Nonparametric Item Response Theory in Action

Junker, B.W.; Sijtsma, K.

Published in:
Applied Psychological Measurement

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

Publication date:
2001

Link to publication

Citation for published version (APA):
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

The online version of this article can be found at: http://apm.sagepub.com

Published by:
SAGE Publications
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** (this article cites 52 articles hosted on the SAGE Journals Online and HighWire Press platforms):
http://apm.sagepub.com/cgi/content/refs/25/3/211
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:
1. **LI.** A (possibly multidimensional) latent variable \( \theta \) exists, such that the joint conditional probability of \( J \) item responses can be written as

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

(1)

2. **Monotonicity.** The IRFs \( P_j(\theta) = P(X_j = 1|\theta) \) are nondecreasing as a function of \( \theta \) (or its coordinates, if \( \theta \) is multidimensional).

3. **Unidimensionality.** \( \theta \) 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, Sijsma, Molenaar, & Junker, 1997; Junker & Ellis, 1997; Sijsma & 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 ordinal consistent inferences about a dominant, unidimensional \( \theta \). 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., Donald, 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 $\theta$ 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 $\theta$ 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 $\text{Cov}(X_j, X_j|\theta)$ 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., 1998).
Estimating $\text{Cov}(X_i, X_j | \theta)$ as a function of $\theta$ can be a useful exploratory device, because it can suggest explanations for multidimensionality by showing where along $\theta$ 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 $\theta$ 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., Mieczkowski 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), DMTEST (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 (Gesseroli & de Champlain, 1996; Oshima & Miller, 1992). Related considerations are also found in nonparametric detection of differential item 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, $\text{Cov}(X_i, X_j)$, of each item pair with the maximum possible covariance $[\text{Cov}_{\text{max}}(X_i, X_j)]$. This preserves the margins of the observed $X_i \times X_j$ table. The bound $\text{Cov}_{\text{max}}(X_i, X_j)$ 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 \( \theta \) 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 \( \theta, t \).

Then, the stochastic ordering of the latent variable (SOL) is
SOL implies that $E[\theta | X_+ = k]$ also is nondecreasing in $k$. On average, the increasing total score then is associated with increasing $\theta$ 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 $\theta$ only in the Rasch model, it can be used for ordering $\theta$ 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 $\theta$ 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 $\theta$. 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 $\theta$ 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 $\theta$. Nevertheless, preliminary simulation results (Sijtsma & Van der Ark, 2001) suggest that, in practice, the mismatch of the ordering of $X_+$ and $\theta$ might not be very serious in data stemming from typical choices of item parameters and a normal $\theta$ 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

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

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 has two discussions (Molenar, 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


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


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

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.