# Developments in practical nonparametric IRT scale analysis

# Developments in Practical Nonparametric IRT Scale Analysis

Klaas Sijtsma
Department of Methodology and Statistics, FSW, Tilburg University, P. O. Box 90153, 5000 LE Tilburg, The Netherlands *k.sijtsma@uvt.nl*

**Summary** The present state of nonparametric item response theory is reviewed. First, assumptions and measurement properties implied by these assumptions are discussed. Then, model-data fit methods are briefly reviewed, and quality indices for items and scales are given. Finally, three computer programs are compared with respect to their practical possibilities in the analysis of test and questionnaire data.
**Keywords**: goodness-of-fit methods for NIRT, NIRT software, nonparametric item response theory (NIRT), scalability coefficients

## 1 What is Nonparametric Item Response Theory?

Item response theory (IRT) models are used to analyze the data of $N$ respondents on $J$ items. The result is scales for the measurement of persons and items. A useful division is into parametric and nonparametric IRT models. The former use a particular parametric formulation for describing the data structure whereas the latter use order restrictions for this purpose; see below for more details and Boomsma et al. (2001) for an overview. This paper discusses nonparametric IRT (NIRT) models, and their practical usefulness in constructing scales.

We use the following notation. Let $j$ be an item index, $j = 1, \ldots, J$ (test length); and let $i$ be a person index, $i = 1, \ldots, N$ (sample size). Let $X_j$ be the random variable for the ordered score on item $j$, with realization $x_j$; $x_j = 0, 1$ for dichotomous items, and $x_j = 0, \ldots, m$ for polytomous items with $m+1$ ordered answer categories typical of rating scales. For respondent $i$, the total score on a test is denoted $X_{i+}$ and item scores are denoted $X_{ij}$ to express their dependence on the respondent and the item. Test score $X_+$ is then defined as

$$X_{i+} = \sum_{j=1}^{J} X_{ij}.$$

Let $\theta$ be the latent trait; and let $P(X_j = 1|\theta)$ be the item response function (IRF) for dichotomous items, $P(X_j = x_j|\theta)$ the category response function (CRF) for polytomous items, and $P(X_j \geq x_j|\theta)$ the item step response function (ISRF) for polytomous items.

NIRT has three classes of assumptions:

1. *Dimensionality of measurement.* For multidimensional measurement, NIRT assumes a vector $\theta = (\theta_1, ..., \theta_T)$, and defines response probabilities as $P(X_j = 1|\theta)$, $P(X_j = x_j|\theta)$, and $P(X_j \geq x_j|\theta)$; see Holland and Rosenbaum (1986). For strictly unidimensional measurement, $\theta$ is a scalar. For essentially unidimensional measurement, $\theta$ contains one dominant latent trait and several nuisance traits, whose influence on the response probability vanishes as $J \to \infty$; see Stout (1990). In the IRT family (and in NIRT), unidimensional (UD) models are predominant over multidimensional models. This reflects the practice of test construction where the ideal is a unidimensional test and a single test score. In this paper we assume a unidimensional $\theta$.

2. *Relationship between items.* NIRT models assume that item scores are locally independent (LI). Let $\mathbf{X} = (X_1, \ldots, X_J)$, and $\mathbf{x} = (x_1, \ldots, x_J)$. Then LI means that

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

A consequence of LI is that for two items, $j$ and $k$, their conditional covariance, $Cov(X_j, X_k|\theta)$, equals 0.

3. *Relationship of item score and latent trait.* NIRT models are based on the general notion that the higher $\theta$, the higher the "success" probability on an item measuring $\theta$. This is reflected by the monotonicity (M) assumption. For dichotomous items, assumption M means that $P(X_j = 1|\theta)$ is nondecreasing in $\theta$, and for polytomous items, that $P(X_j \geq x_j|\theta)$ is nondecreasing in $\theta$. Nonparametric and parametric IRT models differ in the formalization of assumption M. In particular, a NIRT model imposes only order restrictions on the IRF or the ISRF; for example, for two arbitrary values $\theta_a$ and $\theta_b$,

$$P(X_j = 1|\theta_a) \leq P(X_j = 1|\theta_b); \text{ whenever } \theta_a < \theta_b.$$

Assumptions UD, LI, and M define the NIRT model of monotone homogeneity; see Sijtsma and Molenaar (2002). A parametric IRT model defines the IRF or ISRF by choosing a specific parametric function; for example, the 3-parameter logistic model defines the IRF as

$$P(X_j = 1|\theta) = \gamma_j + (1 - \gamma_j)\frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}.$$

Here, $\gamma_j$, $\alpha_j$, and $\delta_j$ are lower asymptote, slope, and location parameters of the IRF of item $j$, respectively.

## 2 Measurement Properties of Nonparametric IRT models

1. *Person ordering.* NIRT models for dichotomous items based on the assumptions of UD, LI, and M imply that the ordering on $X_+$ is stochastically the

same as the ordering on $\theta$. This is the stochastic ordering property (Hemker et al., 1997), formalized as

$$P(\theta > c|X_+ = s) \leq P(\theta > c|X_+ = t); \text{ for all } c; \text{ and } 0 \leq s < t \leq J.$$

This property guarantees ordinal person measurement. For NIRT models for polytomous items, Hemker et al. (1997) showed that the stochastic ordering property is not implied. Van der Ark (2001) used simulated polytomous data to demonstrate that in practice for most tests and latent trait distributions the stochastic ordering property held.

2. *Item ordering.* By adding the assumption that the IRFs do not intersect to the assumptions of UD, LI, and M, we obtain the double monotonicity model. Noting that for dichotomous items $E(X_j|\theta) = P(X_j = 1|\theta)$, for all $j$, and using the non-intersection of the $J$ IRFs, it is easily seen that items can be ordered and numbered accordingly, such that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \ldots \leq E(X_J|\theta); \text{ for all } \theta.$$

This is an invariant item ordering; see Sijtsma and Junker (1996). For polytomous item scores, an invariant item ordering is defined in exactly the same way (Sijtsma and Hemker, 1998) as for dichotomous items, with the expectation ranging from 0 to $m$.

## 3 Investigating the Fit of Nonparametric IRT Models to Test Data

Because NIRT models are based on relatively weak assumptions, in general, their fit to data is relatively good; see Sijtsma and Molenaar (2002). Also, because of their flexibility such models provide excellent opportunities for studying item properties. The outcomes give many indications of how to improve a test or questionnaire.

For example, IRFs and ISRFs are estimated using nonparametric regression methods; see Ramsay (1991), Junker and Sijtsma (2000), and Sijtsma and Molenaar (2002). This provides the researcher with accurate information about their shape as evidenced by the data. Using, in contrast, a logistic regression or normal-ogive function would impose an S-shape on the response curves, and a model test would indicate fit or misfit of this function to the data. For the case of misfit, nonparametric regression methods tell the researcher where and how the response curves deviate from the hypothesized shape (Douglas and Cohen, 2001). For example, the estimated curves may show zero or negative discrimination for parts of the distribution of $\theta$ or suggest a bell-shape. Based on this a researcher may decide to reject the misfitting item from the test or choose a more appropriate IRT model.

Also, NIRT has given inspiration to algorithms that investigate the dimensionality of the test; see Mokken and Lewis (1982) and Stout et al. (1996).

Rather than a priori modeling a unidimensional or a multidimensional structure, these algorithms explore the dimensional structure of the data and produce a division of the items into subsets that tap different latent traits. This approach can also be used to check whether a hypothesized item subset structure is supported by the data. These methods provide useful information about the data structure and may help the researcher to decide which items to remove or which subtests to pursue further.

Next, we review some methods for investigating the fit of NIRT models to the data. The general principle is that we study observable consequences of our assumptions of UD, LI, and M.

*Conditional association.* The first observable consequence is conditional association (CA; see Holland and Rosenbaum, 1986). Split vector $\mathbf{X}$ into two disjoint part vectors: $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Define $n_1$ and $n_2$ to be nondecreasing functions in the item scores from $\mathbf{Y}$, and $m$ to be some function of the item scores in $\mathbf{Z}$. Then UD, LI, and M imply CA,

$$Cov[n_1(\mathbf{Y}), n_2(\mathbf{Y})|m(\mathbf{Z}) = z] \geq 0, \text{ for all } z.$$

Examples of CA that clarify its meaning are the following.

1. Split $\mathbf{X}$ into three disjoint part vectors, $\mathbf{X} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Z})$. For example, $\mathbf{Y}_1$, $\mathbf{Y}_2$ may be testlets and $\mathbf{Z}$ may be another part of an exam. Let $Y_{1+}$, $Y_{2+}$, and $Z_+$ be the unweighted sum scores on the three item subsets, respectively. Let $Z_{+(\text{cutoff})}$ be a cutoff score on the items in $\mathbf{Z}$, and let $Z_{\text{cutoff}}$ be a binary pass(1)/fail(0) indicator. Then CA implies, for example,

$$Cov(Y_{1+}, Y_{2+}|Z_{\text{cutoff}} = z) \geq 0, \text{ with } z = 0 \text{ if } Z_+ < Z_{+(\text{cutoff})}; \text{ else } z = 1.$$

That is, in the subgroup of students that failed and the subgroup that passed the subexam $\mathbf{Z}$, the covariance between the two testlet scores must be nonnegative. Negative testlet covariance rejects the model of monotone homogeneity for the complete item set.

2. Ignoring a subgroup structure and studying covariances within the whole group, CA implies,

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

That is, in the whole group all $\frac{1}{2}J(J-1)$ inter-item covariances must be nonnegative. A negative covariance indicates misfit of the model of monotone homogeneity for at least one of the items involved.

3. Define a rest score based on $\mathbf{X}$ as,

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

Then CA implies that,

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

That is, in the subgroup of respondents that have the same restscore $r$, the covariance between items $j$ and $k$ must be nonnegative. This special case of CA proves to be useful for investigating LI, as we discuss next.

*Local independence.* The previous observable covariance property is used to study a nonobservable consequence of LI; that is, $Cov(X_j, X_k|\theta) = 0$. Because $\theta$ is latent, the rest score $R_{(-j,-k)}$ is used as a proxy for $\theta$. This is justified by a consistency result [for $J \rightarrow \infty$; see Stout (1990) for dichotomous items; and Junker (1991) for polytomous items] and the stochastic ordering property (holds for any $J$, but only for dichotomous items; see Hemker et al., 1997). Based on CA, we expect $Cov[X_j, X_k|R_{(-j,-k)}] \geq 0$. The reason is that, for finite $J$, the conditional distribution $f[\theta|R_{(-j,-k)}]$ has positive (non-zero) variance, which explains the positive covariance. It may be noted that for $J \rightarrow \infty$, this variance equals 0, due to the consistency of the restscore for the latent trait; thus, the conditional covariance then equals 0. Zhang and Stout (1999) showed that for finite $J$ and multidimensional $\boldsymbol{\theta}$, the covariance conditional on the restscore is either negative or positive, depending on the dimensionality structure of the item set. Based on this result, Stout et al. (1996) used a genetic algorithm to find the optimal partition of a multidimensional item set into unidimensional subsets.

*Monotonicity.* We define another restscore as

$$R_{(-j)} = \sum_{k \neq j} X_k.$$

Like $R_{(-j,-k)}$, restscore $R_{(-j)}$ estimates $\theta$, which is justified on the basis of the same stochastic ordering and consistency arguments. Then, $P[X_j = 1|R_{(-j)}]$ is used to estimate $P(X_j = 1|\theta)$. Junker (1993) showed that UD, LI, and M imply manifest monotonicity,

$$P[X_j = 1|R_{(-j)}] \text{ non-decreasing in } R_{(-j)}; R_{(-j)} = 0, \ldots, J-1.$$

The conditional probabilities $P[X_j = 1|R_{(-j)} = r]$, for all $r$, can be used to estimate the IRF by means of nonparametric regression. This yields a discrete estimate of the IRF (see Junker and Sijtsma, 2000; Sijtsma and Molenaar, 2002) or, using kernel smoothing methods, a continuous estimate (see Douglas and Cohen, 2001; Ramsay, 1991). Deviations from manifest monotonicity are in conflict with the model of monotone homogeneity.

*Intersection of IRFs.* Because this paper is limited to the model of monotone homogeneity, we refer to Sijtsma and Junker (1996) for a survey of methods for investigating intersections of IRFs for the purpose of establishing invariant item ordering. They discuss methods for assessing whether pairs of IRFs intersect and whether a set of $J$ IRFs contains intersections.

## 4 Quality of scales: Scalability coefficients

Because assumption M also allows IRFs to have flat or almost flat slopes, in practice a scale that satisfies the model of monotone homogeneity may not allow for accurate person ordering. Scalability coefficient $H$ (Sijtsma and Molenaar, 2002) is used to distinguish inaccurate ordinal scales from accurate ones. For two items $j$ and $k$, scalability coefficient $H_{jk}$ is defined as,

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

Here, $Cov(X_j, X_k)_{max}$ is the maximum covariance given fixed marginals of a $2 \times 2$ contingency table for the frequency counts of the scores on items $j$ and $k$. Also, an item scalability coefficient $H_j$ is defined that weights all $J-1$ $H_{jk}$s involving item $j$, and a total scale $H$ that is a weighted average of all $\frac{1}{2}J(J-1)$ coefficients $H_{jk}$.

How should we interpret the $H$ coefficient? First, the model of monotone homogeneity implies

$$0 \leq H_{jk}, H_j, H \leq 1.$$

Keeping the $\theta$ distribution and the IRFs except for IRF slopes constant, these coefficients increase as a function of the slope of the IRFs. It follows that the value of total $H$ gives an indication of the discrimination power of a test; that is, the degree of separation of respondents (or $\theta$s) by means of the items. Thus, $H$ is a measure of the accuracy of ordering respondents by means of total score $X_+$. For practical applications, $H = 0.3$ is considered to be a minimum requirement for accurately ordering persons. Total $H$ can be used to evaluate an item set as an a priori scale, and to select items from an experimental item set into (sub)scales. An item selection algorithm that does this is part of the computer program MSP5 (Molenaar and Sijtsma, 2000).

## 5 Computer programs, and what they do

Finally, we briefly compare the NIRT item analysis programs MSP5 (Molenaar and Sijtsma, 2000), DETECT (Stout et al., 1996), and TestGraf98 (Ramsay, 2000). MSP5 investigates the model of monotone homogeneity through the sign of inter-item covariances, and a discrete estimate of the IRF; DETECT checks LI through inter-item covariances conditional on the restscore; and TestGraf98 uses continuous IRF estimates (Table 1, first panel). Both MSP5 and TestGraf98 allow for investigating an a priori scale (second panel; MSP5: $H$ coefficient, reliability, information on assumption M; TestGraf98: test response function (TRF) and information function; see Ramsay, 2000). MSP5 and DETECT contain excellent algorithms for exploring the dimensionality of an item set, based on different scaling criteria; and

MSP5 also provides several statistics per scale found (second panel). MSP5 provides several methods for investigating invariant item ordering (third panel). Both MSP5 and TestGraf98 allow for scale analyses in meaningful subgroups (fourth panel). TestGraf98, in particular, provides means for investigating differential item functioning. All three programs are suited for handling polytomous items (fifth panel). TestGraf98 is predominantly a graphical item analysis program. MSP5 provides graphs for estimated IRFs/ISRFs and frequency distributions (sixth panel).

**Table 1.** Comparison of three programs for NIRT analysis

| | Programs | | |
|---|---|---|---|
| Aspects Model | MSP5 | DETECT | TestGraf98 |
| UD, LI, M | InterItemCov $\geq 0$ | – | – |
| LI | – | Dimensionality | – |
| M | Discrete IRF Est | – | Continuous IRF Est |
| A priori scale | H/Reliab; IRF info | – | TRF, InformFunc |
| Dimensionality | Good (M) | Very good (LI) | – |
| InfoPerScaleFound | H/Reliab; IRF info | – | – |
| InvarItemOrdering | ItemPair/Overall | – | – |
| Subgroups | Good | – | Good |
| DiffItemFunc | Modest | – | Good |
| Polytomous Items | + | + | + |
| Graphics | Few | None | Many |

The three programs are complementary more than competitive. MSP5 is well suited for model-data fitting. It uses many statistics and indices for this purpose, which provide the user with a wealth of diagnostic information. Moreover, it contains an automated item selection algorithm for selecting unidimensional scales. DETECT has been designed in particular for dimensionality investigation of an item set, and accomplishes this goal excellently. It is a good alternative to factor analysis, in particular when data are dichotomous. TestGraf98 provides an unprecedented tool for the graphical analysis of response functions. It allows for the visual diagnosis of IRFs, option response curves (for multiple choice items, with one correct and several incorrect answer options, which are nominal), and category response functions (for polytomous items). Each program is a valuable tool for NIRT data analysis. Together these three programs make possible the analysis of one's test data at almost any level of granularity.

## References

Boomsma A, Van Duijn MAJ, Snijders TAB (2001) Essays on Item Response Theory. Springer, New York

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

Hemker BT, Sijtsma K, Molenaar IW, Junker BW (1997) Stochastic ordering using the latent trait and the sum score in polytomous IRT models. Psychometrika 62:331–347

Holland PW, Rosenbaum PR (1986) Conditional association and unidimensionality in monotone latent variable models. The Annals of Statistics 14:1523–1543

Junker BW (1991) Essential independence and likelihood-based ability estimation for polytomous items. Psychometrika 56:255–278

Junker BW (1993) Conditional association, essential independence and monotone unidimensional item response models. The Annals of Statistics 21:1359–1378

Junker BW, Sijtsma K (2000) Latent and manifest monotonicity in item response models. Applied Psychological Measurement 24:65–81

Mokken RJ, Lewis C (1982) A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement 6:417–430

Molenaar IW, Sijtsma K (2000) MSP5 for Windows. User's manual. iecProGAMMA, Groningen, The Netherlands

Ramsay JO (1991) Kernel smoothing approaches to nonparametric item characteristic curve estimation. Psychometrika 56:611–630

Ramsay JO (2000) TestGraf. A program for the graphical analysis of multiple choice test and questionnaire data. Department of Psychology, McGill University, Montreal, Canada

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

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

Sijtsma K, Molenaar IW (2002) Introduction to nonparametric item response theory. Sage, Thousand Oaks CA

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

Stout WF, Habing B, Douglas J, Kim H, Roussos L, Zhang J (1996) Conditional covariance based nonparametric multidimensionality assessment. Applied Psychological Measurement 20:331–354

Van der Ark LA (2001) Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models. Tilburg University, The Netherlands

Zhang J, Stout WF (1999) The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika 64:213–249