Nonparametric Item Response Theory in Action

Junker, B.W.; Sijtsma, K.

Published in:
Applied Psychological Measurement

Publication date:
2001

Document Version
Publisher's PDF, also known as Version of record

Nonparametric Item Response Theory in Action: An Overview of the Special Issue

Brian W. Junker and Klaas Sijtsma

Applied Psychological Measurement 2001; 25; 211
DOI: 10.1177/01466210122032028

Nonparametric Item Response Theory in Action: An Overview of the Special Issue

Brian W. Junker, Carnegie Mellon University

Klaas Sijtsma, Tilburg University

Although most item response theory (IRT) applications and related methodologies involve model fitting within a single parametric IRT (PIRT) family [e.g., the Rasch (1960) model or the three-parameter logistic model (3PLM; Lord, 1980)], nonparametric IRT (NIRT) research has been growing in recent years. Three broad motivations for the development and continued interest in NIRT can be identified:

1. To identify a commonality among PIRT and IRT-like models, model features [e.g., local independence (LI), monotonicity of item response functions (IRFs), unidimensionality of the latent variable] should be characterized, and it should be discovered what happens when models satisfy only weakened versions of these features. Characterizing successful and unsuccessful inferences under these broad model features can be attempted in order to understand how IRT models aggregate information from data. All this can be done with NIRT.
2. Any model applied to data is likely to be incorrect. When a family of PIRT models has been shown (or is suspected) to fit poorly, a more flexible family of NIRT models often is desired. These NIRT models have been used to: (1) assess violations of LI due to nuisance traits (e.g., latent variable multidimensionality) or the testing context influencing test performance (e.g., speededness and question wording), (2) clarify questions about the sources and effects of differential item functioning, (3) provide a flexible context in which to develop methodology for establishing the most appropriate number of latent dimensions underlying a test, and (4) serve as alternatives for PIRT models in tests of fit.
3. In psychological and sociological research, when it is necessary to develop a new questionnaire or measurement instrument, there often are fewer examinees and items than are desired for fitting PIRT models in large-scale educational testing. NIRT provides tools that are easy to use in small samples. It can identify items that scale together well (follow a particular set of NIRT assumptions). NIRT also identifies several subscales with simple structure among the scales, if the items do not form a single unidimensional scale.

Basic Assumptions of NIRT

Each NIRT approach begins with a minimal set of assumptions necessary to obtain a falsifiable model that allows for the measurement of persons and/or items, usually on a scale that has, at most, ordinal measurement properties. These assumptions define a NIRT "model." Researchers accustomed to PIRT often think of these assumptions as defining a class containing many familiar PIRT models.

Let $X_1, X_2, \ldots, X_J$ be dichotomous item response variables for J test items, with $x_j \in \{0, 1\}$. The basic assumptions of NIRT are then as follows:

Applied Psychological Measurement, Vol. 25 No. 3, September 2001, 211–220

1. *LI.* A (possibly multidimensional) latent variable θ exists, such that the joint conditional probability of J item responses can be written as

$$P(X_1 = x_1, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = 1|\theta)^{x_j} \left[1 - P(X_j = 1|\theta)\right]^{1-x_j} . \quad (1)$$

2. *Monotonicity.* The IRFs $P_j(\theta) = P(X_j = 1|\theta)$ are nondecreasing as a function of θ (or its coordinates, if θ is multidimensional).
3. *Unidimensionality.* θ takes values in (a subset of) real numbers.

The NIRT model satisfying only LI, monotonicity, and unidimensionality is known as the monotone homogeneity (MH) model (also called the monotone unidimensional latent variable model; Holland & Rosenbaum, 1986; Meredith, 1965; Mokken, 1971; Mokken & Lewis, 1982). The class of IRT models that satisfies these three assumptions—the MH class—includes, for example, the normal ogive models, the Rasch model, and the 3PLM. Defining the MH class in terms of LI, monotonicity, and unidimensionality shows three of the properties that are common and essential to well-known PIRT models (2001). Omnibus tests for MH and related models have recently been proposed (Bartolucci & Forcina, 2000; Yuan & Clarke, 2001).

Much more along these lines is possible: assumptions can be weakened to a point that ordinal measurement still is possible, or more assumptions can be added to produce more-restrictive models with interesting measurement properties (e.g., Hemker, Sijtsma, Molenaar, & Junker, 1997; Junker & Ellis, 1997; Sijtsma & Hemker, 1998). For example, no two of the three NIRT assumptions define a restrictive model for observable data (e.g., Holland & Rosenbaum, 1986; Junker, 1993; Stout, 1990; Suppes & Zanotti, 1981). Although none of these assumptions can be completely eliminated, they can be weakened considerably. Pursuing inference about persons or items under weakened assumptions is a longstanding interest of both PIRT and NIRT research.

In NIRT research, Stout's (1987,1990) concern was simultaneously weakening LI and monotonicity while retaining enough structure to make ordinally consistent inferences about a dominant, unidimensional θ. Retaining LI and unidimensionality, but replacing monotonicity with other smoothness assumptions to obtain nonparametric regression estimates of nonmonotone IRFs also has been studied (Ramsay, 1991). More recently, Zhang & Stout (1999) followed PIRT work (e.g., McDonald, 1997; Reckase, 1997; for more recent developments, see Béguin & Glas, 1998) by defining a compensatory multidimensional class of NIRT models retaining LI and monotonicity, in which various procedures for estimating the number of latent dimensions can be examined. Polytomous generalizations also have been developed (e.g., Junker, 1991; Molenaar, 1997; Nandakumar, Yu, Li, & Stout, 1998).

In North America, applied NIRT research has been inspired by the need for more flexible data analysis and hypothesis testing tools when PIRT methods fail. In Europe (especially the Netherlands and Germany), inspiration has come from using summary statistics justified by NIRT models to perform item scaling analyses in small samples typically encountered in psychological and sociological research. In the former approach, the model is a filter through which item properties become more transparent as inessential features are stripped away. In the latter approach, the model is a criterion against which items are evaluated.

This difference in focus also is present within PIRT research. Early adherents of the Rasch model (e.g., Andrich, 1988; Fischer, 1974; Wright & Stone, 1979) used the model as a criterion for useful measurement and stressed model-data fit, rejecting items if the Rasch model did not fit them. In contrast, adherents of the two-parameter logistic model and the 3PLM (e.g., Bock & Aitkin, 1981; Hambleton, 1989; Lord, 1980) were more inclined to accept these weaker models for describing the

characteristics of items that were not well fitted by the Rasch model, but still contributed positively to measurement accuracy or a better reflection of the latent trait.

These approaches have much in common, however, and apparent differences are neither large nor fundamental in nature. Growing collaboration across the Atlantic serves to further integrate the approaches. For example, although much work on nonparametric estimates of item category response functions and conditional covariances between items given a possibly incomplete latent trait has been pursued by American and Canadian researchers (e.g., Douglas, 1997; Habing & Donoghue, in press; Ramsay, 1991, 1997, 2000; Stout, 1987), European researchers (e.g., Bartolucci & Forcina, 2000; Vermunt, 2001) brought new modeling insights to these problems. On the other hand, although computationally modest methods for model fit and scale construction based on probability inequalities derived under MH and related models have long been pursued in Europe (e.g., Ellis & van den Wollenberg, 1993; Hemker, Sijtsma, & Molenaar, 1995; Mokken, 1971; Molenaar, 1997), similar efforts have been made by Americans and Australians (e.g. Holland & Rosenbaum, 1986; Huynh, 1994; Junker, 1993).

Exploratory Data Analysis and Item and Test Features

Two major themes in NIRT research—(1) nonparametric regression estimates of IRFs and (2) the estimation of conditional covariances between items, given a θ that might or might not be "complete" in the sense that LI holds—have provided a new repertoire of exploratory techniques for situations in which standard PIRT models do not fit well. A PIRT model can be thought of as a kind of "grid" that is stretched over the data. This grid characterizes the general features of item responses so that predictions can be made from them. Model parameters estimated from the data show how the grid bends to conform to the data, but it is only flexible in a limited number of ways. For example, commonly used PIRT models impose a monotonicity assumption on IRFs so that dips or bumps cannot be seen. Instead, they drive discrimination parameter estimates toward zero. NIRT methods provide a grid that is more flexible, enabling assessment of the importance of potential irregularities in the data-generating process.

Ramsay (1991, 1997, 2000) popularized nonparametric estimation of IRFs by proposing relatively easy-to-implement nonparametric regression methods. Related work also has been pursued (Drasgow, Levine, Tsien, Williams, & Mead, 1992; Samejima, 1998). Ramsay's (2000) TESTGRAF98 program provides a straightforward use of nonparametric regression as an exploratory tool for assessing IRF monotonicity for each item response variable X_j, using as a proxy for θ either the total score, $X_+ = \sum_j X_j$ (Ramsay, 1991), or the rest-score, $R_j = X_+ - X_j$ (Junker & Sijtsma, 2000). This methodology could be used, for example, to explore deviations from parametric IRFs when nonparametric tests (e.g., Molenaar & Sijtsma, 2000; Stout, 1990) confirm the MH model, but a specific parametric form (e.g., the two-parameter logistic model) is rejected. Ramsay (1991) applied this approach to identifying possibly defective test items from a large introductory psychology course. Other applications demonstrated only moderate discriminability in two widely used self-report instruments for screening major depressive disorders (Santor, Zuroff, Ramsay, Cervantes, & Palacios, 1995).

In MH models, the conditional covariances $Cov(X_i, X_j|\theta)$ are all zero. However, LI never holds exactly in practice. Substantial NIRT research effort has been devoted to determining when these conditional covariances are far enough from zero to invalidate a simple monotone unidimensional IRT model. Stout's (1987, 1990) conception of essential independence allows conditional covariances to be positive or negative and to vary considerably from one item pair to the next, yet be controlled enough to allow for consistent ordinal measurement of persons. Conditional covariances also play a role in all formal and informal tests and measures of unidimensionality (e.g., Stout et al.,

1996; Stout, Nandakumar, & Habing, 1996). Estimating $Cov(X_i, X_j|\theta)$ as a function of θ can be a useful exploratory device, because it can suggest explanations for multidimensionality by showing where along θ local dependence occurs (Douglas, Kim, Habing, & Gao, 1998).

In this issue, Habing (2001) provides a review of the application of this methodology to the estimation of the entire conditional covariance function, as well as nonparametric regression estimation of IRFs. Conditional covariance estimates using the total score or the rest score as a proxy for θ are subject to biases (e.g., Junker, 1993). Habing briefly reviews basic bootstrap ideas and demonstrates an application of the parametric bootstrap to nonparametric conditional covariance estimates. This application can be used to reduce or eliminate biases and to provide confidence envelopes for the estimates.

Douglas & Cohen's (2001) paper in this issue compares the fit of a parametric IRF model with a nonparametric regression estimate of the same IRF. They use a parametric bootstrap based on a carefully selected parametric approximation to the nonparametric IRF to generate a reference distribution for testing the fit of the maximum likelihood parametric IRFs. Their bootstrapped hypothesis test might be less biased in favor of the PIRT model than other parametric bootstrap techniques (e.g., Gelman, Meng, & Stern, 1996; Stone, 2000). Douglas and Cohen show, using two simulated and two real-testing examples, that the bootstrap provides a powerful adjunct to graphical techniques.

Model-Data Fit and the Explanation of Data Structure

PIRT models typically specify whether one or more dimensions describe the data. To some extent, these models allow the number of dimensions to be subjected to hypothesis testing (e.g., Bartholomew, 1987; Béguin & Glas, 1998; Bock, Gibbons, & Muraki, 1988; Glas & Verhelst, 1995; McDonald, 1997; Reckase, 1997). However, the greater flexibility of NIRT facilitates the assessment of underlying trait dimensionality. When studying dimensionality within a PIRT family, misfit to the shape of the response model can be misinterpreted as an increase in the number of underlying dimensions. The classic example of this is the tendency for traditional linear factor analysis to over-estimate the number of dimensions in dichotomous data (e.g., Miecskowski et al., 1993). Dimensionality estimated apart from parametric features of the response model might be a more fundamental characteristic of the data and less likely to have arisen as a consequence of some other aspect of model-data misfit. This is the motivation for the item selection procedures in the computer programs MSP5 (Mokken, 1971; Molenaar & Sijtsma, 2000), DIMTEST (Nandakumar & Stout, 1993; Stout, 1990), and DETECT (Kim, Zhang, & Stout, 1995; Zhang & Stout, 1999).

Stout et al.'s (1996) conditional covariance-based methods for assessing latent trait dimensionality have been applied to a variety of data sources, including data from the LSAT/LSAC. They also have been extended to the case of polytomous responses (Nandakumar et al., 1998). Stout et al.'s ideas appear in work on dimensionality assessment (Gessaroli & de Champlain, 1996; Oshima & Miller, 1992). Related considerations are also found in nonparametric detection of differential item and subtest functioning (Bolt & Stout, 1996; Douglas, Stout, & DiBello, 1996; Li & Stout, 1996; Shealy & Stout, 1993).

For dichotomous items ($X_j = 0$ or 1, for an incorrect or correct answer, respectively), a theory of scale construction—selecting groups of items that are related in the sense that the MH model is probably appropriate for them—has existed for quite some time (Mokken, 1971; Sijtsma, 1998). The principal tools involved are easy-to-compute adaptations of Loevinger's (1948; Mokken & Lewis, 1982) H coefficient, comparing the marginal covariance, $Cov(X_i, X_j)$, of each item pair with the maximum possible covariance $[Cov_{max}(X_i, X_j)]$. This preserves the margins of the observed $X_i \times X_j$ table. The bound $Cov_{max}(X_i, X_j)$ is obtained by adjusting the table to remove Guttman

errors (Mokken, 1997; Mokken & Lewis, 1982). These methods also have been extended to polytomous items (Hemker et al., 1995; Molenaar, 1991; Sijtsma & Verweij, 1999).

In his paper in this issue, Bolt (2001) discusses a geometric approach to identifying the continuous multidimensional latent structure underlying observable dichotomous item response data, based on Zhang & Stout (1999). Implementation of the method requires circular/spherical multidimensional scaling of average conditional covariances, given appropriate rest scores, in terms of the angles of item discrimination vectors in the subspace perpendicular to a "dominant" latent dimension.

Bolt (2001) compared the method with DIMTEST and related dimension-counting methods. A broad range of simulated and real-data multidimensional latent structures were recovered, including "simple structure" (items can be partitioned into groups that are unidimensional with respect to different latent variables) and "fan" structures (items load to varying degrees on several latent variables at once). Again, computational and graphical methods combine to give a complete data analysis.

Within psychometrics, there is a growing interest in cognitive assessment models (i.e., testing models that attempt to account for and measure the cognitive processes and solution strategies that underlie dichotomous or polytomous item responses). This interest has resulted in the development of many different parametric "componential" IRT models, including the linear logistic test model (Fischer, 1974, 1995; Scheiblechner, 1972), multidimensional latent trait models (Adams, Wilson, & Wang, 1997; Embretson, 1991; Kelderman & Rijkes, 1994), and a multicomponent latent trait model (Embretson, 1985, 1997). Discrete latent structure approaches also have been proposed, including the constrained latent class approach (Haertel & Wiley, 1993) and the general Bayesian inference network approach (e.g., Mislevy, 1996). Various attempts have been made to blend discrete and continuous methodologies (DiBello, Stout, & Roussos, 1995; Tatsuoka, 1995).

Related to this interest in cognitive modeling is person-fit research or appropriateness measurement (e.g., Emons, Meijer, & Sijtsma, in press; Meijer, 1994; Sijtsma & Meijer, 2001). The main interest is in understanding the psychological mechanisms (e.g., test anxiety, lack of concentration) that produce a particular pattern of item scores. Respondents showing misfitting item score patterns might be removed from the item analysis or the information about misfit might be used for interpreting their latent trait estimates.

Junker & Sijtsma's (2001) paper in this issue concerns the role of NIRT methodology in constructing and evaluating cognitive assessment models. They reanalyzed a dichotomized version of "deductive strategy" transitive reasoning data (Sijtsma & Verweij, 1999) by estimating a discrete latent-structure version of Embretson's (1985, 1997) multicomponent model (see also DiBello et al., 1995; Haertel & Wiley, 1993; Tatsuoka, 1995). Junker and Sijtsma show that appropriate versions of monotonicity and LI plausibly hold for these data. They then speculate about whether simple data summaries that are informative about latent attributes (cognitive components) were present or absent in individual students, based on each pattern of responses to the set of transitive reasoning items. Junker and Sijtsma also discuss the translation of useful stochastic-ordering properties from unidimensional NIRT research to their cognitive assessment models.

Measurement of Person and Item Properties

For models assuming MH, it has been shown (Grayson, 1988; Huynh, 1994) that the latent trait θ is stochastically ordered by the unweighted sum of item scores X_+ for dichotomously scored items. Assume two values of X_+, $0 \leq c < k \leq J$, and a fixed value of θ, t.

Then, the stochastic ordering of the latent variable (SOL) is

$$P(\theta > t|X_+ = c) \leq P(\theta > t|X_+ = k) \ , \quad \text{for all } t. \tag{2}$$

SOL implies that $E[\theta|X_+ = k]$ also is nondecreasing in k. On average, the increasing total score then is associated with increasing θ level, as it should be. SOL holds for all dichotomous response models satisfying LI, monotonicity, and unidimensionality. This is surprising: although X_+ is a sufficient statistic for θ only in the Rasch model, it can be used for ordering θ in any MH model, no matter how far the data deviate from the Rasch model. SOL is a useful measurement property for test practitioners who can confidently use X_+ instead of θ for ordering examinees.

Hemker et al. (1997) found that SOL holds for almost none of the familiar ordered polytomous IRT models—parametric or nonparametric. Let X_+ be the Likert score (i.e., the unweighted sum across items of the item category scores). Then, a higher X_+ does not always imply, for example, a higher mean θ. The only known polytomous response models in which SOL is guaranteed to hold are the partial-credit model (Masters, 1982) and special cases of this model, such as the rating scale model (Andrich, 1978).

Thus, from a theoretical point of view, the use of the Likert score for ordering examinees on θ is justified in almost none of the polytomous IRT models. Unless the partial-credit model fits the data, SOL failure poses a serious potential problem for test practitioners who prefer X_+ over θ. Nevertheless, preliminary simulation results (Sijtsma & Van der Ark, 2001) suggest that, in practice, the mismatch of the ordering of X_+ and θ might not be very serious in data stemming from typical choices of item parameters and a normal θ distribution.

Invariant item ordering (IIO) is an important measurement property for ordering items. Whenever J IRFs do not intersect, they can be renumbered such that

$$P_1(\theta) \leq P_2(\theta) \leq \ldots \leq P_J(\theta) \ , \quad \text{for all } \theta. \tag{3}$$

In many testing situations (e.g., intelligence testing, analysis of differential item functioning, person-fit analysis, exploring hypotheses about the order in which cognitive operations are acquired by children), ordering items by difficulty can be helpful for analyzing test data. In each situation, interpretation and analysis is made easier if the items are ordered by difficulty in the same way for every individual taking the test—i.e., the IRFs do not cross. Sijtsma & Junker (1996) developed methods for empirically investigating IIO for dichotomously scored NIRT models, and Sijtsma & Hemker (1998) investigated methods for polytomously scored PIRT and NIRT models. Sijtsma & Junker (1997) applied these methods to scale construction in developmental psychology.

In this issue, Van der Ark (2001) provides an overview of the most popular and relevant polytomous PIRT and NIRT models and measurement properties (e.g., SOL and IIO). Scoring rules for polytomous items (Akkermans, 1998; Van Engelenburg, 1997) also are addressed. Van der Ark provides useful reference tables for finding the appropriate polytomous IRT model when certain measurement properties are desired. His main points are illustrated with data from five polytomous items measuring strategies for coping with industrial odors.

Vermunt (2001) focuses on testing monotonicity and other ordering properties of the MH model. He fitted latent class models to data that incorporated the relevant order restrictions. Latent class formulations for PIRT and NIRT models are not new (Croon, 1991; Hoijtink & Molenaar, 1997; Lindsay, Clogg, & Grego, 1991), but Vermunt's proposal accommodates a wider range of NIRT/PIRT models and their specific properties than previously was possible. Vermunt provides parametric bootstrap-based tests of fit for constrained latent class models. He then compares the fit of several PIRT and NIRT models for four polytomous self-report items taken from a biopsychosocial survey.

The Special Issue concludes with two discussions (Molenaar, 2001; Stout, 2001). Both authors have devoted considerable energy to NIRT research and have also contributed to a variety of important advances in PIRT and related methods.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logits model. *Applied Psychological Measurement, 21,* 1–23.

Akkermans, L. M. W. (1998). *Studies on statistical models for polytomously scored items.* Doctoral dissertation, University of Twente, Enschede, The Netherlands.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–574.

Andrich, D. (1988). *Rasch models for measurement.* Newbury Park CA: Sage.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* New York: Oxford University Press.

Bartolucci, F., & Forcina, A. (2000). A likelihood ratio test for MTP_2 within binary variables. *Annals of Statistics, 28,* 1206–1218.

Béguin, A. A., & Glas, C. A. W. (1998). *MCMC estimation of multidimensional IRT models* (Research Report No. 98-14). Enschede, The Netherlands: University of Twente, Department of Education and Data Analysis.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an E-M algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bolt, D. (2001). Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement, 25,* 244–257.

Bolt, D., & Stout, W. F. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23,* 67–95.

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,* 315–332.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale NJ: Erlbaum.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62,* 7–28.

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement, 25,* 234–243.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23,* 129–151.

Douglas, J. A., Stout, W. F., & DiBello, L. V. (1996). A kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics, 21,* 333–363.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1992). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143–165.

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. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 195–218). New York: Academic Press.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (in press). Comparing the empirical and the theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement.*

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests* (Introduction to psychological test theory). Bern, Switzerland: Huber.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131–156). New York: Springer-Verlag.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–760.

Gessaroli, M. E., & de Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to

a set of items. *Journal of Educational Measurement, 33,* 157–179.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer-Verlag.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53,* 383–392.

Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement, 25,* 221–233.

Habing, B., & Donoghue, J. (in press). Local dependence assessment for exams with polytomous items and incomplete item-examinee layouts. *Journal of Educational and Behavioral Statistics.*

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Fredriksen & R. J. Mislevy (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale NJ: Erlbaum.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 201–220). New York: Macmillan.

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. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Hoijtink, H. & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171–189.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait models. *Annals of Statistics, 14,* 1523–1543.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59,* 77–79.

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. *Annals of Statistics, 21,* 1359–1378.

Junker, B. W., & Ellis, J. L. (1997). A characterization of monotone unidimensional latent variable models. *Annals of Statistics, 25,* 1327–1343.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24,* 65–81.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

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. F. (1995). *A new index of dimensionality—DETECT.* Unpublished manuscript.

Li, H.-H., & Stout, W. F. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647–677.

Lindsay, B., Clogg, C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association, 86,* 96–107.

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). *Application of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer.

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, The Netherlands.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psychometrika, 30,* 419–440.

Miecskowski, T. A., Sweeney, J. A., Haas, G., Junker, B. W., Brown, R. P., & Mann, J. J. (1993). Factor composition of the Suicide Intent Scale. *Suicide and Life Threatening Behavior, 23,* 37–45.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416.

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

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–368). 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.

Molenaar, I. W. (1991). A weighted Loevinger *H*-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 37,* 97–117.

Molenaar, I. W. (1997). Nonparametric methods 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. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25,* 295–299.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows* [Computer program]. Groningen, The Netherlands: ProGAMMA.

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

Nandakumar, R., Yu, F., Li, H.-H., & Stout, W. F. (1998). Assessing unidimensionality of polytomous data, *Applied Psychological Measurement, 22,* 99–115.

Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement, 16,* 237–248.

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

Ramsay, J. O. (1997). A functional approach to modeling test data. In W. J. van der Linden & R. K. Hambleton (Eds), *Handbook of modern item response theory* (pp. 381–394). New York: Springer.

Ramsay, J. O. (2000). *TESTGRAF98: A program for the graphical analysis of multiple choice test and questionnaire data* [Computer program]. Available from http://www.psych.mcgill.ca/faculty/ramsay/ramsay.html.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago: University of Chicago Press.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds), *Handbook of modern item response theory* (pp. 271–286). New York: Springer.

Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristics of discrete item responses. *Psychometrika, 63,* 111–130.

Santor, D. A., Zuroff, D. C., Ramsay, J. O., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment, 7,* 131–139.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben [The learning and solving of complex reasoning items]. *Zeitschrift für experimentelle und angewandte Psychologie, 3,* 476–506.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Sijtsma, K. (1998) Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3–31.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63,* 183–200.

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

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

Sijtsma, K., & Van der Ark, L. A. (2001). Progress in NIRT analysis of polytomous item scores: Dilemmas and practical solutions. In A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 297–318). New York: Springer-Verlag.

Sijtsma, K., & Verweij, A. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement, 23,* 55–68.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic for IRT models. *Journal of Educational Measurement, 37,* 58–75.

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 unidimen-

sionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Stout, W. F. (2001). Nonparametric item response theory: A maturing and applicable measurement modeling approach. *Applied Psychological Measurement, 25,* 300–306.

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

Stout, W. F., Nandakumar, R., & Habing, B. (1996). Analysis of latent dimensionality of dichotomously and polytomously scored test data. *Behaviormetrika, 23,* 37–65.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48,* 191–199.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale NJ: Erlbaum.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25,* 273–282.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory.* Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.

Vermunt, J. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25,* 283–294.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Yuan, A., & Clarke, B. (2001). Manifest characterization and testing of certain latent traits. *Annals of Statistics, 29(3).*

Zhang, J., & Stout, W. F. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129–152.

Acknowledgments

Most of the papers in this Special Issue, including the discussion papers and this introduction, were presented at a symposium on NIRT held at the July 1999 European meeting of the Psychometric Society in Lüneburg, Germany.

Authors' Addresses

Send requests for reprints or further information to Brian W. Junker, Department of Statistics, 232 Baker Hall, Carnegie Mellon University, Pittsburgh PA 15213, U.S.A.; or Klaas Sijtsma, Department of Methodology and Statistics FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: brian@stat.cmu.edu; k.sijtsma@kub.nl.