Nonparametric IRT in Action: An Overview of the Special Issue

Brian W. Junker
Carnegie Mellon University
Pittsburgh PA
Brian@stat.cmu.edu

Klaas Sijtsma
Tilburg University
Tilburg, NL
K.Sijtsma@kub.nl

December 22, 2000

Abstract

Nonparametric item response theory (NIRT) denotes several related methodological frameworks, for working with item response data without fully committing to any particular well-known family of parametric item response theory (PIRT) models. NIRT helps us to understand what is *common* among, and *essential* about, the many PIRT models that have been proposed or used in practice. NIRT also provides a flexible set of nonparametric data analysis tools to work with, when it is known (or suspected) that the usual PIRT models will not fit the data. Also, NIRT provides computationally modest methods to develop new questionnaires or measurement instruments, without the numbers of examinees and items that are typical of large-scale educational testing applications. NIRT methodology spans a broad range of interests in item response modeling, from unidimensional scaling of persons and items in clinical and developmental psychological settings, to establishing the dimensionality of the latent space, to the assessment of speededness and other effects in large scale testing, to appropriateness measurement and cognitive assessment. NIRT is not merely a theoretical exercise but rather, as illustrated by the many applications and examples throughout this special issue, an active approach to practical IRT and related scaling problems.

Key words: cognitive modeling, conditional covariance, dimensionality research, item ordering, nonparametric item response theory, person ordering, response functions.

1 Introduction

Nonparametric item response theory (NIRT) denotes several related methodological frameworks, for working with item response data without fully committing to any particular well-known family of parametric item response theory (PIRT) models. Although most applications of IRT and related methodologies involve model fitting within a single PIRT family, such as the Rasch (1960) or three-parameter logistic (Lord, 1980) models, the number of NIRT researchers and research publications has been growing in recent years. Given the wide variety and flexibility of PIRT models, what is attractive about the NIRT approach? We see three broad motivations for the development and continued interest in NIRT.

First, as a theoretical and foundational matter, researchers—and practicioners—may wish to understand what is *common* among, and *essential* about, the many parametric IRT and IRT like models that have been proposed or used in practice. To satisfy this desire, we might try to characterize modeling features, such as local independence, monotonicity of item response curves, or unidimensionality of the latent variable, that are common across many families of PIRT models, and perhaps ask what happens when data satisfy only weakened versions of these properties. We might also try to characterize the inferences that are or are not successful under these broad modeling features, as a way of understanding how IRT models aggregate information from data to tell us about latent constructs. Finally we might establish, from this research, certain "desiderata" that new families of NIRT and PIRT models should satisfy, to be broadly useful.

Second, as a practical matter, it is easy to see that any model we apply to data is likely wrong before we start. In the many situations in which a family of PIRT models has been shown (or is suspected) to fit poorly, we may wish to have a more flexible family of NIRT models, perhaps based on nonparametric smoothing ideas, to fall back on. Such NIRT families have been used, for example: to assess violations of local independence due to both trait related (e.g. latent variable multidimensionality) and testing context related (e.g. speededness and question wording) influences; to clarify questions about the sources and effects of differential item functioning; to provide a flexible context in which to develop methodology for establishing the most appropriate number of latent dimensions underlying a test; and to serve as alternatives in tests of fit for PIRT models.

Third, especially in psychological and sociological research, when it is necessary to develop a new questionnaire or measurement instrument, there often are not the numbers of examinees and items that have facilitated fitting complicated PIRT models in large-scale educational testing. Here NIRT can provide tools that are easy to use in small samples, to identify items that scale together well, in the sense that they follow a particular set of unidimensional NIRT assumptions, and to identify several subscales with simple structure among the scales, if the items do not form a single unidimensional scale.

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

To give a feel for one NIRT model, or class of models, and suggest ways in which this class may be further modified in NIRT research, let us consider dichotomous item response variables $X_1, \ldots, X_J$ for J test items, with values $x_j \in \{0, 1\}$. Much NIRT research takes as a starting point the basic assumptions

Local Independence (LI): There exists a (possibly multidimensional) latent variable θ, such that the joint conditional probability of the J item responses x collected in $\boldsymbol{X}$ can be written as

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

Monotonicity (M): The item response functions (IRF's) $P_j(\theta) = P(X_j = 1|\theta)$ are nondecreasing as a function of θ (or its coordinates, if θ is multidimensional); and

Unidimensionalty (UD): θ takes values in (a subset of) the real line.

The NIRT model (or class) satisfying only LI, M and UD, is known as the monotone homogeneity (MH) model (Mokken, 1971; Mokken & Lewis, 1982), and has also been studied under the name monotone unidimensional latent variable model (Holland & Rosenbaum, 1986; see also Meredith, 1965). This simple NIRT class includes, for example, the normal ogive models, the Rasch model and the two- and three-parameter logistic models. Defining the MH class in this way already begins to answer the foundational question raised in the first motivation for NIRT research give above: LI, M and UD are among the properties that are common among, and essential about, these well-known PIRT families. Of course, much more along these lines is possible (e.g., Hemker, Sijtsma, Molenaar & Junker, 1997; Junker & Ellis, 1997; Sijtsma & Hemker, 1998). Omnibus tests for MH and related models have recently been proposed by Bartolucci and Forcina (in press) and Yuan and Clarke (in press).

It is well known (e.g. Suppes & Zanotti, 1981, Holland & Rosenbaum, 1986; Stout, 1990; Junker, 1993) that no two of the three assumptions LI, M, and UD define a restrictive model for observable data. Although none of these three assumptions may be completely eliminated, they may all be weakened considerably, and pursuing inference about persons or items under weakened forms of these assumptions is also a longstanding interest of both PIRT and NIRT research. In NIRT research, work beginning with Stout (1987, 1990) focuses on weakening LI and M simultaneously while retaining enough structure to make ordinally consistent inferences about a dominant, unidimensional θ. Work beginning with Ramsay (1991) retains LI and UD but replaces assumption M with other smoothness assumptions to obtain nonparametric regression estimates of non-monotone IRF's. More recently, Zhang and Stout (1999) follow a long line of PIRT work (e.g. as reviewed by McDonald, 1997; and Reckase, 1997; see also Béguin & Glas, 1998, for more recent developments) in defining a compensatory multidimensional class of NIRT models retaining LI and M, in which various procedures for estimating the number of latent dimensions can be examined. Polytomous generalizations are also natural (e.g. Junker, 1991; Molenaar, 1997; Nandakumar, Yu, Li & Stout, 1998).

Broadly speaking, applied NIRT research centered in North America has been inspired by the second motivation for NIRT outlined above (providing more flexible data analysis and hypothesis testing tools, when PIRT methods fail), and applied NIRT work centered in Europe—especially the Netherlands and Germany—has taken inspiration from the third motivation (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 the properties of the items become more transparent as inessential features are stripped away. In the latter approach, the model is a criterion against which the items are evaluated. It is peculiar that this difference in focus is also present in the parametric IRT tradition, in which 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- and three-parameter logistic models (e.g., Bock & Aitkin, 1981; Lord, 1980; Hambleton, 1989) 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, especially as collaborations across the Atlantic have grown, and the differences outlined here are neither large nor fundamental in nature. 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), recently European researchers such as Vermunt (2001) and Bartolucci and Forcina (in press) have brought new modeling insights to these problems. On the other hand, while computationally modest methods for model-fit and scale construction based on probability inequalities derived under MH and related models have long been pursued by European researchers (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; Junker, 1993; Grayson, 1988; Huynh, 1994), and more computationally intensive scale construction methods has been a central theme in the work of Stout and his colleagues (e.g. Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996).

Five of the six main articles in this special issue of *Applied Psychological Measurement* are by authors who have received Ph.D.'s since 1995, or are otherwise new to NIRT research. All of the papers offer fresh perspectives on current issues in NIRT, and suggest—with theory, applications, and references to other literature—where NIRT is headed in the near future. The remainder of this introductory paper explains how these six contributions to this special issue fit into various areas of NIRT research. In Section 2 we discuss model-based exploratory data analysis and the analysis of item and test features using nonparametric regression methods (Habing, 2001; Douglas & Cohen, 2001); in Section 3 we turn to model-data fit and the explanation of observable data structure in terms of continuous (Bolt, 2001) and discrete (Junker & Sijtsma, 2001) latent structure. Finally, in Section 4 we consider the measurement of item and person properties using summary statistics justified by the MH model outlined above. Van der Ark (2001) surveys and extends the relevant NIRT modeling theory, and Vermunt (2001) suggests a general approach to testing various aspects of this theory.

Most of the papers in this special issue, including earlier versions of this introductory paper and the two discussion papers appearing at the end of the issue (Molenaar, 2001; Stout, 2001), were presented at a symposium on nonparametric IRT held at the European meeting of the Psychometric Society in Lüneburg, Germany, in July 1999. Since then, all of the authors and anonymous reviewers have made tremendous efforts to bring this special issue to press, for which we offer our sincerest thanks. We hope that the contributions in this issue shed sufficient light on the practical usefulness of nonparametric IRT, to broaden understanding and accessibility of the field, and to stimulate many fertile applications of nonparametric IRT in both psychological and educational testing in the future.

2 Exploratory Data Analysis, Item and Test Features

Two major themes in NIRT research involve nonparametric regression estimates of IRFs, and the estimation of conditional covariances between items, given a latent variable θ that may or may not be "complete" in the sense that LI holds. Both enterprises have provided a new repertoire of exploratory techniques for the researcher who suspects or believes that standard PIRT models will not fit well. PIRT models typically provide a kind of grid that we stretch over the data, to characterize general features of the item responses or to make predictions from them. Model parameters estimated from the data show how the grid bends to conform to the data, but the grid is only flexible in a limited number of ways. For example, commonly-used parametric IRT models impose a monotonicity assumption on IRF's so that dips or bumps in the IRF's cannot be seen; they will merely drive discrimination parameter estimates toward zero. Nonparametric methods provide a grid that is more flexible in more directions than the parametric grid, enabling us to assess the importance of these and other potential irregularities in the data-generating process.

Ramsay (1991, 1997, 2000) greatly popularized nonparametric estimation of item response functions, by proposing relatively easy-to-implement nonparametric regression methods; related work has also been pursued by Samejima (1998) and Levine (e.g. Drasgow, Levine, Tsien, Williams

& Mead, 1992). Ramsay's (2000) TestGraf 98 program provides a fairly straightforward way of using nonparametric regression as an exploratory tool for assessing monotonicity of the IRF for each item response variable X_{j_0}, using either the total score $X_+ = \sum_j X_j$ (Ramsay, 1991) or the rest-score $R_{j_0} = X_+ - X_{j_0}$ (Junker & Sijtsma, 2000) as a proxy for θ. This methodology could be used, for example, to explore deviations from parametric IRF form when nonparametric tests (e.g. Molenaar & Sijtsma, 2000; Stout, 1990) confirm the MH model but a specific parametric form such as the 2PLM is rejected. Ramsay (1991) applies the technique to identify possibly-defective test items from a large introductory psychology course; other applications include the work of Santor, Zuroff, Ramsay, Cervantes, and Palacios (1995), demonstrating only moderate discriminability in two widely-used self-report instruments used to screen for major depressive disorder.

Under LI, M and UD, the conditional covariances $\mathrm{Cov}\,(X_i, X_j|\theta)$ are all zero. However in practice, local independence never holds exactly. 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. Conditional covariances are at the heart of Stout's (1987, 1990) conception of essential independence—which allows these covariances to be positive or negative and to vary wildly from one item pair to the next, but still be controlled enough to allow for consistent ordinal measurement of subjects—and they play a role in all of the formal and informal tests and measures of unidimensionality developed by Stout's research group (e.g. Stout et al., 1996; Stout, Nandakumar & Habing, 1996). Estimating the entire conditional covariance curve, $\mathrm{Cov}\,(X_i, X_j|\theta)$, as a function of θ, can also be a useful exploratory device, since it can suggest explanations for multidimensionality by showing where along the θ axis local dependence occurs (Douglas, Kim, Habing & Gao, 1998).

Brian Habing's lead-off article in this special issue of *Applied Psychological Measurement* provides an accessible review, not only of nonparametric regression estimation of IRF's, but also of the application of this methodology to the estimation of the entire conditional covariance function. Estimates of conditional covariance using either total score or rest score as a proxy for θ are subject to biases that are difficult to avoid (e.g. Junker, 1993). After also briefly reviewing some basic bootstrap ideas, Habing (2001) demonstrates a clever application of the *parametric* bootstrap to *nonparametric* conditional covariance estimates, to reduce or eliminate these biases, and to provide confidence envelopes for the estimates. The method is illustrated with a refinement of Douglas et al.'s (1998) study of test speededness using conditional covariance functions.

The contribution of Jeff Douglas and Alan Cohen to this special issue continues the nonparametric regression and parametric bootstrap themes, comparing 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-chosen parametric approximation to the nonparametric IRF, to generate a reference distribution for testing the fit of the maximum likelihood *parametric* IRF's. Their bootstrapped hypothesis test may be less biased in favor of the PIRT model than other parametric bootstrap techniques (e.g. Stone, 2000; Gelman, Meng & Stern, 1996, and the accompanying discussion). This provides a powerful adjunct to graphical techniques, as Douglas and Cohen (2001) show using two simulated and two real educational testing examples.

3 Model-Data Fit and the Explanation of Data Structure

Parametric models typically specify whether one or more dimensions describe the data, and to some extent 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), but the greater flexibility of NIRT can facilitate the assessment of underlying trait dimensionality. One problem with studying dimensionality within a parametric family is that 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 (for a relatively recent example, see Miecskowski, Sweeney, Haas, Junker, Brown, & Mann, 1993). Dimensionality that is estimated apart from parametric features of the response model may be thought to be a more fundamental characteristic of the data, 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 program MSP (Molenaar & Sijtsma, 2000; earlier mainframe versions go back to Mokken 1971), for Stout's (1990) DIMTEST (also, see Nandakumar & Stout, 1993) and DETECT (Kim, Zhang, & Stout, 1995; Zhang & Stout, 1999) procedures, and for similar nonparametric measures of trait dimensionality.

Stout's conditional covariance based methods for assessing latent trait dimensionality (Stout et al., 1996, provide a recent summary) have recently been extended to the case of polytomous responses (Nandakumar, Yu, Li & Stout, 1998), and have been applied to a variety of data sources, including data from LSAT/LSAC (e.g., Stout et al., 1996). The ideas also figure into others' work on dimensionality assessment, including Oshima and Miller (1992) and Gessaroli and de Champlain (1996). Related considerations have spawned productive research in nonparametric detection of differential item and subtest functioning as well (Shealy & Stout, 1993; Bolt & Stout, 1996; Douglas, Stout, & DiBello, 1996; and Li & Stout, 1996).

For dichotomous items ($X_j = 0$ or 1 indicating incorrect or correct answer) a theory of scale construction—selecting groups of items that hang together well in the sense that the MH model is probably appropriate for them—has existed at least since Mokken (1971; see also Sijtsma, 1998). The principal tools of that theory are easy-to-compute adaptations of Loevinger's H coefficients, comparing the marginal covariance $\text{Cov}\,(X_i, X_j)$ of each item pair with the maximum covariance $\text{Cov}_{max}(X_i, X_j)$ possible, preserving the margins of the observed $X_i \times X_j$ table. The bound $\text{Cov}_{max}(X_i, X_j)$ is obtained by adjusting the table to remove Guttman errors (Molenaar, 1991); and indeed the original formulas for the H coefficients were expressed as ratios of Guttman errors (Mokken, 1997). These methods have also been extended to polytomous items, as illustrated by Hemker, Sijtsma and Molenaar (1995), and more recently by Sijtsma and Verweij (1999).

In this special issue, Dan Bolt discusses a geometric approach to identifying the continuous multidimensional latent structure underlying observable dichotomous item response data, based on theoretical work on Zhang and 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. The method, which Bolt (2001) compares with DIMTEST and related dimension-counting methods, seems to be remarkably successful at recovering a broad range of simulated and real-data multidimensional latent structures, including both "simple structure" in which the items can be partitioned into groups that are unidimensional with respect to different latent variables, as well as certain "fan" structures in which items load to varying degrees on several latent variables at once. Again, computational and graphical methods combine to give a complete data analysis.

More broadly within psychometrics, there is a growing interest in cognitive assessment models, that is, 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 (especially Embretson, 1991; Kelderman & Rijkes, 1994; Adams, Wilson, & Wang, 1997, in this context), and the multicomponent latent trait model (Embretson, 1985, 1997). Discrete latent structure approaches have also been proposed, including the constrained latent class approach of Haertel and Wiley (1993) and the general Bayesian inference network approach (e.g. Mislevy, 1996). Various attempts to blend discrete and continuous methodologies include Tatsuoka (see Tatsuoka, 1995, for a good overview) and Dibello, Stout and Roussos (1995). Related to this interest in cognitive modeling is person fit research or appropriateness measurement (e.g., Meijer, 1994; Sijtsma & Meijer, in press; Emons, Meijer, & Sijtsma, in press) where the main interest also is with understanding the psychological mechanisms that produced a particular pattern of item scores. Such psychological mechanisms may be test anxiety or lack of concentration. The purpose may be to remove respondents showing aberrant item score patterns from the item analysis or to use the information about aberrance for interpreting their latent trait estimates.

Brian Junker and Klaas Sijtsma contribute a thought-piece to this special issue, on the role of nonparametric IRT methodology in constructing and evaluating cognitive assessment models. They return to a dichotomized version of the "deductive strategy" transitive reasoning data of Sijtsma and Verweij (1999), and show, by estimating a discrete latent-structure version of Embretson's multicomponent model (which has much in common with the models of Haertel and Wiley, Tatsuoka, and Dibello et al.), that appropriate versions of monotonicity and local independence plausibly hold for this data. They go on to speculate about simple data summaries that are informative about latent attributes or cognitive components present or absent in individual students, based on each pattern of responses to the set of transitive reasoning items, and they discuss the translation of useful stochastic ordering properties from unidimensional NIRT research to their cognitive assessment models.

4 Measurement of Person and Item Properties

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

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

SOL implies that $E[\theta|X_+ = k]$ is also nondecreasing in k, so that on average increasing total score is associated with increasing latent variable level, as it should be. SOL holds for all dichotomous-response models satisfying the three assumptions of MH: LI, M and UD. This is quite a surprising property: even though only in the Rasch model is X_+ a sufficient statistic for θ, X_+ can be used for ordering θ in any MH model, no matter how far the deviation from the Rasch model. SOL is a useful measurement property for test practitioners who can confidently use the well known and simple X_+ for ordering examinees instead of the less well known and more complex θ.

Another surprising result (Hemker et al., 1997) is that SOL holds for almost *none* of the familiar ordered polytomous IRT models, parametric or nonparametric. If we take X_+ to be the Likert score, i.e. the unweighted sum across items of the item category scores, then a higher X_+ does not always imply a higher mean θ, for example. 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, the failure of the SOL property in polytomous IRT poses a serious potential problem for test practitioners who prefer a simple statistic like X_+ over the more complex θ. Nevertheless, preliminary simulation results (Sijtsma & Van der Ark, in press) suggest that, in practice, the mismatch of the ordering of X_+ and θ may not be very serious in data stemming from typical choices of item parameters and a normal θ distribution.

A measurement property which is important for ordering items is invariant item ordering (IIO). Whenever J IRF's do not intersect, they can be renumbered such that the following ordering is obtained:

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

In many testing situations, ordering the items by difficulty is helpful in analyzing the testing data; examples include intelligence testing, analysis of differential item functioning, person-fit analysis, and exploring hypotheses about the order in which cognitive operations are acquired by children. 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 item response functions do not cross. Research into methods for empirically investigating IIO for NIRT models for dichotomous item scores (Sijtsma & Junker, 1996) and into IRT models, both nonparametric and parametric, for polytomous item scores (Sijtsma & Hemker, 1998) provides yet another example of how NIRT research can provide results which are important for IRT in general. Sijtsma and Junker (1997) apply these methods to scale construction in developmental psychology.

Andries Van der Ark's (2001) contribution to this special issue of *Applied Psychological Measurement* provides an encyclopedic overview of most popular polytomous IRT models, parametric and nonparametric, and measurement properties, such as SOL and IIO, that are relevant for IRT models in general. He also addresses scoring rules for polytomous items (see Akkermans, 1998; Van Engelenburg, 1997) and guides the reader through the multitude of models, properties, and their combinations, and presents a reference card, which can be used for finding the appropriate polytomous IRT model when certain measurement properties are desired for the application envisaged. Main points are illustrated with data from five polytomous items measuring subjects' strategies for coping with industrial odors.

Jeroen Vermunt provides the sixth contribution to this special issue. He focuses on testing monotonicity and other ordering properties of the MH model, by fitting latent class models to the data that incorporate the relevant order restrictions. Latent class formulations for parametric and nonparametric IRT models are not new (Lindsay, Clogg & Grego, 1991; Croon, 1991; Hoijtink & Molenaar, 1997), but Vermunt's (2001) proposal in principle accomodates a wider range of NIRT (and PIRT) models and their specific properties than was possible before. Bringing one theme in this special issue full circle, Vermunt (2001) indicates parametric bootstrap based tests of fit for the the constrained latent class models he considers. His methodology is applied to compare the fits of several parametric and nonparametric IRT models for four polytomous self-report items taken from a biopsychosocial survey.

5 Closing Thoughts

This special issue closes with short discussions by two of the major forces in modern item response theory and psychometrics, Ivo Molenaar and Bill Stout. Both have devoted considerable energy to NIRT research, but they have also contributed to a variety of important advances in parametric IRT and related methods.

NIRT methodology spans a broad range of interests in item response modeling, from unidimensional scaling of persons and items in clinical and developmental psychological settings, to establishing latent space dimensionality, to the assessment of speededness and other effects in large scale testing, to appropriateness measurement and cognitive assessment. NIRT can be quite successful as the sole source of tools in test development and analysis, but it is also often a useful and insight-generating adjunct to standard parametric methods. NIRT is not merely a theoretical exercise but rather, as illustrated by the many applications and examples throughout this special issue, an active approach to IRT and related scaling problems.

References

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

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Akkermans, L. M. W. (1998). *Studies on statistical models for polytomously scored items.* Ph.D. dissertation. Enschede, Netherlands: Department of Education and Data Analysis, University of Twente.

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 Publications Inc.

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

Bartolucci, F., & Forcina, A. (in press). A likelihood ratio test for MTP_2 within binary variables. *Annals of Statistics*.

Béguin, A. A. & Glas, C. A. W. (1998). MCMC estimation of multidimensional IRT models. Research Report 98-14. Enschede, Netherlands: Department of Education and Data Analysis, University of Twente.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM 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, xx,* **xxx–xxx.**

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. *The 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. Chapter 15, pp. 361–389, in Nichols, P. D., Chipman, S. F. & Brennan, R. L. (eds). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Erlbaum.

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

Douglas, J. and Cohen, A. (2001). Nonparametric ICC estimation to assess fit of parametric models. *Applied Psychological Measurement, xx,* **xxx–xxx.**

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 IR model. *Psychometrika, 58,* 417–429.

Embretson, S. E. (1985). Multicomponent latent trait models for tests 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 and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer Verlag.

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. Chapter 8, pp. 131–156 in Fischer, G. H. and Molenaar, I. W. (1995). *Rasch models: foundations, recent developments, and applications.* New York: Springer-Verlag.

Gelman, A., Meng, X.-L. & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies (Disc: pp. 760–807). *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). Statistical tools for nonparametric local dependence assessment. *Applied Psychological Measurement, xx,* **xxx–xxx.**

Habing, B. and 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. and Wiley, D. E. (1993). Representations of ability structures: implications for testing. Chapter 14 in Fredriksen, N. and Mislevy, R. J. (eds.) (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. (1989). Principles and selected applications of item reponse theory. Chapter 4, pp. 201–220 in Linn, R. L. (ed.) (1989). *Educational measurement, Third Edition.* 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. and 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. (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 ssumptions, and connections with nonparametric IRT. *Applied Psychological Measurement, xx,* **xxx–xxx.**

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.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, Inc. Hillsdale, NJ.

McDonald, R. P. (1997). Normal-ogive multidimensional model. Chapter 15, pp. 257–269 in van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

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.* Amsterdam, NL: Vrije Universiteit Amsterdam.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100.

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. Chapter 20, pp. 351–368 in in W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern psychometrics* (pp. 369–380). New York: Springer Verlag.

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. Chapter 21, pp. 369–380 in W. J. Van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern psychometrics* (pp. 369–380). New York: Springer Verlag.

Molenaar, I. W. (2001). Discussion. *Applied Psychological Measurement, xx,* **xxx–xxx.**

Molenaar, I. W. and Sijtsma, K. (2000). *MSP for Windows.* Groningen: iecProGAMMA.

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.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* (Expanded edition, 1980.) University of Chicago Press. Chicago, Illinois.

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. Chapter Chapter 22, pp. 381–394 in van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Ramsay, J. O. (2000). TestGraf 98: A program for the graphical analysis of multiple choice test and questionnaire data. Available for download at `http://www.psych.mcgill.ca/faculty/ramsay.html`; program description obtained 30 October 2000.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. Chapter 16, pp. 271–286 in van der Linden, W. J. and Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

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. and 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 AngewandtePsy-*

chologie, 3, 456–506.

Shealy, R. & Stout, W. F. (1993). A Model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF 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,* 3–32.

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. (in press). The person response function as a tool in person-fit research. *Psychometrika.*

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

Stout, W. F. (2001). Discussion. *Applied Psychological Measurement, xx,* **xxx–xxx.**

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, and R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Lawrence Erlbaum Associates.

van der Ark, L. A. (2001). An overview of relationships in polytomous IRT and some applications. *Applied Psychological Measurement, xx,* **xxx–xxx.**

Van Engelenburg, G. (1997). *On psychometric models for polyto- mous items with ordered categories within the framework of item response theory.* Unpublished Ph.D. dissertation. Amsterdam, NL: University of Amsterdam.

Vermunt, J. (2001). On the use of restricted latent class models for defining and testing (non-)parametric IRT models. *Applied Psychological Measurement, xx,* **xxx–xxx.**

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

Yuan, A., & Clarke, B. (in press). Manifest characterization and testing of certain latent traits. *Annals of Statsitics*.

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