Psychological Methods
2004, Vol. 9, No. 3, 354–368

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.3.354

Analyzing Psychopathology Items: A Case for Nonparametric Item Response Theory Modeling

Rob R. Meijer and Joost J. Baneke
University of Twente

The authors discuss the applicability of nonparametric item response theory (IRT) models to the construction and psychometric analysis of personality and psychopathology scales, and they contrast these models with parametric IRT models. They describe the fit of nonparametric IRT to the Depression content scale of the Minnesota Multiphasic Personality Inventory—2 (J. N. Butcher, W. G. Dahlstrom, J. R. Graham, A. Tellegen, & B. Kaemmer, 1989). They also show how nonparametric IRT models can easily be applied and how misleading results from parametric IRT models can be avoided. They recommend the use of nonparametric IRT modeling prior to using parametric logistic models when investigating personality data.

Recently, several authors have introduced and discussed the advantages of applying item response theory (IRT) models (e.g., Embretson & Reise, 2000) to construct personality[1] scales and to explore the structure of personality data sets. For example, Waller, Tellegen, McDonald, and Lykken (1996) contrasted the use of IRT with principal-component factor analysis, and Reise and Waller (2003) discussed the choice of an IRT model to analyze psychopathology-test data. That is, they compared the fit of the two-parameter and three-parameter logistic models (PLMs) on 15 unidimensional factor scales from the Minnesota Multiphasic Personality Inventory—Adolescent (MMPI–A; Butcher et al., 1992). Most studies apply *parametric* IRT models (in particular, the 2PLM and 3PLM) to investigate the quality of personality and psychopathology tests (e.g., Panter, Swygert, Dahlstrom, & Tanake, 1997; Robie, Zickar, & Schmitt, 2001; Steinberg, 1994; Waller, Thompson, & Wenk, 2000).

The aim of the present study was to illustrate the usefulness of nonparametric IRT (NIRT) to construct and to analyze psychopathology and personality scales and tests. In our opinion, the use of NIRT has been underexposed in the recent personality literature (for an exception, see Santor & Ramsay, 1998). We show that these models are very suitable to exploring the psychometric properties of personality data. Interesting in this context is a study by Chernyshenko, Stark, Chan, Drasgow, and Williams (2001), who explored the use of NIRT modeling in personality measurement. Chernyshenko et al. fitted the 2PLM, the 3PLM, a graded response model, and Levine's nonparametric maximum-likelihood formula scoring models to dichotomous and polytomous data of the Sixteen Personality Factor Questionnaire (Conn & Rieke, 1994). They concluded that the nonparametric model provided the best fit of the models considered. Chernyshenko et al. and also Reise and Waller (2003) concluded that the response process underlying personality measurement is less well-understood than the response process in the cognitive domain. This being the case, we argue that using NIRT models based on exploring the simple covariance structure between items and based on nonparametric regression will lead to useful information that (a) can be interpreted very easily by practitioners, (b) avoids forcing the data into a structure they sometimes do not have, and (c) is easily obtained through the use of very user-friendly software programs.

In this article, we show how NIRT may help to avoid misleading results obtained from parametric IRT models and we argue that nonparametric solutions are already available for problems that exist when one is investigating the data structure using parametric IRT models. We are not arguing for the overall replacement of parametric by nonparametric models. Parametric IRT models lead to point estimates of the latent trait and sometimes, in the case of the Rasch (1960) model, to interval scales for measuring re-

Rob R. Meijer, Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, the Netherlands; Joost J. Baneke, Department of Communication Studies, University of Twente.

Correspondence concerning this article should be addressed to Rob R. Meijer, Faculty of Behavioral Sciences, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE, Enschede, the Netherlands. E-mail: r.r.meijer@utwente.nl

[1] In this study, the term *personality* also implies *psychopathology.*

spondents. Such scales can be very convenient, for example, for comparing the results from different tests selected from the same item bank or for the study of change. However, we do think that the emphasis on parametric IRT modeling in clinical assessment may sometimes lead to unnecessarily complicated publications and may sometimes even lead to bad measurement practice. We are not the first to make this observation. There are some excellent publications that discuss the usefulness of NIRT (e.g., Junker & Sijtsma, 2001; Santor, Ramsay, & Zuroff, 1994), but the influence of these publications in clinical and personality assessment has been very modest. Furthermore, we do not pretend to explore the full range of techniques that NIRT modeling has at its disposal. We apply a number of useful methods to explore the data structure of personality tests, and we restrict ourselves to techniques by which we can illustrate how current problems raised in the recent parametric IRT literature can be solved. For more detailed information about different nonparametric fit methods, we refer readers to Ramsay (2000), Stout (1990), and Sijtsma and Molenaar (2002).

This study is organized as follows. First, we introduce the basic principles of parametric IRT and NIRT in relation to the analysis of personality data. In particular we focus on recent results discussed in Reise and Waller (2003), because their results suggest that besides the use of parametric IRT models, NIRT models are useful to analyze personality data. Second, we introduce nonparametric fit methods for two NIRT models without going into technical detail. Third, we illustrate the use of NIRT with empirical data from the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) Depression (DEP) content scale. Finally, we discuss directions for future research in this area.

IRT and Personality Measurement

Parametric IRT Models

Fundamental to IRT is the idea that psychological constructs are latent, that is, not directly observable, and that knowledge about these constructs can only be obtained through the manifest responses of persons to a set of items. IRT explains the structure in the manifest responses by assuming the existence of a latent trait (θ) on which persons and items have a position. IRT models allow the researcher to check whether the data fit the model. The focus in this article is on IRT models for dichotomous items. Thus, one response category is positively keyed (given item score 1), whereas the other is negatively keyed (assigned item score 0). For ability and achievement items these response categories usually reflect the correct and incorrect answers, respectively; for personality and attitude items these response categories usually are labeled as *agree–disagree* or *true–false*.

Most IRT models assume unidimensionality and a specified form for the item response function (IRF) that can be checked empirically. *Unidimensionality* means that all the items in the test measure the same latent trait with the result that persons can be ordered on a linear scale. Related to unidimensionality is the assumption of *local independence*, which holds that the responses in a test are statistically independent conditional on θ. Thus, local independence is evidence for unidimensionality if the IRT model contains person parameters on only one dimension. Furthermore, it is assumed that the probability of endorsing an item is monotonically nondecreasing in θ.

Applying IRT to personality measurement, the probability of endorsing an item g ($g = 1, \ldots, k$) is a function of θ and characteristics of the item. This conditional probability $P_g(\theta)$ is the IRF. It is the probability of a positive response (i.e., *agree* or *true*) among persons with the latent trait value θ. In parametric IRT, $P_g(\theta)$ often is specified using the 1PLM, 2PLM, or 3PLM. The 3PLM (Lord & Novick, 1968, chapters 17–20) is defined as

$$P_g(\theta) = c_g + \frac{(1 - c_g)\exp[a_g(\theta - b_g)]}{1 + \exp[a_g(\theta - b_g)]}, \qquad (1)$$

where a is the item discrimination, b the item location, and c the pseudochance level parameter. The item location b is the point at the trait scale where $P_g(\theta) = 0.5(c + 1)$. The greater the value of the b parameter, the greater the trait value that is required to endorse the item, and, thus, the less popular the item. Less popular items are located to the right or the higher end of the θ scale; popular items are located to the left of the θ scale. When the trait levels are transformed so their mean is 0 and their standard deviation is 1, the values of b vary typically from about -2 (very popular) to $+2$ (very unpopular). The a parameter is proportional to the slope of the IRF at the point b on the trait scale. In practice, a ranges from 0 (flat IRF) to 2 (very steep IRF). Items with steeper slopes are more useful for separating examinees near a trait level θ. The pseudochance level parameter c (ranging from 0 to 1) is the probability of a 1 score for low-ability examinees (that is, as $\theta \to -\infty$). The 2PLM can be obtained by setting $c_g = 0$ for all items, and the 1PLM can be obtained by setting a_g to a constant for all items. In the 2- and 3PLM the IRFs may cross, whereas in the 1PLM the IRFs do not cross.

In personality measurement the 2PLM (or its polytomous extension, the graded response model) is often applied because it is assumed that persons do not guess on the items. However, in a recent study by Reise and Waller (2003) new insights were obtained about the psychometric characteristics of psychopathology data. They compared the fit of the 2PLM and 3PLM on 15 unidimensional factor scales from the MMPI–A item pool. Unidimensionality was investigated using item-level factor analysis, and monotonicity was investigated by inspecting individual item-endorsement proportions against raw-score scales. Relying on chi-square fit statistics as a criterion, they found that the difference in fit between the two models was negligible and that the correlation between the estimated trait levels under both models was uniformly greater than .99. An unexpected finding was that 10% to 30% of the items had substantial lower asymptote parameters ($c > .10$) when the scales were scored in the pathology or nonpathology directions, respectively. The lower asymptote parameters greater than .10 were due to an upper asymptote smaller than 1 when the scales were scored in the nonpathology direction (reversed keying). Reise and Waller argued that the height of the asymptote parameters was attributable to item content ambiguity possibly caused by item-level multidimensionality. For persons at one end of the latent trait scale the item performed well, whereas for persons at the other end of the trait scale the item was ambiguous and undiscriminating. As an example, they discussed Item 142, "feel better than ever,"[2] which is keyed false on the MMPI–A depression scale. They speculated that an adolescent who is depressed will likely respond *false* to this item but that it is unclear how an individual who is not depressed (low trait level) should respond. As Reise and Waller concluded, "individuals without depression can easily think of at least one time in their lives when they felt better than" (p. 176) when filling in the MMPI–A. They concluded that direction of scoring can critically affect IRT analysis in the sense that (a) when one is fitting a parametric model, a nonzero or non-one asymptote may be overlooked, and (b) the item parameters and the test information are not symmetrical for positively versus negatively scored items.

They suggested using a 4PLM (with an extra parameter for the upper asymptote) to characterize responses to noncognitive items, so that an upper bound can be estimated that is smaller than one. This model is given by

$$P_g(\theta) = c_g + \frac{(d_g - c_g)\ \exp[a_g(\theta - b_g)]}{1 + \exp[a_g(\theta - b_g)]}, \qquad (2)$$

where d_g is the upper asymptote parameter. The idea is that even persons with an extreme position on the latent trait will not have a probability of one of agreeing with an item. Instead of using a 4PLM, one can use several nonparametric alternatives as we show below. First, however, we introduce NIRT and discuss some of the fit methods by which nonparametric assumptions can be investigated.

NIRT Models

Although parametric models are used in many IRT applications, nonparametric models and methods are becoming more popular (Cliff & Keats, 2003; Sijtsma & Molenaar, 2002; Stout, 1990). For a comprehensive review of NIRT, see Sijtsma (1998), and for an analysis of cognitive data comparing nonparametric and parametric IRT see, for example, Meijer, Sijtsma, and Smid (1990). In this study, we analyzed the data by means of the Mokken (1971) model of monotone homogeneity (MMH), which is based on estimating covariances between items, and by means of a nonparametric regression model (Ramsay, 2000). Furthermore, we validated some of the results using the program DIMTEST (e.g., Stout et al., 1996). We use these models because they are popular NIRT models (e.g., Mokken, 1997; Sijtsma, 1998) and because user-friendly computer programs are available to operationalize these models, including MSP5 for Windows for the Mokken model (Molenaar & Sijtsma, 2000) and TESTGRAF (Ramsay, 2000) to operationalize nonparametric regression.

Mokken Model

The MMH proposed by Mokken (1971, 1997; see also Molenaar, 1997) assumes unidimensional measurement and an increasing IRF as a function of θ. An important difference between the MMH and the 2PLM and 3PLM is that the IRFs for the MMH need not be of the logistic form. This difference makes the MMH less restrictive for empirical data than logistic models. The MMH allows the ordering of persons with respect to θ using the unweighted sum of item scores (total score). In many personality-testing applications, it often suffices to know the order of persons on a personality trait, for example, in forensic and clinical assessment when measuring the level of depression for referring persons to treatment. Therefore, the MMH is an attractive model for two reasons. First, ordinal measurement of persons is guaranteed when the model applies to the data, and second, the model is not as restrictive with respect to empirical data as the 2- and 3PLM and thus can be used in situations in which these models do not fit the data. Although many psychologists use the sum of item scores or some transformation of it (e.g., *T* scores) without using any IRT model, they do not investigate and thus do not know if they can rank order persons according to their total score. They simply assume that this is the case. This is what Torgerson (1958) called "measurement by fiat" (p.

[2] Throughout the whole study, we use a paraphrased item content for the MMPI and MMPI–A items.

22). Using NIRT, we first investigate if a model applies to the data before we use the total score to rank order persons. Investigating the fit of the model also has the advantages that items can be identified that do not contribute to the rank ordering of persons and that item-score patterns can be identified that are the result of unexpected answering behavior (Meijer, 2003).

Mokken (1971, 1997) also proposed the model of double monotonicity that allows an invariant item ordering across the range of θ values. However, because this model did not describe our data very well, we did not apply it in this study. The usefulness of nonlogistic IRFs is illustrated in Figure 1. Under the MMH model the IRFs may be of the logistic 2PLM form, but they may also be described by linear or exponential equations. IRFs that conform to the MMH model are useful items in the sense that they order persons according to their trait value.

Investigating monotonicity in the MMH. Mokken (1971, 1997) proposed to investigate monotonicity using the scalability coefficient H_{gh} for pairs of items (g, h), the

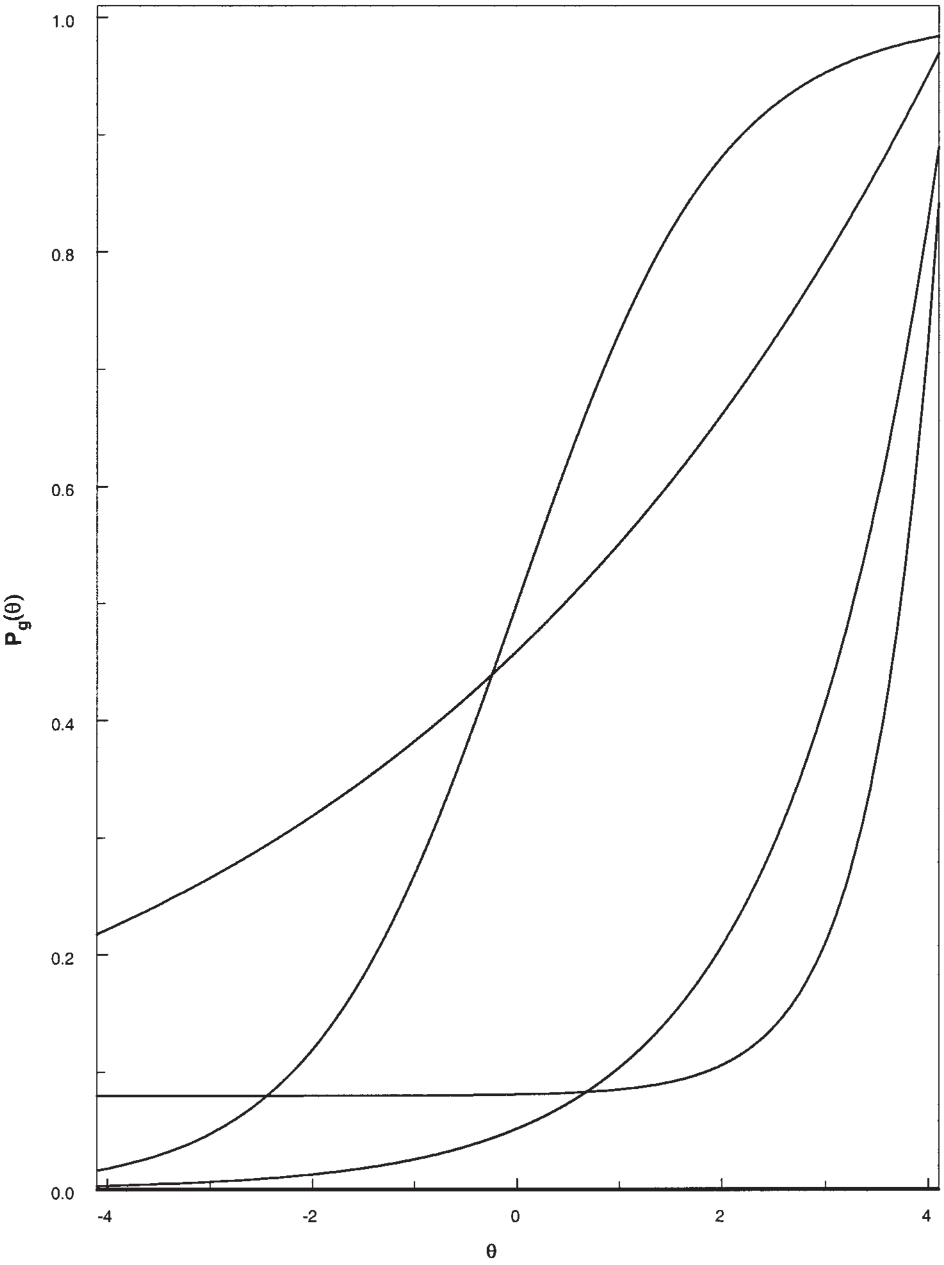

Figure 1. Item response functions that conform to the model of monotone homogeneity.

scalability coefficient H_g for an item with respect to other items in the test, and the scalability coefficient H for the total set of items in the test. These coefficients are given by

$$H_{gh} = \frac{Cov(X_g, X_h)}{Cov_{\max}(X_g, X_h)}, \qquad (3)$$

$$H_g = \frac{\sum_{h \neq g} Cov(X_g, X_h)}{\sum_{h \neq g} Cov_{\max}(X_g, X_h)}, \qquad (4)$$

and

$$H = \frac{\sum\sum_{g<h} Cov(X_g, X_h)}{\sum\sum_{g<h} Cov_{\max}(X_g, X_h)}, \qquad (5)$$

where *Cov* denotes the covariance and X_g and X_h denote the item scores *g* and *h*, respectively. The *H* coefficient is due to Loevinger (1948), although she did not devise the H_g and H_{gh} coefficients. Mokken (1971) used *H* for scale construction and showed that it is useful for measuring the extent to which observed data approach the Guttman (1950) model and that it fulfills this role better than some alternative indices.

Mokken (1971) showed that *H* is a strictly increasing function of the variance of the total score (X_+). Under the MMH, higher positive *H* values reflect higher discrimination power of the items, and as a result, more confidence in the ordering of respondents by means of X_+ (see also De Koning, Sijtsma, & Hamers, 2002). Items with high H_g discriminate well in the group in which they are used. Thus, we can use H_g as a nonparametric analogue to the *a* parameters from logistic IRT models such as the 2PLM and the 3PLM. In practice, *H* and H_g values are between 0 and 1, with H_g values close to 0 implying nearly horizontal IRFs and *H* values close to 1 implying step functions according to the deterministic Guttman (1950) model. For practical test construction purposes, Mokken (1971, p. 185) recommended using $H = .3$ as a lower bound—that is, $.3 \leq H < .4$ denotes a weak scale, $.4 \leq H < .5$ denotes a medium scale, and $H \geq .5$ denotes a strong scale.

Investigating unidimensionality in the MMH. Several nonparametric procedures have been proposed to investigate dimensionality of test data. We first use a relatively simple procedure that is incorporated in MSP5. The results obtained from MSP5 are compared with results obtained from DIMTEST (Stout et al., 1996). The procedure in MSP5 has the advantage that it is in agreement with measurement practice in personality measurement to form facet scales, as we explain below.

For investigating the unidimensionality of an item set, MSP5 contains an automated item selection procedure, based primarily on the interitem covariances and the strengths of the relationship between items and the latent trait(s) as expressed by the item H_g coefficients.[3] Based on such information, clusters of related items measuring a common θ may be identified. The program contains a "bottom-up" procedure that starts by selecting the pair of items for which (a) H_{gh} is significantly larger than 0 and (b) H_{gh} is the largest among the coefficients for all possible item pairs. Then a third item *j* is selected that (c) correlates positively with the items already selected, (d) has an H_j coefficient that is larger than 0, and (e) has an H_j coefficient that is larger than a user-specified value *C*. The program continues to select items as long as items are available that satisfy Conditions c, d, and e. The end result may be one or more item clusters that each tap a different latent trait or latent-trait composite. The substantive interpretation of the clusters is done on the basis of the content of the clustered items and the substantive knowledge one has about the data structure. There are at least three reasons to consider this search algorithm when analyzing personality data.

First, this search algorithm can be used to form homogeneous clusters or facet scales in personality measurement. Facets are item sets with similar content that measure a relatively narrow construct and display high interitem correlations (Reise, Waller, & Comrey, 2000). As Reise et al. discussed, many popular personality measures make extensive use of facets. Facets are often used to build tests consisting of higher order dimensions. We can very simply identify "key" items based on expert opinions or H_{gh} values. Then, on the basis of this, we can build facets, possibly in combination with different lower bound values for the scalability coefficient *H*.

Second, this algorithm can be used to select items that provide sensitive measurement—or equally reliable measurement—across the full range of the trait continuum. To construct a scale with high reliability in a particular trait range, one simply chooses highly discriminating items with item difficulties that span the desired range on the θ continuum. In general, in personality measurement one wants to have high measurement precision across a wide range of trait scores, although on some personality constructs it is difficult to obtain items across the whole range of item difficulties. Theoretical research has shown that items selected in the bottom-up procedure discriminate well across a wide range of item difficulties (Sijtsma & Molenaar, 2002).

[3] This selection procedure does not explicitly use violations of local independence when conditioning on the total score to detect multidimensionality as in the DIMTEST procedure.

Third, Chernyshenko et al. (2001) and Reise and Waller (2003) have suggested that test and item multidimensionality may be the cause of misfit of logistic IRT models to personality data. Hemker, Sijtsma, and Molenaar (1995) showed by means of a simulation study that if multidimensionality is suspected in an empirical data set, well-chosen lower bound values can be used effectively to detect the unidimensional scales. They recommended running the search algorithm several times with varying lower bounds between $C = .0$ and $C = .55$. The typical pattern of results with multidimensional data for varying lower bound C is that with increasing C the following stages can be observed: (a) most or all items are in one scale; (b) two or more unidimensional scales are formed; and (c) two or more smaller scales are formed, and several items are rejected. Hemker et al. indicated that the results from the second stage (when two or more unidimensional scales are formed) should be taken as the final result. With unidimensionality, the typical pattern of results with increasing C is (a) most or all items are in one scale; (b) one smaller scale is found; and (c) one or a few scales are found, and several items are rejected. They recommended in this case considering the scales from the first stage as final. Although they did not consider item multidimensionality, they noted that it is reasonable to assume that because of the correlations between the underlying traits such items will be positively correlated. In this respect the selection algorithm can be used to identify clusters of unidimensional items, which may greatly improve insights into the structure of personality data.

Unidimensionality can also be investigated using DIMTEST. DIMTEST is based on detecting violations of local independence between item pairs when conditioning on the total score that is used as an estimate of θ. When DIMTEST is applied, two subtests of items should be specified from the k items on the test. The first group of M items (Assessment Test1; AT1) consists of items that are dimensionally homogeneous (as determined either by expert opinion or on the basis of a statistical technique such as factor analysis or cluster analysis). The second group of M items (Assessment Test2; AT2) from the $k - M$ items is chosen to be as similar as possible in difficulty level to the first set of items and as dimensionally similar to the remaining items not included in the first subtest. The remaining $k - 2M$ items form the partitioning subtest (PT), on the basis of which persons are partitioned into subgroups according to their total scores. DIMTEST calculates for each subgroup a standardized difference between two variance estimates

$$(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)/SE(\hat{\sigma}_1^2 - \hat{\sigma}_2^2), \tag{6}$$

where $\hat{\sigma}_1^2$ is the actual observed variance of the AT1 or AT2 scores and $\hat{\sigma}_2^2$ is the generalized binomial estimated variance of the AT1 or AT2 scores based on the assumption of unidimensionality. If the test is unidimensional, the standardized difference for AT1 will be the same as for AT2. If not, the standardized difference of AT1 will be larger than AT2. The T statistic calculated by DIMTEST is based on the difference between the AT1 and AT2 standardized differences summed across all the PT groups. This statistic is asymptotically normally distributed with mean 0 and variance 1; values larger than the upper $100(1 - \alpha)$ percentile indicate multidimensionality with alpha denoting the Type I error level. In our analysis we determined the items of AT1 and AT2 on the basis of the default option in DIMTEST that uses factor loadings for selecting items. For recent developments in the context of DIMTEST, see Stout, Froelich, and Gao (2001).

Nonparametric Regression

In nonparametric regression an IRF is estimated without assuming a logistic form as in the parametric logistic IRT models. There are at least two ways of doing this. One possibility is to use kernel smoothing (e.g., Eubank, 1988). Another possibility is to use isotonic regression estimation (Barlow, Bartholomew, Bremmer, & Brunk, 1972). In this study we used the kernel-smoothing technique. Lee and Douglas (2002) compared isotonic regression with kernel smoothing and found, in general, similar results with respect to the estimation of an IRF (also see Rossi, Wang, & Ramsay, 2002, for recent developments in this area). A practical advantage of kernel smoothing is that there is a program, TESTGRAF (Ramsay, 2000), by which it is possible to estimate $P_g(\theta)$. We do not give any technical details about smoothing techniques here; the interested reader is referred to Ramsay (1991) and Cook and Weisberg (1999). Instead of fitting a parametric function to the entire set of data, such as the 2PLM or 3PLM logistic function, using least squares or maximum likelihood, kernel smoothing takes a weighted average at each point of the IRF; the weights are determined by the kernel function. The user-specified bandwidth value h controls the trade-off between bias and sampling variation. Low values of h yield estimated functions with large variance and small bias, and high values of h yield estimated functions with small variance but large bias. Generally, the bottom line is to choose a bandwidth minimizing the mean-square error, which is the sum of the variance and the squared bias. A rule of thumb is to choose a bandwidth $h = 1.1N^{-1/5}$, where N equals the number of observations—in our case, the sample size.

An empirical example of the use of TESTGRAF in the context of personality measurement and using depressive self-ratings can be found in Santor et al. (1994); these authors evaluated item gender bias and response-option weights on the Beck Depression Inventory (Beck & Steen, 1993). It should be realized that smoothing can be affected by the bandwidth and that one should be careful in choosing a bandwidth, especially when there are not many observations. Note that kernel smoothing does not enforce monotonicity in θ. However, we use kernel smoothing to investigate the form of the IRFs.

Measurement Precision in NIRT

TESTGRAF contains several options to investigate measurement precision conditional on the latent trait. We use a plot of the average item information. The item information for dichotomous items equals

$$I_g(\theta) = \left(\frac{dP_g(\theta)}{d\theta}\right)^2 \Big/ P_g(\theta)[1 - P_g(\theta)], \qquad (7)$$

where $P_g(\theta)$ is defined in Equation 2. In TESTGRAF θ is not estimated numerically, but instead a monotone transformation of the total score or rest score, denoted $R_{(g)}$, is used. $R_{(g)}$ is defined as the total score on the items in the test minus the score on item g. Item and option response functions are estimated using the smoothing technique discussed above. For further details, see Ramsay (2000).

Advantages of Using NIRT to Analyze Personality Data

Several arguments can be given for applying NIRT to personality data. We first present two arguments, and then we illustrate these arguments by analyzing data of the MMPI–2 Depression content scale.

The first argument is that NIRT does not impose a specific form on the IRF. A good illustration of what may go wrong when a specific logistic IRF (such as the 2PLM IRF) is fitted to data that do not have a logistic structure was given by Reise and Waller (2003). They fitted both the 2- and 3PLM to MMPI–A data and found that the IRFs fitted the 3PLM and the 2PLM equally well in terms of root-mean-square error. However, they showed that fitting the 2PLM on 3PLM data resulted, for some items, in lower discrimination parameters for the 2PLM than for the 3PLM. This difference in discrimination parameters was the result of the lower asymptote and thus was an artifact of the models. When an item has a significant lower asymptote, the estimated item discrimination parameter will be smaller in the 2PLM relative to the estimated item discrimination in the 3PLM. The lower item-discrimination parameter for the 2PLM resulted in the IRF's fitting both the 2PLM and the 3PLM equally well. Similar findings were described when an item had a non-one upper asymptote. Reise and Waller therefore also suggested using a 4PLM with an additional parameter for the upper asymptote as given in Equation 2. A drawback of the 4PLM, however, is that it is not easy to estimate this additional parameter and no computer programs are available to do this. Fox (in press) proposed a method to estimate the upper asymptote using Markov chain Monte Carlo methods (see also Fox & Glas, 2003), but this procedure requires complex calculation techniques that are difficult for nonspecialists to understand and the method assumes a logistic IRF.

Instead, determining the IRF directly from the data, such as, for example, is possible using TESTGRAF, or estimating the discrimination using the H_g coefficient using MSP5 can be very illuminating because the researcher obtains information about the quality of the data without forcing the data to conform to a logistic IRT model. This argument has a broader implication, namely, that it is often better to use a simple and flexible model to make inferences about the data and to locally check assumptions. Several researchers have emphasized this point (Junker & Sijtsma, 2001; Molenaar, 2001, 2004; Santor & Ramsay, 1998). Besides, statistics such as item–total correlations and factor loadings do not take into account how item performance may vary across levels of the latent trait (such as depression). Analytical techniques based on NIRT are ideal instruments for evaluating how item performance may change as a function of the latent trait.

Not all IRFs have a logistic form. Using NIRT modeling will draw the researcher's attention to this phenomenon. Items that are not modeled efficiently with a logistic IRT model may still be useful items within a particular range of the underlying latent trait or within particular samples. Santor and Ramsay (1998) noted that parametric models assume that characteristics of the parameters hold for the entire sample, which is not very likely. Observations in less dense regions of the distribution will generally be fitted less well than observations in more dense regions. In psychopathology research persons in the extreme regions of the distribution are often of interest. Therefore, accurately modeling data in these regions of the sample requires careful consideration.

A second argument is that although item characteristics are not estimated parametrically, several easy-to-interpret statistics (such as the H coefficients or the endorsement proportions) give information about the characteristics of the IRFs and the quality of the data that are invariant under reversed score keying. These measures warn the researcher against the idea that the quality of a test is independent of the population of interest. This is useful information in personality and psychopathology testing, in which measurement instruments are often constructed using information from the general population but are then applied to discriminate between persons in specific populations (e.g., those with mental retardation). Furthermore, the nonparametric methods can be used with relatively small sample sizes (say, 300–400 persons; see Molenaar, 2001).

A final remark concerns determining the unidimensionality of a data set in parametric IRT and NIRT. In parametric IRT, item factor analysis is often used to establish unidimensionality (e.g., Reise & Waller, 2003). In the factor analytic literature for dichotomous item scores, instead of using the product–moment correlation, the tetrachoric correlation is often used because of the ceiling effect. However, tetrachoric correlations tend to be biased in the presence of nonzero lower and non-one upper asymptotes (e.g., Carroll, 1983), and there is a need to correct tetrachoric correlations. In the NIRT literature it has therefore been suggested (e.g., Nandakumar & Stout, 1993; Sijtsma & Molenaar, 2002) that nonparametric approaches to assessing unidimensionality should be preferred over parametric approaches. Recently, however, computer programs such as MicroFACT 2.0 (Waller, 2000) have begun correcting for the bias in the lower asymptote, and Reise and Waller (2003) also suggested exploring tetrachoric corrections for non-one upper asymptotes. In Mokken scale analysis, the selection of items circumvents the problem of an upper and lower asymptote because the *H* coefficient is used as a criterion for including items in a scale: *H* is a weighted sum of covariances normed against the weighted sum of maximum possible covariances given the $P_g(\theta)$s. The ceiling effect of the product–moment correlation is then absent. Besides, items with asymptotes substantially different from 0 and/or 1 will probably be rejected as not being very discriminating.

Method

Data–Instruments

Data were analyzed from the official Dutch translation of the MMPI–2 DEP scale (Derksen, de Mey, Sloore, & Hellenbosch, 1993). This scale consists of 33 items measuring different levels of depression. The data were collected as part of a larger battery of tests administered to criminal and psychiatric patients in The Netherlands. A sample of 439 persons was available with a mean age of 32.5 years (*SD* = 8.7); 69% were male. The complete MMPI–2 consists of 567 items; we used the original numbering of the items in the MMPI–2. Furthermore, our sample was much smaller than the sample used in the Reise and Waller (2003) study; however, this sample size suffices to illustrate how we can use nonparametric IRT to analyze personality data. In Table 1 the paraphrased item content is given, together with the factor loadings estimated using MicroFACT (Waller, 2000); the item discrimination parameters under the 2PLM, 3PLM, and the 3PLM with reversed keying (3PLM–R) estimated using BILOG–MG (Mislevy & Bock, 1990); and the H_g coefficients. We use these measures to illustrate the similarities and differences between parametric and NIRT analyses below.

We used MSP5 for Windows (Molenaar & Sijtsma, 2000) to conduct a Mokken scale analysis. The search procedure in MSP5 was used to investigate dimensionality, and the *H* values were used to investigate monotonicity. The graphs obtained by TESTGRAF (Ramsay, 2000) were used to investigate monotonicity and the existence of lower and upper asymptotes, using rest scores. Also, we used these graphs to investigate the specific form of the IRF. Furthermore, we used DIMTEST (e.g., Stout et al., 1996) to obtain additional information about the dimensionality of the data.

Results

The mean total score on the DEP scale was 12.93 (*SD* = 7.76). This mean score was significantly larger than the score obtained in the norming sample for the MMPI, which was a representative sample from the normal Dutch population with *M* = 4.30 (*SD* = 4.15). The endorsement proportions ranged from .10 (Items 234 and 303) through .82 (Item 56). The reliability, estimated using Cronbach's alpha, equaled .85.

Investigating Monotonicity

The *H* coefficient for the whole scale equaled .39. In the last column of Table 1, the H_g values are given. Closer inspection of these items showed that, in particular, Item 52, "I have not lived a proper life" (H_{52} = .16), and Item 246, "My sins are unforgivable" (H_{246} = .10), had low H_g values. In contrast, items that discriminated very well between low- and high-scoring persons were Item 215, "I often worry," with H_{215} = .60, and Item 56, "Wish I could be as happy as others," with H_{56} = .58. Because both factor loadings and H_g values are sensitive to the discrimination parameter of the IRF, we expect that high (low) factor loadings go together with high (low) H_g values. Factor loadings equal to or larger than .50 always resulted in H_g values larger than .30 (see Table 1). Factor loadings lower than .50 resulted 6 out of 8 times in H_g values lower than .30.

Figure 2 shows the IRFs from TESTGRAF for some of the items of the DEP scale that we discuss here. We first inspect the IRFs of Items 52, 246, 215, and 82. Item 52 hardly discriminates across the range of rest scores. For almost all persons the chance of endorsement of this item is between .4 and .8. Thus, a person with, say, $R_{(52)} = 5$ has a relatively high probability of endorsing this item, whereas a person with, say, $R_{(52)} = 20$ has a relatively low probability. This item is not very useful to separate persons with $4 \leq R_{(52)} \leq 24$. Inspecting the IRF of Item 246 shows that this item discriminates over a small range of high-scoring persons (higher than, say, $R_{(246)} = 20$). Such an item will only be informative in populations of individuals with se-

Table 1
Paraphrased Item Content, Factor Loadings, Discrimination Parameters, and Scalability Coefficients for the Depression Content Scale

Item	Paraphrased item content	Factor loading	$\hat{a}$ 2PLM	$\hat{a}$ 3PLM	$\hat{a}$ 3PLM–R	H_g
3	Restful sleep	.76	1.15	1.16	1.15	.49
9	Life is interesting	.41	0.69	0.72	0.71	.27
38	Sometimes I couldn't "get going"	.65	1.00	0.98	1.01	.43
52	I have not lived a proper life	.19	0.32	0.96	0.63	.16
56	Wish I could be as happy as others	.81	1.33	1.35	1.35	.58
65	I feel blue	.70	1.19	1.17	1.21	.39
71	I will not achieve anything	.59	0.78	0.75	0.74	.41
75	Life is worthwhile	.50	0.90	0.91	0.91	.42
82	I regret things afterwards	.49	0.88	0.87	0.85	.29
92	I don't care what happens	.56	0.64	0.65	0.65	.35
95	Happy most of the time	.66	1.17	1.15	1.18	.43
130	Think I'm no good	.62	1.19	1.15	1.15	.45
146	I cry easily	.37	0.34	0.32	0.57	.20
215	I often worry	.83	1.57	1.62	1.58	.60
234	I am doomed	.80	1.22	1.24	1.25	.55
246	My sins are unforgivable	.28	0.22	0.25	0.45	.10
277	Feel lonely often	.65	1.29	1.30	1.30	.42
303	Wish I were dead	.75	0.78	1.25	0.85	.49
306	No one cares about me	.45	0.83	0.83	0.84	.33
331	I am pessimistic	.67	0.91	0.93	0.98	.45
377	I am not happy about myself	.66	0.93	0.91	0.92	.43
388	I very seldom have the blues	.70	1.19	1.16	1.19	.48
399	You cannot make any plans for the future	.38	0.79	1.15	0.89	.26
400	Don't care about anything	.67	1.17	1.13	1.13	.43
411	At times I think I am no good at all	.70	1.21	1.19	1.20	.44
454	Future seems hopeless to me	.68	0.87	0.82	0.84	.44
506	I have recently considered killing myself	.55	0.85	0.86	0.86	.34
512	I experienced a great loss	.41	0.66	0.65	0.62	.30
515	Most happy when alone	.70	1.17	1.18	1.17	.48
520	Think about killing myself	.55	0.91	0.92	1.15	.35
539	I do not want to solve my problems	.59	0.91	0.90	0.82	.37
546	I think about death lately	.65	1.04	1.02	1.02	.45
554	Give up life	.69	1.05	1.04	1.07	.45

Note. Item discrimination is represented by $\hat{a}$, and the scalability coefficient for an item with respect to other items in the test is represented by H_g. PLM = parameter logistic model.

vere depression. This item also nicely illustrates a remark by Reise and Waller (2003, p. 180) that "Within a group of highly depressed individuals, very few persons will manifest all key symptoms. . . . Consequently, researchers who fit only the 3PLM . . . might miss this fact because the 3PLM does not fit a non-one upper asymptote to the IRF." In contrast, the IRF of Item 215, "I often worry" ($H = .60$), shows that this item discriminates well between low- and average-scoring persons but does not discriminate between average- and high-scoring persons.

These results show that depression, as it is measured by Item 246 or Item 215, is not a continuum. Santor and Ramsay (1998) argued that the relation between specific symptoms and overall depressive severity is important to evaluating the continuity of depressive symptoms. Viewing depression as a continuum implies that scores should differentiate individuals with varying degrees of depression across the entire range of θ values and that the probability of observing specific symptoms should increase smoothly for larger values of depression, rather than abruptly at a specific threshold. This is not the case for Item 246. Although it is not impossible to reach the same conclusion using parametric IRT modeling (correctly interpreting the *a*, *b*, and *c* parameters and plotting the estimated IRFs), because of the emphasis on estimating item parameters and not on using local checks this is not easily found. For example, we see

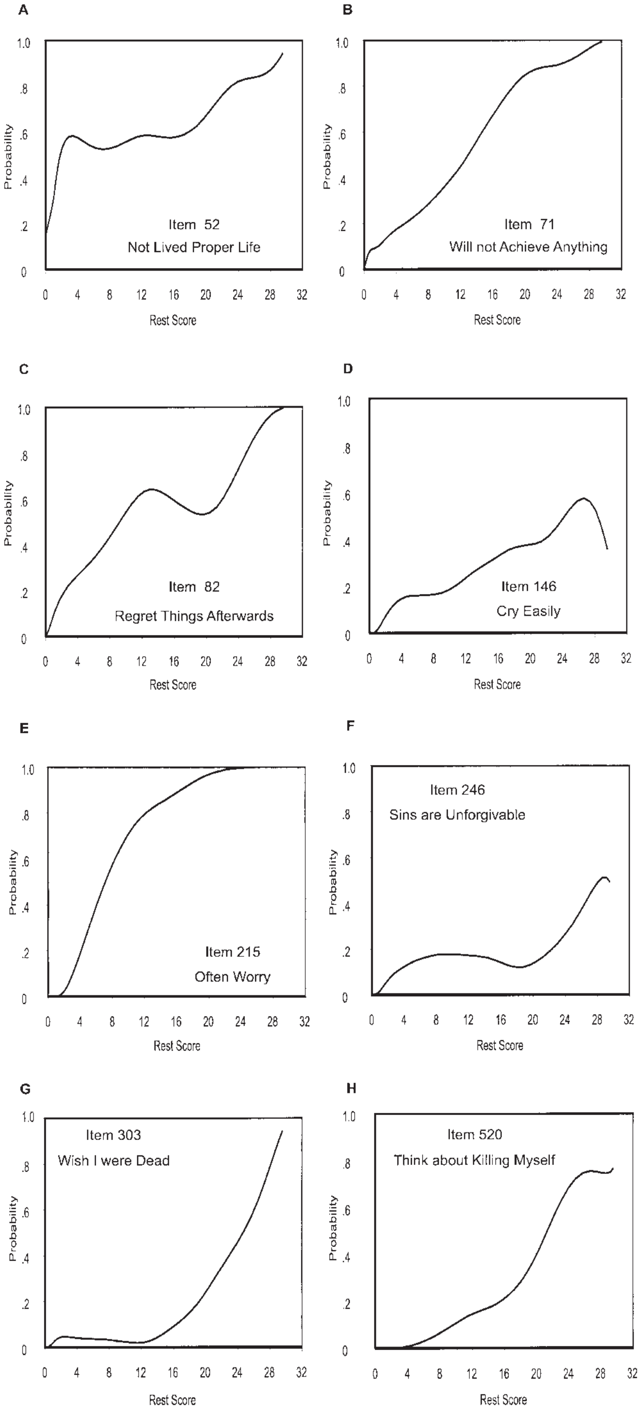

Figure 2. Item response functions for some items of the Depression content scale.

that low H_g values may be the result of different types of IRFs (cf. Items 52 and 246), which is very interesting information to have when one is constructing a personality test. Moreover, in an investigation of unidimensionality using item-factor analysis as is often done in parametric IRT, these kinds of items will probably have been removed before an IRT item calibration analysis is conducted, which prevents the researcher from learning how certain items function across the latent trait range and how these items can possibly be used in a certain range of the latent trait.

Thus, we argue for better data exploration when constructing and revising a personality instrument and before using parametric IRT modeling. In particular, this should enable researchers to better understand how an item and a test are functioning. For example, we learned as a result of this analysis that it is difficult to argue that persons with mild depression will endorse items like Item 246. Another interesting example is Item 82, with $H_{82} = .29$, "I regret things afterwards." Looking at the IRF in Figure 2, we conclude that although this item discriminates between low- and high-scoring persons, for persons between, say, $R_{(82)} = 12$ and $R_{(82)} = 24$ the endorsement probabilities are between .6 and .8. A tentative conclusion is that persons with severe depression are in general scoring higher than persons who are not depressed, but that for persons in the middle of the score range there may be a second trait value that determines the endorsement of this item. In contrast to these three items, note that, for example, the IRF of Item 71, "I will not achieve anything," discriminates well across a broad range of scores.

Investigating monotonicity by inspecting the IRFs revealed that, in particular, Items 146, 246, and 520 (see Figure 2) have non-one upper asymptotes, whereas Items 52 and 399 have a nonzero lower asymptote. A nonzero lower asymptote implies that persons with a low depression score endorse these items with a probability larger than 0, whereas a non-one upper asymptote implies that in our sample of forensic and psychiatric patients those with severe depression (say, $R_{(g)} > 27$) still do not endorse these items with a probability close to one. We speculate that this must be the case in other psychopathology tests. Reise and Waller (2003) mentioned that a non-one upper asymptote may occur when, in the phrasing of the item, words such as *always* or *never* are being used. We found that this is also the case for items that should indicate severe depression such as "My sins are unforgivable" (Item 246) or "Think about killing myself" (Item 520). The non-one upper asymptote of Item 146, "I cry easily," can be explained by noting that a majority of the sample consisted of males.

To illustrate the relative advantage of nonparametric modeling over parametric IRT modeling, we estimated the discrimination parameter of the DEP items under the 2- and 3PLM using BILOG–MG (Mislevy & Bock, 1990) for these items. On the basis of the results of Reise and Waller (2003), we expected that for items with a non-zero lower asymptote the estimated discrimination parameter will be smaller under the 2PLM than under the 3PLM, and for the items with a non-one upper asymptote we expected that the estimated parameters will be similar under the 2PLM and the 3PLM, but lower relative to the 3PLM with reversed keying (3PLM–R). This was confirmed in our data analysis. In Table 1 we depicted the $\hat{a}$ parameters. For Items 52 and 399, the $\hat{a}$ parameters were larger under the 3PLM than under the 2PLM, whereas for Items 146, 246, and 520, the $\hat{a}$s were larger under the 3PLM–R than under the 2PLM and the 3PLM.

For Item 303, "Wish I were dead," different $\hat{a}$ values were obtained under the 2PLM and the 3PLM. Under the 2PLM the item discrimination equaled .78, whereas under the 3PLM it equaled 1.25. Inspecting the IRF of Item 303 in Figure 2 shows that persons within the middle of the score range on the DEP scale do not endorse this item often in this population, whereas it discriminated well between average- and high-scoring persons. Using Mokken scale analysis, we found that $H_{303} = .49$ (see Table 1). Thus, a combination of examining the H coefficient calculated using MSP5 and plotting the IRFs using TESTGRAF led to a good understanding of the functioning of this item and showed that it is useful in discriminating among persons in the upper score range.

In Figure 3 we depict the average item information as a function of the rest score as given by TESTGRAF (solid line). Note that the test is most informative for rest scores in the range $4 \leq R_{(g)} \leq 24$. Furthermore, when using reversed scoring (dashed line in Figure 3), we found that the average item information was symmetrical for positively versus negatively scored items.

Investigating Dimensionality

The search algorithm showed that when $H = .01$ was used as a lower bound, five items (Items 9, 52, 82, 146, and 246) were rejected because of a negative H_g value with one of the other scaled items. We increased the lower bound to $H = .10$, $H = .20$, and $H = .30$ to investigate multidimensionality. The same scale was found for $H = .10$ and $H = .20$, and for $H = .30$ two additional items did not scale with the other items (Items 399 and 512). Thus, no evidence of multidimensionality was found using $H = .1$, .2, or .3 as criteria. In Table 2, the selection order of the items is given. According to the rules of thumb given in Hemker et al. (1995), this scale is unidimensional because the original scale did not separate into different subscales.

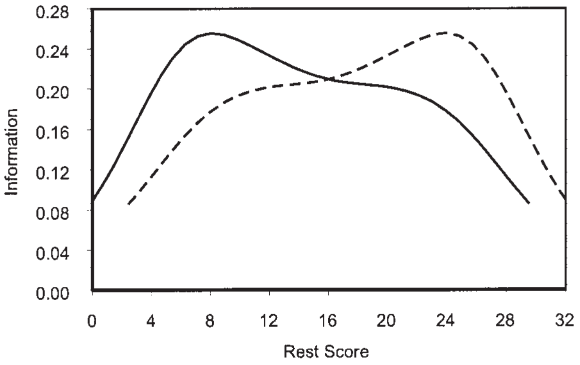

Figure 3. Average item information curves for the Depression content (DEP) scale (solid line) and the DEP scale with reversed scoring (dashed line).

In personality measurement it is informative to select items that discriminate mostly between low and high values of the trait to obtain a shorter test that can be used, for example, in situations in which persons cannot concentrate on a relatively long personality test such as the MMPI. To select those items, we simply choose $H \geq .40$ instead of $H \geq .30$. When we used $H \geq .40$, two additional items were removed from the original scale (Items 506 and 306; see the lower part of Table 2).

Using DIMTEST for the whole scale, we found that $T = 1.51$ (with $p = .064$), which reflects a reasonably unidimensional scale. When we removed Items 9, 52, 82, 146, and 246 that were rejected by MSP using $H \geq .01$, the DIMTEST probability value increased (to $p = .115$).

Discussion

In this study, we explored the use of NIRT to investigate the data structure of personality and psychopathology data. We showed that through these models information can be obtained about the functioning of items that is more difficult to obtain using parametric models. It is our belief that nonparametric models are useful models to explore the data structure, and we showed that "staying close to the data" (i.e., not assuming a specified logistic curve) prevents the researcher from jumping to conclusions. NIRT can also be used to identify items a priori that may require more complex parametric models, such as the 3PLM or the 4PLM. Because NIRT models are more flexible than parametric IRT models, they can more easily describe response behavior that cannot be described by parametric logistic models. NIRT models also do not require complex estimation procedures. For example, the 3PLM requires iterative estimation of its parameters, and these iterations may sometimes converge slowly, converge to a suboptimal value, or depend on large amounts of prior information.

Our analyses showed that it is not necessary to use keyed and reversed-keyed items as was done in the Reise and Waller (2003) study to find that some items have an upper asymptote and that it is sometimes dangerous to use an IRT model that does not fully capture the data structure. Their whole study was a quest to answer the question of which model (2PLM or 3PLM) best described the response behavior to items of a psychopathology test, and they concluded

Table 2
Scalability Coefficients of Two Mokken Search Procedures

Item	H_g
Search procedure $H \geq .3$	
3	.56
546	.51
234	.61
277	.48
215	.68
65	.47
56	.66
303	.52
515	.56
95	.51
454	.49
411	.53
388	.56
554	.52
130	.52
520	.47
331	.52
377	.47
75	.44
400	.48
38	.48
71	.46
539	.39
92	.39
506	.36
306	.35
Search procedure $H \geq .4$	
234	.63
303	.52
539	.39
520	.48
546	.53
92	.41
75	.46
454	.50
554	.54
515	.57
277	.50
400	.50
411	.55
65	.49
377	.48
130	.53
71	.48
331	.53
95	.53
38	.49
3	.57
388	.56
215	.68
56	.66

that an additional parameter is needed to model psychopathology data.

Furthermore, we like the use of relatively "simple" models. An advantage of using a simple model is that it is easier to explain, and that it performs better under replication (Molenaar, 2001). Also, there are easy-to-use NIRT procedures that can be applied to relatively small data sets (say, between 300 and 400 persons). In particular in clinical assessment this is a big advantage because many clinical studies involve small numbers of persons. For example, browsing through Volume 13 (2001) and Volume 14 (2002) of *Psychological Assessment,* we found many studies with sample sizes between 50 and 400 persons. One obvious reason is that in clinical settings it can be difficult to obtain persons from a particular population.

In the 2PLM and the 3PLM, measurement precision is determined conditional on the latent trait using item and test information functions. TESTGRAF also provides the plot of the mean item information function. In personality assessment the use of these information functions is in particular useful because measurement precision is not constant for all persons. Most psychopathology scales differentiate among persons in the high range of the latent trait but are less suited for differentiating persons in the average-to-low range of the trait. By means of item and test information functions, this can be investigated (see e.g., Fraley, Waller, & Brennan, 2000, for an illustration in the context of attachment research).

In our view, IRT models, both parametric and nonparametric, are helpful tools to learn more about the structure of empirical data sets, and a preference for one model over another should be determined by an individual researcher in the specific context of his or her particular research. We thus fully agree with a remark made in Reise and Waller (2003) that "the process of IRT fit assessment and model comparison can be viewed more as a way of learning about item and test functioning than as a process of mere statistical decision making" (p. 180). Using different models and different methods of analysis forces a researcher to think about his or her decisions and also prevents a researcher from relying on the default options found in many statistical packages. Molenaar (2004) provided an excellent discussion on model choice and discussed the importance of replication of results using different models and methods.

We restricted ourselves in this study to 33 items of the DEP scale. A reviewer noted that the NIRT method as discussed in this study may miss the fact that items such as Item 215, "I often worry," may belong to another scale than a depression scale (e.g., an anxiety scale). We expect that if we had chosen more anxiety items, the search algorithm used in MSP5 might have clustered these anxiety items as belonging to one scale. Future research should investigate this question.

Practical Implications

We mention three practical implications of the present findings. First, when researchers analyze personality and psychopathology data, we recommend using nonparametric models to routinely investigate the data structure before applying the 2PLM and 3PLM. Nonparametric models are especially useful when one is constructing a personality test and selecting items from an item pool, that is, in the phase of test construction when a psychologist is learning how items are functioning. Free computer software is available to conduct nonparametric analyses (e.g., Cook & Weisberg, 1999; Ramsay, 2000). By means of these programs, researchers can use regression functions and smoothers to evaluate IRFs.

Second, and related to the first point, in personality and psychopathology research it has been hypothesized (e.g., Widiger, Trull, Clarkin, Sanderson, & Costa, 2002) that there are relations between facet scores and diagnostic criteria. For example, high scores on the revised NEO Personality Inventory (Costa & McCrae, 1994) facet Angry Hostility and low scores on Trust, Straightforwardness, and Compliance are related to a paranoid personality disorder. It is our belief, however, that if the structure of different facets and items within these facets is not well understood, the measurement of higher order constructs may also be problematic. NIRT can help the researcher to investigate the facet structure. For example, the search algorithm we used in this article can be used to select items from a pool of items starting with two items that, according to a domain expert, are highly related. Items that are selected can then be further studied by inspecting nonparametric regression functions.

The third implication is in the area of the measurement of change and emphasizes the importance of the use of both parametric IRT and NIRT models. Both the Reise and Waller (2003) study and the present study showed that some items only discriminate within a particular range of the latent trait. This has important consequences when one is measuring the effects of, for example, therapy, or when one is monitoring persons over time. For example, assume that we are interested in evaluating the effect of a therapy for persons with depression. Furthermore, assume that a person with severe depression becomes less depressed as a result of the therapy. Then, it is very likely that this person will still endorse items such as Item 215, "I often worry," because this item hardly discriminates in the range of average- to high-scoring persons. When the test consists of many items that do not discriminate in the higher trait range, a psychologist not aware of this would erroneously conclude that the therapy had no effect.

On the other hand, when one is monitoring a population with respect to depressive symptoms in, for example, health research, it is very important to have items that discriminate well in the low-to-average trait range (assuming a relatively healthy society). If not, changes in the overall depressive mood in a population will not be detected. Because many psychopathology inventories consist of items that are only informative in the average-to-high trait range, we speculate that many inventories are not very useful in monitoring health outcomes. Using nonparametric Mokken scale analysis, items can be selected that discriminate well in the population of interest. Using MSP5, a researcher can very easily combine information from statistics like H and plots of IRFs, so that items are selected that discriminate well in a particular range of the latent trait. Using TESTGRAF, a researcher can use the item information functions to select such items.

References

Barlow, R. E., Bartholomew, D. J., Bremmer, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions.* London: Wiley.

Beck, A. T., & Steen, R. A. (1993). *Beck Depression Inventory manual.* San Antonio, TX: Psychological Corporation.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *MMPI–A (Minnesota Multiphasic Personality Inventory—Adolescent): Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Carroll, J. B. (1983). The difficulty of a test and its factor composition. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 257–282). Hillsdale, NJ: Erlbaum.

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

Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences.* Mahwah, NJ: Erlbaum.

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

Cook, R. D., & Weisberg, S. (1999). *Applied regression including computing and graphics* [Computer software]. New York: Wiley. Retrieved from http://www.stat.umn.edu/arc

Costa, P. T., Jr., & McCrae, R. R. (1994). *Bibliography for the revised NEO Personality Inventory (NEO–PI–R) and NEO Five-Factor Inventory (NEO–FFI).* Odessa, FL: Psychological Assessment Resources.

De Koning, E., Sijtsma, K., & Hamers, J. H. M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement, 26,* 302–320.

Derksen, J. J. L., de Mey, H. R. A., Sloore, H., & Hellenbosch, G. (1993). *MMPI–2, handleiding bij afname, scoring en interpretatie* [*Manual for the administration, scoring, and interpretation of the MMPI–2*]. Nijmegen, The Netherlands: PEN Test.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression.* New York: Marcel Dekker.

Fox, J.-P. (in press). Markov chain Monte Carlo methods for IRT models with flexible asymptotes. *Journal of Educational and Behavioral Statistics.*

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68,* 169–191.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19,* 337–352.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory [Special issue]. *Applied Psychological Measurement, 25*(3).

Lee, Y. S., & Douglas, J. (2002). *Application of isotonic regression in item response theory.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45,* 507–530.

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

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8,* 72–87.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG user's guide* [Software manual]. Chicago: Scientific Software.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague, The Netherlands: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York: Springer-Verlag.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer-Verlag.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25,* 295–299.

Molenaar, I. W. (2004). About handy, handmade and handsome models. *Statistica Neerlandica, 88,* 1–20.

Molenaar, I. W., & Sijtsma, K. (2000). MSP5 for Windows, a program for Mokken scale analysis for polytomous items. Groningen, The Netherlands: iec ProGAMMA.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41–68.

Panter, A. T., Swygert, K. A., Dahlstrom, W. G., & Tanake, J. S. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment, 68,* 561–589.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611–630.

Ramsay, J. O. (2000). TestGraf. A program for the graphical analysis of multiple-choice tests and questionnaire data [Computer software and manual]. Retrieved from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8,* 164–184.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287–297.

Robie, C., Zickar, M. J., & Schmitt, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14,* 187–207.

Rossi, N., Wang, X. H., & Ramsay, J. O. (2002). Nonparametric item response function estimates with the EM algorithm. *Journal of Educational and Behavioral Statistics, 27,* 291–317.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10,* 345–359.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analysis of the Beck depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6,* 255–270.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3–32.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology, 66,* 341–349.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55,* 293–325.

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–375). New York: Springer-Verlag.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. M. (1996). Conditional covariance-based multidimensionality assessment. *Applied Psychological Measurement, 20,* 331–354.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Waller, N. G. (2000). MicroFACT (Version 2.0) [Software manual]. St. Paul, MN: Assessment Systems.

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.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5,* 125–146.

Widiger, T. A., Trull, T. J., Clarkin, J. F., Sanderson, C., & Costa, P. T., Jr. (2002). A description of the *DSM–IV* personality disorders and dimensions of personality. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 89–99). Washington, DC: American Psychological Association.

Received March 13, 2003
Revision received October 3, 2003
Accepted February 9, 2004 ■