LSAC RESEARCH REPORT SERIES

Simple Nonparametric Checks for Model Data Fit in CAT

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

Law School Admission Council
Computerized Testing Report 01-04
December 2005

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 development and operational use of computerized adaptive testing (CAT) depends heavily on the availability of item response theory (IRT) models. This fact is not surprising, because one of the key features of IRT models is separate parameters for the properties of the items and the ability of the examinee. Due to this feature, the examinee's ability level can be estimated from different sets of items, and items can be selected to be optimal at ability estimates. Recent developments in nonparametric IRT, however, suggest that techniques derived from this field may also contribute to the improvement of the psychometric quality of CAT. An important feature of these techniques is that they are based on very weak assumptions and therefore nearly always apply.

The aim of this paper is to investigate the possible use of nonparametric IRT in CAT. In particular, we explore the use of some simple diagnostics statistics derived from nonparametric IRT. As shown by numerical examples, these statistics are able to check basic properties of item response functions such as nondecreasing probability of success as the ability of the examinees increases and whether the ordering of the item difficulties in the pool is the same for each examinee. Checks of the latter property are important to finding out, for instance, if the ability measured by the CAT is curriculum-independent.

Introduction

In this paper, the usefulness of several nonparametric checks is discussed in a computerized adaptive testing (CAT) context. Although there is no tradition of nonparametric scalability in CAT, it can be argued that scalability checks can be useful to investigate, for example, the quality of item pools.

Although IRT models are strongly embedded in the development and construction of CAT, the development of CAT is strongly related to parametric as opposed to nonparametric IRT modeling. This is not surprising because one of the key features of a CAT is the item selection procedure on the basis of an estimated latent trait from a calibrated item pool. Parametric IRT models enable the separate estimation of item and person parameters and, thus, facilitate this process enormously. The recent developments in nonparametric IRT, however, also suggest that techniques and statistics used in this IRT field may contribute to the development and improvement of the psychometric quality of a CAT. Investigating nonparametric IRT modeling may also help us to gain insight into the assumptions underlying CAT and may help to unify IRT modeling.

The aim of this paper is to investigate the possible use and application of nonparametric IRT in a CAT. Moreover, we will explore the use of some simple diagnostics using nonparametric IRT in a CAT. A CAT can be used in a number of ways. It can be used, for example, for entry testing, certification testing, or growth assessment. When a CAT is used as a diagnostic instrument, statistics that would pinpoint areas of difficulty would be extremely useful for a teacher who wants to individualize instruction in the classroom (e.g., Kingsbury & Houser, 1999). Students have all sorts of different problems in keeping up with the curriculum, and it is sometimes very difficult for a teacher to determine content areas in which the student might be having problems. In this paper, we explore the use of some diagnostic statistics. In fact, several of these methods may be an effective first step to getting an impression of test results and parts of the curriculum that are not well-understood by the examinee.

First, we will discuss two often-used nonparametric IRT models and discuss the assumptions underlying IRT modeling. Second, we will discuss some of these assumptions in a CAT context. Finally, we will discuss some statistics that may be used to check the quality of data.

IRT Models

Parametric IRT Models

In IRT, the probability of obtaining a correct answer on item g ($g = 1, ..., k$) is explained by the latent trait value (θ) and the characteristics of the item (Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997). The conditional probability $\pi_g(\theta)$ is the item response function (IRF). Further, we define $\mathbf{X} = (X_1, ..., X_k)$ and a realization $\mathbf{x} = (x_1, ..., x_k)$. IRT often assumes that the item scores are locally independent,

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = L(\theta) = \prod_{g=1}^{k} \pi_g(\theta)^{x_g} [1 - \pi_g(\theta)]^{1 - x_g}. \quad (1)$$

For any cumulative probability distribution of θ, F(θ), θ can be integrated out, which yields

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{g=1}^{k} \pi_g(\theta)^{x_g} [1 - \pi_g(\theta)]^{1 - x_g} \, dF(\theta). \tag{2}$$

In order to have testable restrictions on the distribution of **X**, specific choices for $\pi_g(\theta)$, for F(θ), or for both have to be made. Whereas F(θ) sometimes is chosen to be normal, $\pi_g(\theta)$ often is specified using the 1-, 2-, or 3-parameter logistic model (1, 2, 3-PLM). The 3-PLM is defined as

$$\pi_g(\theta) = \gamma_g + \frac{(1 - \gamma_g)\exp[\alpha_g(\theta - \delta_g)]}{1 + \exp[\delta_g(\theta - \delta_g)]}, \tag{3}$$

where γ_g is the lower asymptote [γ_g is the probability of a 1 score for low-ability examinees ($\theta \to -\infty$)]; α_g is the item discrimination parameter; and δ_g is the item location parameter. The 2-PLM can be obtained by fixing $\gamma_g = 0$ for all items; and the 1-PLM or Rasch model can be obtained by fixing $\alpha_g = 1$ for all items.

Nonparametric IRT Models

The interest in nonparametric item response theory (IRT) models is growing as demonstrated by recent publications in this area (Sijtsma, 1998, for a comprehensive review). In some articles, characteristics of nonparametric IRT models are specified (e.g., Mokken & Lewis, 1982; Stout, 1987; van der Linden, 1998) other articles are concerned with goodness-of-fit statistics (e.g., Mokken & Lewis, 1982; Sijtsma & Meijer, 1992; Rosenbaum, 1987), or present examples of empirical studies in which nonparametric models are applied (e.g., de Jong, 1984).

Nondecreasing Item Response Functions

In nonparametric IRT models, the particular form of an IRF is not specified. By putting order restrictions on the IRF, testable consequences can be investigated.

Nonparametric IRT models put order restrictions on the IRF, but refrain from a parametric definition of the IRF (Sijtsma, 1998). An example of order restriction is that $\pi_g(\theta)$ is a nondecreasing function of θ (Junker, 1993; Mokken & Lewis, 1982; Stout, 1990). Let items be indexed by g and h (g, h = 1, ..., k), and examinees be indexed by i and j (i, j = 1, ..., n). For two arbitrarily chosen values of the latent trait, say $\theta_i < \theta_j$,

$$\pi_g(\theta_i) \leq \pi_g(\theta_j). \tag{4}$$

The nonparametric IRT model based on Equations 1 and 4 is the monotone homogeneity model (MHM; Mokken & Lewis,1982); also see Holland and Rosenbaum (1986) and Ellis and van den Wollenberg (1993).

An important result from the literature is that, when monotonicity in θ holds for a set of IRFs, the covariance between the items should be positive. This covariance property using different methods of proof can be found in Mokken (1971), Holland (1981), and van der Linden (1998). Mokken used the notion of similarly ordered functions. Let π_g denote the proportion of examinees responding correctly to item g and let π_{gh} denote the proportion of persons responding correctly on both items g and h. The covariance between the scores on items g and h is defined as $\sigma_{gh} = \pi_{gh} - \pi_g\pi_h$. Assuming that $g > h$ implies $\pi_g < \pi_h$, the maximum covariance is obtained if $\pi_{gh} = \pi_g$ and $\sigma_{gh}(\max) = \pi_g(1 - \pi_h)$.

Assuming $i < j$ implies $\theta_i < \theta_j$, two IRFs, $\pi_g(\theta)$ and $\pi_h(\theta)$, are similarly ordered when for each pair of function values, θ_i and θ_j,

$$[\pi_g(\theta_i) - \pi_g(\theta_j)][\pi_h(\theta_i) - \pi_h(\theta_j)] \geq 0. \tag{5}$$

Because of the monotonicity in θ under the MHM it is immediately clear that all IRFs are similarly ordered. It can be shown (Mokken, 1971) that for two similarly ordered functions,

$$\int_{\theta} \pi_g(\theta)\pi_h(\theta) \geq \int_{\theta} \pi_g(\theta) \int_{\theta} \pi_h(\theta). \tag{6}$$

Integrating across θ yields $\pi_{gh} \geq \pi_g \pi_h$ or in terms of covariance

$$\sigma_{gh} = \pi_{gh} - \pi_g \pi_h \geq 0. \tag{7}$$

When $\pi_g(\theta)$ or $\pi_h(\theta)$ are constant over the population, $\pi_{gh} = \pi_g \pi_h$, the value expected for π_{gh} under the hypothesis of marginal independence of the two item scores. Note that for a set of items in this case Equation 2 equals

$$P(\mathbf{X} = \mathbf{x}) = \prod_{g=1}^{k} \pi_g^{x_g} [(1 - \pi_g)]^{1 - x_g}. \tag{8}$$

Although from a point of view of measurement theory the hypothesis of marginal independence represents unsuccessful measurement, it may serve as a baseline against which the probability of an item score pattern under the MHM may be compared. One may argue, for example, that under the MHM the probability of some score patterns should be higher than under marginal independence.

The practical importance of the MHM model is that it implies that the latent trait θ is stochastically ordered (s.o.) by the unweighted sum of scores on the J items from the test, $X = \sum X_j$. That is, for two values of X_+, say s and t, and any value of θ, say c,

$$P(\theta > c \mid X_+ = s) \leq P(\theta > c \mid X_+ = t) \text{ for all } s < t.$$

Nonintersecting Item Response Functions

Another example of an order restriction is that the IRF's are nonintersecting (Mokken & Lewis, 1982; Rosenbaum, 1987 a, b). This order restriction can be formalized as follows: For a finite set of k items, all measuring the same unidimensional latent trait θ, we assume that the items can be ordered and numbered such that

$$\pi_1(\theta) \geq \pi_2(\theta) \geq \ldots \geq \pi_k(\theta), \text{ for all } \theta. \tag{9}$$

Equation 5 is an order restriction on the k IRFs of the test. The IRFs do not intersect, which means that the ordering of the probabilities is the same except for possible ties, for all values of θ. If items have an ordering as in Equation 5, they have a latent scale or invariant item ordering (IIO) (Sijtsma & Junker, 1996). Mokken (1971) referred to Equation (9) as the model of double monotonicity (MDM). Important is that both the MHM and the MDM are stochastic versions of the deterministic Guttman model. The Guttman model (Guttman, 1944, 1950) is defined by

$$\theta < \delta_g \Leftrightarrow \pi_g(\theta) = 0$$

and

$$\theta \geq \delta_g \Leftrightarrow \pi_g(\theta) = 1.$$

The Guttman model thus excludes a correct answer on a relatively difficult item h and an incorrect answer on an easier item g by the same examinee: $X_h = 1$ and $X_g = 0$, for all $g < h$. Such item score combinations (0,1) are called *errors* or *inversions*. Item score patterns (1,0) are permitted and are known as *Guttman patterns* or *conformal* patterns.

Note that the MHM model is a special case of the often used three and two parameter logistic models. Both the MHM model and the logistic models assume nondecreasing IRFs, but the MHM model refrains from the logistic function.

Nonparametric IRT and CAT

Nondecreasing Item Response Functions

Based on the property of stochastic ordering of the total score in θ and the property of positive covariances between the items, several test and item fit statistics have been proposed for paper-and-pencil tests. To be able to use these statistics in a CAT, it is necessary to investigate when the property of s.o. applies in a CAT.

Stochastic ordering of the response variable in the latent trait does not, however, automatically follow for adaptive testing. For a CAT, van der Linden (1998) showed that any nondecreasing function of the response vector stochastic ordering in the latent trait only applies for a fixed-length adaptive test from a 1-PLM item pool when maximum information is used for the item selection and expected a posteriori estimation is used for the ability parameter. Then, the response variables are s.o. in the latent trait and it is thus possible to check for positive covariances between the items selected in the test. It is important to realize that this does not apply for the items in the item pool. That is, stochastic ordering only holds for a realization of a test.

Given these restrictions, it may be helpful to check for positive covariances between the items for examinees that obtain the same set of items (test) in a CAT research program. Checking the covariances may be helpful to check if the items are in agreement with the model. For example, when items are often used, they may become known, and the unidimensional structure under the model is violated, which will be reflected in the covariance structure underlying the items. This analysis may especially be helpful when item pools are relatively small and numbers of persons are relatively small. This may be the case when items are calibrated and not many students are available due to motivation problems in large scale testing, or when CAT is used in small applications, for example in classroom testing or diagnostic testing.

This check may also be used in testlet-based computer adaptive testing where a testlet is a group of items developed as a unit, for example, covering the same content or based on a common reading passage. In a testlet-based CAT, there is a fixed number of predetermined paths that an examinee may follow (Wainer & Kiely, 1987). In a CAT based on testlets, the first testlet selected is based on an initial estimate of θ. After each testlet θ is estimated, the next testlet selected is the testlet with maximum information at the updated $\hat{\theta}$. The CAT ends when the precision of $\hat{\theta}$ is adequate or when a certain number of testlets is administered. Examinees with the same item score patterns may be compared.

Nonintersecting Item Response Functions

Although it is not often discussed in the selection procedure of a CAT, when selecting items from an item pool it is assumed that the item difficulty ordering is the same for each examinee. This is the case because in a CAT the next item is selected based on the current θ estimate and it is assumed that the item ordering is the same for each person. If this is not the case, it may lead to inefficient use of the item pool. This assumption of IIO in a CAT can be checked by methods proposed by Mokken (1971) and Rosenbaum (1984) although these methods assume a complete item-by-person matrix, which is not realistic in a CAT.

Diagnostic Checks

On the person level, invariant item ordering statistical tests are available that compare subtest scores on subsets of items that have different mean item difficulties. Different subtests may be compared to investigate consistency of answering behavior, or preknowledge of items of subgroups of items. Well-known nonparametric statistics may be used to control for individuals that do not have the assumed alternation of correct and incorrect responses.

For example, the method proposed by Sijtsma & Meijer (2001) can be used to obtain information on which parts of the item score pattern deviant behavior is present. This may help the researcher to interpret the type of aberrant behavior. For example, many unexpected item scores at the beginning of the test may indicate careless behavior (Trabin & Weiss, 1983). Sijtsma & Meijer (2001; see also Trabin & Weiss, 1983) discussed a statistical method in which an item score pattern is divided into k subsets of items, so that A_1 contains the m easiest items, A_2 contains the next m easiest items, and so on. Let X_{+e} denote the total score on the easiest subtest and let subsets A_g of m increasingly more difficult items be collected in disjoint and mutually exclusive vectors Y_g such that $Y = (Y_1, Y_2, ..., Y_G)$. Consider newly defined vectors $A_{(g)}$, each of which contains two adjacent subsets A_g and A_{g+1} : $Y_{(1)} = (Y_1, Y_2)$, $Y_{(2)} = (Y_2, Y_3)$, ..., $(Y_{(G-1)}) = (Y_{(G-1)}, Y_G)$. The statistical method also applies to each pair in $Y_{(g)}$. A useful question in CAT research is whether the number of correct answers on the less difficult items is exceptionally low given the total score and the subtest scores. To test this, Sijtsma and Meijer (2001; see also Rosenbaum, 1984) showed that for each pair, a conservative bound (denoted ϑ) based on the hypergeometric distribution can be calculated for the probability that a

person has at most $X_{+e} = x_{+e}$ 1s on the easiest item subset. If for a particular pair ϑ is lower than, say 0.05, then the conclusion is that the total score on the first subset in this pair is unlikely given that all items in the first subset are easier than the items in the second subset. Although this method is based on item response functions that are nonintersecting, Sijtsma & Meijer (2001) showed that the results are robust against violations of this assumption. Note that items may be ordered according to their item difficulty δ or to the item proportion correct score π.

This method can easily be implemented in CAT research by ordering the items for each person from easy to difficult according to their item difficulty from IRT or the item proportion correct score. Another strategy is to stratify the item bank in easy, medium, and difficult items and then compare scores of the examinees on easy and difficult items. This method may be interesting comparing different subject matters especially.

This method can also be combined when subgroups of items are used as when testlets are used in a CAT (testlets, again, are item clusters such that when an examinee gets one item he/she also gets a number of the other items from the cluster). Item score patterns of examinees with the (almost) same item scores can be checked. As discussed above, for these patterns, item covariances should be positive (item-fit) and also the expected correct answers should be in agreement with the observed answers.

A second diagnostic method can be used to investigate whether an examinee has difficulty in getting started; the examinee's response on the first items may be incorrectly answered due to start-up problems and not due to lack of knowledge. As a fit statistic, the number of correct answers may be taken on the first a items

$$W = \sum_{g=1}^{a} X_g. \tag{10}$$

Note that when all items have the same item response functions, W is binomially distributed. When the items have different item response functions, as is usually the case, the distribution of W follows the generalized binomial distribution (Lord, 1980, p. 45; Kendall & Stuart, 1969, 5.10). This distribution cannot be written in a simple form but the mean equals

$$\mu = \sum_{g=1}^{a} \pi_g \tag{11}$$

and the variance

$$\sigma^2 = \sum_{g=1}^{a} \pi_g (1 - \pi_g). \tag{12}$$

However, the generalized binomial distribution can be approximated (Lord, 1980, p. 45) by the binomial distribution using $\overline{\pi}$, the mean proportion correct score across items

$$\binom{a}{x} \overline{\pi}^x (1 - \overline{\pi})^{a-x}. \tag{13}$$

The major difference between the generalized binomial distribution and the binomial approximation in (13) is that the variance of the generalized binomial distribution (σ^2) is always less than the binomial variance using mean π. That is, the variance of the binomial using mean π is $a\overline{\pi}^x(1 - \overline{\pi})^{a-x}$. Let σ_π^2 be the variance between the π values; then the difference between the two variances is

$$\sigma^2 = a\overline{\pi}(1 - \overline{\pi}) - a\sigma_\pi^2 \tag{14}$$

Note that when $\overline{\pi} = \pi$ for all items, the variances for the binomial distribution and the generalized binomial distribution are the same. When items are selected with π values that are similar, the binomial and the generated binomial distributions are similar. Note that although the proportion correct score is dependent on the latent trait distribution, this is not a problem as long as one is interested in a specific population as we often will be in classroom diagnostic testing.

To illustrate this, suppose that we are interested in checking whether the number correct score on the first items on an a test is unexpectedly low given the π values of the items. We can use the same ordering as the order in which the items are presented or we can use the order according to the item difficulty. As an example, we use the π-values of the first four items of a simulated CAT with $\pi_1 = .69$, $\pi_2 = .87$, $\pi_3 = .93$, and $\pi_4 = .81$. Suppose now an examinee generates an item score pattern (0, 0, 0, 1), then according to (13) the probability of occurrence, denoted η, of such a pattern is 0.018. Furthermore, $a\sigma_\pi^2 = .042$ and $a\bar{\pi}(1 - \bar{\pi}) = 0.577$. Thus, the difference of the variance of the generated binomial and the approximation with the binomial with mean π values is small. An obvious drawback of this method is that the number correct score is not taken into account. Therefore, this method is only useful when the researcher first selects examinees with respect to the total score and then selects a subset of the test for which it would be very strange to have a lot of zero scores. For example, we used this method to detect examinees with many incorrect scores in the beginning of the test.

A third method is particularly based on the alternations of correct and incorrect scores on a CAT and is based on the runs test. Suppose an examinee responds to an adaptive test of length k. Let the vector $X = (X_g, ..., X_k)$ be the dichotomous item score vector. Let n_0 denote the number of incorrect responses, n_1 the number of correct responses, and r the total number of runs in a specific response vector. A run is defined as the number of consecutive one scores or zero scores. The probability distribution of *TNR*, the total number of runs of $n = n_0 + n_1$ is known (e.g., Gibbons and Chakraborti, 1992, pp. 72–73) and can be used to classify an item score pattern as normal or aberrant. Because few runs may indicate aberrant response behavior, the significance probability of the observed response vector is the probability of r runs or less.

An alternative is to use the length of the longest run. Let r_{vw} denote the number of runs of type $v = 0, 1$ which are of length $w = 1, ... , n_v$. Let l denote the length of the longest run observed in the response pattern U. Longer runs are more improbable; thus, the significance probability of the random variable *LLR*, the length of the longest run of $n = n_0 + n_1$ objects, is the probability of getting at least one run of length l or more of either type 1 or 0. This probability was derived by Mosteller (1941).

Examinees can be classified as nonfitting when the significance probability of the total number of runs and/or the length of the longest run is smaller than some predefined significance level α; that is, $p^*(TNR) < \alpha$ and/or $p^*(LLR) < \alpha$.

An Example

Five datasets consisting of 10,000 fitting adaptive item score vectors were constructed at five different θ levels: $\theta = -2, -1, 0, 1$, and 2. An item bank of 400 items fitting the 2-PLM with $a_i \sim N(1.0, .2)$ and $b_i \sim U(-3, 3)$ was used to simulate adaptive item score patterns with fixed test length $N = 30$ items.

An adaptive item score pattern was simulated as follows. First, the true θ of a simulee was drawn from $N(0; 1)$ truncated to the interval [–2.5, 2.5]. Then, the item with maximum Fisher information (given $\theta = 0$) was selected as the first item (e.g., Hambleton & Swaminathan, 1985, pp. 101–124). To simulate the answer (1 or 0), a random number, y, was drawn from a uniform distribution on the interval [0, 1]. When $y < P(\theta)$, the response to item I was set to 1 (correct response); the response was set to 0, otherwise (incorrect response). The next three items selected also had maximum Fisher information for $\theta = 0$. Based on the responses to the first four items, $\hat{\theta}$ was obtained. The next item selected was the item with maximum information given $\hat{\theta}$. $P(\theta)$ was then computed for the fifth item, and a response was simulated. Again, θ was estimated, and then the next item was selected based on maximum information given $\hat{\theta}$. This procedure was repeated until the test contained $N = 30$ items.

To simulate aberrant answering behavior, we assumed that an examinee answered the test using two different θ levels. When an examinee has a different θ value during the first half of the test than during the second half, that examinee is said to have "noninvariant ability" (Klauer, 1991). Carelessness and fumbling can cause noninvariant θs. Response vectors were simulated for a two-dimensional value of θ. It was assumed that during the first half of the test an examinee had a different θ value than during the second half. Datasets containing response vectors with a two-dimensional θ were simulated by drawing the first ability value, θ_1, from the standard normal distribution truncated [–2.5, 2.5], and the second ability value was determined as $\theta_2 = \theta_1 + r$, where $r = 2$ was used. Thus, during the first half of the test, $P(\theta_1)$ was used and, during the second half, $P(\theta_2)$ was used to simulate the responses to the items. An example of response behavior with $r = 2$ is a warming-up effect in the first half of the test.

Detection rates for all four methods discussed were determined.

Results

In Table 1, the detection rates for ϑ, η, and *TNR* and *LLR* are given. It can be seen that ϑ has the highest overall detection rate. Also, the detection rate increases when θ increases. This can be explained by the fact that the probability of correct answers increased in θ.

TABLE 1
Detection rate using the cumulative hypergeometric ϑ, the generalized binomial distribution, η, TNR, and LLR, for invariant ability simulees

θ	α	ϑ	η	*LLR*	*TNR*
–2	0.05	0.017	0.003	0.091	0.092
	0.10	0.062	0.006	0.099	0.104
–1	0.05	0.096	0.012	0.096	0.145
	0.10	0.174	0.040	0.174	0.220
0	0.05	0.302	0.048	0.102	0.426
	0.10	0.451	0.112	0.251	0.423
1	0.05	0.450	0.134	0.350	0.531
	0.10	0.652	0.256	0.452	0.412
2	0.05	0.548	0.347	0.583	0.473
	0.10	0.798	0.449	0.678	0.560

Discussion

Depending on the application envisaged, a standard remark can be attached to the test scores if the pattern fits the model. This can be fully automated and is very easy to implement. In fact, an example of a personnel "validity index" is implemented in the New Hampshire Educational and Improvement Assessment program, a statewide paper-and-pencil assessment program administered to grades 3, 6, and 10. This index is used to communicate about the regularity of item score patterns of an examinee to parents. If a pattern is suspicious, this is communicated by a remark on the paper that the irregularity may be caused by lack of motivation, illness, etc. Fit of item score patterns to a test model may even be crucial in high stakes testing when examinees may question the fairness of the test model to select items. Item and person fit may then be used to back up decisions made.

References

De Jong, A. (1984). *Over psychiatrische invaliditert (On psychiatric disability)*. Unpublished doctoral dissertation, Rijksuniversiteit, Groningen.

Ellis, J. L., & van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika, 58*, 417–429.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference*. New York: Marcel Dekker.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

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. Claussen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46*, 79–92.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait variables, *The Annals of Statistics, 14*, 1,523–1,543.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378.

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics*, vol. 1, (3rd ed.). New York: Hafner.

Kingsbury, G. G., & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 93–115). Mahwah, NJ: Erlbaum.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*, 535–547.

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417–430.

Mosteller, F. (1941). Note on an application of runs to quality control charts. *Annals of Mathematical Statistics, 12*, 228–231.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425–435.

Rosenbaum, P. R. (1987a). Comparing item characteristic curves. *Psychometrika, 52*, 217–233.

Rosenbaum, P. R. (1987b). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40*, 157–168.

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

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79–105.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149–157.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191–207.

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

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

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 83–108). New York: Academic Press.

van der Linden, W. J. (1998). Stochastic order in dichotomous item response models for fixed, adaptive, and multidimensional tests. *Psychometrika, 63*, 211–226.

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

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–202.