LSAC RESEARCH REPORT SERIES

Investigating the Quality of Items in CAT Using Nonparametric IRT

Rob R. Meijer
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 04-05
March 2004

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The quality of the items in an item pool is an important determinant of the success of the computer adaptive testing (CAT) program. A mathematical model called item response theory (IRT) is used as the basis for many CAT programs, and statistics derived through IRT are among those that may be used to investigate the quality of the items in the item pool. Among the IRT models, a family of approaches referred to as nonparametric (NIRT) models are useful to investigate the quality of the items and response data because they are not based on strong functional assumptions and enable the use of informative data exploration techniques.

The aim of the present study is to illustrate the usefulness of NIRT for designing good item pools, a problem for which the solutions are still in their infancy. I show how the use of NIRT is very suitable for exploring the structure of CAT data. Particularly I explored the use of NIRT for analyzing the covariance structure between items as well as the (nonparametric) regression of item scores on total scores. It is shown that this use of NIRT leads to useful information, which (1) can be interpreted very easily by practitioners, (2) avoids forcing the data into a structure they sometimes do not have, and (3) is easily obtained through the use of very user-friendly software programs.

Abstract

I discuss the applicability of nonparametric item response theory (IRT) models to the quality of item pool development in the context of CAT, and I contrast these models with parametric IRT models. I also show how nonparametric IRT models can easily be applied and how misleading results from parametric IRT models can be avoided. I recommend the use of nonparametric IRT modeling to routinely investigate the quality of item pools.

Introduction

In item pool development for computer adaptive testing (CAT), the quality of items is an important issue. Both classical and item response theory (IRT) based indices can be used to investigate the quality of the items. However, when exploring the quality of individual and subsets of items, nonparametric IRT (NIRT) models are also useful to investigate data quality, because these models are not based on relatively strong assumptions (such as logistic regression functions) and use data exploration techniques that can be very informative concerning the functioning of individual items (e.g., Santor & Ramsay, 1998). The application of NIRT can be of help to design good item pools. The research concerning building and maintaining good item pools is still in its infancy (Veldkamp, van der Linden, & Ariel, 2003). The aim of the present study is to illustrate the usefulness of NIRT to develop and to analyze item pools. In my opinion, the use of nonparametric IRT has been underexposed in the recent CAT literature. I show that these models are very suitable to explore the structure of CAT data. A study by Chernyshenko, Stark, Chan, Drasgow, and Williams (2001), who explored the use of nonparametric IRT modeling is a good example. Chernyshenko et al. fitted the 2-PLM, the 3-PLM, a graded response model, and Levine's nonparametric maximum likelihood formula scoring models on dichotomous and polytomous data. They concluded that the nonparametric model provided the best fit of the models considered. Because a test can never be better than permitted by the item pool from which it is assembled, exploration of data quality is very important. This being the case, I argue that using nonparametric IRT models based on exploring the simple covariance structure between items and based on nonparametric regression will lead to useful information that (1) can be interpreted very easily by practitioners, (2) avoid forcing the data in a structure they may not have, and (3) is easily obtained through the use of very user-friendly software programs.

In this study I show how NIRT may help to avoid misleading results obtained from parametric IRT models, and I argue that nonparametric solutions are already available for problems that exist when using parametric IRT models to investigate the data structure. I definitely do not want to argue for the overall replacement of parametric by nonparametric models. Parametric IRT models lead to point estimates of the latent trait and to interval scales for measuring respondents. Such scales can be very convenient, for example, for comparing the results from different tests selected from the same item bank. But I do think that the emphasis on parametric IRT modeling in educational measurement may sometimes lead to unnecessarily complicated papers and sometimes even to bad measurement practice. I am not the first to make this observation. Some excellent papers discuss the usefulness of NIRT (e.g., Junker & Sijtsma, 2001; Santor, Ramsay, & Zuroff, 1994), but the influence of these papers is very modest. Furthermore, I do not pretend to explore the full range of techniques that nonparametric IRT modeling has at its disposal. I apply a number of useful methods to explore the data structure of CAT, and I restrict myself to techniques by which I can illustrate how current problems raised in the recent parametric IRT literature can be solved. For more

detailed information about different nonparametric fit methods, see Ramsay (2000), Stout (1990), and Sijtsma and Molenaar (2002).

This study is organized as follows. First, I introduce the basic principles of parametric and nonparametric IRT. Second, I introduce nonparametric fit methods of two nonparametric IRT models without going into technical detail. Third, I illustrate the use of nonparametric IRT to simulated data from a CAT item pool. Finally, I discuss directions for future research in this area.

IRT and CAT

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 on which persons and items have a position. IRT models allow the researcher to check if the data fits the model. The focus in this article is on IRT models for dichotomous items. Thus, one response category is positively keyed (item score 1), whereas the other is negatively keyed (item score 0); for ability and achievement items, these response categories usually reflect the correct and incorrect answers, respectively.

Most IRT models assume unidimensionality and a specified form for the IRF, which can be checked empirically. Unidimensionality means that the latent space that explains the person's test performance is unidimensional. Related to unidimensionality is the assumption of local independence, which states that responses to the items 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.

In IRT, the probability of endorsing an item g ($g = 1, ..., k$) is a function of a person's latent-trait value θ and characteristics of the item. This conditional probability $P_g(\theta)$ is the item response function (IRF). It is the probability of a positive response (i.e., "agree" or "true") among persons with the latent trait value θ. Item characteristics that are often taken into account are the item discrimination (a), the item location (b), and the pseudo-chance level parameter (c). The item location b is the point at the trait scale where $P_g(\theta) = 0.5(c + 1)$. Thus when $c = 0$, $b = 0.5$. The greater the value of the b parameter, the greater the trait value that is required for an examinee to have a 50% chance of endorsing the item; thus the less popular the item. Hard items are located to the right or the higher end of the θ scale; easier 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 easy) to +2 (very difficult). 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 with a trait level θ near b. The pseudo-chance level parameter c (ranging from 0 to 1) is the probability of a 1 score for low-ability examinees (that is, as $\theta \to -\infty$).

In parametric IRT, $P_g(\theta)$ often is specified using the 1-, 2-, or 3-parameter logistic model (1-, 2-, 3-PLM). The 3-PLM (Lord & Novick, 1968, chaps. 17–20) is defined as

$$P_g(\theta) = c_g + \frac{(1 - c_g)\exp\left[a_g(\theta - b_g)\right]}{1 + \exp\left[a_g(\theta - b_g)\right]}. \quad (1)$$

The 2-PLM can be obtained by setting $c_g = 0$ for all items; the 1-PLM or Rasch (1960) model can be obtained by additionally setting $a_g = 1$ for all items. In the 2- and 3-PLM, the IRFs may cross, whereas in the Rasch model the IRFs do not cross.

In educational measurement, the 2- or 3-PLM is often applied. However, in a recent study by Reise and Waller (2003) new insights were obtained about what may go wrong when these models are applied to data that do not fit these models. They compared the fit of the 2-PLM and 3-PLM on 15 unidimensional factor scales. Unidimensionality was investigated using item-level factor analysis, and monotonicity was investigated by inspecting individual item endorsement proportions against raw score scales. Relying on χ^2 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 $r = .99$. An unexpected finding was that 10% to 30% of the items had substantially ($c > .10$) lower asymptote parameters when the scales were scored in the pathology or non-pathology directions. 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. Reise and Waller suggested using a 4-parameter logistic model (4-PLM) (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. The idea is that even persons with an extreme position on the latent trait will not have a probability of one in answering an item correctly. Instead of using a 4-PLM, there are several nonparametric alternatives as I show below. First, however, I introduce nonparametric IRT and discuss some of the fit methods by which nonparametric assumptions can be investigated.

Nonparametric IRT

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

Mokken model

The MHM proposed by Mokken (1971; 1997; see also Molenaar, 1997) assumes unidimensional measurement and an increasing IRF as function of θ. However, unlike parametric IRT, the IRF is not parametrically defined. Thus, an important difference between the MHM and the 2-PLM and 3-PLM is that the IRFs for the MHM need not be of the logistic form. This difference makes MHM less restrictive to empirical data than logistic models. The MHM allows the ordering of persons with respect to θ using the unweighted sum of item scores. Therefore, the MHM 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 3-PLM and thus can be used in situations where these models do not fit the data.

Investigating monotonicity in the MHM. Mokken (1971; 1997) proposed to use the scalability coefficient H_{gh} for pairs of items (g, h), the 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. The H coefficients can be interpreted as statistics for slopes of IRFs relative to the spread of the total score X_+ in the group under consideration (De Koning, Sijtsma, and 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 equivalent for the a parameters from logistic IRT models such as the 2-PLM and the 3-PLM.

The item coefficient H_g is defined as the ratio of the sum of all $k - 1$ covariances of fixed item g and the other items $h (h \neq g)$ in the numerator and the sum of $k - 1$ corresponding maximum covariances in the denominator. The H coefficient for k items is the ratio of all item pair covariances in the numerator and all maximum item pair covariances in the denominator. Mokken (1971) showed that H is a strictly increasing function of the variance of X_+. Under the MHM, 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_+. In practice, H and Hg 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 $H_g = .3$ as a lower bound.

Investigating dimensionality in the MHM. Several nonparametric procedures have been proposed to investigate dimensionality of test data. I first use a relatively simple procedure that is incorporated in MSP5. The results obtained from MSP5 will be compared with results obtained from DIMTEST (Stout et al., 1996).

For investigating the dimensionality of an item set, MSP5 contains an automated item selection procedure based primarily on the inter-item covariances and the strengths of the relationship between items and the latent trait(s) as expressed by the item H_g coefficients. 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 (1) H_{gh} is significantly larger than 0, and (2) H_{gh} is the largest among the coefficients for all possible item pairs. Then a third item j is selected that (3) correlates positively with the items already selected, (4) has an H_j coefficient that is larger than 0, and (5) 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 3, 4, and 5. The end result may be one or more item clusters that each tap another 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 CAT data. First, this search algorithm is an excellent tool to form homogeneous clusters of items. These clusters may be compared to testlets, which are item sets with similar content that measure a relatively narrow construct and display high inter-item correlations. Testlets are often used to build tests consisting of higher order dimensions. Interesting in this respect is that in item pool management we can very simply identify key items based on expert opinions or H_{gh} values on the basis of which we can build testlets, 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, which is discussed by Waller et al. (1996) as one of the advantages of IRT in the context of parametric IRT modeling. 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. Theoretical research has shown that items are selected in the bottom up procedure that discriminate well across a wide range of item difficulties (Sijtsma & Molenaar, 2002).

Third, in several studies (e.g., Chernyshenko et al., 2001) it has been suggested that test and item-multidimensionality may be the cause of misfit of logistic IRT models. It is interesting that Hemker, Sijtsma, and Molenaar (1995) showed by means of a simulation study that, if multidimensionality is suspected in an empirical dataset, well-chosen lower bound values can be used effectively to detect the unidimensional scales. They recommend 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: (1) most or all items are in one scale; (2) two or more unidimensional scales are formed; and (3) two or more smaller scales are formed and several items are rejected. Hemker et al. indicated that the results from the second stage should be taken as the final stage. With unidimensionality, the typical pattern of results with increasing c is: (1) most or all items are in one scale; (2) one smaller scale is found; and (3) one or a few scales are found and several items are rejected. They recommend in this case to consider 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.

Nonparametric Regression

In nonparametric regression, an IRF is estimated without assuming a logistic form as in the parameter 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 used isotonic regression estimation (Barlow, Bartholomew, Bremmer, & Brunk, 1972). In this study, I 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. A practical advantage of kernel smoothing is that the software program TestGraf (Ramsay, 2000) is able to estimate $P_g(\theta)$. I will not give any technical details here. Instead of fitting a parametric function to the entire set of data, such as the 2-PLM or 3-PLM logistic function using least squares or maximum likelihood, kernel smoothing takes a weighted average at each point; the weights are determined by the kernel function. Given independent variable θ and the gth response variable, X_g, the estimated regression function $P_g(\hat{\theta})$ is

$$P_g(\hat{\theta}) = \frac{\sum_{g=1}^{k} K\left(\frac{\theta - \hat{\theta}_j}{h}\right) x_{g,f}}{\sum_{g=1}^{k} K\left(\frac{\theta - \hat{\theta}_j}{h}\right)}$$

where $K(x)$ is the kernel function, h is the bandwidth, and f indexes the k observations. Although different functions for $K(x)$ can be chosen, often the Gaussian kernel function

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp - x^2 / 2$$

is used. 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 with large bias. Generally, the bottom line is to choose a bandwidth minimizing the mean-squared 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 my case,

the sample size. An empirical example of the use of TestGraf can be found in Santor, Ramsay, and Zuroff (1994).

Advantages of nonparametric IRT in item pool management

Several arguments can be given for applying NIRT in item pool management. I first present three arguments and then I illustrate these arguments by analyzing some simulated data.

The first argument is that NIRT does not impose a specific form on the item response function. A good illustration of what may go wrong when a specific logistic IRF (such as the 2-PLM IRF) is fitted to data that do not have a logistic structure is given in Reise and Waller (2003). They fitted both the 2-and 3-PLM to MMPI-A data and found that the IRFs fitted the 3-PLM and the 2-PLM equally well in terms of root mean squared error. However, they showed that fitting the 2-PLM on 3-PLM data resulted in lower discrimination parameters fitting the 2-PLM than fitting the 3-PLM. This difference in discrimination parameter was the result of the lower asymptote, thus an artifact of the models. When an item had a significant lower asymptote under the 2-PLM, the IRF of the 3-PLM resulted in a steeper IRF than the IRF of the 2-PLM. The lower item discrimination parameter for the 2-PLM resulted in the IRF fitting both the 2-PLM and the 3-PLM equally well. Similar findings were described when an item had a non-one upper asymptote. Reise and Waller also suggested using a 4-PLM with an additional parameter for the upper asymptote. A drawback of the 4-PLM, however, is that it is not easy to estimate this additional parameter, and there are no computer programs available to do this. J. P. Fox (personal communication, December, 9, 2002) proposed a method to estimate the upper asymptote using Markov Chain Monte Carlo methods, but this method requires complex calculation techniques which are difficult to understand for non-specialists, and it assumes a logistic IRF.

As just seen, imposing a specific parametric model may result in misleading information when the assumptions are not met. Instead, determining the IRF directly from the data using TestGraf or estimating the discrimination coefficient, H_g, using MSP can be very illuminating, because the researcher obtains information about the quality of the data without forcing the data to comply 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; 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 nonparametric item response theory are ideal instruments for evaluating how item performance may change as a function of the latent trait.

Thus, not all IRFs have a logistic form. Using nonparametric IRT modeling will draw the researcher's attention to this phenomena. Items which 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) note 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 than observations in more dense regions. Therefore, accurately modeling data in these regions of the sample should be considered carefully.

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. Besides, 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 item pool management in which measurement instruments are often constructed using information from normal populations but applied to discriminate between persons in specific populations (e.g., developmentally disabled). Furthermore, the nonparametric methods can be used on relatively small sample sizes of 300–400 persons (see Molenaar, 2001).

A third argument is related to factor analysis. In parametric IRT, item factor analysis is often used to establish unidimensionality (e.g., Reise & Waller, 2003). In the factor analysis 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 overestimate the strength of the relationship between items. There is a need to correct tetrachoric correlations for the guessing effect that occurs on multiple choice exams, because a non-zero lower asymptote biases tetrachoric correlations. Reise and Waller also suggest exploring tetrachoric corrections for upper asymptotes that do not equal one. Sijtsma and Molenaar (2002) argue, however, that the selection procedure of the items in the Mokken scale analysis circumvents this problem, because it uses the H-coefficient 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.

Method

Data/Fit Methods

First, data were simulated according to the 3-PLM. θ was drawn from a standard normal distribution and the item parameters of 15 items were used from the calibrated item pool from LSAT, with relatively extreme a, b, or c values. These items are not often selected, so the methods discussed above can be helpful to investigate the functioning of these items. Note that the item score patterns are generated on the basis of the 3-PLM and thus fit this model. In the first simulation study, it was my aim to show how nonparametric IRFs look for items with different item configurations used in the LSAT item pool.

Second, I analyzed empirical data of a computer-based examination used by a large testing agency to select persons. This test consists of 20 four-choice items; a sample of 213 persons was available.

I used MSP5 for Windows (Molenaar & Sijtsma, 2000) to conduct a Mokken scale analysis and TestGraf (Ramsay, 2000) to obtain graphs of the response functions. The search procedure in MSP5 was used to investigate dimensionality, and the *H*-values were used to investigate monotonicity. Furthermore, I used DIMTEST (e.g., Stout et al., 1996) to obtain additional information about the dimensionality of the data. DIMTEST is a nonparametric testing procedure to investigate the dimensionality of a set of items. It is based on detecting violations of local independence between item pairs when conditioning on the total score, which is used as an estimate of θ. When applying DIMTEST, two subtests of items should be specified from the k items of the test. The first group of M items (Assessment Test 1, AT1) consists of items that are dimensionally homogeneous (either determined by expert opinion or on the basis of a statistical technique like factor analysis or cluster analysis). The second group of M items (Assessment Test 2, AT2) from the $k - M$ items are 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 comprise the partitioning subtest (PT) on the basis of which the persons are partitioned into subgroups according to their total scores. DIMTEST calculates for each assessment subtest the difference in observed variance of the AT1 and AT2 scores with the variance of the binomial model. 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 subgroups. This statistic is asymptotically normally distributed with mean 0 and variance 1. Values larger than the upper 100 $(1 - \alpha)$ percentile indicate multidimensionality, with α denoting the type I error. In my analysis, I determined the items of AT1 and AT2 on the basis of the default option in DIMTEST that uses factor loadings for selecting items.

The graphs obtained by TestGraf were used to investigate monotonicity and the existence of lower and upper asymptotes. In addition, I used these graphs to investigate the specific form of the IRF.

Results

In Table 1, the H_g values are given. Note that almost all items had low H_g values. This can be explained by the fact that all items had non-zero lower asymptotes with the result that the discrimination power is reduced. Figure 1 shows the item response functions of items 1, 2, and 15 obtained by TestGraf.

TABLE 1
H_g values for CAT

Item	H_g
1	.45
2	.16
3	.17
4	.18
5	.19
6	.21
7	.13
8	.22
9	.23
10	.19
11	.21
12	.31
13	.16
14	.25
15	.31

It is very illuminating to inspect the IRFs for items 1, 2, and 15 (Figure 1). Inspecting the curve of item 2 shows that this item hardly discriminates between $X_+ = 8$ and $X_+ = 21$. One could argue that this item is not a good item in the sense that, when we see knowledge as a continuum, it implies that scores should differentiate individuals with varying degrees of knowledge across the entire range of θ values. It also implies that the probability of endorsement should increase smoothly for larger values of the test, rather than abruptly at a specific threshold. This is not the case for item 2. Now consider the IRF of item 15. This is an item with a non-zero lower asymptote, and it only discriminates between total scores between 7 and 16. In a CAT, this item will probably not be selected because of its relatively low item discrimination power. However, for the range of X_+ between $X_+ = 8$ and $X_+ = 13$ the item discriminates well between persons. The confidence intervals are also smallest in this area, and the measurement is most reliable. It is possible 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, note that a low discrimination parameter (low H_g-values) may be the result of different types of IRFs, which is very interesting information when constructing a CAT. In contrast to these items, item 1 discriminates well across the score range.

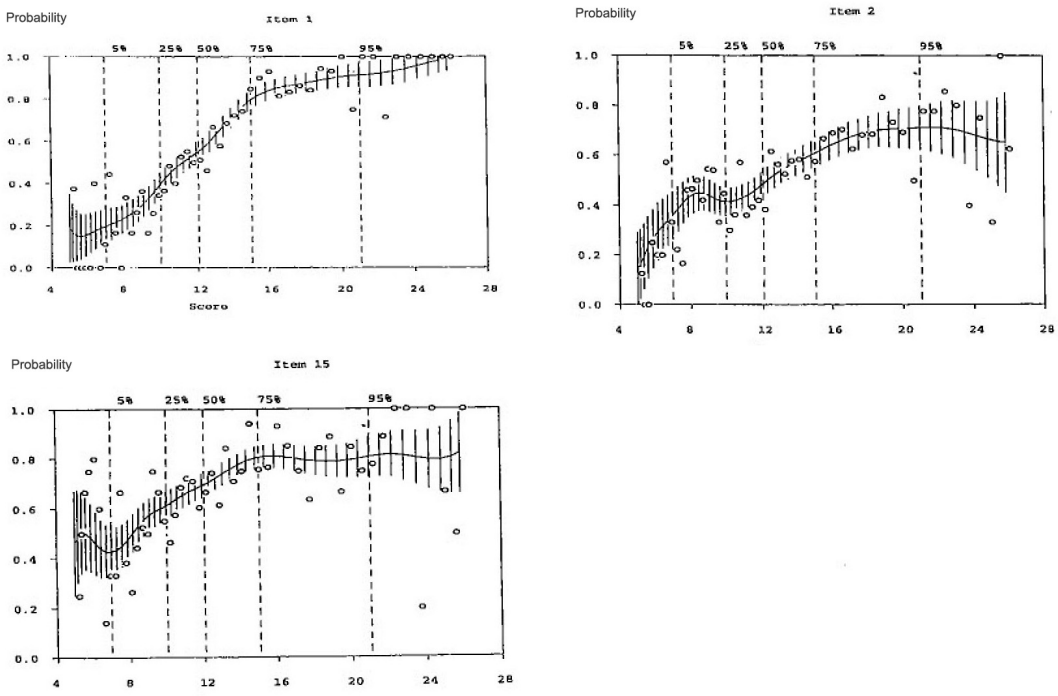

FIGURE 1. *Item response functions for simulated data using the 3-PLM*

Now consider the IRFs of some items in the computer-based examination. In Figure 2, some examples are given. Inspecting the graph of item 6 it can be seen that in the middle range of the score distribution this item hardly discriminates between persons. Thus, a person with a low score will have a relatively high probability of endorsing this item, whereas a person with a high score will have a relatively low probability (i.e., lower than 1). Inspecting the curve of item 20 it can be seen that this curve does not comply with the logistic model and even levels off at the end of the score range.

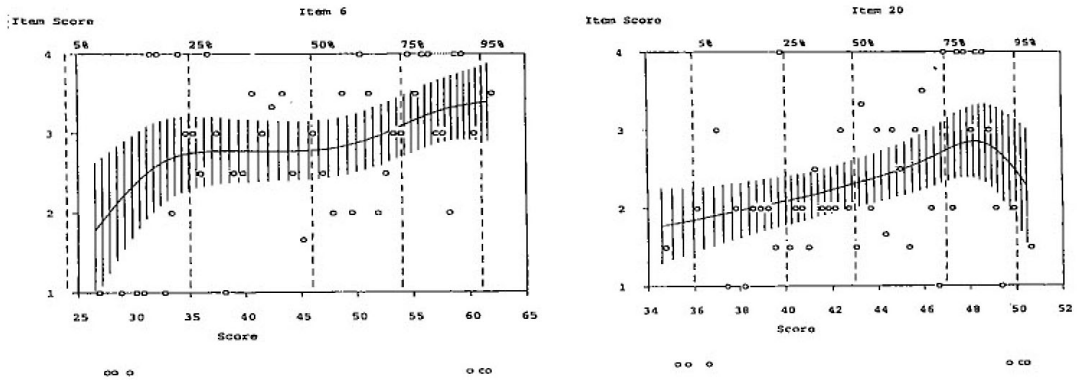

FIGURE 2. *Two item response functions from a computer based test*

Note that by investigating 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 us 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, I argue for a better data exploration when constructing and revising a measurement instrument and before using parametric IRT modeling, in particular because it enables us to better understand how an item and a test is functioning.

Investigating multidimensionality, I found that the 20 item computer test fell into two subscales with 7 and 8 items each, whereas 5 items were not scalable. Using DIMTEST for the whole scale, I found that $T = 6.34$ with $p = .003$, which also point at a multi-dimensionality.

Discussion

In this study, I explored the use of nonparametric IRT to investigate the data structure of data that are usually described by logistic parametric IRT models. I showed that through these models, information can be obtained about the functioning of items that is more difficult to obtain using parametric models. In my view, IRT models (both parametric and nonparametric) are helpful tools to learn more about the structure of empirical data sets, and the preference for one model over another should be made by an individual researcher in the specific context of his or her particular research.

It is my belief that nonparametric models are useful models to explore the data structure, and I showed that staying close to the data (i.e., not assuming a specified logistic curve) prevents the researcher from jumping to conclusions. Furthermore, I 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 procedures that can be applied to relatively small datasets (say, between 300 and 400 persons). Future research may focus on the development and application of goodness-of-fit tests in the nonparametric area to CAT (see e.g., Hart, 1997).

References

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

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: Lawrence Erlbaum.

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.

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

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

Hart, J. D. (1997). *Nonparametric smoothing and lack of fit tests*. New York: Springer.

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 annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

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.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: 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., & Sijtsma, K. (2000). *MSP5 for windows, a program for Mokken scale analysis for polytomous items*. Groningen: iec ProGAMMA, The Netherlands.

Ramsay, J. O. (2000). *TestGraf. A program for the graphical analysis of multiple choice tests and questionnaire data*. Unpublished manuscript, McGill University, Montreal, Quebec, Canada.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8(2)*, 164–184.

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 analyses 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.

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

Veldkamp, B. P., van der Linden, W. J., & Ariel, A. (2003). Mathematical programming approaches to test item pool design. *Advances in Psychological Research* (pp.93–108). New York: Nova Science Publishers.

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.