Nonparametric item response theory models

Sijtsma, K.

Published in:
Encyclopedia of Social Measurement

Publication date:
2005

Nonparametric Item Response Theory Models

Klaas Sijtsma
Tilburg University, Tilburg, The Netherlands

Glossary

invariant item ordering An ordering of items that is the same for each value of the latent trait scale.
monotone homogeneity model A benchmark model within nonparametric item response theory that assumes that all items in a test measure the same latent trait, that the relationship between the item score and the latent trait is monotone, and that the test procedure is free of influences on test performance other than the latent trait.
nonparametric item response theory (NIRT) A version of item response theory that assumes that the relationship between the item score and the latent trait is limited only by order restrictions but is otherwise free.
parametric item response theory (PIRT) A version of item response theory that assumes that the relationship between the item score and the latent trait is defined by a parametric function, such as the logistic or the normal ogive.
stochastic person ordering An ordering of individuals by means of a simple sum score that reflects in a probabilistic way the person ordering on the latent trait scale.

Nonparametric item response theory is a family of item response models for ordinal person and item measurement. The distinctive feature that makes an item response model nonparametric is that in a test either each item response function or the set of all item response functions is restricted by some monotonicity condition, without specifying a parametric family of monotone functions such as the logistic. Instead, item response functions are estimated from the test data and the hypothesized monotonicity condition is evaluated. Several models have been proposed and several methods and software packages are available to evaluate the fit of a model to the data. Nonparametric models are primarily data-oriented in that they study features of the data necessary to obtain ordinal scales for people and sometimes, items as well. Ordinal scales are useful in applications such as selecting the best applicants for a job or identifying the worst-performing students for remedial teaching.

Measurement Using Nonparametric Item Response Theory Models

Tests consist of well-chosen collections of items—exercises, tasks, questions, and statements—that are used to measure different aspects of a hypothetical construct, for example, arithmetic ability, spatial orientation, knowledge of national history, and introversion. A hypothetical construct is a theoretical structure that explains the relationships among a particular set of behaviors. Nonparametric item response theory (NIRT) models are statistical methods that are used to analyze the item response data collected in a sample of individuals to find out whether the items can be considered to be indicators of the same hypothetical construct. If the answer is affirmative, NIRT models provide an ordinal scale for the theoretical construct of interest. This scale is called the latent trait scale. A rank ordering of individuals allows for statements such as "For this expensive follow-up course, we will admit only the 10 students having the highest arithmetic scores" and "For this job, we will hire the candidate with the highest scale score on general knowledge."

In addition to a person scale, a successful NIRT analysis also provides information on the quality of the individual items and the whole test as a measurement instrument for a particular theoretical construct. Information on individual items may reveal two things: (1) Whether the item sharply distinguishes people with relatively low latent trait values from others with relatively high latent trait values, and (2) whether the item

measures the same latent trait as the other items in the test. These issues relate to the well-known issues of reliability and validity, respectively. In the phase of test construction, an item that does not distinguish people well or clearly measures a latent trait other than the other items may be removed from the test and replaced by a better one, in order to improve the overall measurement quality of the test.

Assumptions of Nonparametric Item Response Theory Models

Common Assumptions

Methodologically, the latent trait represents an operationalization of the hypothetical construct by means of the collection of items in the test. Semantically, the term latent trait is used to summarize the psychological influences that drive the responses of individuals to each of the items in a test. NIRT assumes that during testing a person's latent trait value is not affected by practice effects, such as those due to learning and development, or flaws in item construction that produce structural dependencies between items. Also, items are related to the latent trait in a way that is specified by the particular NIRT model. This is the common context of most NIRT models. Some models alternatively assume a more complex underlying latent trait structure or formalize practice or training effects during testing that affect test performance. Although potentially interesting and important, these models are not the core of NIRT.

Let X_j be the random variable for the score on item j, and let this score be $x_j = 0, 1, \ldots, m$. The assumption of unidimensionality (UD) means that the relationships between J items in the test can be explained by one common latent trait, that is denoted θ. The assumption of local independence (LI) means that given the latent trait value the probability of a score x_j on item j, $P(X_j = x_j \mid \theta)$, is independent of the scores on the other $J-1$ items in the test. That is, given θ for a vector of J item score random variables, $\mathbf{X}$, and its realization, $\mathbf{x}$, we have:

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{j=1}^{J} P\big(X_j = x_j \mid \theta\big). \qquad (1)$$

An implication of LI is that for any pair of items, say j and k, their conditional covariance equals 0; that is, $\mathrm{Cov}(X_j, X_k \mid \theta) = 0$.

The third assumption defines the relationship between the item score, X_j, and the latent trait, θ, known as the response function; this is the conditional probability, $P(X_j = x_j \mid \theta)$. NIRT models typically specify order restrictions on this relationship but no other restrictions. For simplicity, we assume that items are scored dichotomously, with $X_j = 0, 1$. These scores may, for example, indicate that the answer was incorrect (score 0) or correct (score 1). Later on, we return to the general case of $m+1$ ordered item scores. The probability $P_j(\theta) = P(X_j = 1 \mid \theta)$ is known as the item response function (IRF). A simple assumption that specifies an order relation on the IRF assumes that it is a monotone nondecreasing function. That is, for item j and two fixed arbitrary values of θ, denoted θ_a and θ_b:

$$P_j(\theta_a) \le P_j(\theta_b) \quad \text{whenever } \theta_a < \theta_b, \text{and for all } j. \qquad (2)$$

This is the monotonicity (M) assumption. Examples of IRFs that satisfy the assumption M are given in Fig. 1A. The assumptions of UD, LI, and M together define the NIRT model, known as the monotone homogeneity model and introduced in 1971 by Mokken. This model can be seen as a benchmark within NIRT.

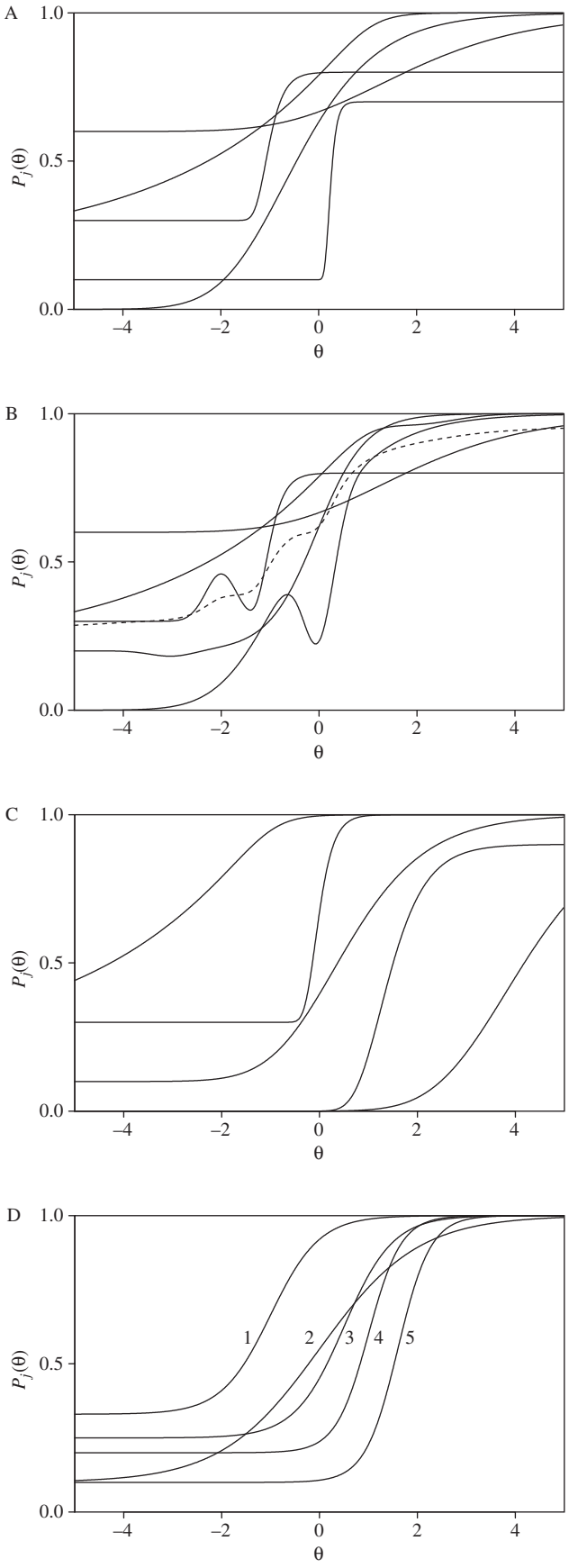
Typical of NIRT is the research into relaxations of the assumptions of UD, LI, and, M that seek to restrict the data as little as possible while maintaining important measurement properties such as the ordinal scale for individuals. For example, in 1990 Stout introduced the relaxation of strict unidimensionality to essential unidimensionality, which assumes one dominant latent trait and several nuisance traits whose influence on all statistical properties of the test vanishes for large J (in fact, $J \to \infty$). Another relaxation is that of LI to essential independence, which allows some conditional interitem covariances to be positive or negative while in the long run ($J \to \infty$) the mean across all absolute item pair covariances equals 0. A third example is weak monotonicity, which says that the mean of J IRFs is an increasing function in θ. This mean represents the average response to the test and is known as the test response function. Weak monotonicity does not restrict the individual IRFs as long as their mean is increasing; this means that assumption M is dropped at the individual item level. See Fig. 1B for an example of weak monotonicity. Junker showed in 1993 that none of the three assumptions, UD, LI, and M, can be dropped entirely and still leave enough structure in the data for ordering individuals consistently on a dominant latent trait.

Additional Assumptions

Whereas this and other work are aimed at ordinal person measurement under weak (or the weakest possible) assumptions, models that have more restrictions have been defined for studying item properties. For example, it may be assumed that the J IRFs do not intersect; that is, if for some θ_0 we know for items j and k that $P_j(\theta_0) < P_k(\theta)$, then

$$P_j(\theta) \le P_k(\theta) \quad \text{for all } \theta \quad \text{and} \quad \text{all } j, k; j \ne k. \qquad (3)$$

This is the assumption of invariant item ordering (IIO), which says that the J items have the same ordering by



response probability for all θs, with the possible exception of some θs for which ties may exist; see Fig. 1C for examples of nonintersecting IRFs that also satisfy assumption M. An interesting NIRT model that is defined by UD, LI, M, and IIO is the double monotonicity model, which was introduced by Mokken in 1971.

Models that have even more restrictions have been proposed by Sijtsma and Hemker in 1998 for items that have polytomous scoring ($m \geq 2$), and that imply that for J items an ordering,

$$E(X_1 \mid \theta) \leq E(X_2 \mid \theta) \leq \cdots \leq E(X_J \mid \theta) \qquad \text{for all } \theta \quad (4)$$

exists after the appropriate renumbering of the items. For polytomous items, this ordering of expected conditional item scores defines the concept IIO. Because for dichotomous items $E(X_j \mid \theta) = P_j(\theta)$, the item ordering for polytomous items captures IIO for dichotomous items as a special case.

Finally, local homogeneity is an example of an assumption that is needed when it is assumed that an NIRT model holds in each subgroup from the population of interest. Ellis and Van den Wollenberg showed in 1993 that it is necessary to distinguish between latent trait values θ and individuals when it is assumed that the response probability $P_j(\theta)$ is a within-person expectation of a propensity distribution of the item score for that person. Then different individuals with the same θ value must have the same $P_j(\theta)$ and no other person differences can influence this probability. This is important, for example, in differential item functioning research.

Parametric Item Response Theory Models

Parametric item response theory (PIRT) models are different from NIRT models in that they assume a specific parametric IRF, such as a normal ogive or a logistic function. An example is the three-parameter logistic model, defined as:

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \qquad (5)$$

Figure 1 (A) Five IRFs satisfying assumption M. (B) Two IRFs (solid curves) that satisfy assumption M, three IRFs that are not monotone increasing, and the test response function that is monotone increasing (dashed curve). (C) Five nonintersecting IRFs that satisfy assumption M. (D) Five IRFs under the three-parameter logistic model (parameter values: $\gamma_1 = 0.33$, $\gamma_2 = 0.10$, $\gamma_3 = 0.25$, $\gamma_4 = 0.20$, $\gamma_5 = 0.10$; $\delta_1 = -0.10$, $\delta_2 = 0.00$, $\delta_3 = 0.50$, $\delta_4 = 1.00$, $\delta_5 = 1.60$; $\alpha_1 = 2.00$, $\alpha_2 = 1.00$, $\alpha_3 = 2.00$, $\alpha_4 = 3.00$, $\alpha_5 = 3.00$).

Here, γ_j is the lower asymptote for $\theta \to -\infty$, α_j is a slope parameter, and δ_j is a location parameter; see Fig. 1D for examples of three-parameter logistic IRFs. For a data matrix produced by N individuals who responded to J items, the likelihood may be solved for each of these item parameters and the latent trait parameter. The resulting estimates give information on the probability (γ_j) that someone with a low scale value correctly solves item j (or gives an affirmative response); the item's location on the scale (δ_j) sometimes called its difficulty; the item's potential to distinguish between people with low and high scale values (α_j, called the discrimination power) to the left and the right of location δ_j; and the individuals' scale values (θ). Because NIRT models impose order restrictions on the IRFs but do not impose a logistic or another parametric restriction, they do not have likelihood functions from which these parameters can be solved. Nevertheless, they also provide information on the latent trait and the item parameters; however, they use other statistics and parameters, to be discussed shortly.

Measurement Properties of Nonparametric Item Response Theory Models

Person Measurement

The monotone homogeneity model is a measurement model for individuals. It implies an ordinal person scale in a stochastic ordering sense. Let test performance be summarized in a total score:

$$X_+ = \sum_{j=1}^{J} X_j \tag{6}$$

and let x_{+a} and x_{+b} be an arbitrarily chosen pair of realizations of X_+ such that $x_{+a} < x_{+b}$. Further, let t be an arbitrary value of θ. Then, the monotone homogeneity model for dichotomous items implies that:

$$P(\theta \geqslant | X_+ = x_{+a}) \leq P(\theta t | X_+ = x_{+b}). \tag{7}$$

An implication of this stochastic ordering property is $E(\theta | X_+ = x_{+a}) \leq E(\theta | X_+ = x_{+b})$. Thus, the higher the total score X_+, the higher the θ value. In an NIRT context, Grayson showed in 1988 that the observable total score X_+ replaces latent trait θ for ordinally measuring individuals.

For polytomous items, Hemker, Sijtsma, Molenaar, and Junker showed in 1997 that the monotone homogeneity model, defined by UD, LI, and a monotonicity assumption on response function $P(X_j \geq x_j | \theta)$ for $x_j = 1, \ldots, m$ (for $x_j = 0$, this probability trivially equals 1) does not imply the stochastic ordering property. Van der Ark showed in 2002 that for most realistic tests and distributions of θ an ordering of individuals on X_+ reflects an ordering on θ, but sometimes with reversals for adjacent X_+ values. Reversals of scores this close are unimportant for the practical use of tests because they represent only small differences with respect to the latent trait. For the polytomous NIRT model based on UD, LI, and M, and for NIRT approaches to polytomous items (and dichotomous items as a special case) based on weaker assumptions, X_+ is a consistent estimator of θ. This means that for infinitely many polytomous items the ordering of individuals using X_+ gives the exact ordering on θ, and Junker showed in 1991 that this result is true for several versions of polytomous NIRT models.

Item Measurement

The double monotonicity model is a measurement model for both individuals and items. Because it is a special case of the monotone homogeneity model, it has the same stochastic ordering and consistency properties for person measurement as that model. In addition, it implies an IIO, discussed, for example, by Sijtsma and Molenaar in 2002; see Eq. (4). An IIO implies that for expected item scores in the whole group:

$$E(X_1) \leq E(X_2) \leq \cdots \leq E(X_J). \tag{8}$$

These expectations can be estimated from sample data using the mean item score:

$$\bar{X}_j = J^{-1} \sum_{i=1}^{N} X_{ij}, \qquad j = 1, \ldots, J. \tag{9}$$

If the double monotonicity model fits the data, these sample means are then used to estimate the ordering of the expected conditional item scores, $E(X_j | \theta)$, $j = 1, \ldots, J$.

The IIO property can be used in several kinds of test applications in which it is important that individuals or subgroups have the same item ordering. For example, person-fit analysis and differential item functioning analysis of real data are better understood if an IIO can be hypothesized for the population, and results that deviate at the level of individuals or subgroups can be interpreted relative to this ordering. Also, in other research an IIO can be the null hypothesis when it is assumed that, for example, the items reflect ascending developmental stages and the equality of this ordering can be tested between age groups. Intelligence tests such as the Wechsler Intelligence Scale for Children use starting and stopping rules that assume an IIO—children of a particular age group start at an item that is easy for them (assuming that the previous items are of trivial difficulty), then are administered items in ascending difficulty ordering, and stop when they fail on several consecutive items (assuming that they would fail also on the next items that are even more difficult).

Fitting Nonparametric Item Response Theory Models to Test Data

If the monotone homogeneity model or a more relaxed version of it fits the test data, an ordinal person scale based on X_+ ordering is implied. If the double monotonicity model fits, not only is an ordinal person scale implied but also an IIO. The question is how to investigate the fit of these models to the test data. NIRT implies two properties of observable variables that are the basis of a variety of methods for investigating model-data fit.

Conditional Association

The first property, introduced by Holland and Rosenbaum in 1986, is conditional association. In principle, if a subgroup of individuals is selected from the population of interest on the basis of their performance on a subset of the items from the test, then within this subgroup the covariance between two sum scores based on another item subset must be nonnegative. The item scores may be dichotomous, polytomous, or continuous. A simple example to start with is that all individuals are selected, thus ignoring a subgroup structure altogether. Then all covariances between two items j and k must be nonnegative [$\mathrm{Cov}(X_j, X_k) \geq 0$] and negative covariances give evidence of model-data misfit.

A procedure for selecting one or more unidimensional item subsets from a larger item pool uses item scalability coefficients, denoted H_j, based on this nonnegative covariance property. The outcome of this procedure (implemented in the computer program MSP5 introduced in 2000 by Molenaar and Sijtsma) is one or more subsets of items that each measure another latent trait with items that have discrimination power with a lower bound defined by the researcher. Discrimination power is expressed for items by the scalability coefficients H_j and for the total test by the scalability coefficient H.

A more complex example is the following. $J-2$ items, not including items j and k, are used to define a sum score $R_{(-j,-k)} = \sum_{h \neq j,k} X_h$. Then within subgroups of individuals based on values r of $R_{(-j,-k)}$, conditional association means that $\mathrm{Cov}(X_j, X_k \mid R_{(-j,-k)} = r) \geq 0$, for all values r. This is the basis of other procedures (implemented in the computer programs DETECT and HCA/CCPROX, and discussed by Stout *et al*. in 1996) that try to find a subset structure for the whole test that approximates local independence as good as possible. The optimal solution best represents the latent trait structure of the test data.

Manifest Monotonicity

The second property, introduced in 1993 by Junker, is manifest monotonicity. This property can be used to investigate whether an IRF, $P_j(\theta)$, is monotone nondecreasing, as assumption M requires. To estimate the IRF for item j, first a sum score on $J-1$ items excluding item j, $R_{(-j)} = \sum_{k \neq j} X_k$, is used as an estimate of θ and then the conditional probability $P[X_j = 1 \mid R_{(-j)} = r]$ is calculated for all values r of $R_{(-j)}$. Given the monotone homogeneity model, the conditional probability $P[X_j = 1 \mid R_{(-j)}]$ must be nondecreasing in $R_{(-j)}$; this is manifest monotonicity. For real test data, manifest monotonicity is investigated for each item in the test, and violations are tested for significance; see Fig. 2A for an example of a discrete estimate of the IRF that violates assumption M. The program MSP5 can be used for estimating such discrete response functions and testing violations of monotonicity for significance. The program TestGraf98 made available by Ramsay in 2000, estimates continuous response functions using kernel smoothing and provides many graphics. The theory underlying this program was discussed by Ramsay in 1991.

Manifest monotonicity is also basic to the investigation whether the IRFs of different items are nonintersecting, as the assumption of IIO requires. To investigate IIO for the items j and k, the sign of the difference of the conditional probabilities $P[X_j = 1 \mid R_{(-j,-k)}]$ and

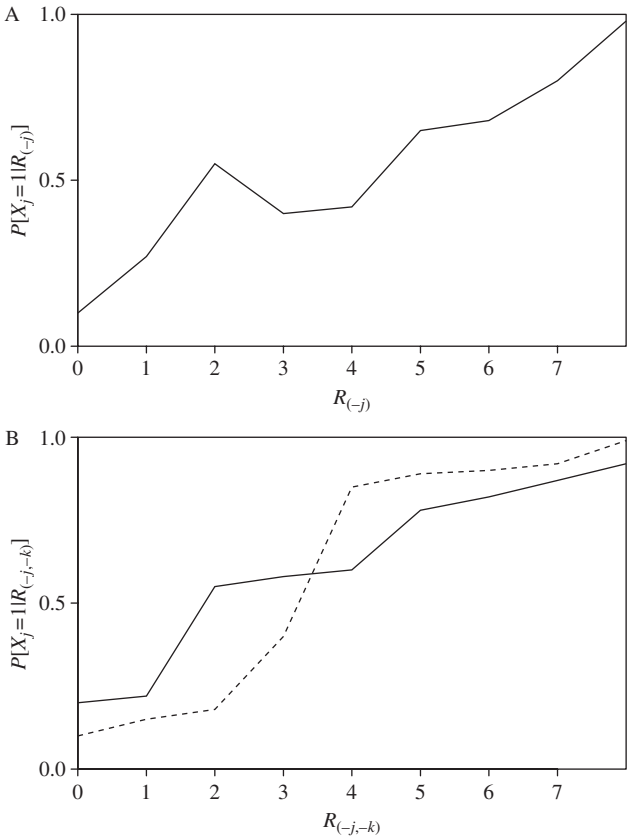


Figure 2 (A) Discrete estimate of an IRF that violates assumption M. (B) Two discrete estimates of IRFs that are intersecting.

$P[X_k = 1 | R_{(-j,-k)}]$ can be determined for each value r of the sum score $R_{(-j,-k)}$ and compared with the sign of the difference of the sample item means ($\bar{X}_j$ and $\bar{X}_k$) for the whole group. Opposite signs for some value r of $R_{(-j,-k)}$ indicate intersection of the IRFs and are tested against the null hypothesis that $P[X_j = 1 \,|\, R_{(-j,-k)} = r] = P[X_k = 1 \,|\, R_{(-j,-k)} = r]$ in the population (which means that these probabilities are the same but that their ordering is not opposite to the overall ordering). See Fig. 2B for an example of two observed IRFs that give evidence of intersecting IRFs in the population. This and other methods for investigating an IIO have been discussed and compared by Sijtsma and Molenaar in 2002. The program MSP5 can be used for investigating IIO.

For evaluating the monotonicity assumption of the response functions of a polytomous item, $P(X_j \geq x_j \,|\, \theta)$ for $x_j = 1, \ldots, m$, and the nonintersection of the response functions of different polytomous items, $P(X_j \geq x_j \,|\, \theta)$ and $P(X_k \geq x_k \,|\, \theta)$, several of these methods are also available in MSP5. Theoretical research to further support the underpinnings of these methods is in progress.

Nonparametric Item Response Theory Models for Nonmonotone Response Functions

For some latent traits and particular item types, a monotone nondecreasing response function cannot adequately describe the probability of an affirmative response. For example, as part of a questionnaire that measures the attitude toward the government's crime-fighting policy people may be asked to indicate whether they think that a 6-month prison term is an adequate sentence for burglary. Both opponents and proponents of long prison sentences may have a low probability of giving an affirmative response to this item, but for opposite reasons, and people having a moderate attitude may have higher probabilities. An NIRT model that successfully describes item scores produced this way thus has to assume that the relationship between the item score and the latent trait first increases and then decreases after a certain latent trait value or interval. In an NIRT context, such a response function could look like the irregular (solid) curve in Fig. 3. The NIRT monotonicity assumption now could be something like: First the IRF increases monotonically, then for some value θ_0 it reaches a maximum value $P_j(\theta_0)$ or a θ range in which it has a constant value, and then it decreases monotonically. Such an order restriction may be compared with a smooth parametric response function (dashed curve in Fig. 3) from a hypothetical parametric model, defined as:

$$P(X_j = 1 \,|\, \theta) = \frac{\lambda_j \exp[q(\theta - \delta_j)]}{1 + \exp[q(\theta - \delta_j)]} \qquad (10)$$

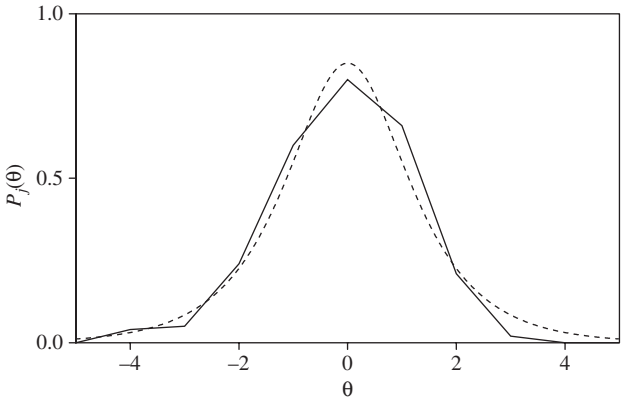


Figure 3 Irregular nonparametric IRF (solid curve) and smooth parametric IRF (dashed curve; $\lambda_j = 1.7$; $\delta_j = 0.0$) for preference data.

with $q = 1$ if $\theta - \delta_j < 0$, and otherwise $q = -1$; and $0 < \lambda_j < 2$. It may be noted that the nonparametric model encompasses the parametric model as a special case.

In 1992, Post studied the theoretical foundation of NIRT models for nonmonotone response functions and also derived methods for investigating model-data fit. Van Schuur proposed in 1984 a scaling procedure for selecting the best fitting items. Due both to their mathematical complexity and to the scarceness of real data that require nonmonotone response functions, NIRT models for nonmonotone response functions have not gained the popularity of the other NIRT models discussed here.

Practical Applications of Nonparametric Item Response Theory Models

NIRT models are particularly useful for the construction of ordinal scales for person and item measurement. They have proven their usefulness in many fields of applied research, such as psychology (nonverbal intelligence, induction reasoning, and tiredness from workload), sociology (attitudes toward abortion and machiavellism), political science (trustworthiness of inhabitants of foreign countries and political efficacy), marketing research (recency, frequency, and monetary value of purchase applied to market segmentation), and health-related quality-of-life research (quality of life for cancer patients and genital sensations and body image after surgery). Each of the scales for these properties allows the ordering of individuals and groups of individuals and, if the double monotonicity model fits, the ordering of items. What are the typical contributions of NIRT to the analysis of test data?

Item Quality

Compare the monotone homogeneity model with the three-parameter logistic model. An item analysis using the monotone homogeneity model investigates the IRFs by means of discrete estimates $P[X_j = 1 \mid R_{(-j)}]$ (in MSP5) or continuous estimates (in TestGraf98), and these estimates provide information on how IRFs may deviate from assumption M. For example, estimated curves may show zero or negative discrimination for parts of the distribution of θ (Fig. 2A) or even suggest that the IRF is bell-shaped (Fig. 3). When the item discriminates weakly for the lower and the middle part of the θ distribution and also has a low $P_j(\theta)$ for those θs, this may suggest that the item would be more appropriate in a test for a higher θ group. Bell-shaped IRFs suggest a nonmonotone NIRT model. An analysis using the three-parameter logistic model fits the S-shaped curve (Fig. 1D) to such items, which has the effect of driving the slope parameter α_j to 0 instead of giving diagnostic information about the misfit. Thus, instead of stretching a grid over the data that bends only in some directions but fails to detect other deviations, as is typical of a PIRT approach, an NIRT approach is more data-oriented in that it catches most of the peculiarities of the IRFs. When assumption M is satisfied, NIRT models use the item mean $E(X_j)$ and the item scalability coefficient H_j to replace location δ_j and slope α_j from PIRT, respectively. Items with empirical IRFs that do not have the typical logistic S-shape but that satisfy assumption M and have item scalability coefficients H_j that are reasonably high may be included in a test because they contribute to reliable person measurement.

Dimensionality

PIRT models usually fit a unidimensional model to the data, and multidimensional PIRT models may be used when the data are suspected to be driven by multiple latent traits. Fitting models yield useful parameter estimates that can be the basis for building item banks, equating scales, and adaptive testing. Within NIRT, algorithms have been produced that explore the data for the optimal dimensionality structure using assumption LI (e.g., program DETECT) or assumption M (program MSP5). Thus, NIRT explores the data more than PIRT, which is typically more oriented toward fitting an *a priori* chosen model. NIRT also imposes restrictions on the data, but compared to PIRT these restrictions are weaker, which renders NIRT more data-oriented and exploratory than PIRT.

NIRT models may be used because often little is known about the hypothetical construct, and a model that forces little structure onto the data, while maintaining ordinal measurement scales, may be a wise choice to start with. Given its exploratory orientation, NIRT could then be used as a precursor to PIRT by exploring the dimensionality structure of the data before a more restrictive hypothesis is tested by means of a PIRT model. Also, when starting with a PIRT model, instead, that does not fit the test data, an NIRT model can then be an alternative to fit to the data. The result of a fitting NIRT model is an ordinal scale for individuals (and items). Depending on the purpose of the test, such a scale is useful for selecting the best applicants for a job, the best students for a follow-up course, or the lowest-scoring pupils for remedial teaching or, in scientific research, for establishing relationships of the test score to other variables of interest.

See Also the Following Articles

Item Response Models for Nonmonotone Items • Item Response Theory

Further Reading

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

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

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.

Holland, P. W., and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Ann. Statist.* **14,** 1523–1543.

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. *Ann. Statist.* **21,** 1359–1378.

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis.* De Gruyter, Berlin.

Molenaar, I. W., and Sijtsma, K. (2000). *MSP5 for Windows: User's Manual.* iecProGAMMA, Groningen, The Netherlands.

Post, W. J. (1992). *Nonparametric Unfolding Models: A Latent Structure Approach.* DSWO Press, Leiden, The Netherlands.

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

Ramsay, J. O. (2000). *A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data.* Department of Psychology, McGill University, Montreal.

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

Sijtsma, K., and Molenaar, I. W. (2002). *Introduction to Nonparametric Item Response Theory.* Sage, Thousand Oaks, CA.

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

Stout, W. F., Habing, B., Douglas, J., Kim, H., Roussos, L., and Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Appl. Psychol. Meas.* **20,** 331–354.

Van der Ark, L. A. (2002). Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models. *Psychometrika* (in press).

Van Schuur, W. H. (1984). *Structure in Political Beliefs: A New Model for Stochastic Unfolding with Applications to European Party Activists.* CT Press, Amsterdam.