Tilburg University

Methodology review

Sijtsma, K.

Published in:
Applied Psychological Measurement

DOI:
[10.1177/01466216980221001](https://doi.org/10.1177/01466216980221001)

Publication date:
1998

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, 22(1), 3-31. <https://doi.org/10.1177/01466216980221001>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Applied Psychological Measurement

<http://apm.sagepub.com>

Methodology Review: Nonparametric IRT Approaches to the Analysis of Dichotomous Item Scores

Klaas Sijtsma

Applied Psychological Measurement 1998; 22; 3

DOI: 10.1177/01466216980221001

The online version of this article can be found at:
<http://apm.sagepub.com/cgi/content/abstract/22/1/3>

Published by:

<http://www.sagepublications.com>

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: <http://apm.sagepub.com/cgi/alerts>

Subscriptions: <http://apm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations <http://apm.sagepub.com/cgi/content/refs/22/1/3>

Methodology Review: Nonparametric IRT Approaches to the Analysis of Dichotomous Item Scores

Klaas Sijtsma
Tilburg University, The Netherlands

This paper reviews developments in nonparametric item response theory (NIRT). Historically, Mokken's (1971) item response theory (IRT) model and procedure of scale analysis have been the starting point for NIRT models; the developments that Mokken's work have inspired and those that originated independently are discussed here. Attention is focused on results pertaining to dichotomously scored items. The usefulness of new theoretical results for practical test construction is explained. NIRT is especially useful when the purpose of test application is to order persons, items, or both. The assumptions inherent in several NIRT models, model-data fit methods, item selection procedures,

and reliability methods are discussed. NIRT models often allow more items into a test than stronger models, yielding higher reliability and better coverage of the latent trait. Moreover, the theoretical results from NIRT are often relevant to parametric IRT. Examples include conditional association and manifest monotonicity, as well as recent results for polytomous NIRT models that concern the ordering of persons, which also hold for several parametric IRT models. *Index terms:* automated item selection, essential unidimensionality, invariant item ordering, model of double monotonicity, model of monotone homogeneity, Mokken model, nonparametric item response theory, ordinal person measurement.

During the development of item response theory (IRT) over the past three decades, nonparametric IRT (NIRT) approaches have played a relatively modest, albeit persistent role. The item response function (IRF) is central to both parametric IRT and NIRT approaches. The IRF is defined as the relation between the probability of obtaining a score of 1 (the keyed response) on a dichotomous item and the latent trait (θ). Unlike parametric IRT models, NIRT models do not assume a particular parametric form for the IRF. A typical assumption of a NIRT model is that the IRF is nondecreasing in θ . Given this order constraint, any form of the IRF is acceptable.

NIRT has provided several new insights that are interesting for IRT. Methods for investigating whether a NIRT model fits the data are often easy to apply in a model-data fit analysis. Although a parametric definition of the IRF is avoided, NIRT allows person or item orderings on the θ scale using simple summary statistics, such as the total test score and the mean item score, respectively. From a practical point of view, NIRT models can be useful in the analysis of test data, especially if a researcher is not willing to assume a particular parametric form for the IRFs.

For these reasons, it may be useful for researchers, both at the theoretical and practical level, to increase their knowledge of NIRT. This paper provides a critical overview of the developments in this field. Because one paper cannot do justice to the more detailed developments, a comprehensive reference list is provided.

The focus here is almost exclusively on NIRT for dichotomous items. Much attention is given to matters of practical test and questionnaire construction. The theoretical and practical developments often are placed in a wider framework of other NIRT and parametric IRT approaches. Polytomous NIRT is not surveyed because the developments are rather new (e.g., Hemker, 1996;

Junker, 1991; Molenaar, 1997) and more time is needed to resolve the problems that are involved in these more complex models (e.g., Hemker, Sijtsma, Molenaar, & Junker, 1996, 1997; Junker, in press; Scheiblechner, 1995).

NIRT researchers pursue models that allow certain measurement applications with relatively weak assumptions. For example, if the purpose is to order persons on a scale, models that allow ordinal person measurement are needed. Given this measurement goal, the measurement of persons on an interval scale or the scaling of items are of secondary interest and may not be incorporated into the model selected. Thus, rather than modeling individual IRFs in terms of monotonicity or order restrictions (Ellis & Van den Wollenberg, 1993; Mokken & Lewis, 1982), assumptions may involve monotonicity restrictions on the mean of the IRFs, leaving the individual IRFs free within the boundaries set by this restriction (Junker, 1993; Stout, 1990). Note that a monotonicity restriction on the mean IRF allows the possibility that an individual IRF is locally decreasing, as long as this decrease is compensated for by increases in other IRFs at the same θ values. Other assumptions may weaken the assumption of local independence by the introduction of a mild form of multidimensionality in which one trait dominates the others. Consequently, conditioning on the dominant trait does not result in all covariances being exactly 0 (Stout, 1990).

The NIRT approaches advocated by Cliff & Donoghue (1992) and Ramsay (1991) are different from other NIRT models because they concentrate on observable variables to make inferences about test performance and avoid the explicit formulation of sets of assumptions and their consequences. For example, Ramsay (1991, 1995) discussed methods for estimating IRFs and response functions for each incorrect option to assess the quality of dichotomously scored multiple-choice items, but without assuming and testing an underlying latent structure, as defined by NIRT models. Alternative approaches to ordinal test theory other than NIRT were proposed by Cliff (1977) and Schulman & Haden (1975).

Origins and Applications of NIRT Models

Meredith (1965) was one of the first authors to discuss an approach to IRT that is now characterized as NIRT. The first comprehensive treatment was by Mokken (1971). The core ideas of his monograph were published by Mokken & Lewis (1982). Mokken's work was further developed by Stokman & Van Schuur (1980); Lewis (1983); Roskam, Jansen, & Van den Wollenberg (1986); Mokken, Lewis, & Sijtsma (1986); Sijtsma & Molenaar (1987); Croon (1991); Ellis (1994); Meijer, Sijtsma, & Molenaar (1995); Hemker, Sijtsma, & Molenaar (1995); and Sijtsma & Junker (1996). Independent of this research, nonparametric conceptualizations of IRT were also proposed and studied by Holland (1981); Rosenbaum (1984, 1987a); Stout (1987, 1990); Ramsay (1991); Junker (1991, 1993); and Douglas (1997). Somewhat different but related nonparametric approaches to the analysis of test data have been proposed by Meredith (1965), Cliff (1979), and Cliff & Donoghue (1992).

Mokken's methods have been applied to the practical construction of tests and questionnaires in a variety of research areas, using the programs MOKKEN SCALE (Niemöller, 1980), MOKKEN TEST (Niemöller & Van Schuur, 1980), and MSP (Molenaar, Debets, Sijtsma, & Hemker, 1994). These applications include developmental psychology (Kingma & Ten Vergert, 1985; Verweij, 1994; Verweij, Sijtsma, & Koops, 1996), political science research (Stokman, 1977), social psychology (Van der Veen, 1992), medical science (Ringdal & Ringdal, 1993; Weijmar Schultz & Van de Wiel, 1991), psychiatry (De Jong, 1984; De Jong & Molenaar, 1987), and intelligence (Laros & Tellegen, 1991). More recently, applications of NIRT were introduced by Stout (1987, 1990) using the data analysis procedures DIMTEST (Stout, 1987, 1990) and DETECT (Kim, Zhang, & Stout, 1996). TESTGRAF (Ramsay, 1993) is also available for the nonparametric estimation of IRFs according to a method proposed by Ramsay (1991).

The following topics are discussed here: (1) the two basic NIRT models proposed by Mokken (1971), which include the monotone homogeneity model (MHM) and the double monotonicity model (DMM); (2) further elaborations on the theory and testing of the MHM; (3) a scaling procedure for the MHM; (4) practical methods of item selection, proposed by Mokken (1971), Roussos & Stout (1995), and Kim et al. (1996); (5) observable consequences and fit investigation of the DMM; and (6) reliability estimation for individual item scores and total scores. Several sections include critical comments on developments, results, and missing links in NIRT research.

Two Basic NIRT Models

Fundamental to IRT is the idea that psychological constructs are latent (i.e., not directly observable) and that knowledge about these constructs can only be obtained through the manifest responses of persons to a set of tasks or items. These tasks or items, and the responses provided by an examinee, do not coincide with the construct of interest. 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.

For example, suppose θ is spatial orientation ability. A set of 20 items could consist of drawings of three-dimensional geometric objects that have to be rotated mentally to a prescribed position, and responses are scored as either correct or incorrect. Each examinee can be characterized by a particular θ level, whereas each item requires a particular θ level to be solved with a given probability. In other words, persons vary with respect to θ level and items vary with respect to difficulty, and possibly with respect to other characteristics such as discrimination. In a similar example, attitudes toward euthanasia could be assessed using a set of five statements about particular aspects of euthanasia or situations relevant to the issue. Responses could indicate whether a person endorses a statement or not. Persons and items can then be characterized by their positions on the attitude continuum.

IRT models are defined by sets of assumptions. These assumptions pertain to the number of θ s measured by a test or questionnaire and the relation between the observable responses and θ . An IRT model will be called nonparametric if it places only monotonicity or order restrictions on the IRFs, and refrains from a parametric definition of the IRF such as logistic or normal ogive. Note that in statistics the term nonparametric usually refers to the absence of assumptions with respect to distributions of variables (Lehmann, 1986, p. 107). NIRT could also be called nonparametric in this sense, but because most parametric IRT models also refrain from such assumptions, nonparametric will be used only in reference to properties of the IRFs.

Mokken (1971, 1997) and Mokken & Lewis (1982) proposed two NIRT models that can be considered basic to NIRT. Before discussing these models, notation will be defined. Let N be the size of a sample of persons who have responded to a set of k dichotomously scored items. Items are indexed g and h . Persons and items are measured on a scale that represents θ . Measurement values of persons are denoted θ and items are denoted δ . One response category is positively keyed (item score 1) with respect to θ , whereas the other is negatively keyed (item score 0); for ability and achievement items these response categories are usually designated "correct" and "incorrect," respectively. For personality, attitude, and opinion items, either category may be designated positive or negative, respectively, and is usually labeled "agree" or "disagree," "yes" or "no," "does apply" or "does not apply," and so forth, depending on the wording of the item.

The Monotone Homogeneity Model

The first model proposed by Mokken (1971) was the MHM. This model is defined by three assumptions. The first assumption is unidimensionality. This means that all test items measure the same θ or the same amalgamation of θ s so that persons are represented by one scalar parameter.

The assumption allows items to measure unique characteristics, and item responses are subject to random measurement error.

The second assumption is local independence. Let X be a random vector of k binary item score variables, and let x be a realization of X . The conditional joint probability of observing some score vector x is denoted $P(X = x|\theta)$. Let x_g and x_h be realizations of the binary item score variables X_g and X_h , respectively. The conditional univariate probability of observing an item score x_g is denoted $P(X_g = x_g|\theta)$; likewise, for item h . Local independence means that the conditional joint probability is the product of k conditional univariate probabilities,

$$P(X = x|\theta) = \prod_{g=1}^k P(X_g = x_g|\theta). \quad (1)$$

If the covariance between two item scores is the result of the variation of θ in the population, holding θ constant results in a 0 covariance [$\text{Cov}(X_g, X_h|\theta) = 0$]. This assumption of local independence is related to, but not identical to, unidimensionality (Lord & Novick, 1968, p. 361).

Unidimensionality and local independence can be represented as follows. Let the conditional probability of correctly answering item g , $P(X_g = 1|\theta)$, be abbreviated $P_g(\theta)$. The cumulative distribution of θ is denoted $G(\theta)$. Then the multivariate distribution of X can be written as

$$P(X = x) = \int_{\theta} \prod_{g=1}^k P_g(\theta)^{x_g} [1 - P_g(\theta)]^{1-x_g} dG(\theta). \quad (2)$$

Holland & Rosenbaum (1986; also Suppes & Zanotti, 1981) noted that Equation 2 does not restrict the distribution of X in any way, as long as there are no restrictions on $P_g(\theta)$, $G(\theta)$, or both. Usually, in IRT the restrictions are on $P_g(\theta)$.

The restriction on $P_g(\theta)$ is the third assumption of the MHM. This assumption states that for each item in a test the probability of obtaining a score of 1 is a monotonely nondecreasing function of θ . This function is the IRF. Let θ_i and θ_j denote fixed values of θ ; then,

$$P_g(\theta_i) \leq P_g(\theta_j), \text{ for all } \theta_i \leq \theta_j. \quad (3)$$

When applied to the spatial orientation ability example, this assumption says that the higher the θ level on spatial orientation, the higher the probability of responding correctly to an item that is a measure of spatial orientation. Likewise, for the euthanasia example, the more legalization of euthanasia is favored, the higher the probability that a positively worded item measuring this attitude is endorsed. A graphical display of IRFs for the MHM is given in Figure 1a.

The MHM restricts the structure of empirical data, but is not restrictive enough to allow estimation of θ and δ [however, see Mokken & Lewis (1982) and Lewis (1983) for the estimation of a modified θ parameter]. The MHM allows the ordering of persons with respect to θ by using the number of items on which a person obtains a score of 1. In practice, the unweighted sum of the item scores, denoted X_+ , is used to estimate the ordering of persons. This estimation is justified because the total score has a monotone likelihood ratio (MLR) in θ , given the three assumptions of the MHM (Grayson, 1988; Huynh, 1994). More specifically, for two realizations of X_+ , denoted s and t , MLR means that the likelihood ratio

$$\frac{P(X_+ = t|\theta)}{P(X_+ = s|\theta)} \quad (4)$$

is nondecreasing in θ whenever $s \leq t$.

MLR implies two types of stochastic ordering (Junker, 1993, proof of Lemma 4.1; see also Lehmann, 1959, p. 74; and Hemker, Sijtsma, Molenaar, & Junker, 1997). The first type of stochastic ordering states that the probability of having at least x_{+0} items correct is nondecreasing in θ ; that is,

$$P(X_+ \geq x_{+0} | \theta_i) \leq P(X_+ \geq x_{+0} | \theta_j), \text{ for all } \theta_i \leq \theta_j. \quad (5)$$

An implication of Equation 5 and also of Equation 3 (Mokken, 1971, pp. 128–129; Hemker et al., 1996) is

$$E(X_+ | \theta_i) \leq E(X_+ | \theta_j), \text{ for all } \theta_i \leq \theta_j. \quad (6)$$

Note that in Equations 5 and 6, the conditioning is on the unknown parameter θ . However, in practice X_+ is observed and a conclusion about θ is required. The second type of stochastic ordering that holds under the MHM accommodates this situation. Let θ_0 be an arbitrary value of θ ; then the probability that θ is larger than θ_0 is nondecreasing in the total (number-correct or number-keyed) score X_+ . That is,

$$P(\theta \geq \theta_0 | X_+ = s) \leq P(\theta \geq \theta_0 | X_+ = t), \text{ for all } s \leq t. \quad (7)$$

Equation 7 implies

$$E(\theta | X_+ = s) \leq E(\theta | X_+ = t), \text{ for all } s \leq t. \quad (8)$$

Equation 7 is extremely important for practical measurement using a weak IRT model such as the MHM, because it justifies the use of the total score X_+ to order persons on the θ scale (Hemker et al., 1997).

In the context of NIRT for unfolding or scaling of items measuring preferences, which is characterized by bell-shaped or single-peaked IRFs, Post & Snijders (1993; see also Post, 1992) obtained MLR and stochastic ordering properties, based on items rather than total scores, that are related but not equivalent to Equations 4 and 7 (Post, 1992, p. 36, A4; p. 42, Theorem 2). In fact, they demonstrated the equivalence of a stochastic ordering property like Equation 7 and an MLR property like Equation 4 when $X_+ = s$ is replaced by $X_h = 1$ and $X_+ = t$ by $X_g = 1$. Then the MLR property becomes a monotone traceline ratio (MTR) property that says

$$\frac{P(X_g = 1|\theta)}{P(X_h = 1|\theta)} \quad (9)$$

is nondecreasing in θ . Other MLR results relevant to NIRT were discussed by Rosenbaum (1984, 1985).

The assumption of nondecreasing IRFs (Equation 3) allows the possibility that an IRF is constant across an interval on the θ scale. In this interval, all persons have the same probability of receiving a score of 1. Note that this probability is the conditional expected item score [$P_g(\theta) = E(X_g|\theta)$]. For example, if Persons A and B are both located inside this interval, they cannot be ordered on the basis of their expected item scores. If one or both of them are located outside the interval, however, they can be ordered using the expected item scores. Therefore, except for possible ties due to such intervals, for each item the same ordering of persons using the expected item score is obtained. This is also true for expected total scores based on any subset of items from the domain that complies with the MHM (refer to Equation 6). Note that if Person A has the higher θ value, his/her probabilities of obtaining 1s are at least as high as Person B's on each item from the test. Thus, any item Person B would answer correctly, Person A would answer correctly with higher probability. If at least one IRF were not monotonically nondecreasing, this would generally not be true.

The Double Monotonicity Model

The second model proposed by Mokken (1971) was the DMM, a special and more restrictive case of the MHM. The same three assumptions that are made for the MHM are also true for the DMM; however, a fourth assumption for the DMM is that the IRFs do not intersect. Suppose that k items are ordered and numbered such that for the conditional probabilities of obtaining the positive response,

$$P_1(\theta) \leq P_2(\theta) \leq \dots \leq P_k(\theta), \text{ for all } \theta. \quad (10)$$

This is the assumption of invariant item ordering (IIO; Sijtsma & Junker, 1996). Equation 10 says that, except for possible ties, the ordering of the items is identical for all θ values. This is also true for each subgroup from the population for which Equation 10 holds. Sijtsma & Junker (1996) noted that the MTR property (Equation 9) is similar but not equivalent to the IIO property (Equation 10).

Rosenbaum (1987a) stated that item g is uniformly more difficult than item h if $P_g(\theta) \leq P_h(\theta)$ for all θ . Sijtsma & Junker (1996) argued that the assumption of IIO simply says that items can be arranged so that each is uniformly more difficult than the next (see Equation 10). If Equation 2 is adopted as a representation for the item responses and items are ordered by uniform relative difficulty [Rosenbaum, 1987b; i.e., equivalent to IIO (Equation 10)], then the items have a latent scale, using the terminology of Rosenbaum (1987b). Figure 1b shows IRFs for four items under the DMM.

IIO may be desirable in intelligence testing, differential item functioning (DIF) research, and person-fit research. In intelligence testing, IIO facilitates a simple form of adaptive testing allowing children to skip the items that are either too difficult or too easy (e.g., see Laros & Tellegen, 1991). Item orderings that are different in relevant groups may be indicative of bias, or DIF, against one of the groups (Hulin, Drasgow, & Parsons, 1983, p. 162). IIO on the person level is required in several approaches for the assessment of the fit of item score patterns per person (Meijer, Molenaar, & Sijtsma, 1994; Rosenbaum, 1987b).

In general, the property of IIO is relevant in research for which items are expected to have the same ordering across θ . Examples include the measurement of the developmental construct of transitive reasoning (Sijtsma & Junker, 1997; Verweij, 1994) and the attitude reflecting the commitment to political activity (Mokken, 1971). In the context of theory testing, Ellis (1993) discussed the relation between psychological theory and the validity of a theory at the person level, and invariance properties of covariance structures.

The position of the items on the θ scale is expressed by δ , the value of θ at which $P_g(\theta) = .5$ (Mokken, 1971, p. 174). This choice may differ from parametric IRT models, such as the logistic models, where the conditional probability that corresponds to δ lies halfway between the lower and upper asymptotes of the IRF. Sijtsma & Junker (1996) explained how that choice would lead to contradictions in the nonparametric DMM.

Like the latent person parameter θ , the latent item parameter δ cannot be estimated numerically. However, the item ordering according to δ can be replaced, except for possible ties, by the ordering according to the proportion of the population that gave a positive answer to item g (Mokken, 1971, p. 118, 126). This proportion is denoted π_g and is defined as

$$\pi_g = \int_{\theta} P_g(\theta) dG(\theta). \quad (11)$$

According to Equation 10, for $g < h$, $P_g(\theta) \leq P_h(\theta)$ for all θ ; thus, it follows that $\pi_g \leq \pi_h$. Note that in the MHM, the ordering by π_g would only provide information at the group level and not at the θ level. Figure 1a shows that the ordering of the conditional probabilities is invariant only in the intervals of θ in which the IRFs do not intersect.

The DMM might be considered a nonparametric counterpart of unidimensional, locally independent, parametric IRT models with nondecreasing IRFs that do not cross. Models of this type include the Rasch (1960) model (see also Fischer & Molenaar, 1995), the one-parameter normal ogive model (Hambleton & Swaminathan, 1985, p. 35), and a variation of the three-parameter logistic model (3PLM) with the same slope across the items and item locations and lower asymptotes oppositely ordered across the items. The IRF of the more familiar Rasch model can be defined as

$$P_g(\theta) = \frac{\exp(\theta - \delta_g)}{1 + \exp(\theta - \delta_g)}. \quad (12)$$

Figure 1c shows that the IRFs of the Rasch model are translations of each other along the θ scale. As a result, any set of Rasch IRFs is also a set of DMM IRFs. In general, however, DMM IRFs cannot be transformed into Rasch IRFs [refer to Levine (1970) for more specific results]. For example, the DMM allows IRFs with asymptotes that are not 0 and 1, whereas the Rasch model excludes such IRFs. In addition, DMM IRFs need not be symmetrical with respect to the point $(\delta, .5)$, which is the inflection point of the Rasch IRF.

The θ scale of the Rasch model has an interval metric (Fischer, 1995, p. 21, Theorem 2.2; also p. 31). For example, the definition of the model in Equation 12 implies a difference scale, and a

reparameterization such as $\xi = \exp(\theta)$ and $\epsilon = \exp(-\delta)$ implies a ratio scale (Fischer, 1974, p. 217). Due to the dependence of the scale properties on the parameterization, the interpretation of score differences on the θ scale is related to the model formulation selected and not to differences in equal units of a psychological trait. In fact, like the DMM, the Rasch model leads to ordinal interpretation when measurement values are compared. Note that in Equation 12 a score difference, $\theta_i - \theta_j$, equals the difference of two log-odds or, equivalently, a difference of logits (Hambleton & Swaminathan, 1985, pp. 57–58). The odds of θ_i producing a positive response to item g , denoted O_{ig} , is equal to $O_{ig} \equiv \exp(\theta_i - \delta_g)$; thus,

$$\theta_i - \theta_j = \ln(O_{ig}) - \ln(O_{jg}) \equiv \text{logit}[P_g(\theta_i)] - \text{logit}[P_g(\theta_j)] . \quad (13)$$

Equation 13 shows that score differences are not in terms of equal units of a psychological trait (e.g., spatial orientation or introversion), and the comparison of examinees on these traits using θ is ordinal rather than interval. If applications such as equating and adaptive testing are pursued, numerical estimates of θ and δ using the appropriate sufficient statistics are convenient, and the Rasch model or another parametric IRT model may be used, such as the two-parameter logistic model (2PLM; Fischer, 1974, p. 207) and the one-parameter logistic model (OPLM; Verhelst & Glas, 1995; this is a model with location parameters, δ , and imputed slope indices rather than slope parameters).

A theoretical study of a model for dichotomous, polytomous, and continuous response variables that resembles the DMM was presented by Scheiblechner (1995; see also Junker, in press, for critical comments). For dichotomous variables, this model is equivalent to the DMM. For polytomous variables, Scheiblechner's (1995, p. 285, Definition) model is different from the DMM for polytomous items proposed by Molenaar (1997; Hemker et al., 1995; see Sijtsma & Hemker, in press). For continuous response variables, the DMM has not been defined. A theoretical and practical comparison of the MHM, the DMM, and the Rasch model was provided by Meijer, Sijtsma, & Smid (1990; refer also to De Gruijter, 1994).

Further Elaborations on the Theory and Testing of the MHM

Theoretical Developments

Lewis (1982, 1983), Stout (1987, 1990), and Junker (1993) have studied IRT models that elaborate on the MHM and that allow person ordering. The NIRT formulation proposed by Lewis (1982) is characterized by the assumption of relative consistency, which is weaker than item-free person ordering using expected conditional item scores and is the minimal assumption for a labeled ordinal trait scale. This scale is a modification of θ and may be loosely interpreted as a scale of ordered equivalence classes.

Stout (1987, 1990) conjectured that essential unidimensionality might be a more realistic alternative for strict unidimensionality, which is assumed in almost all IRT models, including the parametric models. Essential unidimensionality means that the items predominantly measure the same trait, but that some of the items may also measure other traits. This would mean that conditioning on the dominant trait θ does not lead to 0 covariances between all item pairs (Junker, 1991; Stout, 1990). For infinitely large sets of items, however, essential unidimensionality is characterized by essential independence (EI) rather than local independence. To define EI, define a vector $\theta = (\theta, \theta_1, \theta_2, \dots)$ that contains all traits relevant to the item set, in the sense that they produce nonzero covariances between at least one pair of items. The trait θ is dominant. EI is then defined (Stout, 1990) as

$$\frac{\sum_{1 \leq g < h \leq k} |\text{Cov}(X_g, X_h | \theta)|}{\binom{k}{2}} \rightarrow 0 \text{ if } k \rightarrow \infty. \quad (14)$$

Another assumption that relaxes the MHM is that the proportion-correct “true score” function, or the mean IRF, is nondecreasing in θ (Stout, 1987). This is the assumption of weak monotonicity (WM). If $\theta_i \leq \theta_j$ in each pair of corresponding elements so that $\theta_i \leq \theta_j$, $\theta_{i1} \leq \theta_{j1}$, and so on, then under WM

$$\sum_{g=1}^k P_g(\theta_i) \leq \sum_{g=1}^k P_g(\theta_j). \quad (15)$$

This assumption replaces the monotonicity assumption with respect to individual items. As a result, the property of an item-free person ordering using expected conditional item scores is lost. Together, Equations 14 and 15 define essential unidimensionality.

A strengthening of WM is local asymptotic discrimination (LAD). LAD means that for every latent trait from θ and for every value θ_0 of a θ there exists $\epsilon_{\theta_0} > 0$, such that for every θ_{0*} close to θ_0

$$\frac{1}{k} \sum_{g=1}^k \frac{P_g(\theta_{0*}) - P_g(\theta_0)}{\theta_{0*} - \theta_0} \geq \epsilon_{\theta_0} > 0, \text{ for all } k. \quad (16)$$

LAD implies that the mean IRF from the test is sufficiently discriminating locally for each θ . Stout (1990) showed that θ can be estimated consistently on an ordinal scale under the assumptions of EI and LAD. Junker (1991) obtained this result for polytomous items in a nonparametric model formulation using these assumptions. Note that LAD entails both a relaxation (i.e., the sum of IRFs rather than individual IRFs) and a strengthening of the MHM (i.e., strict increasingness rather than nondecreasingness) compared with the monotonicity assumption (Equation 3).

Ellis & Van den Wollenberg (1993) discussed an implication of the “stochastic subject” formulation (Holland, 1990; Lord & Novick, 1968, pp. 29-31) of the probability $P_g(\theta)$. This formulation illustrates the idea that a particular person has varying responses across independent replications of the same test or item. Thus, $P_g(\theta)$ is a within-person probability. Holland (1990) also discussed the random sampling formulation of $P_g(\theta)$ in which $P_g(\theta)$ is the expected proportion of persons with fixed θ who obtain a score of 1 on item g . Shealy & Stout (1993) discussed a NIRT model based on a random sampling formulation for test bias and DIF.

Ellis & Van den Wollenberg (1993) showed that the assumption of local homogeneity is an integral part of any IRT model that follows the stochastic person interpretation of $P_g(\theta)$. Given the mathematical subtlety of the arguments, local homogeneity for a vector X with all item score variables, a vector θ , and a person v with a particular θ value [denoted $\theta(v)$] can be characterized as

$$P[X \leq \mathbf{x} | \theta(v)] = P[X \leq \mathbf{x} | \theta]. \quad (17)$$

This means that for person v the probability of $X \leq \mathbf{x}$ is equal to the probability of any person with trait value θ (Ellis & Van den Wollenberg, 1993, p. 420). Local homogeneity implies that an IRT model holds in all subpopulations of the population under consideration; the reverse is also true, that is, subpopulation invariance of an IRT model implies local homogeneity (Ellis & Van den

Wollenberg, 1993, Theorem 3). Thus, the MHM holds in all subpopulations if the assumption of local homogeneity is added. The MHM, with the assumption of local homogeneity added, is referred to as the *homogeneous monotone IRT model* (Ellis & Van den Wollenberg, 1993).

An additional assumption of NIRT was discussed by Ellis & Junker (in press; see also Junker & Ellis, 1997). Their theoretical framework is more general than the MHM because it allows discrete and continuous variables, as well as combinations of these variables. Their model formulation adds the assumption of *tail-measurability* to the three MHM assumptions. Tail-measurability means that θ can be estimated consistently from an infinite sequence, or tail, of observed item score variables with any finite subset missing. It was shown that these four assumptions are equivalent to two conditions that involve only observable item score variables, but not θ . The first condition is conditional association (Holland & Rosenbaum, 1986; see also Equation 19, discussed below in another context). The second condition is vanishing conditional dependence, which replaces local independence. The conditioning is on a sequence of observable item score variables, rather than θ . This property holds if the number of items conditioned on approaches infinity. It was further shown that the consistently estimated θ has ordinal measurement properties.

Observable Consequences From the MHM and Its Elaborations

Results for covariances. From the basic MHM and DMM, observable consequences can be derived that allow investigation of whether these models can represent the data structure. A fundamental result (Holland & Rosenbaum, 1986; Mokken, 1971, p. 120) is that, in the population, the covariance between items g and h is nonnegative:

$$\text{Cov}(X_g, X_h) \geq 0, \text{ for all } g, h. \quad (18)$$

Ellis & Van den Wollenberg (1993, p. 425) refer to Equation 18 as "pairwise nonnegative association." This result plays an important role in the scaling procedure proposed by Mokken (1971, pp. 190–193).

Equation 18 is a special case of a result that was proven by Rosenbaum (1984) and that leads to additional ways to test the MHM. Let Y and Z be vectors containing the scores on disjoint subsets of items. Let $m(Z)$ be a function of the item scores in Z and let $n_1(Y)$ and $n_2(Y)$ be coordinate-wise nondecreasing functions of the item scores in Y . Further, z is a fixed value of $m(Z)$. Then, under the assumptions of the MHM,

$$\text{Cov}[n_1(Y), n_2(Y) | m(Z) = z] \geq 0, \text{ for all } z. \quad (19)$$

This is called *conditional association* of Y in Z . If $m(Z)$ is the complete space of Z in the sense that the complete population is selected, and if $n_1(Y) = X_g$ and $n_2(Y) = X_h$, then Equation 19 specializes to Equation 18. Equation 19 is important because it points to several observable consequences of the MHM. For example, Equation 19 shows that the covariance between X_g and X_h is nonnegative in each group with a fixed total score on the other $k - 2$ items in the test. Rosenbaum (1984) provided several other examples of useful choices of subsets of items and also discussed statistical tests for testing hypotheses about conditional association. Note that Equation 19 generalizes a result by Holland (1981). Results for associated random variables (i.e., nonnegative covariance) were provided by Esary, Proschan, & Walkup (1967). See Lehmann (1966) for three successively stronger definitions of positive dependence of random variables.

Ellis (1994, chap. 5) proposed three statistical tests for the null hypothesis that all covariances between items are non-negative (Equation 18) versus the alternative hypothesis that there are systematic deviations from this expectation (e.g., Ellis, 1994, chap. 5, p. 8). He also discussed adaptations of these tests for conditional association (Equation 19) by groupwise testing and combining test results afterward, and by a pooled covariance matrix. Simulation results showed that

one of the tests always performed at least as well as the other two. Ellis (1994, chap. 5, pp. 23–24) noted that further research is needed to resolve problems with computational limitations, with the requirement of large sample sizes to produce sufficiently accurate estimates, and with the small number of items for which the tests are presently useful.

Necessary and sufficient conditions for the homogeneous monotone IRT model (Ellis & Van den Wollenberg, 1993) for dichotomous items are (1) that the item pool is experimentally independent (i.e., local independence for persons, rather than values of θ), and (2) that Equation 18 holds for all subpopulations from the population of interest (Ellis & Van den Wollenberg, 1993, Theorem 7). Provided that experimental independence holds, this condition can be empirically tested by examining the sign of the covariances between all item pairs in all possible subpopulations. Due to practical limitations, this must be restricted to meaningful subpopulations. Even under these circumstances, however, the choice of appropriate subgroups likely to reveal negative covariances that contradict the model is often not straightforward.

Equation 19 (Rosenbaum, 1984) provides another way to test the MHM, given that appropriate disjoint subsets of items can be found for particular applications. This also applies to the Ellis & Van den Wollenberg (1993) results concerning convenient choices of subpopulations, under the restriction that experimental independence can be assumed to hold in the particular application. In addition, a stochastic person formulation must be adopted.

Results for IRFs. Junker (1993) showed that an observable consequence of the MHM is that the regression of the score on item g , X_g , on the total score from the other $k - 1$ items (the *rest score*) is nondecreasing. Junker & Sijtsma (1997) explain why the regression of the item score on the total score on all k items, X_+ , need not be monotone under the MHM and, therefore, why it is not an adequate function for studying the IRF. Let the total score on $k - 1$ items be denoted $R_{(g)}$, so that $R_{(g)} \equiv X_+ - X_g$. Further, let r_1 and r_2 be values of $R_{(g)}$ such that $r_1 < r_2$. Junker (1993, Prop. 4.1) showed that, given unidimensionality, local independence, and monotonely nondecreasing IRFs,

$$P[X_g = 1 | R_{(g)} = r_1] \leq P[X_g = 1 | R_{(g)} = r_2]. \quad (20)$$

This is what Junker (1993) called *manifest monotonicity*. Note that an alternative sum score (R') could be a score on a single item h ($h \neq g$) or on any subset of the set of $k - 1$ items without g that are in agreement with the MHM. $R_{(g)}$ is a very convenient choice, however, because it is based on all items except item g and thus often leads to a more reliable subdivision of the sample than choices of R' that use fewer items. The item-rest score regression can be seen as a discrete estimate of the IRF.

A procedure for the nonparametric estimation of the IRFs was discussed by Ramsay (1991) who, after rank ordering persons on X_+ or a conveniently weighted sum of the item scores, used kernel smoothing methods to estimate the option response function, $P_{gm}(\theta)$, where m is the m th option of a multiple-choice item with a keyed option and distractors. If m is the correct option, $P_g(\theta)$ is estimated. Confidence bands for the IRFs can also be estimated. Ramsay (1991, p. 615) expressed some preference for person ranking based on items for which evidence exists of monotonely nondecreasing IRFs, selected on the basis of preliminary analyses. Because Junker (1993) and Junker & Sijtsma (1997) found disadvantages to using the total score X_+ and advantages for using $R_{(g)}$, a modification of Ramsay's procedure that uses $R_{(g)}$ rather than X_+ might be worth considering.

A related procedure for examining the dimensionality of the data, also based on kernel smoothing, was proposed by Ramsay (1995). This procedure was combined with principal component analysis to study the dimensionality among items. These smoothing methods involve an assumption about the person distribution. The program TESTGRAF (Ramsay, 1993) is available for estimating option

response functions. Note that the methods proposed by Ramsay (1991) predominantly estimate IRFs from the data without assuming an explicit NIRT model (e.g., the MHM) to guide the estimation. Conversely, the estimates can be used to verify the assumption of nondecreasingness in a NIRT model.

A Scaling Procedure for the MHM

A Scalability Measure for the MHM

The deterministic Guttman (1944, 1950) model is often treated as a predecessor of IRT (Hambleton & Swaminathan, 1985, pp. 29, 35; Hulin et al., 1983, pp. 15–17; Nunnally, 1978, pp. 72–75; Rost, 1996, pp. 104–109). The Guttman model is defined by

$$\theta < \delta_g \leftrightarrow P_g(\theta) = 0, \quad (21)$$

and

$$\theta \geq \delta_g \leftrightarrow P_g(\theta) = 1. \quad (22)$$

The IRFs of the Guttman model are unit step functions (see Figure 1d). The model, therefore, excludes a correct answer on a relatively difficult item and an incorrect answer on an easier item by the same person: $X_g = 1$ and $X_h = 0$, for all $g < h$. Such combinations are called “errors.” Thus, if Person A has a higher θ value than Person B, given the Guttman model Person A would answer all the items correctly that Person B would answer correctly, but Person A would also answer other, more difficult items correctly. Probabilistic IRT models, such as the MHM, DMM, and the logistic models, allow Guttman errors [Roskam & Jansen (1984) demonstrated that the Rasch model (Equation 12) is the probabilistic analogue of the Guttman model]. The exact formulation of a model restrains the degree to which these error patterns are allowed. In each model, the null model of horizontal IRFs, which implies global independence, constitutes a lower bound in the sense that it represents a case of trivial measurement. Note that for horizontal IRFs, the expected total score is equal for all θ values. In particular, for $\theta_i = \theta_j$,

$$E(X_+|\theta_i) = \sum_{g=1}^k E(X_g|\theta_i) = \sum_{g=1}^k E(X_g|\theta_j) = E(X_+|\theta_j). \quad (23)$$

This means that observed score differences between persons are entirely due to measurement error. A reliable person ordering based on total score is, thus, impossible.

To evaluate the practical usefulness of the MHM, Mokken (1971, pp. 148–153) proposed a scalability coefficient known as the H coefficient (see also Cliff, 1977, who denoted H as c_{13} ; Loevinger, 1948; Molenaar, 1991, who provided a generalization of H to polytomous items). H can be used to express the position of the data between the extremes of the mean interitem covariance equal to 0 and the Guttman model.

Let F_{gh} denote the observed error frequency for items g and h . Let E_{gh} denote the expected error frequency, assuming that the responses to the items are globally independent across persons and that the marginals of the 2×2 crosstable of bivariate frequencies of items g and h are fixed. For items g and h , the scalability coefficient H_{gh} is defined as

$$H_{gh} = 1 - \frac{F_{gh}}{E_{gh}}. \quad (24)$$

The definition of an error assumes only that $\pi_g \leq \pi_h$ [in case of an equality, the (0,1) and (1,0) cells have the same frequency]; these are the item means in the entire population. Thus, there is no assumption that the item ordering is the same for each θ (i.e., IIO), which would be characteristic of the DMM (see Equation 10).

Mokken (1971, p. 151) showed that the H_{gh} coefficient equals the ratio of the covariance between items g and h and their maximum covariance given the marginal distributions of g and h :

$$H_{gh} = \frac{\text{Cov}(X_g, X_h)}{\text{Cov}(X_g, X_h)^{\max}}. \quad (25)$$

In addition, let F_g and E_g denote the sum of the frequencies F_{gh} and E_{gh} ($h = 1, \dots, k; h \neq g$), respectively, and let F and E be the sums of the F_{gh} and the E_{gh} across all item pairs, respectively. Then the scalability coefficient of a single item g , H_g , with respect to the other items in the test, and the scalability coefficient for a set of k items, H , are defined respectively as

$$H_g = 1 - \frac{F_g}{E_g}, \quad (26)$$

and

$$H = 1 - \frac{F}{E}. \quad (27)$$

Interpretation of H

Mokken (1971, p. 150) showed that for k items with nondecreasing IRFs, $0 \leq H \leq 1$, where $H = 0$ if and only if at least $k - 1$ IRFs are constant functions of θ . Although formally in agreement with the MHM, this condition constitutes a trivial case of unsuccessful measurement because persons are ordered by at most one item (refer to Equation 23 for k constant IRFs). Therefore, the assumption of nondecreasing IRFs is rejected if H (H_{gh} or H_g) has a negative value. Mokken et al. (1986) argued that the larger H is, the more confidence can be held in the ordering of persons on the θ scale using the total score. H , therefore, reflects the degree to which persons can be accurately ordered using their total score on k items.

Roskam et al. (1986) noted that a positive H or H_g is not sufficient for IRFs to be nondecreasing or for the MHM to hold. Indeed, a low positive H_g may be caused by (1) an almost flat IRF with a positive slope, (2) an irregular IRF that jumps up and down across θ , (3) an irregular IRF that appears single-peaked, or (4) an irregular IRF that appears bimodal. Such cases are less likely to occur when H has a higher value. High positive values tend to be characteristic of IRFs that are increasing, although there may be small local decreases across θ . Roskam et al. [1986; also Jansen (1982a) and Mokken et al. (1986)] noted that given the MHM, H is an increasing function of the variance of θ (see also Mokken, 1971, p. 151), the spread of the item difficulties, or the slopes of the IRFs, keeping the other two of these conditions constant. However, a particular value of H_g does not provide specific information about any of these factors. Although experience has shown that a high H_g value is usually indicative of items with steeply increasing IRFs and small, practically unimportant deviations from nondecreasingness, the manifest monotonicity should be investigated to obtain more specific information about the shape of the IRFs (Junker, 1993).

For a discussion on the advantages and disadvantages of using the H coefficient in the context of the MHM, see Jansen (1982a, 1982b); Jansen, Roskam, & Van den Wollenberg (1982, 1984); Mokken et al. (1986); Molenaar (1982a, 1982b), Sijtsma (1984, 1986); and Roskam et al. (1986). This discussion has stimulated much research on NIRT (e.g., Ellis, 1994; Hemker, 1996; Meijer, 1994; Sijtsma, 1988).

Definition of a Scale

Mokken (1971, p. 184) defined a scale as a set of items for which (1) $\text{Cov}(X_g, X_h) > 0$, for all item pairs; and (2) $H_g \geq c > 0$, where c is a constant, for all items. The first condition implies $H_{gh} > 0$; the second implies $H \geq c$ (Mokken, 1971, p. 153). Note that the definition of a scale contains the model requirement of positive covariances in the first part and the practical requirement of having moderately, or more strongly, discriminating items (i.e., IRFs with at least moderately steep slopes) in the second part. Mokken (1971, p. 185) considered a set of items as a “weak” scale if $.3 \leq H < .4$; a “medium” scale if $.4 \leq H < .5$; and a “strong” scale if $H \geq .5$. A set of items with $H < .3$ was considered to be unscalable for practical purposes.

The requirement of $H_g \geq .3$ is not part of the MHM because the exact lower bound of $H_g = .3$ cannot be derived from the three assumptions of the MHM. Rather, $H_g \geq .3$ constitutes a practical requirement that is conveniently combined with the model requirement that $\text{Cov}(X_g, X_h) > 0$ in constructing a scale. For test construction, every scaling procedure needs such practical requirements in addition to theoretical requirements.

For example, all unidimensional IRT models allow horizontal or nearly horizontal IRFs. However, when a test is constructed that must be highly precise in a particular region on the θ scale, items should be selected that have the highest discrimination in that region. That is, items are selected with high information values (Lord, 1980, chap. 5) so that the standard error of $\hat{\theta}$ is acceptably small in the region of interest. These items usually are a subset of the items that are in agreement with the model requirements and thus, in practice, items that agree with the model may be rejected from the test.

Another example is classical test theory, a test theory model that defines methods for estimating the reliability of test scores. To increase reliability, however, the researcher often selects items that have corrected item-total correlations of at least .3 (Nunnally, 1978, pp. 263, 282) or some other accepted lower bound. Note that such requirements, as with high information or high H_g values, are imposed as a result of practical considerations. They cannot be derived from the models of interest. These practical requirements are not in conflict with the model. They help to select the best items from the set that comply with the model. Whether items comply with the model is an issue that must be investigated independently using other methods for testing the model.

Methods of Item Selection

Item Selection According to Mokken's Definition of a Scale

Based on the definition of a scale, an automated bottom-up item selection procedure was developed (Mokken, 1971, pp. 190–193) that is part of the computer program MSP (Molenaar et al., 1994). Suppose an item pool consists of K items. The starting set for scale construction by default consists of the two items with the largest significant H_{gh} , but may also consist of a kernel of meaningful items specified by the researcher. In each consecutive step, one item is added to the already selected set (k_{sel} items) in such a way that the common H coefficient of the $k_{sel} + 1$ items thus selected is the largest, given all possible choices, from the remaining $K - k_{sel}$ items. Item selection continues as long as the definition of a scale is satisfied. After one scale has been constructed, the procedure continues selecting items from the unselected set to construct a second scale, and so forth. Throughout the selection process, the significance level used to test whether H and H_g are 0 against the alternative of positive coefficients is progressively adapted to protect against chance capitalization. MSP also provides the additional option of making items selected in previous scales available for selection in scales to be formed later.

It is often found in practice that H tends to decrease during the item selection process. This decrease is due to the variation in the quality of the items in the initial item pool. In each selection

step, the best item is selected from the remaining items and added to the scale. This redistribution of items lowers the quality of the unselected item set. The quality of the selected item set is also lowered by the addition of this same item, because it is worse than any of the previously selected items. Hypothetically, if test length were not an issue for other quality measures such as reliability, the best scale in terms of H would thus consist of the first two items selected. In practice, however, the quality of the scale in terms of H is attenuated with the selection of more items. If the initial item pool is small and the items have strongly varying quality, item selection might result in short, and often unreliable and invalid, scales.

Sijtsma & Prins (1986) studied the order in which items are selected from an item pool and the trend of H during this process. Data were simulated using the OPLM and the 2PLM, and a normal distribution of θ . Item difficulties were equidistant between -2 and 2 . Six levels of K were completely crossed with five levels of the standard deviation of θ . Each dataset was analyzed with the item selection procedure using a lower bound $c = 0$. This is the lowest c value allowed by the MHM.

The first two items selected were the easiest and the most difficult; the third item was of medium difficulty; the fourth and fifth items were halfway between the easiest and the medium difficulty items, and between the medium difficulty and most difficult items, respectively. With each next item, a gap on the difficulty scale was filled in such a way that the end result was always a set of approximately equidistant items across the entire range of item difficulties.

With Rasch items (i.e., items with the same discrimination values), H decreased rapidly with the first five items selected but stabilized with the selection of more items. The final H depended primarily on the standard deviation of θ and secondarily on the number of items. When an item set also contained relatively weakly discriminating items (i.e., a relatively flat slope), these items were selected last, causing H to decrease sharply. Relatively highly discriminating items (i.e., a relatively steep slope) were selected as part of the first few items.

Selecting among items with approximately the same discriminations and considerable variation in difficulty produces a long scale of equidistant items, thus allowing precise measurement for most of the population, with the possible exception of the extremely low and high θ s. In general, precise measurement of extreme θ s is problematic with most tests and, in principle, is best solved using an adaptive testing strategy (Wainer, 1990; Weiss, 1985).

The retrieval of unidimensional item sets, consisting of five-point Likert items, from multidimensional item sets was investigated by Hemker et al. (1995). Although this study did not involve dichotomous items, some of the results are discussed here because it seems highly plausible that results for dichotomous items would not be much different than for polytomous items. Data were simulated using the graded response model (Samejima, 1969) and a multivariate normal distribution for the θ s. For two and three traits, respectively, completely crossed designs were analyzed with six levels of correlation between traits, three levels of mean discrimination per item (note that, with five response categories an item has four category response functions), and two levels of variation of discrimination parameters per item. Most importantly, there was not a unique lower bound, c , or a unique range of lower bounds that showed whether the item set was multidimensional.

A relatively simple strategy for investigating the dimensionality of an item set is to conduct the item selection strategy several times, starting with $c = 0$ or a small positive value such as $c = .05$. Very small c values reveal whether there are scales with a correlation close to 0, because items from different scales often have H_{gh} values close to 0. If the scales have higher correlations, these H_{gh} values are positive; with a small c , the items are all selected into the same scale. Next, the traditional values of $c = .3$, $.4$, and $.5$ can be used. A more fine-grained analysis would increase c by $.05$ in each consecutive computer run. Values higher than $.6$ often do not yield information

about the dimensionality because the H_{gh} values of most item pairs are smaller than this c value and, as a result, most items are not selected. Hemker et al. (1995) found that multidimensional and unidimensional item sets can be identified by different patterns of scale results that are found across several computer runs using varying values of c .

Roussos and Stout's Item Selection for Constructing Unidimensional Scales

Stout (1987) proposed the DIMTEST procedure to investigate the dimensionality of a set of items in the EI, WM framework (Stout, 1990; see Equations 14 and 15). For the DIMTEST procedure to have good statistical power for detecting multidimensionality, the user must split the item set into two sets believed to be dimensionally distinct (Nandakumar & Stout, 1993). This subdivision can be based on expert judgment of the underlying traits required to solve the items, or on a formalized procedure such as cluster analysis. Factor analysis has also been used for this purpose (Nandakumar, 1993; Stout, 1987).

Proposals for the use of convenient proximity measures in a hierarchical cluster analysis (HCA), which should construct unidimensional clusters, were made by Roussos & Stout (1995). Suppose items g and h are studied. Let the rest score $R_{(gh)}$ be the total score on the items except items g and h : $R_{(gh)} \equiv X_+ - X_g - X_h$. Roussos & Stout (1995) found that using one measure based on the correlation and another based on the covariance between dichotomous items (i.e., conditioned on rest score groups and then averaged across these groups) was most effective in retrieving the dimensionality structure in simulated multidimensional IRT data under certain variations of HCA. Roussos, Stout, & Marden (1994) demonstrated, using empirical data, how DIMTEST can be used after a first clustering of the items has been obtained using HCA.

The actual investigation of the dimensionality of the item set using DIMTEST is done by calculating two total score variances for one hypothetical unidimensional item set in each group with the same total score on another reference item set. This second item set is as distinct as possible from the first set in terms of dimensionality. As indicated above, this has been established using HCA or factor analysis. Without going into further technical details, note that under unidimensionality in the complete test both variance estimates should be equal. However, if the two item sets indeed measure different traits, the two variance estimates will be significantly different. A standard normal test statistic T , which combines the differences between the two variance estimates across all total score groups on the reference test, is used to test the hypothesis of unidimensionality.

Hattie, Krakowski, Rogers, & Swaminathan (1996) evaluated the DIMTEST procedure using factor analysis based on tetrachoric correlations, rather than HCA, to construct a subdivision of the items. T was used to evaluate essential unidimensionality. Based on simulated data, they concluded that DIMTEST is not very sensitive to correlations between latent traits and variation in item location and guessing parameters, or to the method used to produce the subdivision of the items, such as factor analysis or nonlinear factor analysis. These are good outcomes for a procedure that investigates dimensionality. Nandakumar & Stout (1993) noted that DIMTEST was sensitive to item discrimination. Note that in a two-dimensional setting it would be expected that highly discriminating items would provide more statistical power than items with low discrimination.

Hattie et al. (1996) concluded that DIMTEST should not be used when the data can be explained by partially compensatory multidimensional IRT models, in which a decrease in one trait can only be compensated for by a large increase in another trait to produce the same response probability [see Hattie et al. (1996), for further discussion]. They also noted that DIMTEST assesses only essential unidimensionality, but cannot identify the number of dimensions when the data are multidimensional. In general, however, DIMTEST detected deviations from unidimensionality and, at the

same time, was robust against influences from many other sources. Goa & Stout (1996) studied a modification of DIMTEST that simplifies the item subdivision. This led to better applicability of DIMTEST to smaller item sets (e.g., $k < 25$).

Both the Mokken procedure and the Roussos and Stout HCA procedure discussed above rely heavily on cluster analysis to produce scales. The Mokken procedure uses the covariance between items, and the item scalability coefficient H_g with a conveniently selected lower bound, to assess inclusion of an item in a cluster. Thus, inclusion depends to a large extent on the shape of the IRF: IRFs should preferably be nondecreasing and have at least moderate slopes. Hemker et al. (1995) suggested how the procedure may be used to detect unidimensional scales. The Roussos and Stout HCA procedure uses measures that are based on the covariance or correlation between items in rest score groups, capitalizing on local independence (or the lack of it). Thus, inclusion in a cluster largely depends on whether an item measures the same trait as the items already selected in a cluster.

Another difference concerns the type of cluster analysis used. HCA selects clusters of items that are as close as possible according to a particular distance measure. Initially, each item is a cluster. Then, in the first step, the two items that are closest are clustered; in the second step, two other items may be clustered or a third item is joined with the first pair; in each consecutive step, the two clusters that are closest at that stage are joined; and in the final step, all items are clustered. With empirical data, the solution considered most appropriate in terms of dimensionality is a matter of judgement. For Mokken's procedure, in the first step the two items that have the highest H_{gh} are clustered, and in each consecutive step the item from the unselected subset, which is closest to the already selected cluster in terms of H_g , is joined with this cluster. A second cluster can be formed from the remaining items after the first cluster has been completed, and so forth.

There are many similarities and differences between the item selection procedures advocated by Roussos and Stout and by Mokken. In particular, because these approaches investigate different assumptions and practical requirements, it may be useful to use both models to test data in order to conduct a complete test of an underlying NIRT model.

Such a comparative study might also include another method for studying unidimensionality, proposed by Holland & Rosenbaum (1986) and compared with DIMTEST by Nandakumar (1994). This method is based on Equation 19 and uses the conditional covariance between X_g and X_h , given the total score on the other $k - 2$ items in the test. Nandakumar (1994) found that, in contrast to DIMTEST, the statistical test based on this conditional covariance had little power to detect multidimensionality. This failure was attributed to the contamination of the rest score, $R_{(gh)}$, and it was argued that a conditioning variable $m(\mathbf{Z})$ should be used that is more strongly related to one particular latent trait.

Finding Unidimensional Scales According to Kim, Zhang, & Stout (1996)

More recently, Kim et al. (1996) proposed an index for the lack of unidimensionality, denoted DETECT, and a procedure to identify that partition $\mathcal{P}$ of the available k items into disjoint and exhaustive subsets that best represents the underlying dimensionality structure (see Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). The best partition has the highest DETECT value and is denoted DETECT_{max}. For the unidimensional case, DETECT is close to 0. DETECT_{max} is a measure of the amount of departure from unidimensionality or, equivalently, the amount of multidimensionality. The partition $\mathcal{P}$ that produces DETECT_{max} provides a count of the number of dimensions.

Note that the amount of departure from unidimensionality, or the amount of multidimensionality, are rather vague notions: For example, a high DETECT_{max} could mean that there are two weakly correlated dimensions each measured by large sets of items, whereas a low DETECT_{max} could mean

that there is one dominant dimension measured by many items and five other dimensions that correlate highly and are each measured by only a few items. Thus, DETECT_{max} does not indicate the number of dimensions. Kim et al. (1996) provided guidelines for the interpretation of the amount of multidimensionality in four ordered numerical DETECT_{max} categories labeled unidimensional, weak, moderate, and strong.

DETECT is based on the covariances of pairs of items, g and h , conditional on the rest score $R_{(gh)}$ that serves as an estimate of the composite $(\theta_1, \theta_2, \dots)$ in the case of multidimensionality. For each fixed item pair (g, h) , conditional covariances are combined across all values of $R_{(gh)}$ without items g and h , yielding an estimate of Cov_{gh} . These covariance estimates are averaged across item pairs for a particular item partition $\mathcal{P}$. Given multidimensionality, Kim et al. (1996) argued that Cov_{gh} will be positive if items measure the same trait and negative if they measure different traits. If two items, g and h , are in the same cluster defined by $\mathcal{P}$, Cov_{gh} receives a weight of 1; if they are in different clusters, Cov_{gh} receives a weight of -1 . The more partition $\mathcal{P}$ represents the true dimensionality, the higher the value of DETECT, and DETECT_{max} is reached if $\mathcal{P}$ exactly represents true dimensionality. If weighting of covariances is denoted by $\omega_{gh} = -1, 1$, DETECT is defined as

$$\text{DETECT}(\mathcal{P}) = \frac{2}{k(k-1)} \sum_{g,h} \omega_{gh} (\widehat{\text{Cov}}_{gh} - \overline{\text{Cov}}). \quad (28)$$

The mean covariance serves as a correction of positive bias that is caused primarily by conditioning on $R_{(gh)}$. Note that this conditioning represents a special case of conditional association in Equation 19. The theoretical basis of DETECT was discussed by Zhang & Stout (1995).

Kim et al. (1996) discussed a heuristic that reduces the number of possible partitions, $\mathcal{P}$, and leads to an approximation to DETECT_{max}. Simulations showed that, in general, (1) DETECT obtains its maximum at a number of item clusters equal to the number of latent traits, (2) DETECT usually correctly assigns items to dimensions, and (3) DETECT_{max} decreases as the correlation between latent traits increases. The above comparison of item selection procedures might also include DETECT.

Comments on Item Selection Procedures

Standardized item selection procedures can be of great help to researchers because the investigation of the fit of any IRT model often requires a large number of tests and checks. The pattern of results is often complex and decisions that must be made when interpreting the results frequently lead to further discussion (Meijer et al., 1990). It may be asked, however, whether it is a good idea to replace the researcher by a standardized procedure, because the computer is blind to the content of the items and has no knowledge about the history of a test and the characteristics of the population for which it is intended.

Researchers should never blindly accept the outcome of any computerized item selection procedure. In fact, the results by Hemker et al. (1995) for the Mokken procedure demonstrated that the use of various lower bounds for H in consecutive computer runs may reveal more about the dimensionality of the data than using only the MSP default $c = .3$. Furthermore, using different meaningful starting sets of items is recommended. This strategy assumes that often there is no unique outcome of item selection or model testing, and that other, informal criteria may be used to reach a final decision.

Observable Consequences and Testing of the DMM

There is no automated item selection procedure available that selects items according to their fit to the DMM. Mokken (1971, pp. 132–133) proposed verifying nonintersection of IRFs (IO; Equation

10) using visual inspection of the ordering properties of the rows and columns of the $P(++)$ and the $P(--)$ matrices. The $P(++)$ matrix contains the joint proportions, $\pi_{gh}(11)$, of the persons with both items g and h correct. The $P(--)$ matrix contains the joint proportions, $\pi_{gh}(00)$, of the persons who provided negative answers to both items. Assume that the k items are ordered according to increasing π_g along the marginals of these matrices. Given this arrangement, the DMM implies that the rows and columns of the $P(++)$ matrix are nondecreasing and that the rows and columns of the $P(--)$ matrix are nonincreasing. Thus, given the DMM, for all $f \neq g, h$,

$$\pi_{gf}(11) \leq \pi_{hf}(11) , \quad (29)$$

and

$$\pi_{gf}(00) \geq \pi_{hf}(00) . \quad (30)$$

Joint proportions with respect to (01) and (10) patterns can also be defined and corresponding inequalities can be derived:

$$\pi_{gf}(01) \geq \pi_{hf}(01) , \quad (31)$$

and

$$\pi_{gf}(10) \leq \pi_{hf}(10) . \quad (32)$$

It can be shown (Mokken, 1971, p. 133; Sijtsma & Junker, 1996) that Equation 29 holds if and only if Equation 31 holds, and Equation 30 holds if and only if Equation 32 holds. Thus, it is sufficient to examine Equations 29 and 30 only by inspecting $P(++)$ and $P(--)$.

The information from $P(++)$ and $P(--)$ is related to the investigation of the intersection of IRFs in the following way: Consider Equations 29 and 32 (Equation 32 is equivalent to Equation 30; Equation 30 is used in practice). After the division of both sides of Equation 29 by π_f , and both sides of Equation 32 by $1 - \pi_f$, the following expressions are obtained, respectively:

$$P(X_g = 1|X_f = 1) \leq P(X_h = 1|X_f = 1) , \quad (33)$$

and

$$P(X_g = 1|X_f = 0) \leq P(X_h = 1|X_f = 0) . \quad (34)$$

These expressions show that in the groups with item f correct and item f incorrect, respectively, items g and h have the same difficulty ordering. From the MLR result for the MHM (Equation 4; Grayson, 1988), it can be deduced that θ is stochastically ordered by X_f in any MHM, thus including the DMM, which is a special case of the MHM. As a result, the mean θ in the group with $X_f = 0$ is smaller than in the group with $X_f = 1$. This means that Equation 33 provides information about the ordering of items g and h for relatively high θ values, whereas Equation 34 provides that information for relatively low θ values. Intersections of the IRFs for items g and h that occur in a region on the θ scale that corresponds approximately to the interval between $E(\theta|X_f = 0)$ and $E(\theta|X_f = 1)$ can be detected because there will be a different difficulty ordering of these items in the two subgroups.

Note that for items g and h , conditioning is done consecutively on each of the other $k - 2$ items (X_f ; $f \neq g, h$). If the difficulties of these $k - 2$ items vary considerably, each conditioning divides the total group into pairs of subgroups that vary by their composition over these items f . Thus, using each of these $k - 2$ items amounts to the investigation of the entire range of the θ scale in the population of interest for intersections of the IRFs for items g and h .

Finally, note that a division of the total group with respect to θ on the basis of only one item score, X_f , will be unreliable. Because only two subgroups are formed, however, the method will have relatively high power to detect reversals of the item ordering across subgroups.

Several other methods have been proposed for investigating the intersection of IRFs without assuming a particular parametric form for these functions (Croon, 1991; Rosenbaum, 1987a, 1987b). Sijtsma & Junker (1996) pointed out that the method using $P(++)$ and $P(--)$ is a special case of two general principles that both suggest alternative methods for investigating the intersection of IRFs. Without reiterating the technical details [refer to Sijtsma & Junker (1996)], these two principles can be phrased in the following way. First, Equations 29 and 30, which describe the orderings in $P(++)$ and $P(--)$, can be considered as special cases of similar sets of inequalities that pertain to q -variate joint probabilities of the form

$$\pi_{g1g2, \dots, gq} = P(X_{g1} = 1, X_{g2} = 1, \dots, X_{gq} = 1), \quad (35)$$

and similar q -variate joint probabilities for item scores equal to 0. Assume a numbering and an ordering of the k items from the test such that a higher index corresponds to an easier item for all θ . For a given q , it can be shown, by applying the notion of functions increasing in transposition (Sijtsma & Junker, 1996; or decreasing in transposition, Rosenbaum, 1987b), that particular strings of inequalities of q -variate joint probabilities should hold given that the IRFs of all k items do not intersect or the k items have IIO. Sijtsma & Junker (1996) pointed out that the inspection of strings of q -variate joint probabilities probably does not lead to a method for the investigation of the intersection of IRFs that is of much practical use for, say, $k \geq 10$, because (1) a very large number of orderings must be inspected, and (2) the sample frequencies necessary to estimate the q -variate proportions will often be very small or 0.

For an outline of the second principle, consider the conditional probabilities in Equations 33 and 34. These inequalities can be described by a more general theorem proposed by Rosenbaum (1987a) that can be phrased as follows. Let $\mathbf{Z}$ be the vector with $k - 2$ item score variables, excluding X_g and X_h , and let $m(\mathbf{Z})$ be any function of $\mathbf{Z}$. It can be shown that if the IRFs of items g and h do not intersect, then for any function $m(\cdot)$

$$P[X_g = 1 | m(\mathbf{Z}) = z] \leq P[X_h = 1 | m(\mathbf{Z}) = z]. \quad (36)$$

Equation 36 says that the ordering of items g and h is the same in each subgroup defined by a particular value z of the function $m(\cdot)$ of the $k - 2$ remaining item scores. An obvious choice for $m(\mathbf{Z})$ is the total score $R_{(gh)}$ on the $k - 2$ items in $\mathbf{Z}$ [refer to Molenaar (1997) for a further discussion of this possibility, and to Sijtsma & Junker (1996) for an alternative choice of $m(\cdot)$]. For the present purposes, it suffices to observe that if $m(\mathbf{Z}) = X_f$, then Equations 33 and 34, which are strongly related to Equations 29 and 30 in $P(++)$ and $P(--)$, follow after the substitution of $X_f = 1$ and $X_f = 0$.

To summarize, note that Mokken's method using $P(++)$ and $P(--)$ is part of a larger family of methods for investigating the intersection of IRFs. More research concerning the power of different methods is needed, and alternative methods could be included in this research.

A totally different approach (Sijtsma & Meijer, 1992) to investigating nonintersection of k IRFs uses scalability coefficients such as H . Again, the technical details are beyond the scope of this paper. The most convenient way to introduce scalability coefficients is by applying the H coefficient and the H_g coefficient to the transposed $N \times k$ data matrix, $\mathbf{X}_{Nk}^T$. The roles of persons and items are interchanged, but the formal structure of the equations is the same. Let X_v denote a person score, rather than an item score. For person v , this variable has k observations (1s and 0s for correct and incorrect answers on each of the k items from the test). The covariance between the k scores

of person v and the k scores of person w can be expressed as an expectation using the joint and univariate proportion-correct true scores.

Adopting this idea, the H coefficient applied to the transposed data matrix X_{Nk}^T can be defined as

$$H^T = \frac{\sum_{v \neq w} \text{Cov}(X_v, X_w)}{\sum_{v \neq w} \text{Cov}(X_v, X_w)^{\max}}, \quad (37)$$

with maximum covariances determined by the expected proportion-correct scores. See Cliff (1977) for a related coefficient denoted c_{p3} .

A coefficient is defined for one person in relation to the other $N - 1$ persons as

$$H_v^T = \frac{\sum_{w \neq v} \text{Cov}(X_v, X_w)}{\sum_{w \neq v} \text{Cov}(X_v, X_w)^{\max}}. \quad (38)$$

If k items have nonintersecting IRFs, $0 \leq H^T \leq 1$, with equality to 0 if and only if the k IRFs coincide with the exception of at most one θ value. It can be argued that given nonintersection of the IRFs, small H^T values are characteristic of IRFs that are close together and perhaps partly coincide. On the basis of research using simulated data, Sijtsma & Meijer (1992) suggested guidelines for the use of H^T and the H_v^T in practical test construction. Practical experience is needed to draw more definite conclusions about the usefulness of this approach. Several comments made by Roskam et al. (1986) about the H coefficient may also be relevant for H^T .

Comments on the Investigation of the DMM

Using H^T and H_v^T does not suggest an automated item selection procedure because H^T cannot be dissected into item contributions. A procedure that would be completely symmetrical with respect to using H and H_g would select persons on the basis of their H_v^T values, rather than items. Selection of persons is conceptually different from item selection because it would accept the item set as given and select the optimal person subgroup with respect to the item set. H^T and H_v^T should alternatively be used as global methods for the evaluation of sets, rather than individual items. The practical use of the $P(++)$ and $P(--)$ information and the information from other methods, such as Rosenbaum's, was discussed by Sijtsma & Junker (1996). Meijer (1994) studied H_v^T as a person-fit measure.

Some measurement theorists and practitioners have the impression that the DMM is the only Mokken model (e.g., Gillespie, TenVergert, & Kingma, 1988; Henning, 1976; Moorer & Suurmeijer, 1994; Niemöller & Van Schuur, 1983). Although the literature on Mokken's IRT approach explicitly mentions two models, the misinterpretation seems to be rather persistent. The weaker MHM is a useful model if the purpose of a test or questionnaire is to order persons or groups. Note that the ordering of persons for selection purposes is important in psychology and education, whereas the ordering of groups may be of interest in sociology and political science. In these cases, there is no need to use a stronger model such as the DMM. However, if IIO is considered necessary in addition to person ordering, the stronger DMM should be satisfied. The adoption of a more restrictive model will often lead to a loss of items, resulting in a decrease in the reliability of the total score (Meijer et al., 1990).

Reliability of Item Scores and Total Scores

Not every item that complies with the MHM, the DMM, or other IRT models, such as the logistic models, is useful in a test or questionnaire. This will be considered first for parametric IRT models and then for NIRT models. In terms of statistical information, denoted $I_g(\theta)$ for item g , the contribution of X_g to the maximum likelihood estimation of θ is

$$I_g(\theta) = \frac{\left[\frac{\partial P_g(\theta)}{\partial \theta} \right]^2}{P_g(\theta)[1 - P_g(\theta)]} . \quad (39)$$

The item contributes more, in terms of precision to the estimation of θ , the higher the first derivative (i.e., the slope) of the IRF at a particular θ value. A similar effect is obtained by a smaller error variance in the denominator of $I_g(\theta)$.

Maximum likelihood estimates of θ and the item parameters, and estimates of the item information function $I_g(\theta)$, can be obtained under a specific parametric IRT model. For example, substitution of the Rasch IRF from Equation 12 for $I_g(\theta)$ in Equation 39 yields

$$I_g(\theta) = P_g(\theta)[1 - P_g(\theta)] = \frac{\exp(\theta - \delta_g)}{[1 + \exp(\theta - \delta_g)]^2} . \quad (40)$$

This function can be estimated using the maximum likelihood estimates of θ and δ_g . One advantage of a parametric IRF is that measurement precision can be determined as a function of θ . This is done by taking the reciprocal of $I_g(\theta)$, yielding the asymptotic covariance matrix with $\text{Var}(\hat{\theta}|\theta)$. Thus, it is possible to select items so as to maximize precision for particular θ values (Van der Linden & Boekkooi-Timminga, 1989).

An approach to local measurement accuracy comparable to the information function approach has yet to be designed within the NIRT context. Meijer et al. (1995) argued that in a NIRT approach, information about the slope of the IRF is summarized by the item reliability coefficient (Mokken, 1971, p. 143). This coefficient expresses the degree to which the scores of a group of persons obtained in an administration of item g can be repeated in an independent replication of item g , denoted g' . Assume that $P_g(\theta) \equiv P_{g'}(\theta)$. Let $\pi_{gg'} \equiv \pi_{gg}$ be the proportion of the population with correct answers on two independent replications of item g . Further, note that $\pi_g = \pi_{g'}$. The reliability of item g is the correlation between X_g and $X_{g'}$, denoted ρ_g , and defined as

$$\rho_g = 1 - \frac{\pi_g - \pi_{gg}}{\pi_g(1 - \pi_g)} = \frac{\pi_{gg} - \pi_g^2}{\pi_g(1 - \pi_g)} , \quad (41)$$

(Mokken, 1971, p. 143). Because independent replications of item g are usually unavailable, a direct estimate of π_{gg} cannot be obtained. Mokken (1971, pp. 142-147) proposed two methods for approximating π_{gg} that use information from item g , the next easier item $g + 1$, and the next more difficult item $g - 1$. Both methods are based on the following rationale.

Note that π_{gg} is defined as

$$\pi_{gg} = \int_{\theta} P_g(\theta) P_g(\theta) dG(\theta) . \quad (42)$$

Next, one of the factors in the integrand is replaced by a linear function of one conditional probability, $P_{g-1}(\theta)$ or $P_{g+1}(\theta)$ [Mokken's Method 1], or two conditional probabilities, $P_{g-1}(\theta)$ and $P_{g+1}(\theta)$ [Mokken's Method 2]:

$$\tilde{P}_g(\theta) = a + bP_{g-1}(\theta) + cP_{g+1}(\theta); \quad a, b, c \text{ are constants.} \quad (43)$$

The choice of the constants a , b , and c is discussed elsewhere (Mokken, 1971, pp. 142-147). Substitution for one of the factors $P_g(\theta)$ in the integrand, and then integration, yields

$$\tilde{\pi}_{gg} = a\pi_g + b\pi_{g-1,g} + c\pi_{g,g+1}. \quad (44)$$

This approximation to π_{gg} can be estimated from observable quantities. Insertion of the estimate of the approximation to π_{gg} in ρ_g yields an estimate of the item reliability. Mokken (1971) proposed these methods for estimating total score reliability. Molenaar & Sijtsma (1984; Sijtsma & Molenaar, 1987) proposed a method denoted MS that was intended to further improve the estimation of total score reliability. All methods were also generalized to polytomous item response data (Molenaar & Sijtsma, 1988).

Meijer et al. (1995) found that item reliability can be estimated almost without bias if MS is used. In addition, MS has an acceptably small standard error for $N \geq 300$, which is a reasonable lower bound for the sample size for most test and questionnaire construction research. The results were generally consistent when one of the assumptions (nonintersecting IRFs) of the method was violated. Item reliability can be used in the process of test construction; it replaces the discrimination parameter (a parameter that is monotonically related to the first derivative of the IRF in the inflection point of a logistic or normal ogive curve) used in parametric IRT models (Meijer et al., 1995; Meredith, 1965).

To assess the accuracy of the measurement of persons using the total score on k items, X_+ , the reliability of this total score is estimated using the same rationale as for item reliability. The equation for total score reliability is

$$\rho_{X_+X'_+} = 1 - \frac{\sum (\pi_g - \pi_{gg})}{\sigma^2(X_+)}, \quad (45)$$

where $\sigma^2(X_+)$ is the variance of X_+ . After the π_{gg} s have been estimated, they are inserted in Equation 45.

For total score reliability, Sijtsma & Molenaar (1987) found that Mokken's methods and MS had negligible bias, whereas classical lower bound methods such as Cronbach's (1951) α and Guttman's (1945) λ -2 performed relatively poorly. The standard error of all methods, including the lower bounds, was acceptable for realistic sample sizes.

The estimated total score reliability can be used to estimate a standard error of measurement if the test is intended for individual diagnostic use. If the purpose is to compare groups with respect to their mean total scores (e.g., on a questionnaire about euthanasia), total score reliability may be lower and the standard error of measurement is of little interest.

Comments on Reliability Estimation

The estimation and use of item reliability is not a common practice in test and questionnaire construction research. Like item reliability, H_g is an increasing function of the slope of the IRF. Therefore, it can be questioned whether using item reliability in addition to H_g contributes to knowledge about individual items.

For two independent replications of item g , there is remarkable equivalence between the reliability of item g and the H coefficient. Note that $\text{Cov}(X_g, X_{g'}) = \pi_{gg} - \pi_g^2$. The scalability coefficient $H_{gg'}$ for two independent replications thus is

$$H_{gg'} = \frac{\text{Cov}(X_g, X_{g'})}{\text{Cov}(X_g, X_{g'})^{\max}} = \frac{\pi_{gg} - \pi_g^2}{\pi_g(1 - \pi_g)}. \quad (46)$$

This is exactly the reliability of item g : $H_{gg'} = \rho_g$. Item reliability can be interpreted as the scalability of an item with respect to an identical item (i.e., another item that measures the same latent trait with exactly the same IRF).

This result shows that item reliability is related to, but not identical to, the concept of scalability. Item reliability indicates how well item performance can be repeated under similar circumstances. To estimate ρ_g , information is used from one or two neighboring items in the difficulty ordering that replace a real replication of item g . How to use ρ_g and H_g in a complementary way is a topic for future investigation.

Finally, note that information about measurement accuracy obtained using item reliability (Equation 41) is not as fine-grained as similar information obtained using the item information function (Equation 39). Given the absence of the information function in NIRT, reliability is a necessary and well-accepted alternative.

Discussion

One defining characteristic of NIRT is its effort to get the most out of the weakest possible assumptions. For example, the MHM (Mokken & Lewis, 1982) implies a person ordering in the sense of a stochastic ordering of θ by the observable sum score X_+ (Equation 7). By implication, this is also true for parametric models that are special cases of the MHM, such as the Rasch model, the 2PLM, and the 3PLM. If person ordering is the only objective of a test application, then the MHM will be useful because it will probably allow more items to be included in the test than the stronger models (Meijer et al., 1990). As a result, depending on the characteristics of the additional items included in the test, X_+ may be more reliable and the latent trait better measured. These practical results may also appeal to test constructors and measurement practitioners. Stout's (1987, 1990) essentially unidimensional model based on WM may be interpreted as a step further in weakening the assumptions. Yet, using Stout's model still allows enough information to be obtained to order persons. NIRT should not, and cannot, always replace stronger IRT models. For example, test equating and adaptive testing are more easily implemented in a parametric IRT context; the same is true for testing hypotheses about cognitive processes underlying item responses (Embretson, 1991; Kelderman & Rijkes, 1994).

More importantly, the definition of NIRT models, and the derivation of testable consequences from their assumptions, lead to theoretical results that are also potentially important for parametric IRT. For example, conditional association (Equation 19) and manifest monotonicity (Equation 20) not only hold for the MHM, but by implication (and with minor modifications due to strict increasingness of the IRFs) also hold for the Rasch model, the 2PLM, the 3PLM, the normal ogive models, and the OPLM. If conditional association or manifest monotonicity fail, none of these parametric models holds. Another example is Ellis & Van den Wollenberg's (1993) local homogeneity assumption (Equation 17), which is important for a better understanding of DIF because the MHM model, and each parametric special case, will hold only in all subpopulations if local homogeneity is part of that model. A third example is the monotone likelihood ratio property (Equation 4; Grayson, 1988) of the MHM that also holds for the 2PLM and 3PLM. By implication, in these models θ is stochastically ordered by X_+ ; however, in these models X_+ is not a sufficient statistic for θ as it is in the Rasch model.

The choice of topics discussed here could easily have been broadened. For example, the work of Van der Flier (1980, 1982) and Meijer (1994; Meijer et al., 1994) addressed person-fit evaluation in a nonparametric framework using the pattern of item scores on k dichotomous items from a test or questionnaire. Shealy & Stout (1993) discussed a multidimensional NIRT model for the assessment of test bias and DIF. Van Schuur (1984) addressed the analysis of dichotomous preference data using a nonparametric unfolding procedure that was inspired by the Mokken scaling procedure. A fundamental model formulation underlying this procedure was provided by Post (1992) and Post & Snijders (1993).

The extension of the NIRT approach to polytomous item scores (Hemker, 1996; Hemker et al., 1995; Junker, 1991; Molenaar, 1982c, 1986, 1997) also deserves full attention in future NIRT research. The following example may further clarify how NIRT can provide fundamental results that are useful for IRT in general.

Hemker et al. (1996, 1997) found that, unlike the dichotomous MHM and DMM, none of the existing polytomous NIRT models has the MLR property (Equation 4; Hemker et al., 1996) or the weaker stochastic ordering property in Equation 7 (Hemker et al., 1997) for the total score X_+ based on the sum of k polytomous item scores. This negative result holds for the polytomous versions of the MHM and DMM (Molenaar, 1997), and also for a polytomous NIRT version (Hemker et al., 1996, 1997) of the partial credit model (Masters, 1982). These are theoretical results, and preliminary results based on simulated data have indicated that a stochastic ordering of θ by X_+ may be possible for typical choices of item parameters and person distributions; that is, a reasonable spread of the item locations, an average discrimination per item, and a bell-shaped distribution of θ . Much is unknown, however, and additional research is needed.

One solution to these problems would be to resort to parametric polytomous IRT for analyzing the data. However, the negative results for MLR and stochastic ordering were not typical of NIRT and were also found for the studied parametric IRT models, such as the graded response model (Samejima, 1969) and the generalized partial credit model (Muraki, 1992). MLR and stochastic ordering hold, however, for the partial credit model (Masters, 1982) and its special case, the rating scale model (Andrich, 1978). The absence of ordering properties, such as MLR and stochastic ordering of θ by X_+ , thus seems to be problematic in most polytomous models, and clearly goes beyond NIRT. This is one of the instances for which the study of NIRT has produced theoretical results that are relevant in a much wider IRT context.

References

- Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika*, 42, 375–401.
- Cliff, N. (1979). Test theory without true scores? *Psychometrika*, 44, 373–393.
- Cliff, N., & Donoghue, J. R. (1992). Ordinal test fidelity estimated by an item sampling model. *Psychometrika*, 57, 217–236.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, 44, 315–331.
- De Gruijter, D. N. M. (1994). Comparison of the nonparametric Mokken model and parametric IRT models using latent class analysis. *Applied Psychological Measurement*, 18, 27–34.
- De Jong, A. (1984). *Over psychiatrische invaliditeit* (On psychiatric disability). Unpublished doctoral dissertation, Rijksuniversiteit, Groningen.
- De Jong, A., & Molenaar, I. W. (1987). An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *Journal of Psychiatric Research*, 21, 137–149.
- Douglas, J. (1997). Joint consistency of nonparametric item characteristic curves and ability estimation. *Psychometrika*, 62, 7–28.
- Ellis, J. L. (1993). Subpopulation invariance of patterns in covariance matrices. *British Journal of Mathematical and Statistical Psychology*, 46, 231–254.

- Ellis, J. L. (1994). *Foundations of monotone latent variable models*. Nijmegen, The Netherlands: NICI, Catholic University of Nijmegen.
- Ellis, J. L., & Junker, B. W. (in press). Tail-measurability in monotone latent variable models. *Psychometrika*.
- 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.
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56, 495-515.
- Esary, J. D., Proschan, F., & Walkup, D. W. (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, 38, 1466-1474.
- Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* [Introduction to psychological test theory]. Bern: Huber.
- Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15-38). New York: Springer.
- Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.
- Gillespie, M. W., TenVergert, E. M., & Kingma, J. (1988). Secular trends in abortion attitudes: 1975-1980-1985. *The Journal of Psychology*, 122, 323-341.
- Goa, F., & Stout, W. F. (1996). *Using resampling to eliminate DIMTEST estimation bias*. Manuscript submitted for publication.
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, 53, 383-392.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282.
- 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.
- Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, 20, 1-14.
- Hemker, B. T. (1996). *Unidimensional IRT models for polytomous items, with results for Mokken scale analysis*. Unpublished doctoral dissertation, Utrecht University, The Netherlands.
- 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.
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, 61, 679-693.
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, 62, 331-347.
- Henning, H. J. (1976). Die Technik der Mokken-Skalenanalyse [The technique of Mokken scale analysis]. *Psychologische Beiträge*, 18, 410-430.
- Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika*, 46, 79-92.
- Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, 55, 577-601.
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, 14, 1523-1543.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood IL: Dow Jones-Irwin.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, 59, 77-79.
- Jansen, P. G. W. (1982a). De onbruikbaarheid van Mokschaalanalyse [On the uselessness of Mokken scale analysis]. *Tijdschrift voor Onderwijsresearch*, 7, 11-24.
- Jansen, P. W. G. (1982b). Homogenitätsmessung mit Hilfe des Koeffizienten H von Loevinger: Eine kritische Diskussion [Measuring homogeneity by means of Loevinger's coefficient H: A critical discussion]. *Psychologische Beiträge*, 24, 96-105.
- Jansen, P. G. W., Roskam, E. E. Ch. I., & Van den Wollenberg, A. L. (1982). De Mokschaal gewogen [Weighing the Mokken scaling procedure]. *Tijdschrift voor Onderwijsresearch*, 7, 31-42.
- Jansen, P. G. W., Roskam, E. E. Ch. I., & Van den Wollenberg, A. L. (1984). Discussion on the usefulness of the Mokken procedure for nonparametric scaling. *Psychologische Beiträge*, 26, 722-735.
- Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, 56, 255-278.
- Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, 21, 1359-1378.

- Junker, B. W. (in press). Some remarks on Scheiblechner's treatment of ISOP models. *Psychometrika*.
- Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *The Annals of Statistics*, 25, 1327–1343.
- Junker, B. W., & Sijtsma, K. (1997). *Latent and manifest monotonicity in item response models*. Manuscript submitted for publication.
- Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, 59, 149–176.
- Kim, H. R., Zhang, J., & Stout, W. (1996). *A new index of dimensionality–DETECT* (Internal Report). Urbana-Champaign: University of Illinois, Dept. of Statistics.
- Kingma, J., & Ten Vergert, E. M. (1985). A nonparametric scale analysis of the development of conservation. *Applied Psychological Measurement*, 9, 375–387.
- Laros, J. A., & Tellegen, P. J. (1991). *Construction and validation of the SON-R 5 1/2-17, the Snijders-Oomen non-verbal intelligence test*. Groningen, The Netherlands: Wolters-Noordhoff.
- Lehmann, E. L. (1959). *Testing statistical hypotheses*. New York: Wiley.
- Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, 37, 1137–1153.
- Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Chapman & Hall.
- Levine, M. V. (1970). Transformations that render curves parallel. *Journal of Mathematical Psychology*, 7, 410–443.
- Lewis, C. (1982). *Developments in nonparametric ability estimation* (Heymans Bulletin HB-82-609-EX). Groningen, The Netherlands: Rijksuniversiteit Groningen.
- Lewis, C. (1983). Bayesian inference for latent abilities. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 224–251). San Francisco: Jossey-Bass.
- 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. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Meijer, R. R. (1994). *Nonparametric person fit analysis*. Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam.
- Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, 18, 111–120.
- Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. *Applied Psychological Measurement*, 19, 323–335.
- 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.
- Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika*, 30, 419–440.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.
- 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.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6, 417–430.
- Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement*, 10, 279–285.
- Molenaar, I. W. (1982a). De beperkte bruikbaarheid van Jansens kritiek [On the limited usefulness of Jansen's criticisms]. *Tijdschrift voor Onderwijsresearch*, 7, 25–30.
- Molenaar, I. W. (1982b). Een tweede weging van de Mokkenschaaal [A second weighing of the Mokken scaling procedure]. *Tijdschrift voor Onderwijsresearch*, 7, 172–181.
- Molenaar, I. W. (1982c). Mokken scaling revisited. *Kwantitatieve Methoden*, 3(8), 145–164.
- Molenaar, I. W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën [An exercise in item response theory for three ordered answer categories]. In G. F. Pikemaat & J. J. A. Moors (Eds.), *Liber amicorum Jaap Mulwijk* (pp. 39–57). Groningen, The Netherlands: Econometrisch Instituut.
- Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multi-category items. *Kwantitatieve Methoden*, 12(37), 97–117.

- 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.
- Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift voor Onderwijsresearch*, 9, 257–268.
- Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden*, 9(28), 115–126.
- Molenaar, I. W., Debets, P., Sijtsma, K., & Hemker, B. T. (1994). *User's manual for MSP*. Groningen, The Netherlands: iecProGAMMA.
- Moorer, P., & Suurmeijer, T. P. B. M. (1994). A study of the unidimensionality and cumulativeness of the MOS short-form general health survey. *Psychological Reports*, 74, 467–470.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement*, 17, 29–38.
- Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement*, 31, 17–35.
- Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, 18, 41–68.
- Niemöller, B. (1980). *Mokken scale. STAP User's Manual, Volume 4*. Amsterdam: University of Amsterdam Press.
- Niemöller, B., & Van Schuur, W. H. (1980). *Mokken test. STAP User's Manual, Volume 4*. Amsterdam: University of Amsterdam Press.
- Niemöller, K., & Van Schuur, H. (1983). Stochastic models for unidimensional scaling: Mokken and Rasch. In D. McKay, N. Schofield, & P. Whitley (Eds.), *Data analysis and the social sciences* (pp. 120–170). London: Frances Pinter.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Post, W. J. (1992). *Nonparametric unfolding models: A latent structure approach*. Leiden University, the Netherlands: DSWO Press.
- Post, W. J., & Snijders, T. A. B. (1993). Nonparametric unfolding models for dichotomous data. *Methodika*, 7, 130–156.
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611–630.
- Ramsay, J. O. (1993). *TESTGRAF: A computer program for nonparametric analysis of testing data*. Unpublished manuscript, McGill University.
- Ramsay, J. O. (1995). A similarity-based smoothing approach to nondimensional item analysis. *Psychometrika*, 60, 323–339.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Ringdal, G. I., & Ringdal, K. (1993). Testing the EORTC Quality of Life Questionnaire on cancer patients with heterogeneous diagnosis. *Quality of Life Research*, 2, 129–140.
- Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49, 425–435.
- Rosenbaum, P. R. (1985). Comparing distributions of item responses for two groups. *British Journal of Mathematical and Statistical Psychology*, 38, 206–215.
- 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.
- Roskam, E. E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. In E. Degreef & J. van Buggenhout (Eds.), *Trends in mathematical psychology* (pp. 293–307). Amsterdam: Elsevier Science Publishers.
- Roskam, E. E., Van den Wollenberg, A. L., & Jansen, P. G. W. (1986). The Mokken scale: A critical discussion. *Applied Psychological Measurement*, 10, 265–277.
- Rost, J. (1996). *Test Theorie. Test Konstruktion* [Test theory. Test construction]. Bern: Huber.
- Roussos, L., & Stout, W. (1995, April). *Simulation study of the effectiveness of using new proximity measures with hierarchical cluster analysis to detect dimensionality structure in simulated data*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Roussos, L. A., Stout, W. F., & Marden, J. I. (1994). *Analysis of the multidimensional simple structure of standardized tests using DIMTEST with hierarchical cluster analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.
- Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika*, 60, 281–304.
- Schulman, R. S., & Haden, R. L. (1975). A test theory model for ordinal measurements. *Psychometrika*, 40, 455–472.

- Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale NJ: Erlbaum.
- Sijtsma, K. (1984). Useful nonparametric scaling: A reply to Jansen. *Psychologische Beiträge*, 26, 423–437.
- Sijtsma, K. (1986). Another note on the usefulness of Mokken scaling. *Psychologische Beiträge*, 28, 425–432.
- Sijtsma, K. (1988). *Contributions to Mokken's non-parametric item response theory*. Amsterdam: Free University Press.
- Sijtsma, K., & Hemker, B. T. (in press). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*.
- Sijtsma, K., & Junker, B. W. (1996). A survey of the theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, 49, 79–105.
- Sijtsma, K., & Junker, B. W. (1997). Invariant item ordering of transitive reasoning tasks. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 100–110). Münster, Germany: Waxmann Verlag.
- 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., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, 52, 79–97.
- Sijtsma, K., & Prins, P. M. (1986). Itemselectie in het Mokken model [Item selection in the Mokken model]. *Tijdschrift voor Onderwijsresearch*, 11, 121–129.
- Stokman, F. N. (1977). *Roll calls and sponsorship: A methodological analysis of third world group formation in the United Nations*. Leyden: Sijthoff.
- Stokman, F. N., & Van Schuur, W. H. (1980). Basic scaling. *Quality & Quantity*, 14, 5–30.
- 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.
- Stout, W. F., Habing, B., Douglas, J., Kim, H., Rousos, L., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331–354.
- Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, 48, 191–199.
- Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets & Zeitlinger.
- Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13, 267–298.
- Van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika*, 54, 237–247.
- Van der Veen, G. (1992). *Principes in praktijk. CNV-leden over collectieve acties* [Principles into practice. Labor union members on means of political pressure]. Kampen, The Netherlands: Kok.
- Van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists*. Amsterdam: CT Press.
- Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–237). New York: Springer-Verlag.
- Verweij, A. C. (1994). *Scaling transitive inference in 7–12 year old children*. Unpublished doctoral dissertation, Vrije Universiteit, Amsterdam.
- Verweij, A. C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development*, 19, 219–238.
- Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale NJ: Erlbaum.
- Weijmar Schultz, W. M. C., & Van de Wiel, H. B. M. (1991). *Sexual functioning after gynecological cancer treatment*. Unpublished doctoral dissertation, Rijksuniversiteit, Groningen.
- Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, 53, 774–789.
- Zhang, J., & Stout, W. (1995, April). *Theoretical results concerning DETECT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Acknowledgments

Thanks to Bas T. Hemker, Brian W. Junker, Rob R. Meijer, Ivo W. Molenaar, Wijbrandt van Schuur, William F. Stout, and an anonymous reviewer for their critical comments on previous drafts of this paper.

Author's Address

Send requests for reprints or further information to Klaas Sijtsma, Vakgroep Methoden en Technieken van Onderzoek, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: k.sijtsma@kub.nl.