Nonparametric item response theory and related topics
Sijtsma, K.; Meijer, R.R.

Published in:
Handbook of statistics 26

Publication date:
2007

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Nonparametric Item Response Theory and Special Topics

Klaas Sijtsma and Rob R. Meijer

Abstract
This chapter discusses the ideas and assumptions of nonparametric IRT and its merits and drawbacks in analyzing real test and questionnaire data. The place of nonparametric IRT in IRT in general is reviewed and nonparametric IRT is compared to parametric IRT. Procedures for assessing the dimensionality of item response data are explained and compared, and several methods are discussed that evaluate the shape of response functions of items, item steps and response options of items. It is discussed how to investigate whether item response functions are invariantly ordered across the latent variable, and why this is important in practical test usage. Latent class models are discussed as discrete alternatives for continuous latent variable IRT and as model fitting tools in a nonparametric IRT context. The strength of nonparametric IRT is illustrated with recent developments in person-fit analysis using an order-restricted person-response function together with graphical and statistical tools for drawing inferences from such functions about someone’s item-response data.

1. Place of nonparametric models in item response theory
Two ways of positioning nonparametric IRT (NIRT) relative to parametric IRT (PIRT) are that NIRT (1) encompasses PIRT by providing a broader theoretical framework from which IRT in general is better understood; and (2) primarily is an exploratory toolkit for the analysis of test and questionnaire data whereas PIRT primarily deals with confirmatory data analysis by fitting specific models to data. This chapter illuminates these two functions of NIRT, and discusses many recent developments. We first introduce some notation, definitions and assumptions.

Let the item score random variable of item \( j \) be denoted \( X_j \) with realization \( x_j \), \( j = 1, \ldots, J \). The item score can be dichotomous with binary scoring, usually for incorrect/correct answers, such that \( X_j \in \{ 0, 1 \} \), or ordered polytomous, such that \( X_j \in \{ 0, \ldots, m_j \} \), with \( m_j + 1 \) being the number of ordered answer categories in, for example, a rating scale. Throughout, we assume that \( m_j = m \); thus, each item in the test has the same number of ordered answer categories. Assume that responses to items are
driven by one latent variable, denoted $\theta$, or several latent variables, collected in vector $\theta = (\theta_1, \ldots, \theta_Q)$, with $q = 1, \ldots, Q$ indexing latent variables. For simplicity, assume one latent variable unless mentioned otherwise.

Both NIRT and PIRT – together known as item response theory (IRT) – model the probability of having a particular score on item $j$ as a function of $\theta$. For binary item scoring, IRT models probability $P(X_j = 1|\theta) = P_j(\theta)$, which is known as the item response function (IRF). For polytomous items, IRT models probability $P(X_j = x_j|\theta)$, known as the category response function (CRF). Other response functions may be modeled, such as $P(X_j \geq x_j|\theta)$. This function is known as the item step response function (ISRF). The ISRFs and the CRFs are related through

$$P(X_j \geq x_j|\theta) = \sum_{y=x_j}^{m} P(X_j = y|\theta).$$

Reversely, we have that

$$P(X_j = x_j|\theta) = P(X_j \geq x_j|\theta) - P(X_j \geq x_j + 1|\theta).$$

Other interesting definitions of response functions exist (e.g., Tutz, 1990; Van der Ark et al., 2002) but are not relevant for this chapter.

The reason for calling one class of IRT models nonparametric and another parametric, is that the former only puts order restrictions on response functions and the latter assumes a specific parametric function, such as the logistic or the normal ogive. For example, several NIRT models typically assume a positive monotone relationship between the binary item score and the latent variable; that is, for all $\theta$ pairs, assume that

$$P_j(\theta_a) \leq P_j(\theta_b), \text{ whenever } \theta_a < \theta_b, \text{ and } j = 1, \ldots, J;$$

whereas a PIRT model such as the 2-parameter logistic model, in addition to monotonicity requires this relationship to be logistic with a location parameter, $\delta_j$, and a slope parameter, $\alpha_j$, for each item, such that

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

Another difference, albeit gradual and not as a matter of principle, is that NIRT methods aim primarily at exploring the dimensionality of $\theta$ whereas PIRT models often posit a specific form of dimensionality and then test the fit of the model that hypothesizes this dimensionality to the data. These differences should give the reader some intuition why NIRT (1) is more general than PIRT and (2) as a data analysis tool primarily is exploratory and PIRT confirmatory.

1.1. Assumptions of NIRT and PIRT models

Three classes of assumptions play a central role both in NIRT and PIRT. IRT models are defined by a specific choice from one or more of these classes, for example, with respect to the number of latent variables or the kind of restriction on the response function.
1.1.1. Dimensionality of the data
A distinction can be made between one latent variable and multiple latent variables. Corresponding with this distinction are the assumptions of unidimensionality (UD) and multidimensionality, and strictly unidimensional IRT models and multidimensional IRT models.

1.1.2. Relationship between items
Most IRT models assume that the $J$-variate distribution of the item scores conditional on $\theta$ – with $\theta$ as a special case – is the product of the $J$ marginal conditional distributions of the separate item scores. This assumption is known as local independence (LI). For individual $v$, define vector $X_v = (X_{vj})$ with realization $x_v = (x_{vj})$, $j = 1, \ldots, J$; then, assumption LI is defined as

$$P(X_v = x_v | \theta) = \prod_{j=1}^{J} P(X_{vj} = x_{vj} | \theta). \tag{5}$$

An important implication of LI is that the conditional covariance between the scores on items $j$ and $k$ is 0; that is

$$\text{Cov}(X_j, X_k | \theta) = 0, \quad \text{for all pairs } j, k, \text{ and } j \neq k. \tag{6}$$

Reversely, this property, also known as weak LI, does not imply LI (Junker, 1993) because Eq. (6) only deals with bivariate relationships which may be 0 while multivariate relationships are not.

In practical NIRT data analysis, one tries to find the dimensionality of $\theta$ which is adequate for those data, such that LI is satisfied as good as possible. Stout (2002) noted that in real data weak LI is much easier to evaluate than LI, because with weak LI one only needs to evaluate conditional covariances while multivariate relationships between items can be ignored. For realistic $J$ the number of multivariate relationships beyond the $\frac{1}{2}J(J - 1)$ conditional covariances is huge and in real data many combinations of scores are not observed (Sijtsma and Junker, 1996). Stout (2002) concluded that, although mathematically not identical, in practice LI and weak LI are so much alike – in fact, it is difficult to imagine that Eq. (6) is satisfied but multivariate relationships deviate significantly from 0 – that the better accessible weak LI property may be used for dimensionality analysis in real data. This topic is discussed extensively in Section 2.1.2.

Assume that vectors with item scores of $N$ individuals are collected in a data matrix $X_{N \times J} = (X_1, \ldots, X_N)^T$, and that the respondents have been sampled i.i.d. (independently and identically distributed) from the cumulative multivariate distribution $G(\theta)$, so that

$$P(X_{N \times J} = x_{N \times J} | \theta) = \prod_{v=1}^{N} \prod_{j=1}^{J} P(X_{vj} = x_{vj} | \theta). \tag{7}$$

Notice that Eq. (7) is the general likelihood equation of any IRT model that assumes i.i.d. sampling and LI (Eq. (5)). By integrating over $\theta$, the multivariate distribution of
\(X_{N \times J}\) is obtained; that is, for \(G(\theta)\) we have that

\[
P(X_{N \times J} = x_{N \times J}) = \prod_{v=1}^{N} \left( \prod_{j=1}^{J} \int P(x_{vj} = x_{vj} | \theta) \, dG(\theta) \right).
\]

Modeling the item response data collected in a sample of \(N\) respondents involves restricting the response functions on the right-hand sides of Eqs. (5) and (8) and the latent variable distribution \(G(\theta)\) in Eq. (8).

1.1.3. Relationship of item score and latent trait

The most important distinction is between a nonparametric restriction on the response function as in Eq. (3) and a parametric definition as in Eq. (4). The model that is based on assumptions UD, LI and Eq. (3), which is also known as assumption M, is known as the monotone homogeneity model (Mokken, 1971; also, see Sijtsma and Molenaar, 2002). The model based on UD, LI and Eq. (4) is the 2-parameter logistic model (e.g., Baker and Kim, 2004). It may be noted that, compared with the IRFs of the 2-parameter logistic model the nonparametric IRFs (1) need not be S-shaped, and need not even be smooth; (2) can have several inflexion points, and may be constant over several ranges of \(\theta\); and (3) can have lower asymptotes or minimum values greater than 0 and upper asymptotes or maximum values smaller than 1.

Inserting in the likelihood (Eq. (7)) the conditional probabilities, \(P_j(\theta_v)\) from the 2-parameter logistic model (Eq. (4)) if \(x_{vj} = 1\) was observed in cell \((v, j)\) of \(X_{N \times J}\), and \(Q_j(\theta_v) = 1 - P_j(\theta_v)\) if \(x_{vj} = 0\) was observed; and solving for the item parameters, \(\alpha_j\) and \(\delta_j, j, \ldots, J\), and the latent variable values \(\theta_v, v = 1, \ldots, N\), amounts to joint maximum likelihood (ML) estimation of the model parameters. Insertion of these conditional probabilities in Eq. (8), choosing a distribution for \(G(\theta)\), such as the multivariate normal, integrating \(\theta\) out of the equation, and solving for the item parameters, is known as marginal maximum likelihood estimation. Inserting the marginal ML estimates of the item parameters thus obtained in the likelihood in Eq. (7) and solving for \(\theta\) yields ML estimates of the person parameters. Other estimation methods exist in addition to joint ML and marginal ML (e.g., Baker and Kim, 2004), but are not discussed here.

Restricting response probabilities in Eq. (8) by, for example, assumption M as reflected in Eq. (3) does not lead to ML estimates of latent variable and item parameters as in Eq. (4), simply because these parameters do not appear in the likelihood of the model. However, these order restrictions are sufficient to restrict the data, \(X_{N \times J}\), such that observable consequences can be derived that can be used to evaluate the fit of the model to the data. Moreover, together with UD and LI, assumption M can be shown to imply an ordinal scale for persons which can be estimated from the data and, if a restriction on the ordering of the IRFs is added as a fourth assumption, also implies a scale for items. Another approach to evaluating NIRT models comes from order-restricted latent class analysis (LCA) modeling. This approach replaces the continuous latent variable by an ordered, discrete latent variable, and restricts the ordering of response probabilities conditional on these latent classes. Bayesian methods (Hoijtink and Molenaar, 1997; Van Onna, 2002) and ML methods (Vermunt, 2001) are used to evaluate the fit of the model to the data, and allocate respondents to one of the ordered latent classes of the fitting model.
1.1.4. Other assumptions in NIRT

Eq. (8) defines a latent structure that underlies the multivariate distribution of the item-response data in $X_{N \times J}$, but Suppes and Zanotti (1981) showed that it is always possible, for finite $J$, to construct a unidimensional random variable, $\xi$, which is not necessarily latent, such that the item score variables in $X_{N \times J}$ satisfy LI (Eq. (5)). Thus, any data set can be "forced" to satisfy Eq. (8) meaning, reversely, that Eq. (8) does not restrict the data and in that sense does not define a practically useful IRT model. Other results exploring the boundaries of what is minimally required to have an identifiable IRT model are presented by Junker (1993) and Ellis and Junker (1997); also see Junker and Ellis (1997). Ignoring the many subtleties in their work, what matters here is to note that if one restricts the dimensionality of the data and assumes LI, an assumption like M (Eq. (3)) or a parametric restriction like Eq. (4) is needed to have an IRT model that restricts the data.

In general, it is true that PIRT models restrict the data in $X_{N \times J}$ more than NIRT models. For example, the IRF of the 2-parameter logistic model (Eq. (4)) is a special case of the order-restricted IRF of the monotone homogeneity model (Eq. (3)). Given that both models assume UD and LI, it follows that the 2-parameter logistic model is a special case of the monotone homogeneity model. Many comparable relationships can be found within pairs of PIRT and NIRT models, but within some pairs the hierarchy is not unequivocal (e.g., De Koning et al., 2002) or may even be reversed (e.g., Sijtsma and Hemker, 1998). Thus, while it is true in general that NIRT is less restrictive than PIRT it is not true that each NIRT model is less restrictive than each one of the PIRT models.

One of the themes of NIRT is in what degree assumptions may be relaxed such that particular measurement properties still are implied while no parametric restrictions are needed. Stout (1987, 1990) and Junker (1991, 1993) studied the class of essential unidimensional (EU) IRT models. EU models are based on the following assumptions:

- Essential unidimensionality (EU); one dominant latent variable, $\theta$, and $Q$ nuisance variables, collected in $\theta = (\theta, \theta_1, \ldots, \theta_Q)$, are assumed much as in factor analysis where a distinction is sometimes made between common factors and unique factors;
- Essential independence (EI); for infinitely many items the mean of the inter-item covariances conditional on $\theta$ equals 0; thus,

$$\frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} |\text{Cov}(X_j, X_k|\theta)| \to 0, \quad \text{for } J \to \infty. \quad (9)$$

- Weak monotonicity (WM); weak monotonicity replaces the stronger assumption M, by assuming that the mean of $J$ IRFs, also known as the test response function, is monotone in each of the $\theta_q$s from $\theta$. That is,

$$J^{-1} \sum_{j=1}^{J} P_j(\theta_a) \leq J^{-1} \sum_{j=1}^{J} P_j(\theta_b), \quad \text{whenever } \theta_a \leq \theta_b, \quad \text{coordinatewise}. \quad (10)$$
WM does not restrict the individual IRFs, as long as their mean is monotone nondecreasing. This means that assumption M is dropped as a requirement for individual IRFs but that it is active for the test response function under the name of WM.

Stout (1990, Theorem 3.2) showed that for EU models based on EI and WM, that total score

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

(11)

is a consistent ordinal estimator of the dominant $\theta$, and Junker (1991) generalized this result to polytomous EU models, based on EI and ISRFs defined as $P(X_j \geq x_j | \theta)$. The replacement of assumptions UD, LI, and M by assumptions EU, EI, and WM, respectively, implies a more general nonparametric model. As a result, properties implied by UD, LI, and M with respect to stochastically ordering persons on $\theta$ by means of $X_+$ and investigating assumption M in real data are lost (these properties are discussed in Sections 1.2.1 and 2.2.1, respectively). However, the more general EU model is more flexible than, for example, the monotone homogeneity model and thus accommodates more data sets while maintaining attractive measurement properties; see Stout (2002) for a summary.

1.2. Differences between NIRT and PIRT

In this section we further explore the two important characteristics of NIRT, which are that NIRT (1) encompasses PIRT by providing a broader theoretical framework from which IRT in general is better understood; and (2) primarily is an exploratory toolkit for the analysis of test and questionnaire data whereas PIRT primarily deals with confirmatory data analysis by fitting specific models to data. An example of the first characteristic is the scale of measurement in NIRT and an example of the second characteristic is the estimation of response functions and the way these estimates are used in practical data analysis.

1.2.1. Measurement of individuals

The metric of the $\theta$ scale of the 2-parameter logistic model can be derived as follows. Define the odds of a success on item $j$ as $\Omega_j = \frac{P_j(\theta)}{Q_j(\theta)}$, then $\ln \Omega_j = \alpha_j (\theta - \delta_j)$. For two respondents, $v$ and $w$, the log-odds ratio equals

$$\ln \frac{\Omega_{vj}}{\Omega_{wj}} = \alpha_j (\theta_v - \theta_w).$$

(12)

Because NIRT models typically do not parameterize the response function, a $\theta$ metric for differences as in Eq. (12) is not available. Alternatively, NIRT models for dichotomous items based on the assumptions of UD, LI and M imply an ordinal scale in the sense that latent variable $\theta$ is stochastically ordered by total score $X_+$; that is, for any value $t$ of $\theta$ and any pair of total scores, $x_{+a} < x_{+b}$, we have that

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

(13)
This is known as stochastic ordering of the latent variable by the observed sum score (SOL; Hemker et al., 1997; also see Grayson, 1988; Hemker et al., 1996; Huynh, 1994, for a discussion of the more general monotone likelihood ratio property). An implication of Eq. (13) is that

\[
E(\theta|X_{+a} = x_{+a}) \leq E(\theta|X_{+b} = x_{+b}).
\]  

(14)

Thus, in practice any IRT model based on UD, LI, and M implies the ordering of respondents on latent variable \( \theta \) by means of the ordering on the observable sum score \( X_+ \), except for random error. Note that this applies to all NIRT and PIRT models that assume monotonicity for the IRFs in addition to UD and LI. The parametric 2-parameter logistic model (based on UD, LI, and Eq. (4)) and the nonparametric monotone homogeneity model (based on UD, LI, and Eq. (3)) are examples. Other examples of parametric models based on UD and LI and with monotone IRFs are the much used Rasch model or 1-parameter logistic model, which defines IRFs as logistic with \( \alpha_j = 1 \) for all \( j \) (thus, in Eq. (4), \( \alpha_j = 1 \)); and the 3-parameter logistic model, which adds a lower asymptote to the 2-parameter logistic model, denoted \( \lambda_j \) (with \( 0 \leq \lambda_j \leq 1, \ j = 1, \ldots, J \)). Hemker et al. (1997) discussed failure of SOL in many polytomous-item IRT models, and Van der Ark (2005) provided many reassuring robustness results.

The importance of the SOL result is that for ordering individuals on the \( \theta \) scale it suffices to use an IRT model that assumes property M in addition to UD and LI, but not necessarily an increasing logistic or normal ogive IRF. This result has positive consequences, not only for NIRT models but also for PIRT models.

First note that the relatively weak assumptions of NIRT models imply an ordinal scale. Moreover, as a kind of bonus for not asking for an abundance of structure in the data, as a rule NIRT models allow relatively many items in a scale. Even though these items would allow for reliable measurement, several would not be tolerated by the 2-parameter logistic model because their IRFs are not logistic. Among these rejected items would be items with IRFs that are asymmetrical, constant over some ranges of \( \theta \) but steep over others, and have minimum values greater than 0 and maximum values smaller than 1 (see, e.g., Meijer and Baneke, 2004). However, such items contribute to reliable measurement provided that their IRFs are relatively steep locally on \( \theta \). It is important to note that such contributions do not depend on the IRF being logistic, normal ogive, or otherwise parametric, but on assumption M. Ramsay (2000) discusses a method for estimating the item and test information functions in an NIRT context, and Meijer et al. (1995) discuss methods for estimating the reliability of single items based on methods for estimating total score \( (X_+) \) reliability in an NIRT context (Sijtsma and Molenaar, 1987). Also note that among the items not well described by PIRT models may be items that add to the coverage of the variable represented by \( \theta \).

Second, when PIRT models fit to the data at the expense of perhaps some of the items that would be fitted by an NIRT model their numerical \( \theta \) scale is a highly convenient tool for applications such as equating and adaptive testing. However, the SOL property also has advantages for PIRT models because it justifies the use of the number-correct \( X_+ \) score for communicating test results to teachers, parents, students, and also researchers as an intuitively appealing alternative for the inaccessible \( \theta \) scale.
An interesting suggestion for an alternative scale that may replace the arbitrary and difficult to interpret \( \theta \) scale was done by Ramsay (1996); also see Rossi et al. (2002). Basically, they propose to use a scale based on response probabilities, \( P_j(\theta) \), because, unlike \( \theta \) or monotone transformations of \( \theta \), a probability like \( P_{12}(\theta_{Mary}) = 0.74 \) represents a property of Mary relative to Item 12 which is invariant with respect to any scale of measurement. Thus, it lacks the arbitrariness of other scales, like \( \theta \). A curve is proposed which results from plotting the \( J \) response probabilities against one another as a function of \( \theta \). Starting from the scale’s zero point, say, \( \theta_0 \), the position of person \( v \) on this curve can be determined by means of the arc length measure \( s(\theta_v) \) as

\[
s(\theta_v) = \int_{\theta_0}^{\theta_v} \left[ \sum_{j=1}^{J} \left( \frac{dP_j(\theta)}{d\theta} \right) \right]^2 \frac{d\theta}{2}
\]

For more details, see Ramsay (1996) and Rossi et al. (2002).

We conclude by noting that NIRT has considerably contributed to defining scales that are either easy to obtain and understand (\( X_+ \)) or firmly rooted in empirical data [\( s(\theta) \)], and circumvent the complexity and the arbitrariness of the \( \theta \) scale. Clearly, such scales are also advantageous to PIRT.

1.2.2. Model diagnostics

From the perspective of parsimony it is convenient to assume that IRFs are, for example, logistic, and that items can be described using no more than a few parameters. Many tests have been constructed and item banks assembled that generate data that may be approximated by means of unidimensional or multidimensional PIRT models, but in itself there is no compelling reason why real data must exhibit the often hypothesized parametric curvature (Meredith, 1965). Detailed data analysis often reveals IRFs that deviate from this regular shape, with deviations ranging from small and irrelevant oscillations to deep dips and decreases along parts of the \( \theta \) scale. It is interesting to know how PIRT and NIRT models react to such deviations from assumption M. See Meijer and Baneke (2004) for a confrontation of PIRT and NIRT models and an application to data from the Depression content scale of the Minnesota Multiphasic Personality Inventory-2 (Butcher et al., 1989).

Estimating a PIRT model is like stretching a grid over the data which is flexible only in some directions but not in others, and then summarizing the curvature and location of the grid as far as its flexibility allows it. This summary consists of one or a few parameters, which are valid for the data to the degree in which the hypothesized IRFs match the IRFs identified from the data. Obviously, peculiarities in the data may not be caught by logistic or other parametric curves but, if properly revealed, could help the researcher to make useful decisions about the items. For example, assume that the true IRF increases at the low and middle \( \theta \)s, but decreases at the higher \( \theta \)s. The logistic curve that is fitted to the data generated by means of this nonmonotone IRF accommodates this local decrease by a necessarily monotone curve which increases with a relatively flat slope (Junker and Sijtsma, 2001a). Thus, the item is identified as a weakly discriminating item, and may be a candidate for removal from the test.
An interesting question is, whether for given distribution \( G(\theta) \) different models for the IRFs can yield the same distribution of the data, \( P(X_{N \times J} = x_{N \times J}) \) (Eq. (8)). Douglas (2001) showed that there is only one correct IRT model, and that nonparametric methods can consistently estimate the IRFs. This means that if an IRF is found in data analysis of which the shape does not match that predicted by a particular PIRT model, that this PIRT model is incorrect as a model for the data. Based on this result, Douglas and Cohen (2001) used nonparametric estimation of IRFs to study the fit of PIRT models. This was done by finding, given the distribution of \( \theta \) that is used for estimating a particular PIRT model (or that results from the estimation procedure), the IRF from that PIRT model that matches the nonparametrically estimated IRF the closest. The discrepancy between the two functions indexes model misfit, and this was evaluated by means of re-sampling.

In general, NIRT approaches use kernel smoothing or another nonparametric regression method to estimate the IRF or other response functions from the data, and thus catch the monotone and nonmonotone parts of the curve. The value of knowing the whole curve mentioned in the previous example is that not only the dip at the higher \( \theta \)s is identified, but that it is also revealed that this item may actually discriminate well at the lower and middle regions of the \( \theta \) distribution. Even though it fails assumption M, such an item may be included in the test when the researcher has only few items available, most of which discriminate well at the high \( \theta \)s but not at the low \( \theta \)s. Thus, here the item’s strength may outbalance its weakness and the researcher may decide to include the item in the test so as to improve measurement precision at the lower \( \theta \)s. In another context, such as marketing, if many IRFs are found to exhibit this bell shape, the conclusion could be that another model is needed to analyze the data; that is, one that accommodates measurement of preferences by means of unimodal IRFs (e.g., Johnson and Junker, 2003; Post, 1992).

2. Analyzing test data using nonparametric IRT

The relatively weak sets of assumptions on which they are based lend NIRT models a flexibility which provides excellent opportunities for studying item and test properties. Thus, the outcomes on an NIRT analysis may give many clues of how to improve a test or questionnaire. We distinguish outcomes with respect to the dimensionality of the data and outcomes with respect to the fit a unidimensional NIRT model to the data. A sensible data analysis may start with a dimensionality analysis, identifying item subsets that predominantly are driven by one \( \theta \), followed by fitting an NIRT model to each of the dimensionally distinct item subsets so as to ascertain the measurement properties for each. Next, we review some methods for dimensionality analysis and model fitting.

2.1. Dimensionality analysis

At the basis of all NIRT methods for dimensionality analysis is conditional association (CA; Holland and Rosenbaum, 1986). Let vector \( X \) contain ordered polytomous items.
scores. Split $X$ into two disjoint part vectors: $X = (Y, Z)$. Define $f_1$ and $f_2$ to be nondecreasing functions in the item scores from $Y$, which means that $f_1$ and $f_2$ are nondecreasing coordinatewise in $Y$; and define $g$ to be some function of the item scores in $Z$. Then UD, LI, and M imply CA; that is,

$$\text{Cov}[f_1(Y), f_2(Y)|g(Z) = z] \geq 0, \quad \text{for all } z. \quad (16)$$

CA is a powerful tool for investigating the fit of NIRT models to data because there are so many ways in which item subsets and functions of item scores can be defined. We only mention two examples that are of direct relevance to dimensionality analysis under NIRT.

2.1.1. Dimensionality analysis based on nonnegative inter-item covariance

Ignore item subset $Z$, and define $f_1(Y) = X_j$ and $f_2(Y) = X_k$, then CA reduces to nonnegative inter-item covariance within the whole group; that is,

$$\text{Cov}(X_j, X_k) \geq 0, \quad \text{all pairs } j, k; \ j \neq k. \quad (17)$$

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—defined by assumptions UD, LI, and M—for at least one of the items involved.

How is Eq. (17) related to dimensionality analysis? Before we go into this topic, we first define coefficient $H$ for the scalability of a set of $J$ items and coefficient $H_j$ for the scalability of item $j$ with respect to the other $J - 1$ items in the set. Let the maximum possible covariance of items $j$ and $k$, given the marginal distributions of the $(m + 1) \times (m + 1)$ frequency table, be denoted $\text{Cov}(X_j, X_k)_{\text{max}}$; then coefficient $H$ (Loevinger, 1948; Mokken, 1971) is defined as

$$H = \frac{\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \text{Cov}(X_j, X_k)}{\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \text{Cov}(X_j, X_k)_{\text{max}}}, \quad (18)$$

and the item coefficient $H_j$ is defined as

$$H_j = \frac{\sum_{k \neq j} \text{Cov}(X_j, X_k)}{\sum_{k \neq j} \text{Cov}(X_j, X_k)_{\text{max}}}, \quad j = 1, \ldots, J. \quad (19)$$

Note that $\text{Cov}(X_j, X_k)_{\text{max}} > 0$, for all $(j, k)$ pairs, implies that the denominators of $H$ and $H_j$ are positive; and that nonnegative inter-item covariance (Eq. (17)) implies that the numerators are nonnegative; thus, $H \geq 0$ and $H_j \geq 0$, for $j = 1, \ldots, J$. Further, combining this result with $\text{Cov}(X_j, X_k) \leq \text{Cov}(X_j, X_k)_{\text{max}}$, for all $(j, k)$ pairs, we obtain that $0 \leq H \leq 1$ and $0 \leq H_j \leq 1$, for all $j = 1, \ldots, J$. Mokken (1971, p. 152) showed that coefficient $H$ is a positively weighted average of the $H_j$s.

Mokken (1971) (see Mokken and Lewis, 1982; Molenaar, 1997; Sijtsma and Molenaar, 2002, Chapter 5) proposed a bottom-up, sequential item selection procedure that clusters items on the basis of their relationship with the $\theta$ they predominantly measure, using coefficient $H$ as the procedure's loss function. Noting first that concepts like IRF slope are not well defined in an NIRT context, Mokken et al. (1986) then argued that,
keeping the distribution of $\theta$ and the $J$ item locations constant, both $H$ and $H_j$ are increasing functions of the slopes of logistic IRFs, $P_j(\theta)$ (Eq. (4)). Their line of reasoning may be generalized to non-smooth IRFs, $P(X_j \geq x_j|\theta)$ (Van Onna, 2004). Now, assume that all items in the set from which the procedure selects measure the same $\theta$. The procedure selects items one-by-one into a cluster such that, in each selection step, $H$ is maximized given all possible choices from the unselected items at that moment (i.e., at that selection step). Thus, in each step items are selected that have response functions with relatively steep slopes. These are items that hang together well due to the $\theta$ that they have in common. The procedure stops if in the next step none of the unselected items is selected due to an $H_j$ with the selected items that is lower than some user-specified positive lowerbound, $c$.

Next, assume that different subsets of items in the set from which the procedure selects measure different $\theta$s. Then, from the items not selected into the first cluster the procedure tries to select a second cluster, a third, and so on, until no items are left to be clustered. Thus, assumption M forms the basis of the dimensionality investigation: Items that measure the same $\theta$ have steep regressions (IRFs or ISRFs) on that $\theta$ and are singled out from the larger item set and selected into the same cluster. The item selection procedure is implemented in the computer program Mokken Scale analysis for Polytomous items (MSP; Molenaar and Sijtsma, 2000).

2.1.2. Dimensionality analysis based on nonnegative conditional inter-item covariance

For dichotomously scored items define a rest score based on $X$ as,

\[ R_{(-j,-k)} = \sum_{h \neq j,k} X_h, \]  

and let $g(Z) = R_{(-j,-k)}$; and define $f_1(Y) = X_j$ and $f_2(Y) = X_k$. Then CA implies that,

\[ \text{Cov}[X_j, X_k|R_{(-j,-k)} = r] \geq 0, \]

all $r = 0, 1, \ldots, J - 2$; and all $j, k$; $j \neq k$. \hfill (21)

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 dimensionality analysis that is based on investigating weak LI, as we discuss next (see Nandakumar et al., 1998, for generalizations to polytomous items).

Stout (1987, 1990) and his coworkers (Stout et al., 1996; Zhang and Stout, 1999a) proposed methods for investigating weak LI; see Eq. (6). Justified by consistency results derived by Stout (1987) for dichotomous items and Junker (1991) for polytomous items, they estimated $\text{Cov}(X_j, X_k|\theta)$ twice by inserting not only $\hat{\theta} = R_{(-j,-k)}$ but also $\hat{\theta} = X_+$ and then taking the mean of both estimates, thus obtaining smaller bias than had only one of these $\theta$ estimates been used. Estimated conditional covariances equal or close to zero were taken to indicate weak LI. Since these results are based on asymptotics, two realistic situations involving finite numbers of items, one pertaining to unidimensionality and the other to multidimensionality, may be distinguished.
First, assume that all $J$ items measure the same $\theta$, what can one say about $\text{Cov}(X_j, X_k|\hat{\theta})$ with either $\hat{\theta} = R(-j, -k)$ or $\hat{\theta} = X_+$ inserted for $\theta$? For $\hat{\theta} = R(-j, -k)$ or other estimates that do not involve $X_j$ and $X_k$ we have by CA that $\text{Cov}(X_j, X_k|\hat{\theta}) \geq 0$ (Eq. (21)); and for $\hat{\theta} = X_+$ it has been shown that $\text{Cov}(X_j, X_k|\hat{\theta}) < 0$ both for items that agree with the Rasch model (Junker, 1993, p. 1370) but also for i.i.d. coin flips. However, given only the general conditions of UD, LI, and M little is known about results for the sign let alone the magnitude of the conditional covariance, thus rendering this covariance somewhat problematic for investigating weak LI.

Second, when the items measure $Q$ latent variables to a different degree (i.e., some items predominantly measure one latent variable, other items measure another, and so on), then we may assume that $\theta_\alpha$ is a linear combination of these latent variables. The performance on the $Q$ latent variables is estimated by means of scalars $R(-j, -k)$ and $X_+$, that both summarize test performance but ignore multidimensionality. However, Zhang and Stout (1999b) showed that the sign of $\text{Cov}(X_j, X_k|\theta_\alpha)$ [and the estimator $\text{Cov}(X_j, X_k|\hat{\theta}_\alpha)$] provides useful information about the dimensionality of the data. Roughly speaking, the covariance is positive if the two items measure approximately the same latent variable and negative if they clearly measure different latent variables. This sign behavior is the basis of an algorithm called DETECT (Zhang and Stout, 1999a) that splits a given set of items into clusters that together approach estimated weak LI as good as possible given all possible cluster composites.

Assume that several of the latent variables measured by the items are dominant and that others are nuisance variables. Consider an arbitrary partitioning, denoted $\mathcal{P}$, of the set of items into disjoint clusters; and let $d_{jk}(\mathcal{P}) = 1$ if the items $j$ and $k$ are in the same cluster, and $d_{jk}(\mathcal{P}) = -1$ if they are not. The loss function of DETECT, denoted $D_\alpha(\mathcal{P})$, is defined as

$$
D_\alpha(\mathcal{P}) = \frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} d_{jk}(\mathcal{P}) E[\text{Cov}(X_j, X_k|\theta_\alpha)].
$$

(22)

The DETECT procedure tries to find the partitioning of the items, denoted $\mathcal{P}^*$, that maximizes Eq. (22). This is the item clustering that best approximates weak LI, and it is taken as the best description of the dimensionality of the data.

2.1.3. Comparative research

Van Abswoude et al. (2004a) compared MSP and DETECT with respect to their effectiveness in retrieving the true dimensionality from simulated data sets. They concluded that, in general, DETECT was better capable of retrieving a simulated dimensionality structure than MSP. However, DETECT required greater sample sizes to arrive at reliable results and for smaller sample sizes appeared to be more vulnerable to chance capitalization than MSP. It was recommended to use different procedures next to one another and use their strengths – for DETECT its sensitivity to LI and for MSP its sensitivity to IRF steepness – to arrive at better results than are possible using only one procedure.

Other cluster procedures based on nonnegative inter-item covariance and $H$ coefficients (Van Abswoude et al., 2004b) and nonnegative conditional inter-item covariance
and weak LI (Roussos et al., 1998) exist, but are not considered in further detail. Van Abswoude (2004) suggested a cluster approach that combines the unique orientations of DETECT towards dimensionality evaluation and Mokken’s procedure towards evaluation of the response function’s slope into one algorithm, such that for a minimally required quality level (or higher) in terms of item discrimination the optimal clustering in terms of dimensionality is obtained. This research is in progress.

2.1.4. Conditional covariance functions
Douglas et al. (1998) suggested estimating the conditional covariance function, \( \text{Cov}(X_j, X_k|\theta) \), and using its specific form to find out whether only one \( \theta \) drives item performance or whether other latent variables or particular characteristics of the test procedure are also influential. Let \( P_{jk}(\theta) = P(X_j = 1, X_k = 1|\theta) \), and estimate the conditional covariance as

\[
\text{Cov}(X_j, X_k|\hat{\theta}) = \hat{P}_{jk}(\hat{\theta}) - \hat{P}_j(\hat{\theta}) \hat{P}_k(\hat{\theta}). \tag{23}
\]

In Eq. (23), estimate \( \hat{\theta} \) is the percentile rank of rest score, \( R_{(-j,-k)} \) (Eq. (20)). The conditional probabilities are estimated using kernel smoothing (e.g., Douglas, 1997; Ramsay, 1991). The minimum and maximum values of \( \text{Cov}(X_j, X_k|\theta) \) depend on the IRFs, \( P_j(\theta) \) and \( P_k(\theta) \), and are expected to vary across \( \theta \). To eliminate this effect, instead of using the odds ratio to express the relationship between items \( j \) and \( k \), a function of the odds ratio, denoted \( f[\alpha(\theta)] \), is used (for details, see Douglas et al., 1998) which has several desirable properties, two of which are that (1) \( f[\alpha(\theta)] = 0 \) indicates independence, \( f[\alpha(\theta)] = -1 \) maximum negative dependence, and \( f[\alpha(\theta)] = 1 \) maximum positive dependence; and (2) \( \text{sign}(f[\alpha(\theta)]) = \text{sign}([\text{Cov}(X_j, X_k|\theta)]) \).

Function \( f[\alpha(\theta)] \) is used for studying multidimensionality phenomena at the level of item pairs such as caused by a second common latent variable, denoted \( \theta_0 \), when it can be shown that while \( \text{Cov}(X_j, X_k|\theta, \theta_0) = 0 \) by LI, ignoring \( \theta_0 \) is revealed by \( \text{Cov}(X_j, X_k|\theta) > 0 \). Indeed, for an item pair that was suspected of measuring a second latent variable it was found that the estimated function was consistently positive across the whole \( \theta \) range. Another application concerned the investigation of effects of speededness (also, see Habing, 2001), which was argued to be manifest as a violation of pairwise weak LI in areas of \( \theta \) where finishing all items was not accomplished by all respondents.

2.2. Fitting NIRT models to dimensionally distinct item clusters
Once the item set has been divided into dimensionally distinct clusters, each of which represent one dominant latent variable, each cluster is subjected to an item analysis that serves to establish the measurement quality of the scale defined by the items in a cluster. This part of the analysis concentrates on the response functions. Two properties are of particular interest. One is whether the response functions are monotone functions, so that SOL (Eq. (13)) is established, and the other is whether the response functions intersect. This is relevant for establishing an item ordering, as is explained in Section 2.2.2.
2.2.1. Diagnosing monotonicity and other response function shapes

Like restscore \( R(-j, -k) \) (Eq. (20)), restscore \( R(-j) \), defined as

\[
R(-j) = \sum_{k \neq j} X_k, \tag{24}
\]

estimates \( \theta \), a result which is justified on the basis of the same stochastic ordering and consistency arguments. Based on this result, for dichotomous items \( P[X_j = 1| R(-j)] \) is used to estimate \( P(X_j = 1| \theta) \). Junker (1993) showed that UD, LI, and M together imply manifest monotonicity (MM),

\[
P[X_j = 1| R(-j)] \quad \text{nondecreasing in } R(-j); R(-j) = 0, \ldots, J - 1. \tag{25}
\]

Contrary to intuition, a monotonicity result as in Eq. (25) is not obtained when \( R(-j) \) is replaced by \( X_+ \) (Junker and Sijtsma, 2000).

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 is done by estimating, for each value of \( r \), the proportion of the population that have item \( j \) correct, plotting these proportions as a function of \( r \), and then checking visually for MM. This approach, which is known as binning, yields a limited number – at most \( J - 1 \) – of discrete points of the IRF. Each sample deviation,

\[
\hat{P}[X_j = 1| R(-j) = r] > \hat{P}[X_j = 1| R(-j) = s], \quad r < s, \tag{26}
\]

is tested for significance by means of a normal approximation to the binomial test (Molenaar, 1973; Molenaar and Sijtsma, 2000). Karabatsos and Sheu (2004) proposed a Bayesian approach using Markov Chain Monte Carlo simulation to evaluating assumption M for \( J \) items simultaneously. This procedure also gives information about item fit.

Alternatively, kernel smoothing methods may be used to obtain a continuous estimate of the IRF, the ISRFs, and the so-called option response curves, \( P(X_j = x_j| \theta) \), with \( X_j \) nominal, representing the options of a multiple-choice item (Ramsay, 1991, 1997). Jack-knife procedures may be used to estimate confidence envelopes. The program MSP (Molenaar and Sijtsma, 2000) produces discrete estimates of IRFs and ISRFs based on binning and the program TestGraf (Ramsay, 2000) produces quasi-continuous estimates of IRFs, ISRFs, and option response curves.

Rossi et al. (2002) proposed a methodology that uses ML estimation based on EM to obtain the logit transformation of the IRF, denoted \( \lambda_j(\theta) = \ln \Omega_j \), by means of a linear combination of polynomials,

\[
\lambda_j(\theta) = \sum_{k=1}^{K} \beta_j^T \phi(\theta), \tag{27}
\]

in which \( \phi(\theta) = (\phi_1(\theta), \ldots, \phi_K(\theta)) \) are basis functions, so-called B-splines, chosen by the researcher and used to approximate adjacent segments of \( \lambda_j(\theta) \), each weighed by the corresponding coefficient in \( \beta_j = (\beta_1, \ldots, \beta_K) \). These weights are estimated from the data given the choice of basis functions, and they control the smoothness of \( \hat{\lambda}_j(\theta) \). As with kernel smoothing, too much irregularity in the curve cannot be separated from
sampling error, but too much smoothness polishes away systematic and interesting phenomena that are useful to diagnose the curve.

Another counterintuitive result is that Eq. (25) is invalid for polytomous items. Substituting $X_+$ for $R_{(-j)}$ does not solve this problem (result due to Hemker; cited in Junker and Sijtsma, 2000). Thus, probabilities, $P[X_j \geq x_j|R_{(-j)} = r]$, in which $R_{(-j)}$ is the sum of $J - 1$ polytomous variables, may be estimated, plotted, checked, and tested for deviations of monotonicity, but only heuristically. An alternative may be the following (Junker, 1996; Sijtsma and Van der Ark, 2001). Let $D_j$ be a dichotomization of the polytomous item score, $X_j$, at score $x$, such that

$$D_j = 0 \quad \text{if} \quad X_j \leq x, \quad \text{and} \quad D_j = 1 \quad \text{if} \quad X_j > x, \quad j = 1, \ldots, J;$$

and let

$$D_{(-j)} = \sum_{k \neq j} D_k.$$  \hspace{0.5cm} (29)

Using $D_{(-j)}$, the following "manifest monotonicity" result is true for polytomous items,

$$P[X_j \geq x|D_{(-j)}] \quad \text{nondecreasing in} \quad D_{(-j)}; \quad D_{(-j)} = 0, \ldots, J - 1; \quad x = 1, \ldots, m.$$  \hspace{0.5cm} (30)

Eq. (30) is also true if different items are dichotomized at different values of $x_j$. Thus, assumption M may be investigated using a multitude of "manifest monotonicity" results, first, by varying score $x_j$ in $P[X_j \geq x_j|D_{(-j)}]$ and, second, by trying different versions of $D_{(-j)}$. How to make sensible decisions to reduce the abundance of results and draw one conclusion about assumption M with respect to item $j$ is a topic of our present research.

2.2.2. Investigating invariant item ordering

Some applications require that the ordering of the $J$ items in the test by difficulty is the same for each respondent taking the test. A set of $J$ items – dichotomous or polytomous, with $m + 1$ ordered scores for each item – has an invariant item ordering (IIO; Sijtsma and Junker, 1996; Sijtsma and Hemker, 1998) if the items can be ordered and numbered accordingly, such that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_J|\theta), \quad \text{for each} \quad \theta.$$  \hspace{0.5cm} (31)

In this ordering ties may occur for some values or intervals of $\theta$. Integration of the conditional expectation in Eq. (31) across $G(\theta)$ yields an ordering of item means, denoted $\mu_{X_j}$, for $j = 1, \ldots, J$,

$$\mu_{X_1} \leq \mu_{X_2} \leq \cdots \leq \mu_{X_J}. \hspace{0.5cm} (32)$$

If an IIO has been established by one of the methods that is discussed in this section, the ordering of the item mean scores estimated from the sample can be used to estimate the item ordering as it applies to each individual $\theta$ and each subgroup from the population of interest. That is, use $\hat{\mu}_{X_j} = N^{-1} \sum_i X_{ij}$, for $j = 1, \ldots, J$, to estimate the sample mean for each item, order these sample means from small to large, and consider the result to
be the item ordering that is valid for each $\theta$. For dichotomous items, the nonparametric model defined by assumptions UD, LI, M, and IIO is the double monotonicity model (Mokken, 1971).

An example of a situation that requires an IIO is one in which the items are presented to the respondents in a fixed order, usually in ascending difficulty (opposite to the ordering in Eq. (31)), to make respondents feel at ease by presenting the easiest items first. This is common practice in intelligence testing (e.g., Bleichrodt et al., 1984), and the IIO property is applied here both at the individual and at the age-group level. For example, John, who belongs to the second-youngest age group, skips the first five items because these are of trivial difficulty to his age group, and then continues trying items of ascending difficulty until he fails, say, three consecutive items. Because the next items are more difficult, John is expected to fail almost all items of greater difficulty, meaning that the item level no longer matches his ability level. The higher someone’s ability level, the longer it takes to fail three consecutive items. A similar procedure is followed for the other age groups, with each next age group skipping more of the easier items and starting at a higher item difficulty level.

This procedure with starting rules, tailored to the group level, and stopping rules, tailored to the individual level, may be seen as a kind of adaptive testing using a small “item bank”. Within an age group, each respondent starts at the same difficulty level and is administered the same sequence of items until he/she fails at a string of items of which the difficulty exceeds his/her ability level by far. Metaphorically, one might say that this way of testing “scans” a wide neighborhood on the $\theta$ scale in which a respondent’s $\theta$ value is located, while “true” adaptive testing tries to target his/her $\theta$ value with each new item narrowing the neighborhood as quickly as possible, given the present knowledge of $\theta$.

Adaptive testing in general uses fewer items to obtain the same measurement accuracy, but requires huge item banks to attain this goal. The vast majority of test applications in psychology and elsewhere do not have access to huge item banks because (1) they are much too expensive if the application is not large-scale and commercially productive, and (2) the construction of items in tests and questionnaires is more difficult than in an educational environment that is often concerned with the reproduction of knowledge. For example, it is easier to construct items for arithmetic than for the personality domain of Dysphoria, because with arithmetic each new problem poses a new challenge to the student while with Dysphoria each new item is likely to repeat basic questions like “I often feel unhappy”, “I take a gloomy view of things”, and “I am often down in the dumps”. Variations on these basic questions as would be required in an item bank, would yield little more than a reshuffling of words and phrases without providing a better coverage of the latent variable. Thus, we support the point of view that for personality assessment short inventories consisting of the same high-quality items that are presented to everyone are more appropriate than computerized adaptive testing procedures based on item banks (see, e.g., Reise and Henson, 2000). Also, if such short tests have an IIO the quality of the assessment could be improved even further.

Sijtsma and Junker (1996) (also see Sijtsma and Molenaar, 2002, Chapter 6) presented a survey of methods for investigating IIO for dichotomous items. Several methods evaluate pairs of discrete estimates of IRFs. To understand their general principle,
consider the two IRFs in Figure 1, and notice that they intersect at $\theta^*$. Now, suppose that this true situation is unknown and that possible intersection is evaluated in the subgroups with $\theta \leq t_0$ and $\theta > t_0$ — assuming that it would be possible to make this sharp distinction with respect to $\theta$. This evaluation typically takes the form of comparing the sign of the difference of the two item means in each of these subgroups. If the signs are the same, this is taken as evidence of nonintersection of the IRFs; and if they are different, this is taken as evidence of intersection. The latter conclusion would be justified but the former can be highly misleading, as the next example shows.

Let $f(\theta)$ be the density of $\theta$. In the subgroup with $\theta \leq t_0$, which is the subgroup with the relatively low (L) $\theta$ values, denote the item means by $\mu_{jL}$ and $\mu_{kL}$; then we have that

$$\mu_{jL} - \mu_{kL} = \frac{1}{\int_{-\infty}^{t_0} f(\theta) \, d\theta} \int_{-\infty}^{t_0} [P_j(\theta) - P_k(\theta)] f(\theta) \, d\theta.$$  

The sign of this difference is positive because in Figure 1 most of the density mass of $f(\theta)$ is where the difference between the IRFs is positive; that is, to the right of $\theta^*$. For the relatively high (H) $\theta$ subgroup we have, similarly, that

$$\mu_{jH} - \mu_{kH} = \frac{1}{\int_{t_0}^{\infty} f(\theta) \, d\theta} \int_{t_0}^{\infty} [P_j(\theta) - P_k(\theta)] f(\theta) \, d\theta,$$  

which also is positive because the difference of the IRFs is positive for each $\theta$ in the interval of integration. Thus, a subgrouping based on $t_0$ masks the intersection. Obviously,
the most effective subdivision would be obtained for \( t_0 = \theta^* \), but other subdivisions in the neighborhood of \( \theta^* \) would also produce opposite item mean orderings in subgroups. Also, the \( \theta \) continuum may be divided into a greater number of disjoint and exhaustive intervals, using boundary values, \( t_{01}, t_{02}, \ldots \).

In practice, subdivisions are necessarily based on estimates of \( \theta \) and thus cannot make a "clean cut" as with \( t_0 \) in the example. The following trade-off is relevant here. If only two or a small number of subgroups are identified, the subgroup size is large and the method has much power but often in the "wrong" neighborhood on the \( \theta \) scale (Figure 1); that is, far from \( \theta^* \) (see the example in Figure 1). If a large number of subgroups is identified, in principle intersections are more easily spotted but may remain unseen because of little power due to small subsample sizes. Let us now look at the proposed methods.

Let \( A_{(-j, -k)} \) denote a set of \( F \) items not including items \( j \) and \( k \), with \( 1 \leq F \leq J - 2 \); let \( A_{(-j, -k)} \) be defined \textit{ad hoc}; and let

\[
S_{(-j, -k)} = \sum_{h \in A_{(-j, -k)}} X_h
\]

be an unweighed sum score based on the items in \( A_{(-j, -k)} \), with realization \( s \). Sum score \( S_{(-j, -k)} \) is used to define intervals on the \( \theta \) scale. Suppose \( A_{(-j, -k)} = \{X_h\} \), so that \( S_{(-j, -k)} = X_h \), with \( X_h = 0, 1 \). Thus, \( S_{(-j, -k)} \) defines two subgroups with densities of \( \theta \) conditional on \( X_h \), denoted \( f_0(\theta) \) (for \( X_h = 0 \)) and \( f_1(\theta) \) (for \( X_h = 1 \)), which overlap partly, and have means that are ordered.

This ordering follows from the SOL property in Eq. (13) and its implication for expected values in Eq. (14), with \( X_h \) substituted for \( X_+ \). Thus, \( X_h \) and other sum scores may be used to define subgroups that correspond with intervals on \( \theta \) in a stochastic ordering sense. Then inspection of pairs of discrete IRF estimates boils down to checking whether

\[
P[X_j = 1|S_{(-j, -k)} = s] \leq P[X_k = 1|S_{(-j, -k)} = s], \quad \text{for all } s,
\]

and this is done for all \( \frac{1}{2} J(J - 1) \) item pairs in the test. Three possibilities for \( S_{(-j, -k)} \) have been proposed:

- **Restscore method.** Let \( A_{(-j, -k)} \) contain all \( J - 2 \) items. Rosenbaum (1987) suggested using \( S_{(-j, -k)} = \sum_{h \neq j, k} X_h \), usually denoted restscore \( R_{(-j, -k)} \) (Eq. (20)), with realizations \( r = 0, \ldots, J - 2 \). Success probabilities are estimated in each restscore group, and the ordering must be the same across \( r \), as in Eq. (37) but with \( r \) substituted for \( s \). Note that for moderate to large test length \( J \), restscore \( R_{(-j, -k)} \) is a relatively accurate estimate of \( \theta \) and defines many intervals for evaluating ILO, but that subgroup sizes may be small. More accuracy can be gained by joining small adjacent subgroups.

- **Item-splitting method.** Define \( S_{(-j, -k)} = X_h \), with \( X_h = 0, 1 \), for \( h \neq j, k \), as before. For item pair \((j, k)\), Eq. (37) is evaluated once for \( X_h = 0 \) and once for \( X_h = 1 \). Obviously, item-splitting uses a crude division of the \( \theta \) scale, but compensates for that
by doing this for each item $h \neq j, k$. As item $h$ is more difficult, the "fuzzy" boundary between the two groups – due to error – shifts to the right so that the item-splitting method uses several sub-groupings and is able to pick up most intersections of the IRFs. This method is mathematically equivalent with Mokken’s (1971, pp. 132–133) $P(+)P(-)/P(-,-)$ matrices method.

- Restscore-splitting method. Sijtsma and Junker (1996) suggested using $R(-j,-k) \leq r$ and $R(-j,-k) > r$, with $r = 0, \ldots, J - 3$, and evaluating Eq. (37) for each value of $r$ (which is substituted for $s$). This method combines the use of the more reliable sum score, $R(-j,-k)$, for subdivision with the greater group size produced by splitting on the basis of item score $X_h$.

While each of these methods evaluates only item pairs and produces many results for the researcher to combine, coefficient $H^T$ (Sijtsma and Meijer, 1992) provides an overall impression of the degree in which $J$ IRFs intersect. $H^T$ is formally identical to coefficient $H$ in Eq. (18), but reverses the roles of persons and items; thus, it is $H$ applied to the transposed data matrix $X^T = X_{jN}$; hence coefficient $H^T$. Its use in real data analysis was explained by Sijtsma and Meijer (1992). Te Marvelde et al. (2005) generalized $H^T$ to ordered, polytomous item scores. Other work on investigating IIO for polytomous items is scarce, but research is in progress (e.g., Karabatsos and Sheu, 2004; Van der Ark and Bergsma, 2004; also see Sijtsma and Hemker, 1998).

2.3. Ordered latent class approaches to NIRT

Let $\zeta$ denote a discrete version of a latent variable, with values $w = 1, \ldots, W$. LCA models assume that a division of the group of respondents into a limited number of $W$ subgroups or classes, each characterized by an unobserved latent variable value, $\zeta = w$, produces LI,

$$P(X_v = x_v|\zeta = w) = \prod_{j=1}^{J} P(X_j = x_j|\zeta = w).$$ (38)

Because the classes are not defined a priori, but estimated from the data, they are considered to be latent; hence, latent class analysis.

Using the property of conditional probabilities for independent events, $A$ and $B$, that $P(A \land B) = P(B)P(A)$, and applying LI to $P(A)$, we may write the LCA model as (Goodman, 2002; Heinen, 1996, p. 44; McCutcheon, 2002),

$$P(X_v = x_v \land \zeta = w) = P(\zeta = w) \prod_{j=1}^{J} P(X_{vj} = x_{vj}|\zeta = w).$$ (39)

The probability that a randomly chosen respondent produces item-score vector $X = x$, is

$$P(X = x) = \sum_{w=1}^{W} P(\zeta = w) \prod_{j=1}^{J} P(X_j = x_j|\zeta = w).$$ (40)
This equation shows how LCA models the \( J \)-variate distribution of the item scores in terms of latent class probabilities, \( P(\zeta = w) \), and probabilities of having particular item scores \( X_j = x_j \) \((j = 1, \ldots , J)\) given class membership, \( P(X_j = x_j|\zeta = w) \).

The class probabilities and the conditional probabilities can be estimated from the data for several choices of the number of latent classes, \( W \). In practical data analysis, \( W \) often varies between 1 and 5. The parameter estimates for the best-fitting model are used to estimate the discrete distribution of \( \zeta, P(\zeta = w) \), with \( w = 1, \ldots , W \). This distribution can be used together with the conditional probabilities, \( P(X_j = x_j|\zeta = w) \), to assign people to latent classes. For respondent \( v \), this is done using probabilities \( P[\zeta = w|X_v] \), for \( w = 1, \ldots , W \), on the basis of which he/she is assigned to the class that has the greatest subjective probability.

The likelihood of the unrestricted LCA model is

\[
P(X_{N \times J} = x_{N \times J}|\zeta = w) = \prod_{v=1}^{N} \sum_{w=1}^{W} P(\zeta = w) \prod_{j=1}^{J} P(X_{vj} = x_{vj}|\zeta = w).
\]

(41)

This likelihood is unidentified; also, see Eq. (8) for continuous latent variable \( \theta \). Hoijtink and Molenaar (1997) adapted work by Croon (1991) for binary-item, unidimensional, order-restricted LCA models based on assumptions M and IIO to an LCA version of the model of monotone homogeneity – only M, but no IIo – with a restricted number of discrete latent variables, for example, \( \zeta = (\zeta_1, \zeta_2) \). The resulting response probabilities, \( P_j(\zeta) \) are order-restricted; that is,

\[
P_j(\zeta_1 = 1, \zeta_2 = w_2) \leq \cdots \leq P_j(\zeta_1 = W_1, \zeta_2 = w_2), \quad w_2 = 1, \ldots , W_2,
\]

(42)

and

\[
P_j(\zeta_1 = w_1, \zeta_2 = 1) \leq \cdots \leq P_j(\zeta_1 = w_1, \zeta_2 = W_2), \quad w_1 = 1, \ldots , W_1.
\]

(43)

These and other restrictions are used to estimate the model in a Bayesian framework, using Gibbs sampling to simulate the sampling distribution of several test diagnostics for properties based on monotonicity and conditional association, and to reproduce the distribution of \( X_+ \) as good as possible so as to determine the appropriate number of latent classes. This work was generalized by Van Onna (2002) to polychotomous items and ordered ISRFs (Eq. (2)), for the LCA version of assumption M in the monotone homogeneity model,

\[
P(X_j \geq x_j|\zeta_a) \leq P(X_j \geq x_j|\zeta_b),
\]

whenever \( \zeta_a < \zeta_b \); for \( x_j = 1, \ldots , m \); for \( j = 1, \ldots , J \),

(44)

and two versions of the double monotonicity model (all three models were described in more detail by Sijtsma and Hemker, 1998, for continuous \( \theta \)).

Vermunt (2001) used a logit formulation of three basic types of IRT models for polychotomous items, the cumulative probability model, the adjacent category model, and
the continuation ratio model (also, see Hemker et al., 2001, and Mellenbergh, 1995; Van der Ark et al., 2002 provide the hierarchical relationships between the nonparametric versions of these models). For example, for the cumulative probability model, also known as the monotone homogeneity model (Molenaar, 1997) or the nonparametric graded response model (Hemker et al., 1997) – here with discrete ISRF defined in Eq. (44) – define the log odds as,

$$\log \frac{P(X_j \leq x - 1|\xi)}{P(X_j \geq x|\xi)}.$$ (45)

For $\xi_a$ and $\xi_b$ the difference between the log odds is,

$$\log \frac{P(X_j \leq x - 1|\xi_a)}{P(X_j \geq x|\xi_a)} - \log \frac{P(X_j \leq x - 1|\xi_b)}{P(X_j \geq x|\xi_b)} \geq 0, \quad \text{whenever} \ \xi_a < \xi_b,$$ (46)

which is readily checked to be identical to Eq. (44). Thus, Eq. (46) provides an alternative formulation of assumption M for the polytomous-item monotone homogeneity model. Another restriction on the the log odds for items $j$ and $k$, is that,

$$\log \frac{P(X_j \leq x|\xi_j)}{P(X_k \leq x|\xi_k)} \geq 0, \quad \mu_j \leq \mu_k.$$ (47)

Eq. (47) restricts the ordering of the response probabilities, $P(X_j \leq x|\theta)$, across $j$ to be the same, except for possible ties, for each $x$; and for each $\xi$. For the cumulative probability formulation, this restriction assures an IIO (Scheiblechner, 1995; Sijtsma and Hemker, 1998). Models with inequality constraints are estimated using specialized procedures based on ML. Extensions of the models include multidimensionality, local dependence, and covariates.

3. Special topics in NIRT

3.1. Overview

NIRT researchers have contributed to several special topics in psychometrics. Examples are skills assessment (DiBello et al., 1995; McDonnell Hartz, 2002; Junker and Sijtsma, 2001b), unfolding models (Johnson and Junker, 2003; Post, 1992), and a generalization of the $H$ coefficient (Eq. (18)) to a multilevel context (Snijders, 2001; also see Reise et al., 2006). Reviews of NIRT have been provided by Sijtsma (1998), Junker (2001) and Stout (2002). NIRT models have been used on numerous occasions for constructing scales. Sijtsma and Molenaar (2002, pp. 149–150) list many examples of abilities, traits and attitudes for which European psychologists, sociologists, political science researchers, and marketing researchers successfully constructed scales by fitting specific NIRT models to the data. North American NIRT specialists have concentrated more on modeling educational test data, for example, by means of response curve fitting (Ramsay, 1991), dimensionality assessment (Bolt, 2001; Stout et al., 1996), and differential item functioning (DIF) analysis (Shealy and Stout, 1993; Stout, 2002), but Santor et al. (1994) and Santor and Ramsay (1998) used NIRT for personality assessment. Bouwmeester and Sijtsma (2004) employed NIRT measurement
data analysis tools from both sides of the Atlantic for constructing a scale for the Piagetian developmental concept of transitive reasoning.

Many new lines of research suggest themselves as soon as one starts to think about these topics. For example, DIF could be combined with IIO in the study of unidirectional DIF, as defined by Hessen (2003, pp. 22–23), following Shealy and Stout (1993). Consider two meaningful subgroups from the group of interest, and denote one as reference (R) group and the other as focal (F) group. Exactly what it means to identify such groups has been defined mathematically by Ellis and Van den Wollenberg (1993) by means of their local homogeneity concept. Usually, the focal group is suspected to be at a disadvantage when taking the test, often due to language deficiencies or cultural differences when such differences are expected not to play a role in test performance but in fact do. Absence of DIF is defined as

$$pl(O) = pf(O),$$

for all $0$; and for $j = 1, \ldots, J$.

This definition implies that an IRF that is not exactly the same in groups R and F reflects some kind of DIF. Unidirectional DIF means that $pl_j(\theta)$ and $pf_j(\theta)$ do not intersect although they may touch at one or more $\theta$s or intervals of $\theta$; that is, technically these IRFs have an IIO. The methods discussed earlier may be used to investigate IIO for pairs of IRFs (stemming from the same item) because samples from subgroups R and F do not share observations and the IRFs can be treated as they were generated by two different items. Then, after an IIO has been established a measure such as the area between the IRFs can be used to express the amount of unidirectional DIF (also, see Raju, 1988); that is, for

$$pl_j(\theta) \geq pf_j(\theta),$$

we determine

$$AREA = \int_{\theta} [pl_j(\theta) - pf_j(\theta)] f(\theta) d\theta.$$  

The IRFs can be estimated by means of binning or kernel smoothing, or another method that yields smooth estimates, and AREA can be calculated using numerical integration. The bulk of the work in nonparametric DIF analysis has been done by Stout (2002) and his co-workers, and what we did here was tie DIF together with IIO in an attempt to suggest future lines of research.

3.2. Person-fit analysis using the person response function

We finish this chapter by discussing a recent example from our own work in NIRT, in particular person-fit analysis (PFA; e.g., Emons et al., 2005; Meijer, 2003). We believe that this example nicely illustrates the potential of NIRT as a flexible and rich toolkit, by introducing individual diagnosis using the person response function (PRF), using kernel smoothing and jackknifing to obtain estimates of this PRF and corresponding confidence bands, respectively, and testing for deviation from the expected decreasing trend – performance deteriorates as items grow more difficult – by means of logistic regression.
PFA studies individuals’ vectors of $J$ item scores, collected in $X_v$. Given the probabilities of having item $j$ correct given $\theta_v$, and the monotonicity assumption of most IRT models, item-score vectors are more unlikely the larger the number of item-score pairs in which the relatively easy item was failed and the more difficult item was succeeded. Aberrant item-score vectors may be due, for example, to a lack of concentration in high-ability examinees when trying the easiest items and to answer copying to the most difficult items from other examinees by low-ability examinees. Similar to DIF, PFA attempts to uncover multidimensionality underlying test performance. This helps to diagnose test performance by taking advantage of the information available in the $J$ item scores.

Many statistics have been proposed that summarize for each item-score vector, $X_v$ ($v = 1, \ldots, N$), the degree in which an IRT model fits that vector or the degree in which the vector agrees with the vectors produced by a group of people (see Meijer and Sijsma, 2001, for a review; and Karabatsos, 2003, for a comparison of 36 PFA statistics by means of simulated data). A recent development is the comprehensive PFA methodology proposed by Emons et al. (2005). It consists of three stages:

- **Global person-fit test.** Van der Flier’s (1982) global person-fit statistic $ZU3$ – one of the most powerful person-fit statistics identified by Karabatsos (2003) – is used to identify fitting and misfitting item-score vectors. $ZU3$ is a monotone function of the likelihood of the data given an IRT model based on UD, LI, M, and IIO.

- **Graphical person-response function analysis.** The PRF (Sijtsma and Meijer, 2001) provides detailed information about the misfitting item-score vectors that were flagged by $ZU3$. For a respondent $v$, the PRF – which is the “mirror image” of the IRF – gives the probability of a correct response (scored 1) as a function of item difficulty. In this context, item difficulty is defined as the proportion-incorrect on item $j$,

$$\beta_j = 1 - \int_{\theta} P_j(\theta) \, dG(\theta).$$

(51)

For person-response variable, $S_v$, which replaces item score $X_j$, and a hypothetical item domain on which the continuous item difficulty, $\beta$, is defined, which replaces latent variable $\theta$, the PRF is defined as

$$P_v(\beta) \equiv P(S_v = 1|\beta).$$

(52)

This function is nonincreasing when the $J$ IRFs do not intersect (Sijtsma and Meijer, 2001). Eq. (52) is estimated from $J$ item scores by means of kernel smoothing. This is done by rank ordering the items according to increasing $\hat{\beta} = N^{-1} \sum_{v=1}^{N} (1 - X_{vj})$; and letting each item be focal point once, denoted $\beta_0$. The estimate of $P(S_v = 1|\beta_0)$ is obtained by taking a weighted average of the scores, $X_{vj}$, of person $v$ for the focal item and the neighboring items. Weights are defined by the kernel function, $\omega(\beta_j)$, which may be based on the Gaussian or some truncated function, such that,

$$\hat{P}_v(\beta_0) = \sum_{j=1}^{J} \omega(\beta_j)x_{vj}.$$

(53)
More details are provided by Emons et al. (2004). Suppose that a misfitting item-score vector that was flagged by ZU3 has an estimated U-shaped PRF; see Figure 2 for a real-data example based on intelligence test data (Emons et al., 2005). In this example, the local increase occurs at relatively difficult items ($\hat{\beta} \geq 0.63$, approximately), which suggest that the difficult items were easier for this person than the items of moderate difficulty. Confidence bands – here, 90% bands – for this function may be obtained by means of a jackknife procedure (see Emons et al., 2004, for details). It can be seen that for approximately $\hat{\beta} = 0.75$ and higher the estimated PRF exceeds the upper confidence band for the lowest value of the PRF (at approximately $\hat{\beta} = 0.63$). Thus, success probabilities for values $\hat{\beta} > 0.75$ are significantly larger than the minimum success probability at $\hat{\beta} = 0.63$, which is approximately $P_v(\hat{\beta}) = 0.5$. We conclude that Figure 2 gives evidence of person-misfit.

**Local PRF analysis.** Sijtsma and Meijer (2001) and Emons (2003) discuss conservative local statistical tests for evaluating increases in the PRF that take only the part of the PRF into account where the increase occurs. Alternatively, Emons et al. (2004) use logistic regression to model the U-shape (e.g., Figure 2) and other shapes suggested by the estimated PRF. Denote the PRF for respondent $w$ by $P_w(\beta)$, and assume that the items have been ordered from easy to difficult and numbered accordingly. Then, for a U-shaped PRF the logistic regression model is fitted to the item rank numbers; that is

$$\ln \left( \frac{P_w(\beta)}{1 - P_w(\beta)} \right) = \gamma_0 + \gamma_1 j + \gamma_2 j^2. \quad (54)$$

This quadratic model may be fitted against a linear model ($\gamma_2 = 0$; in addition $\gamma_1 = 0$ may be assumed). Emons et al. (2004) discuss the interpretation of results and also other logistic models that accommodate differently shaped PRFs.

The combination of global testing, graphical inspection of the PRF for misfitting-item-score vectors, and local testing of increases found in the PRF and fitting logis-
tic regression models together help to better diagnose the misfit indicated by $ZU3$. Auxiliary information from earlier testing, personality characteristics, personal history and social-economic background may further enhance the interpretation of test performance.

References


