	34	10	8	6	6	2.66	1.83
	Triples	1	82	77	35	16	9	2.56	1.09

across the scales. Essentially, the results can be grouped into two categories. One set of scales showed good parametric fit, while the other set demonstrated a lack of parametric fit. In particular, the Openness to Change and Tension scales appeared to be fit well by dichotomous logistic models. Both 2PL and 3PL models had the majority of χ^2/df ratios for item pairs and triples in the very small to medium range (≤ 3). For example, the mean of the 2PL ratios for the Openness to Change scale was 2.59 for pairs

and 2.83 for triples. The more flexible 3PL model showed little improvement over the simpler 2PL model: The mean ratios were 2.56 for pairs and 2.79 triples. This result indicated that introducing a lower asymptote parameter had little effect on model-data fit and that the 2PL model was appropriate for these scales. The fit of the dichotomous and polytomous MFS models to the Openness to Change scale was somewhat better (2.10 and 2.40 for pairs and 2.17 and 2.23 for triples, respectively), but the improvements over parametric models were relatively minimal. (An explanation for the improved fit is suggested by the fit plots, and is discussed below.)

The data from the second category of 16PF scales (e.g., Sensitivity and Impression Management scales) were not fit well by the logistic models. The majority of χ^2/df ratios for item pairs and triples for the 2PL and 3PL models were in the moderately large to very large range (≥ 3), with means generally above 4.5. The nonparametric MFS models, however, had noticeably smaller ratios that were mostly in the small to medium range. For example, the Impression Management scale had mean χ^2/df ratios for item pairs and triples of 2.49 and 2.71 for the dichotomous MFS model, while the means for the 2PL model were 4.72 and 5.81, respectively. The 3PL model again showed little improvement over the 2PL model (means of 4.68 and 5.76). Together, these results indicate that the parametric IRT models did not fit well.

Fit indices for the polytomous MFS model were very similar to those for the dichotomous MFS model. One might expect that a polytomous model would always fit better because more parameters are estimated. However, increasing the number of model parameters generally leads to larger estimation errors that may offset any improvements in predicting patterns of responses for item pairs and triples. Moreover, the polytomous MFS model makes more specific predictions than the dichotomous MFS model (i.e., it predicts the value of polytomous responses), again introducing greater possibility for error.

In our analyses, the polytomous MFS model utilized more parameters than the dichotomous MFS model but did not show much difference in fit. Evidently, increased estimation error balanced with the improved prediction provided by the polytomous model. Another reason that fit was not improved may be rooted in the data structure of the 16PF. The middle option of the 16PF has relatively low endorsement rates (5 to 10 percent for most items) as compared to the other two options, so relatively small amounts of information are lost during dichotomization. Not surprisingly, the dichotomous MFS model accounted for the data nearly as well as the more complex polytomous MFS model.

Fit plots were used to obtain visual representations of model-data. One fit plot per item was constructed for each of the dichotomous models and three plots per item (one for each response option) were constructed for the SGR and polytomous MFS models. Due to space limitations, we selected two typical 16PF items to illustrate the results. Item 13 from the Openness to Change scale is one of the items from a scale showing good parametric model-data fit. Fit plots for that item are presented in Figures 3 and 4. Item 5 from the Sensitivity scale is typical of items that were not fit well by parametric models; the fit plots are presented in Figures 5 and 6. The vertical lines in each figure describe the approximate 95% confidence intervals for the empirical item/option response functions.

An examination of Figure 3 reveals considerable misfit for the SGR model; namely, there are large discrepancies between the empirical proportions and the estimated option response functions. Note that this systematic error was undetected by the χ^2/df ratio for this item when examined individually (χ^2/df ratio was 0.78), but it became apparent when the item was paired with other similarly misfit items from the Openness to Change scale. The estimated functions for the dichotomous IRT models are

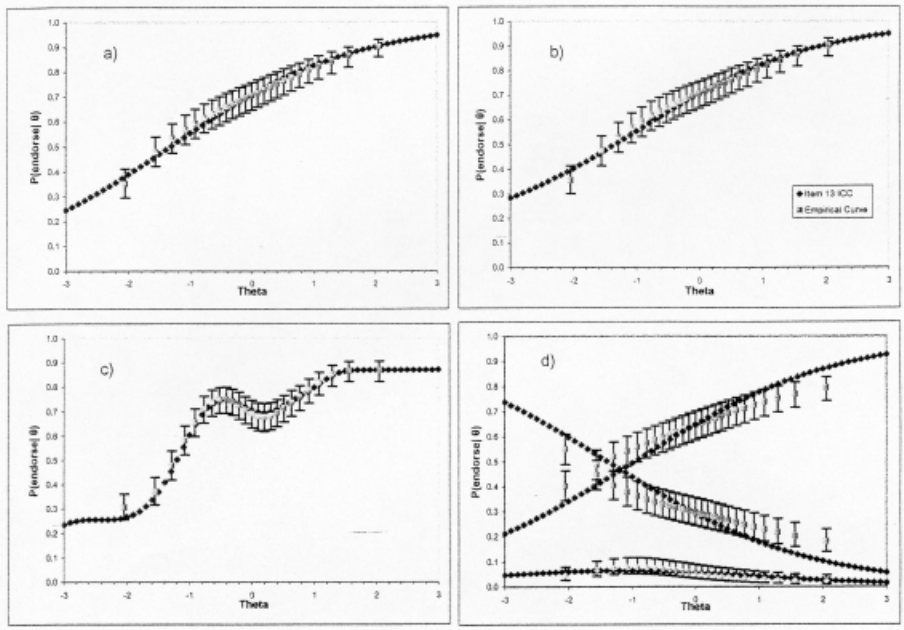

Figure 3
Fit Plots for Item 13 of the Openness to Change Scale of the Fifth Edition of the 16PF: (a) 2PL, (b) 3PL, (c) Dichotomous MFS, (d) SGR

much more satisfactory than those estimated for the SGR model, because they fit the empirical proportions better. Note, however, that both models adequately described the lower tail of the response function, so there was little need for a lower asymptote parameter. The shape of the item response function estimated by the dichotomous MFS model resembled the shape of 2PL and 3PL logistic functions with the exception of a moderate oscillation near $\theta = 0$. Also, note that the 2PL, 3PL, and MFS models had similar patterns of χ^2/df ratios.

Figure 4 presents fit plots for the polytomous MFS model with each response option plotted separately. Visual inspection indicated that the fits between the estimated and empirical curves for the polytomous MFS model were better than those for the SGR model. The shapes of the nonparametric option response functions for the lowest and highest response categories were generally similar to the corresponding SGR functions. However, the middle response option differed: It was bimodal for the MFS model. Further research is needed to determine whether that result was due to estimation error or may be substantively important.

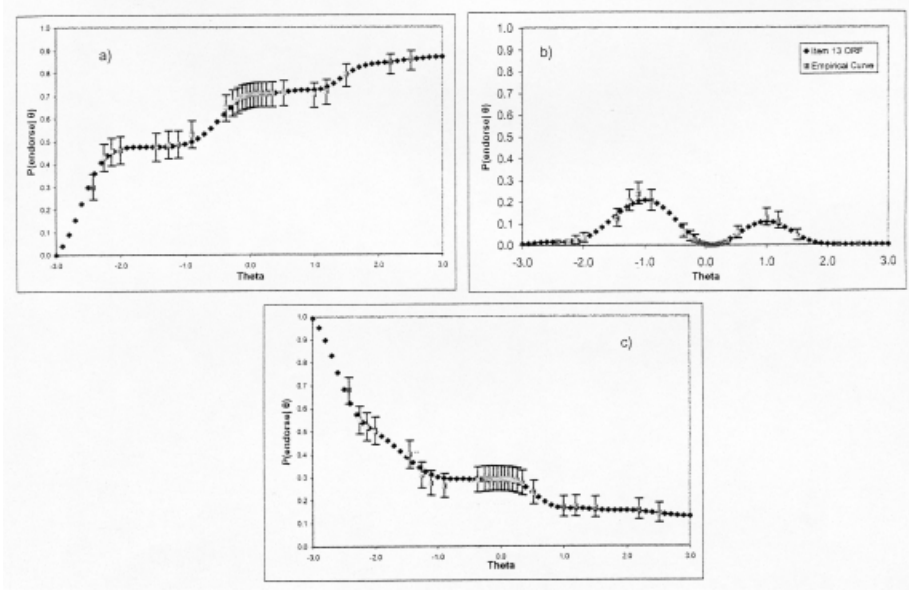

Figure 4
Fit Plots for the Polytomous MFS Model (Item 13 of the Openness to Change Scale of the 16PF): (a) Positively Scored Option, (b) Middle Option, (c) Negatively Scored Option

Figures 5 and 6 present fit plot results for an item that was fit poorly by the parametric models. The fit plots for the dichotomous IRT models are interesting and somewhat surprising. The estimated functions for all three models fit the empirical proportions nearly perfectly. Nonetheless, the χ^2/df ratios for item pairs and triples were large for parametric models: 4.05 and 5.45 for the 2PL and 3.89 and 5.23 for the 3PL. Due to the size of these ratios for pair and triples, we expected the 2PL and 3PL fit plots to exhibit the kind of misfit seen for the SGR model (i.e., we expected the empirical item response function to be consistently above the estimated IRF at some trait levels and below it at other trait levels). Contrary to expectations, the fit plots for 2PL, 3PL and MFS models showed very good fit. The shape of the MFS curve provided some insight about why the χ^2/df ratios for parametric models were large for item pairs and triples and small for single items: Its profound nonmonotonicity suggests that s-shaped item response functions cannot adequately explain response patterns. Thus, the 2PL and 3PL models were capable of modeling single item responses but failed when predictions about patterns of responses for two or three items were needed.

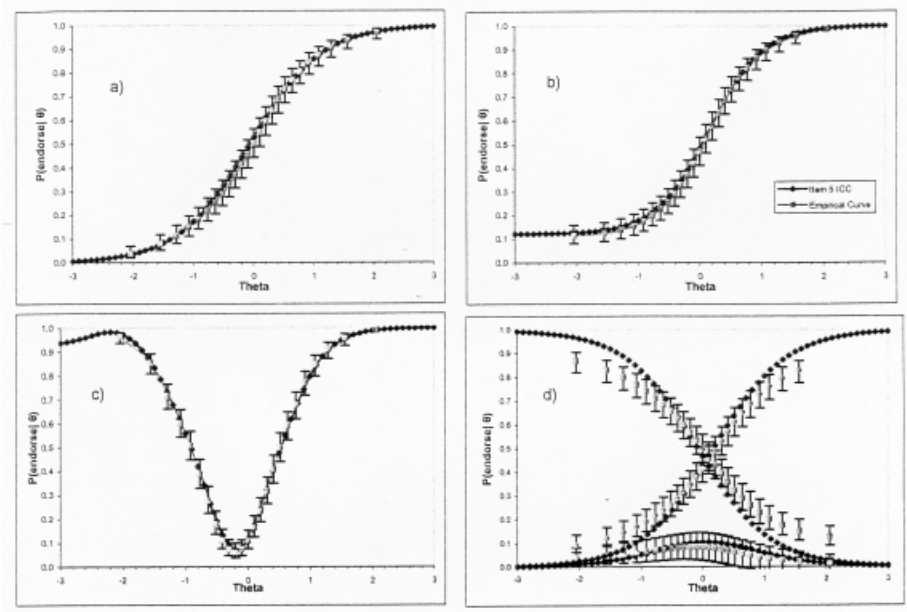

Figure 5
Fit Plots for Item 5 of the Sensitivity Scale of the Fifth Edition of the 16PF: (a) 2PL, (b) 3PL, (c) Dichotomous MFS, (d) SGR

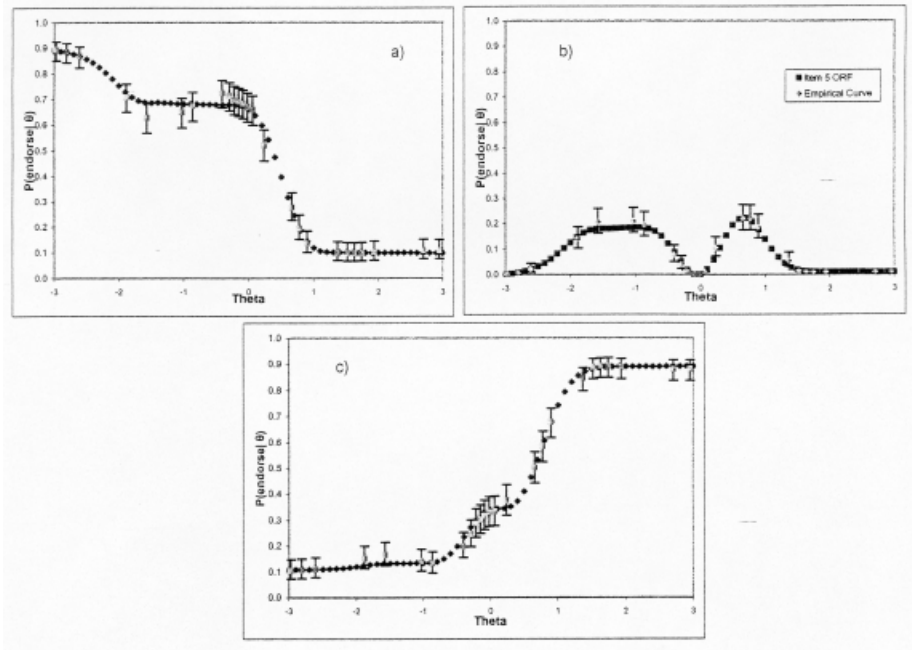

Figure 6
Fit Plots for the Polytomous MFS Model (Item 5 of the Sensitivity Scale of the 16PF): (a) Positively Scored Option, (b) Middle Option, (c) Negatively Scored Option

Turning now to the results for the polytomous analyses, there were substantial discrepancies between the empirical proportions and the estimated option response functions for the SGR model and closer correspondence for the polytomous MFS model. Note that the response function for the most positive option was monotonically increasing in panel C of Figure 6. However, this qualitative feature of the option response function was obtained by sacrificing normality; that is, we did not require latent trait scores to be normally distributed.

Big Five

Unidimensionality. MPA plots of the Big Five scales indicated that they were sufficiently unidimensional for IRT analysis. The CFA goodness of fit indices are reported in Table 2. The GFI, NFI and CFI indices indicate adequate fit of the one factor model for all Big Five scales. The RMSEA

scores for Emotional Stability and Intellect scales, however, were somewhat large.

Model-Data fit. Table 4 contains the adjusted χ^2/df ratios for the three of the Big Five personality scales. None of the parametric models appeared to fit the Big Five data well. Large adjusted χ^2/df statistics for single items, pairs and triples for the SGR model clearly showed that this model did not fit the data. The fit of dichotomous parametric models was also poor, even though the mean χ^2/df statistics for single items was moderate for some scales (e.g., the mean for the Emotional Stability scale was 2.90 for the 2PL model). The adjusted mean χ^2/df statistics for item pairs and triples ranged from 4.29 to 14.22 for the Big Five scales, indicating a considerable lack of parametric fit. These results were different from those obtained in the 16PF investigation where χ^2/df statistics were small for single items and relatively large for item pairs and triples only on some of the scales.

The nonparametric dichotomous MFS model, however, had noticeably smaller ratios for all Big Five scales. The adjusted χ^2/df statistics for item singles were relatively small for this model. Item pairs and triples statistics

Table 4
Frequency, Means and SD of χ^2 /*df* Ratios for Three "Big Five"Scales

Model	Item	Frequency Distribution of Adjusted χ^2 to *df* Ratio						Mean	SD
		<1	1 - <2	2 - <3	3 - <4	4 - <5	>=5		
		Big Five Agreeableness Scale							
2PL									
	Singles	6	0	0	0	0	4	14.22	27.39
	Doubles	7	1	3	1	1	32	12.02	11.85
	Triples	3	5	6	9	14	83	9.81	6.37
3PL									
	Singles	6	0	0	0	0	4	14.19	26.98
	Doubles	8	0	2	2	1	32	11.25	11.41
	Triples	3	4	11	8	13	81	9.32	6.22
MFS									
	Singles	6	2	0	0	1	1	1.93	3.65
	Doubles	25	4	2	2	2	10	2.71	4.16
	Triples	49	9	11	9	5	37	3.18	3.38
SGR									
	Singles	2	0	1	2	0	5	7.30	6.46
	Doubles	1	0	2	5	7	30	7.08	3.91
	Triples	0	3	13	23	20	61	5.76	2.93

Table 4 (cont.)

Model	Item	Frequency Distribution of Adjusted χ^2 to *df* Ratio						Mean	SD
		<1	1 - <2	2 - <3	3 - <4	4 - <5	>=5		
		Big Five Conscientiousness Scale							
2PL									
	Singles	3	1	0	0	1	5	6.72	6.70
	Doubles	4	3	1	6	2	29	7.30	4.76
	Triples	3	6	6	14	14	77	6.89	3.49
3PL									
	Singles	3	0	1	0	1	5	6.65	6.38
	Doubles	3	5	2	3	5	27	6.60	4.31
	Triples	4	8	9	14	10	75	6.34	3.40
MFS									
	Singles	7	2	0	0	0	1	1.13	3.64
	Doubles	25	2	2	2	4	10	3.19	5.29
	Triples	54	7	8	8	8	35	3.36	3.91
SGR									
	Singles	4	0	2	1	0	3	2.73	2.66
	Doubles	2	2	12	4	8	17	4.69	2.53
	Triples	2	30	34	24	12	18	3.18	1.54
		Big Five Emotional Stability Scale							
2PL									
	Singles	5	2	0	0	0	3	2.90	4.27
	Doubles	14	4	8	4	2	13	4.29	6.24
	Triples	20	14	13	21	10	42	5.39	5.66
3PL									
	Singles	6	0	1	0	0	3	3.92	6.19
	Doubles	14	3	4	6	3	15	4.67	7.20
	Triples	21	18	17	10	14	40	5.32	6.08
MFS									
	Singles	8	0	1	0	0	1	0.98	2.32
	Doubles	30	2	1	1	1	10	2.75	5.11
	Triples	51	8	11	10	9	31	3.18	3.68
SGR									
	Singles	5	0	0	1	0	4	4.85	6.26
	Doubles	1	0	1	5	5	33	8.28	6.77
	Triples	0	1	10	25	20	64	6.10	3.72

were close to or below 3.0 indicating good fit of the MFS model to the data. This is surprising because dichotomous MFS was the most complex model for the Big Five items; estimation errors resulting from the small (N = 930) calibration samples were expected to create substantial challenges for cross-validation. Doubtlessly such estimation errors occurred, but the nonparametric model's ability to fit the Big Five data was not substantially affected.

Fit plots for the Big Five results are not presented due to space limitations. Recall that *within* each 16PF scale, the items tended to have similarly shaped MFS response functions. In contrast, there was substantial variation in the shapes of the MFS response functions within the Big Five scales. Whereas some items seemed to conform to the monotonic, s-shape of the logistic function, other items had response functions that were clearly nonmonotonic.

Discussion

The main conclusion from this study is that the issue of fitting IRT models to personality data is more complicated than previously suggested. Our research differed markedly from other investigations of model-data fit because we fit several increasingly complicated IRT models to two personality inventories and used improved statistical and graphical fit procedures. In contrast to previous researchers that focused on fit statistics for item singles, we examined interactions among items by studying item pairs and triples. The results showed that the 2PL and 3PL logistic models did not consistently fit the data from the 16PF or the Big Five. Specifically, some of the 16PF scales were fitted well by the 2PL model but others were not. None of the 16PF or Big Five scales were fitted adequately by the SGR model. The nonparametric MFS models appeared to fit all 16PF and Big Five scales; note, however, that the item response functions differed markedly in their shapes across scales.

Two important questions arise from these findings: (a) why don't traditional logistic models fit personality scales consistently? and (b) what are the effects (if any) of model-data misfit on the applications of IRT in personality measurement? Although more research is clearly needed to adequately answer these questions, we will share some of our thoughts on these topics.

Our first hypothesis about the source of model-data misfit for personality scales was very straightforward. We were hoping to identify a specific type of item that could not be fit by IRT models. As an anonymous reviewer suggested, the negatively keyed items might be problematic. To test this

hypothesis, we created a group of positively keyed items and a group of negatively keyed items for each scale, computed the average chi-square fit statistic for each group, and examined the differences. No noticeable differences were found. For example, the conscientiousness scale of the Big Five inventory had 6 positively keyed items and 4 negatively keyed items. The average adjusted χ^2/df statistic for positive item singles was 8.08 while the same statistic for negatively keyed items was 4.52. Both exceeded the recommended 3.0 level. The adjusted χ^2/df statistics for all pairs of positive and negative items examined separately were 6.94 and 6.48 respectively. This indicates essentially no difference in model-data fit between the two groups of items. Similar results were obtained for other Big Five and 16PF scales.

We also looked at the content of items in search of possible answers. Because the 16PF is a copyrighted instrument, we can only discuss the content of items from the Big Five inventory (public domain), but similar conclusions can be reached for the 16PF data. Let us consider the following items from the Conscientiousness scale of the Big Five inventory: "I pay attention to details," "I make a mess of things," "I like order." All three items showed good fit for the 3PL model at the level of single items (adjusted χ^2/df ranged from 0 to 2.41). However, when we examined patterns of responses to these items, only one pair (item 1 and item 3) fit the model well (adjusted χ^2/df was 0.01); the other two pairs showed considerable misfit (adjusted χ^2/df were 13.37 and 6.68). It is unclear why the local independence assumption for the item "I like order" is violated when it is paired with the item "I make mess of things," but not violated when it is paired with the item "I pay attention to details." Overall, we were unable to identify any group of items or any content problems that would result in consistently better or worse fit by the parametric models.

A reviewer of this manuscript presented an interesting explanation for the above example of misfit. This reviewer noted that one possible explanation of local dependence for the "I like order" and "I make a mess of things" is that these two items assess a "neatness" dimension, while "I pay attention to details" assesses a "vigilance to fine points" dimension. It is quite possible that individuals may endorse the "neatness" items and not endorse the "details" item. The reviewer's comment brings an important point to our discussion. It draws attention to the issue of multidimensionality as the possible cause of misfit. In our article we uncovered a contradiction between the outcomes of the unidimensionality analyses (MPA, CFA) and the results of the two- and three-item chi-square fit statistics. As we noted in the Introduction, the chi-square test is a stronger test for model-data fit because it requires modeling of patterns of responses. When chi-square

statistics are large, two explanations are possible: (a) violation of the unidimensionality assumption and (b) the model fits the data inadequately. It is possible that the chi-square statistics were much more sensitive to violations of unidimensionality than were the MPA and CFA statistics. Note that the RMSEA values in Table 2 indicated only a "fair" fit in the majority of cases, thus, supporting the multidimensionality hypothesis. However, in our previous research we found that cognitive ability scales with similar MPA and CFA statistics did not have large chi-square statistics for the 3PL model. That fact directs our attention to the possibility that traditional logistic models may be inappropriate for personality items.

Our second hypothesis about the source of model-data misfit for personality scales concerns the nature of responses to personality items. Many researchers have argued that individuals might respond differently to personality items than cognitive ability items. Cronbach (1960) was one of the first psychometricians who emphasized the distinction between tests of typical and maximum performance. Maximum performance tests are designed to reflect what an individual "can do" (i.e., can provide a correct answer to an algebra question), whereas typical performance tests are designed to reflect what an individual "will do" or "usually does" (i.e., talks to many different people at parties). Traditionally the typical vs. maximum performance distinction was maintained primarily to separate tests into two domains: the cognitive ability domain and the nonability domain (e.g., personality). Recently, Campbell (1990) and Sackett, Zedeck and Fogli (1988) emphasized that this distinction goes beyond the simple classification issue and involves differences in test responding. While responding to cognitive ability tests, the individual is aware that his/her performance is being monitored and, hence, he/she is motivated to do the task, expend a high level of effort on the task and maintain a high level of effort throughout the measurement period. Clearly, the behavior of the individual is severely restricted by the testing situation. In contrast, individuals have a great array of choices regarding the time on task, level of effort, and persistence of effort while answering typical performance tests. Thus, it is possible that traditional IRT models can model well the constrained responding to maximum performance tests, but cannot model the complexity of responding to typical performance tests. In our study, we found that the more complicated MFS model adequately captured the responses to personality items while logistic models failed to provide such consistency.

Additional arguments regarding the fundamental difference between responding to personality and cognitive ability items can be found in the attitude measurement literature. There has been an increasing number of publications suggesting that the Likert approach to scale construction may

have limitations for noncognitive items (e.g., Andrich, 1988, Roberts, Laughlin, & Wedell, 1999). The Likert procedure assumes that responding to personality items follows a *dominance (or cumulative) response process* (Coombs, 1964); namely, the individual has a high probability of endorsing an item if he/she is located above the item on the underlying continuum. However, several researchers (Andrich, 1996; Roberts et al., 1999) have argued that this assumption might be flawed for attitude items. They have found that participants generally use some type of *ideal point response process*. The premise of this process is that the probability of a person endorsing an item depends on both the location of the person and the position of the statement along the latent trait continuum. People tend to agree with statements having scale values similar to their own, while they tend to disagree with statements having scale values that are *either* more or less extreme. Thus, ideal point models have nonmonotonic item response functions.

We believe that personality items are essentially attitude statements about people's own behavior. Consequently, we are not particularly surprised that logistic models have some problems fitting personality items. The adequate performance of logistic models on some personality scales may be explained by the fact that items were preselected to be so extreme that we were unable to observe individuals with extremely high trait levels who began to agree less with statements because the item did not reflect the extremity of their position on the trait continuum. Scales that were fitted poorly by logistic models may have had less extreme items and more individuals with extreme trait levels. While that did not affect the fit of individual items much, the misfit increased exponentially as pairs and triples of these less extreme items were considered. In our study, the Tension scale of the 16PF was fit well by the 3PL model while, the Sensitivity scale was fit poorly. It may be the case that it was easier to generate extreme items for the Tension scale of the 16PF than for the Sensitivity scale. Similarly, it may be easier to write extreme items for the relatively narrow constructs of the 16PF than for the broad constructs of the Big Five. Currently, we are conducting a series of studies that explore the issue of fit of ideal point models to personality data.

In summary, the fit of the traditional IRT models to personality data is a matter of concern. How much the violation of parametric model assumptions illustrated in this paper affects the applications of traditional IRT models in personality measurement is another question. One interesting result has been found by Stark, Chernyshenko, Chan, Drasgow and Lee (2001) while conducting a differential item functioning (DIF) analysis of the 16PF scales. In that study, two approaches to DIF were used: a parametric approach based on IRT (Lord's chi square; Lord, 1980) and a nonparametric approach (SIBTEST; Stout &

Roussos, 1996). In general, the same items were identified as having DIF by parametric and nonparametric approaches for scales that were fit well by the 3PL model, but *different* items were identified by SIBTEST and Lord's chi-square as having DIF for scales that had model-data misfit. For personnel selection, where DIF studies are used widely as part of test development and validation, disagreement between DIF methods poses a considerable problem. If the items flagged as DIF depend on the analysis then we do not know which items are truly a problem and, thus, we cannot improve our measurement instrument to satisfy today's testing standards. Other problems that result from applying poorly fitted IRT models may involve difficulties in assessing the quality of individual items (e.g., finding highly discriminating items), creating parallel test forms or selecting items during computer adaptive testing. In contrast, selection of an IRT model that provides a good fit would enable personnel psychologists to take full advantage of modern psychometric theory and to improve many aspects of personality measurement.

We feel that future research should focus on two issues. First, from a psychometric viewpoint, more complex models (e.g., MFS) should be considered when modeling existing personality data. Second, other approaches to personality scale construction (e.g., ideal point) should be attempted.

References

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12*, 33-51.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49*, 347-365.

Barrick, M. R. & Mount M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M. R. & Mount, M. K. (1993). Autonomy as a moderator of relationships between the Big Five personality dimensions and job performance: A meta-analysis. *Journal of Applied Psychology, 78*, 111-118.

Becker, P. (1989). *Der Trier Persönlichkeitsfragebogen (TPF) Handanweisung* [The Trier Personality Inventory (TIP) Manual]. Göttingen, West Germany: Hogrefe.

Bentler, P. M. (1990). Comparative fix indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

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

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.

Campbell, J. P. (1990). The role of theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*: *Vol. 2.* (2nd ed., pp. 39-73). Palo Alto: Consulting Psychologists Press.

Chan, K. (1999). *Toward a theory of individual differences and leadership: Understanding the motivation to lead*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Conn, S. & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual.* Champaign, IL: Institute for Personality and Ability Testing.

Cooke, D. J. & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist-Revised. *Psychological Assessment, 9,* 3-14.

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

Cronbach, L. J. (1960). *Essentials of psychological testing*. New York: Harper.

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

Drasgow, F. & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68,* 363-373.

Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory. *Journal of Cross-Cultural Psychology, 2,* 133-148.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429 456.

Flanagan, W., Raju, N. S., & Haygood, J. M. (1998, April). Impression management, measurement equivalence, and personality factors: Can IRT be used to determine the impact of faking. In F. Drasgow (Chair), *Improvements in measurement: Application of item response theory*. Symposium conducted at the 13th annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525-546.

Goldberg, L. R. (1997). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe, Vol. 7*. The Netherlands: Tilburg University Press.

Goldberg, L. R. (1998, March 18). *International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences*. [On-line]. Available HTTP: http//ipip.ori.org/ipip/ipip.html.

Gratias, M. & Harvey, R. J. (1998, April). *Gender and ethnicity-based differential item functioning on the MBTI*. Paper presented at the 13th Annual Conference of the Society of Industrial and Organizational Psychology, Dallas, TX.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hare, R. D. (1991). *The Hare Psychopathy Checklist-Revised.* Toronto, Ontario, Canada: Multi-Health Systems.

Harkness, A. R. & McNulty, J. L. (1994). The Personality Psychopathology Five (PSY-5): Issues from the pages of a diagnostic manual instead of a dictionary. In S. Strack & M. Lorr, (Eds.), *Differentiating normal and abnormal personality.* (pp. 291-315). New York: Springer.

Hattie, J. A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49-78.

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol. 2* (2nd ed., pp. 873-919). Palo Alto: Consulting Psychologists Press.

Hough, L. M. (1996). Personality at work: Issues and evidence. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection.* Hillsdale, NJ: Erlbaum.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 95-108.

Hough, L. M. & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations*. San Francisco, CA: Jossey Bass.

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

Hulin, C. L., Drasgow, F., & Parsons C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Humphreys, L. G. & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10,* 193-205.

Jansen, P. G. & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51,* 69-91.

Jennings, D. & Schmitt, N. W. (1998, April). *Examination of differential item functioning: Subgroups and personality*. Paper presented at the 13th Annual Conference of the Society of Industrial and Organizational Psychology, Dallas, TX.

Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd Ed.). Chicago: SPSS.

Levine, M. V. (1984). *An introduction to multilinear formula score theory.* (Personnel and Training Research Programs, Office of Naval Research, Measurement Series No. 84-4). Arlington, VA: Personnel and Training Research Programs.

Levine, M. V. & Williams, B. A. (1991, May). *An overview and evaluation of non-parametric IRF estimation strategies.* Paper presented at the Office of Naval Research Contractors' Meeting on Model-Based Measurement, Princeton, NJ.

Levine, M. V. & Williams, B. A. (1993). *Nonparametric models for polychotomously scored item responses: Analysis and integration*. Unpublished manuscript.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

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

Mislevy, R. J. & Bock, R. D. (1991). *BILOG user's guide.* Chicago, IL: Scientific Software.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59-71.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679-703.

Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Toquam, J. L., & Wing, A. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. *Personnel Psychology, 43,* 247-276.

Reise, S. P. & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211-233.

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment, 72*, 282-307.

Sackett, P. R., Burris, L. R., & Callahan, C. (1989). Integrity testing for personnel selection: An update. *Personnel Psychology, 42*, 491-529.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482-486.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores.* (Psychometric Monograph No. 18). Iowa City, IA: Psychometric Society.

Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report N0. 79-4). Knoxville: University of Tennessee, Department of Psychology.

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A Festschrift for Frederick M. Lord* (pp. 159-182). Hillsdale NJ: Erlbaum.

Schmit, M. J. & Ryan, A. M. (1997, April). *Specificity of item content in personality tests: An IRT analysis*. Paper presented at the 12th Annual SIOP Conference, St. Louis, MO.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Drasgow, F., & Lee, W. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*, 943-953.

Steiger, J. H. & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. & Roussos, L. (1996). *SIBTEST manual*. Statistical Laboratory for Educational and Psychological Measurement, University of Illinois at Urbana-Champaign.

Tanaka, J. S & Huba, G. J. (1984). Structures of psychological distress: Testing confirmatory hierarchical models. *Journal of Consulting and Clinical Psychology, 52*, 719-721.

Tellegen, A. (1982). *A brief manual for the Multidimensional Personality Questionaire.* Unpublished manuscript, University of Minnesota.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Thissen, D. (1991). *MULTILOG user's guide* (Version 6.0). Mooresville, IN: Scientific Software.

Thissen, D. & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

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

Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. *Journal of Personality, 64,* 545-576.

Williams, B. A. (1986, August). *The shapes of item response functions*. Paper presented at the Office of Naval Research Model-Based Measurement Contractors Conference, Gatlinburg, TN.

Williams, B. A. (1999). *EMPOCC: A computer program for IRT fit plots.* Unpublished manuscript. University of Illinois: Urbana-Champaign

Williams, B. A. & Levine, M. V. (1993). *FORSCORE: A computer program for nonparametric item response theory*. Unpublished manuscript. Department of Educational Psychology, University of Illinois: Urbana-Champaign.

Zickar, M. J. & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71-87.

Zickar, M. & Robie, C. (1998, April). *Modeling faking good on personality items: An item-level analysis*. Paper presented at the 13th Annual Conference of the Society of Industrial and Organizational Psychology, Dallas, TX.

Accepted October, 2000.