Nonparametric item response theory

Sijtsma, K.

Published in:
Encyclopedia of Statistics in Behavioral Science

Publication date:
2005

Nonparametric Item Response Theory Models

Goals of Nonparametric Item Response Theory

Nonparametric item response theory (NIRT) models are used for analyzing the data collected by means of tests and questionnaires consisting of J items. The goals are to construct a scale for the ordering of persons and – depending on the application of this scale – for the items. To attain these goals, test constructors and other researchers primarily want to know whether their items measure the same or different traits, and whether the items are of sufficient quality to distinguish people with relatively low and high standings on these traits. These questions relate to the classical issues of validity (*see* **Validity Theory and Applications**) and reliability, respectively. Other issues of interest are **differential item functioning**, person-fit analysis, and skill identification and cognitive modeling.

NIRT models are most often used in small-scale testing applications. Typical examples are intelligence and personality testing. Most intelligence tests and personality inventories are applied to individuals only once, each individual is administered the same test, and testing often but not necessarily is individual. Another example is the measurement of attitudes, typical of sociological or political science research. Attitude questionnaires typically consist of, say, 5

to 15 items, and the same questionnaire is administered to each respondent in the sample. NIRT is also applied in educational testing, preference measurement in marketing, and health-related quality-of-life measurement in a medical context. See [16] for a list of applications.

NIRT is interesting for at least two reasons. First, because it provides a less demanding framework for test and questionnaire data analysis than parametric item response theory, NIRT is more data-oriented, more exploratory and thus more flexible than parametric IRT; see [7]. Second, because it is based on weaker assumptions than parametric IRT, NIRT can be used as a framework for the theoretical exploration of the possibilities and the boundaries of IRT in general; see, for example, [4] and [6].

Assumptions of Nonparametric IRT Models

The assumptions typical of NIRT models, and often shared with parametric IRT models, are the following:

- *Unidimensionality (UD).* A unidimensional IRT model contains one **latent variable**, usually denoted by θ, that explains the variation between tested individuals. From a fitting unidimensional IRT model, it is inferred that performance on the test or questionnaire is driven by one ability, achievement, personality trait, or attitude. Multidimensional IRT models exist that assume several latent variables to account for the data.
- *Local independence (LI).* Let X_j be the random variable for the score on item $j (j = 1, \ldots, J)$; let x_j be a realization of X_j; and let $\mathbf{X}$ and $\mathbf{x}$ be the vectors containing J item score variables and J realizations, respectively. Also, let $P(X_j = x_j|\theta)$ be the conditional probability of a score of x_j on item j. Then, a latent variable θ, possibly multidimensional, exists such that the joint conditional probability of J item responses can be written as

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

  An implication of LI is that for any pair of items, say j and k, their conditional covariance equals 0; that is, $\mathrm{Cov}(X_j, X_k|\theta) = 0$.
- *Monotonicity (M).* For binary item scores, $X_j \in \{0, 1\}$, with score zero for an incorrect answer and score one for a correct answer, we define $P_j(\theta) \equiv P(X_j = 1|\theta)$. This is the item response function (IRF). Assumption M says that the IRF is monotone nondecreasing in θ. For ordered rating scale scores, $X_j \in \{0, \ldots, m\}$, a similar monotonicity assumption can be made with respect to response probability, $P(X_j \geq x_j|\theta)$.

Parametric IRT models typically restrict the IRF, $P_j(\theta)$, by means of a parametric function, such as the logistic. A well-known example is the three-parameter logistic IRF. Let γ_j denote the lower asymptote of the logistic IRF for item j, interpreted as the pseudochance probability; let δ_j denote the location of item j on the θ scale, interpreted as the difficulty; and let $\alpha_j (>0)$ correspond to the steepest slope of the logistic function, which is located at parameter δ_j and interpreted as the discrimination. Then the IRF of the three-parameter logistic model is

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \quad (2)$$

Many other parametric IRT models have been proposed; see [20] for an overview.

NIRT models only impose order restrictions on the IRF, but refrain from a parametric definition. Thus, assumption M may be the only restriction, so that for any two fixed values $\theta_a < \theta_b$, we have that

$$P_j(\theta_a) \leq P_j(\theta_b). \quad (3)$$

The NIRT model based on assumptions UD, LI, and M is the monotone homogeneity model [8]. Assumption M may be further relaxed by assuming that the mean of the J IRFs is increasing, but not each of the individual IRFs [17]. This mean is the test response function, denoted by $T(\theta)$ and defined as

$$T(\theta) = J^{-1}\sum_{j=1}^{J} P_j(\theta), \text{ increasing in } \theta. \quad (4)$$

Another relaxation of assumptions is that of strict unidimensionality, defined as assumption UD, to

essential unidimensionality [17]. Here, the idea is that one dominant trait drives test performance in particular, but that there are also nuisance traits active, whose influence is minor.

In general, one could say that NIRT strives for defining models that are based on relatively weak assumptions while maintaining desirable measurement properties. For example, it has been shown [2] that the assumptions of UD, LI, and M imply that the total score $X_+ = \sum_{j=1}^{J} X_j$ stochastically orders latent variable θ; that is, for two values of X_+, say $x_{+a} < x_{+b}$, and any value t of θ, assumptions UD, LI, and M imply that

$$P(\theta \geq t | X_+ = x_{+a}) \leq P(\theta \geq t | X_+ = x_{+b}). \quad (5)$$

Reference [3] calls (2) a stochastic ordering in the latent trait (SOL). SOL implies that for higher X_+ values the θs are expected to be higher on average. Thus, (5) defines an ordinal scale for person measurement: if the monotone homogeneity model fits the data, total score X_+ can be used for ordering persons with respect to latent variable θ, which by itself is not estimated.

The three-parameter logistic model is a special case of the monotone homogeneity model, because it has monotone increasing logistic IRFs and assumes UD and LI. Thus, SOL also holds for this model. The item parameters, γ, δ, and α, and the latent variable θ can be solved from the likelihood of this model. These estimates can be used to calibrate a metric scale that is convenient for equating, item banking, and adaptive testing in large-scale testing [20]. NIRT models are candidates for test construction, in particular, when an ordinal scale for respondents is sufficient for the application envisaged.

Another class of NIRT models is based on stronger assumptions. For example, to have an ordering of items which is the same for all values of θ, with the possible exception of ties for some θs, it is necessary to assume that the J items have IRFs that do not intersect. This is called an invariant item ordering (IIO, [15]). Formally, J items have an IIO, when they can be ordered and numbered such that

$$P_1(\theta) \leq P_2(\theta) \leq \cdots \leq P_J(\theta), \text{ for all } \theta. \quad (6)$$

A set of items that is characterized by an IIO facilitates the interpretation of results from differential item functioning and person-fit analysis, and provides the underpinnings of the use of individual starting and stopping rules in intelligence testing and the hypothesis testing of item orderings that reflect, for example, ordered developmental stages. The NIRT model based on the assumptions of UD, LI, M, and IIO is the double monotonicity model [8].

The generalization of the SOL and IIO properties from dichotomous-item IRT models to polytomous-item IRT models is not straightforward. Within the class of known polytomous IRT models, SOL can only be generalized to the parametric partial credit model [20] but not to any other model. Reference [19] demonstrated that although SOL is not implied by most models, it is a robust property for most tests in most populations, as simulated in a robustness study. For J polytomously scored items, an IIO is defined as

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_J|\theta), \text{ for all } \theta. \quad (7)$$

Thus, the ordering of the mean item scores is the same, except for possible ties, for each value of θ. IIO can only be generalized to the parametric rating scale model [20] and similarly restrictive IRT models. See [13] and [14] for nonparametric models that imply an IIO.

Because SOL and IIO are not straightforwardly generalized to polytomous-item IRT models, and because these models are relatively complicated, we restrict further attention mostly to dichotomous-item IRT models. More work on the foundations of IRT through studying NIRT has been done, for example, by [1, 3, 4, 6, and 17]. See [8] and [16] for monographs on NIRT.

Evaluating Model-data Fit

Several methods exist for investigating fit of NIRT models to test and questionnaire data. These methods are based mostly on one of two properties of observable variables implied by the NIRT models.

Conditional-association Based Methods

The first observable property is conditional association [4]. Split item score vector $\mathbf{X}$ into two disjoint

part vectors, $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Define f_1 and f_2 to be nondecreasing functions in the item scores from $\mathbf{Y}$, and g to be some function of the item scores in $\mathbf{Z}$. Then UD, LI, and M imply conditional association in terms of the covariance, denoted by Cov, as

$$\text{Cov}[f_1(\mathbf{Y}), f_2(\mathbf{Y})|g(\mathbf{Z}) = z] \geq 0, \quad \text{for all } z. \quad (8)$$

Two special cases of (8) constitute the basis of model-data fit methods:

Unconditional inter-item covariances. If function $g(\mathbf{Z})$ selects the whole group, and $f_1(\mathbf{Y}) = X_j$ and $f_2(\mathbf{Y}) = X_k$, then a special case of (8) is

$$\text{Cov}(X_j, X_k) \geq 0, \quad \text{all pairs } j, k; j < k. \quad (9)$$

Negative inter-item covariances give evidence of model-data misfit.

Let $\text{Cov}(X_j, X_k)\text{max}$ be the maximum **covariance** possible given the marginals of the cross table for the bivariate frequencies on these items. Reference [8] defined coefficient H_{jk} as

$$H_{jk} = \frac{\text{Cov}(X_j, X_k)}{\text{Cov}(X_j, X_k)\text{max}}. \quad (10)$$

Equation 9 implies that $0 \leq H_{jk} \leq 1$. Thus, positive values of H_{jk} found in real data support the monotone homogeneity model, while negative values reject the model. Coefficient H_{jk} has been generalized to (1) an item coefficient, H_j, which expresses the degree to which item j belongs with the other $J - 1$ in one scale; and (2) a scalability coefficient, H, which expresses the degree to which persons can be reliably ordered on the θ scale using total score X_+.

An item selection algorithm has been proposed [8, 16] and implemented in the computer program MSP5 [9], which selects items from a larger set into clusters that contain items having relatively high H_j values with respect to one another – say, $H_j \geq c$, often with $c > 0.3$ (c user-specified) – while unselected items have H_j values smaller than c. Because, for a set of J items, $H \geq \min(H_j)$ [8], item selection produces scales for which $H \geq c$. If $c \geq 0.3$, person ordering is at least weakly reliable [16]. Such scales can be used in practice for person measurement, while each scale identifies another latent variable.

Conditional inter-item covariances. First, define a total score – here, called a *rest score* and denoted R – based on $\mathbf{X}$ as,

$$R_{(-j,-k)} = \sum_{h \neq j,k} X_h. \quad (11)$$

Second, define function $g(\mathbf{Z}) = R_{(-j,-k)}$, and let $f_1(\mathbf{Y}) = X_j$ and $f_2(\mathbf{Y}) = X_k$. Equation 8 implies that,

$$\text{Cov}[X_j, X_k|R_{(-j,-k)} = r] \geq 0, \text{ all } j, k; j < k; \quad \text{all } r = 0, 1, \ldots, J - 2. \quad (12)$$

That is, in the subgroup of respondents that have the same rest score r, the covariance between items j and k must be nonnegative. Equation 12 is the basis of procedures that try to find an item subset structure for the whole test that approximates local independence as good as possible. The optimal solution best represents the latent variable structure of the test data. See the computer programs DETECT and HCA/CCPROX [18] for exploratory item selection, and DIMTEST [17] for confirmatory hypothesis testing with respect to test composition.

Manifest-monotonicity Based Methods

The second observable property is manifest monotonicity [6]. It can be used to investigate assumption M. To estimate the IRF for item j, $P_j(\theta)$, first a sum score on $J - 1$ items excluding item j,

$$R_{(-j)} = \sum_{k \neq j} X_k, \quad (13)$$

is used as an estimate of θ, and then the conditional probability $P[X_j = 1|R_{(-j)} = r]$ is calculated for all values r of $R_{(-j)}$. Given the assumptions of UD, LI, and M, the conditional probability $P[X_j = 1|R_{(-j)}]$ must be nondecreasing in $R_{(-j)}$; this is manifest monotonicity.

Investigating assumption M. The computer program MSP5 can be used for estimating probabilities, $P[X_j = 1|R_{(-j)}]$, plotting the discrete response functions for $R_{(-j)} = 0, \ldots, J - 1$, and testing violations of manifest monotonicity for significance. The program TestGraf98 [11, 12] estimates continuous response functions using kernel smoothing, and

provides many graphics. These response functions include, for example, the option response functions for each of the response options of a multiple-choice item.

Investigating assumption IIO. To investigate whether the items j and k have intersecting IRFs, the conditional probabilities $P[X_j = 1|R_{(-j,-k)}]$ and $P[X_k = 1|R_{(-j,-k)}]$ can be compared for each value $R_{(-j,-k)} = r$, and the sign of the difference can be compared with the sign of the difference of the sample item means, $\overline{X}_j$ and $\overline{X}_k$, for the whole group. Opposite signs for some r values indicate intersection of the IRFs and are tested against the null hypothesis that $P[X_j = 1|R_{(-j,-k)} = r] = P[X_k = 1|R_{(-j,-k)} = r]$ – meaning that the IRFs coincide locally – in the population. This method and other methods for investigating an IIO have been discussed and compared by [15] and [16]. MSP5 [9] can be used for investigating IIO.

Many of the methods mentioned have been generalized to polytomous items, but research in this area is still going on. Finally, we mention that methods for estimating the reliability of total score X_+ have been developed under the assumptions of UD, LI, M, and IIO, both for dichotomous and polytomous items [16].

Developments, Alternative Models

NIRT developed later than parametric IRT. It is an expanding area, both theoretically and practically. New developments are in adaptive testing, differential item functioning, person-fit analysis, and cognitive modeling. The analysis of the dimensionality of test and questionnaire data has received much attention, using procedures implemented in the programs DETECT, HCA/CCPROX, DIMTEST, and MSP5. Latent class analysis has been used to formulate NIRT models as discrete, ordered latent class models and to define fit statistics for these models. Modern estimation methods such as **Markov Chain Monte Carlo** have been used to estimate and fit NIRT models. Many other developments are ongoing.

The theory discussed so far was developed for analyzing data generated by means of monotone IRFs, that is, data that conform to the assumption that a higher θ value corresponds with a higher expected item score, both for dichotomous and polytomous items. Some item response data reflect a choice process governed by personal preferences for some but not all items or stimuli, and assumption M is not adequate. For example, a marketing researcher may present subjects with J brands of beer, and ask them to pick any number of brands that they prefer in terms of bitterness; or a political scientist may present a sample of voters with candidates for the presidency and ask them to order them with respect to perceived trustworthiness. The data resulting from such tasks require IRT models with single-peaked IRFs. The maximum of such an IRF identifies the item location or an interval on the scale – degree of bitterness or trustworthiness – at which the maximum probability of picking that stimulus is obtained. See [10] and [5] for the theoretical foundation of NIRT models for single-peaked IRFs and methods for investigating model-data fit.

References

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

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

[3] Hemker, B.T., Sijtsma, K., Molenaar, I.W. & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models, *Psychometrika* **62**, 331–347.

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

[5] Johnson, M.S. & Junker, B.W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models, *Journal of Educational and Behavioral Statistics* **28**, 195–230.

[6] Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models, *The Annals of Statistics* **21**, 1359–1378.

[7] Junker, B.W. & Sijtsma, K. (2001). Nonparametric item response theory in action: an overview of the special issue, *Applied Psychological Measurement* **25**, 211–220.

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

[9] Molenaar, I.W. & Sijtsma, K. (2000). *MSP5 for Windows. User's Manual*, iecProGAMMA, Groningen.

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

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

[12] Ramsay, J.O. (2000). *A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data*, Department of Psychology, McGill University, Montreal.
[13] Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP), *Psychometrika* **60**, 281–304.
[14] Sijtsma, K. & Hemker, B.T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models, *Psychometrika* **63**, 183–200.
[15] 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.
[16] Sijtsma, K. & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*, Sage Publications, Thousand Oaks.
[17] Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation, *Psychometrika* **55**, 293–325.
[18] Stout, W.F., Habing, B., Douglas, J., Kim, H., Roussos, L. & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment, *Applied Psychological Measurement* **20**, 331–354.
[19] Van der Ark, L.A. (2004). Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models, *Psychometrika* (in press).
[20] Van der Linden, W.J. & Hambleton, R.K., (eds) (1997). *Handbook of Modern Item Response Theory*, Springer, New York.

KLAAS SIJTSMA