TILBURG UNIVERSITY

Tilburg University

Hierarchically Related Nonparametric IRT Models, and Practical Data Analysis Methods

van der Ark, L.A.; Hemker, B.T.; Sijtsma, K.

Published in:
Latent Variable and Latent Structure Modeling

Publication date:
2002

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
van der Ark, L. A., Hemker, B. T., & Sijtsma, K. (2002). Hierarchically Related Nonparametric IRT Models, and Practical Data Analysis Methods. In G. Marcoulides, & I. Moustaki (Eds.), *Latent Variable and Latent Structure Modeling* (pp. 41-62). Lawrence Erlbaum.

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.

Download date: 13. Jan. 2022

3 Hierarchically Related Nonparametric IRT Models, and Practical Data Analysis Methods*

L. Andries van der Ark[1], Bas T. Hemker[2], and Klaas Sijtsma[1]

[1] Tilburg University, The Netherlands
[2] CITO National Institute, The Netherlands

3.1 Introduction

Many researchers in the various sciences use questionnaires to measure properties that are of interest to them. Examples of properties include personality traits such as introversion and anxiety (psychology), political efficacy and motivational aspects of voter behavior (political science), attitude toward religion or euthanasia (sociology), aspects of quality of life (medicine), and preferences towards particular brands of products (marketing). Often, questionnaires consist of a number (k) of statements, each followed by a rating scale with $m+1$ ordered answer categories, and the respondent is asked to mark the category that (s)he thinks applies most to his/her personality, opinion, or preference. The rating scales are scored in such a way that the ordering of the scores reflects the hypothesized ordering of the answer categories on the measured properties (called latent traits).

Items are indexed $i = 1, \ldots, k$, and item score random variables are denoted by X_i, with realizations $x = 0 \ldots, m$. Such items are known as polytomous items. Because individual items capture only one aspect of the latent trait, researchers are more interested in the total performance on a set of k items capturing various aspects than in individual items. A summary based on the k items more adequately reflects the latent trait, and the best known summary is probably the unweighted total score, denoted by X_+, and defined as

$$X_+ = \sum_{i=1}^{k} X_i. \tag{3.1}$$

This total score is well known from classical test theory (Lord & Novick, 1968) and Likert (1932) scaling, and is the test performance summary most frequently used in practice. Data analysis of the scores obtained from a sample of N respondents, traditionally using methods from classical test theory, may reveal whether X_+ is reliable, and factor analysis may be used to investigate whether X_+ is based on a set of k items measuring various aspects of predominantly the same property or maybe of a conglomerate of properties.

* Parts of this chapter are based on the unpublished doctoral dissertation of the second author.

Item response theory (IRT) uses the pattern of scores on the k items to estimate the latent trait value for each respondent (θ), in an effort to obtain a more accurate estimate of test performance than the simple X_+. For some IRT models, known as Rasch models (e.g., Fischer & Molenaar, 1995), their mathematical structure is simple enough to allow all statistical information to be obtained from the total score X_+, thus making the pattern of scores on the k items from the questionnaire superfluous for the estimation of θ. Some advanced applications of Rasch models (and other IRT models not relevant to this chapter), such as equating and adaptive testing, may still be better off with measurement on the θ scale than on the X_+ scale. Most questionnaires could either use X_+ or θ, as long as the ordering of respondents is the only concern of the researcher, and provided that X_+ and θ yield the same respondent ordering.

This chapter concentrates on *nonparametric* IRT (NIRT) models for the analysis of polytomous item scores. A typical aspect of NIRT models is that they are based on weaker assumptions than most *parametric* IRT models and, as a result, often fit empirical data better. Because their assumptions are weaker, θ cannot be estimated from the likelihood of the data, and the issue of which summary score to use, X_+ or θ, cannot come up here. Since a simple count as in Equation 3.1 is always possible, the following question is useful: When a NIRT model fits the data, does X_+ order respondents on the latent trait θ that could be estimated from a parametric IRT model?

The purposes of this chapter are twofold. First, three NIRT models for the analysis of polytomous item scores are discussed, and several well known IRT models, each being a special case of one of the NIRT models, are mentioned. The NIRT models are the *nonparametric partial credit model* (np-PCM), the *nonparametric sequential model* (np-SM), and the *nonparametric graded response model* (np-GRM). Then, the hierarchical relationships between these three NIRT models is proved. The issue of whether the ordering of respondents on the observable total score X_+ reflects in a stochastic way the ordering of the respondents on the unobservable θ is also discussed. The relevant ordering properties are monotone likelihood ratio of θ in X_+, stochastic ordering of θ by X_+, and the ordering of the means of the conditional distributions of θ given X_+, in X_+. Second, an overview of statistical methods available and accompanying software for the analysis of polytomous item scores from questionnaires is provided. Also, the kind of information provided by each of the statistical methods, and how this information might be used for drawing conclusions about the quality of measurement on the basis of questionnaires is explained.

3.2 Three Polytomous NIRT Models

Each of the three polytomous NIRT models belongs to a different class of IRT models (Molenaar, 1983; Agresti, 1990; Hemker, Van der Ark, & Sijtsma, in

press; Mellenbergh, 1995). These classes, called *cumulative probability models*, *continuation ratio models*, and *adjacent category models*, have two assumptions in common and differ in a third assumption. The first common assumption, called *unidimensionality* (UD), is that the set of k items measures one scalar θ in common; that is, the questionnaire is unidimensional. The second common assumption, called *local independence* (LI), is that the k item scores are independent given a fixed value of θ; that is, for a k-dimensional vector of item scores $\mathbf{X} = \mathbf{x}$,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{k} P(X_i = x|\theta). \tag{3.2}$$

LI implies, for example, that during test taking no learning or development takes place on the first s items ($s < k$), that would obviously influence the performance on the next $k - s$ items. More general, the measurement procedure itself must not influence the outcome of measurement. The third assumption deals with the relationship between the item score X_i and the latent trait θ. The probability of obtaining an item score x given θ, $P(X_i = x|\theta)$, is often called the *category characteristic curve* (CCC) and denoted by $\pi_{ix}(\theta)$. If an item has $m + 1$ ordered answer categories, then there are m so-called *item steps* (Molenaar, 1983) to be passed in going from category 0 to category m. It is assumed that, for each item step the probability of passing the item step conditional on θ, called the *item step response function* (ISRF) is monotone (nondecreasing) in θ. The three classes of IRT models and, therefore, the np-PCM, the np-SM, and the np-GRM differ in their definition of the ISRF.

3.2.1 Cumulative probability models and the np-GRM

In the class of cumulative probability models an ISRF is defined by

$$C_{ix}(\theta) = P(X_i \geq x|\theta) = \sum_{y=x}^{m} \pi_{iy}(\theta). \tag{3.3}$$

By definition, $C_{i0}(\theta) = 1$ and $C_{i,m+1}(\theta) = 0$. Equation 3.3 implies that passing the x-th item step yields an item score of at least x and failing the x-th item step yields an item score less than x. Thus, if a subject has an item score x, (s)he passed the first x item steps and failed the next $m - x$ item steps. The np-GRM assumes UD, LI, and ISRFs (Equation 3.3) that are nondecreasing in θ, for all i and all $x = 1, \ldots, m$, without any restrictions on their shape (Hemker, Sijtsma, Molenaar, & Junker, 1996, 1997).

The CCC of the np-GRM, and also of the parametric cumulative probability models, equals

$$\pi_{ix}(\theta) = C_{ix}(\theta) - C_{i,x+1}(\theta).$$

The np-GRM is also known as the monotone homogeneity model for polytomous items (Molenaar, 1997; Hemker, Sijtsma, & Molenaar, 1995).

A well known parametric cumulative probability model is the graded response model (Samejima, 1969), where the ISRF in Equation 3.3 is defined as a logistic function,

$$C_{ix}(\theta) = \frac{\exp[\alpha_i(\theta - \lambda_{ix})]}{1 + \exp[\alpha_i(\theta - \lambda_{ix})]}, \tag{3.4}$$

for all $x = 1, \ldots, m$. In Equation 3.4, λ_{ix} is the location parameter, with $\lambda_{i1} \leq \lambda_{i2} \leq \ldots \leq \lambda_{im}$, and α_i ($\alpha_i > 0$, for all i) is the slope or discrimination parameter. It may be noted that the slope parameters can only vary over items but not over item steps, to assure that $\pi_{ix}(\theta)$ is nonnegative (Samejima, 1972).

3.2.2 Continuation ratio models and the np-SM

In the class of continuation ratio models an ISRF is defined by

$$M_{ix}(\theta) = \frac{P(X_i \geq x|\theta)}{P(X_i \geq x - 1|\theta)}. \tag{3.5}$$

By definition, $M_{i0}(\theta) = 1$ and $M_{i,m+1}(\theta) = 0$. Equation 3.5 implies that subjects that have passed the x-th item step have an item score of at least x. Subjects that failed the x-th item step have an item score of $x - 1$. Subjects with an item score less than $x - 1$ did not try the x-th item step and thus did not fail it. The probability of obtaining a score x on item i in terms of Equation 3.5 is

$$\pi_{ix}(\theta) = [1 - M_{i,x+1}(\theta)] \prod_{y=0}^{x} M_{iy}(\theta). \tag{3.6}$$

The np-SM assumes UD, LI, and ISRFs (Eq. 3.5) that are nondecreasing in θ for all i and all x. Parametric continuation ratio models assume parametric functions for the ISRFs in Equation 3.5. An example is the sequential model (Tutz, 1990), where

$$M_{ix}(\theta) = \frac{\exp(\theta - \beta_{ix})}{1 + \exp(\theta - \beta_{ix})}. \tag{3.7}$$

In Equation 3.7, β_{ix} is the location parameter. Tutz (1990) also presented a rating scale version of this model, in which the location parameter is linearly restricted. The sequential model can be generalized by adding a discrimination parameter α_{ix} (Mellenbergh, 1995); $\alpha_{ix} > 0$ for all i and x, such that

$$M_{ix}(\theta) = \frac{\exp[\alpha_{ix}(\theta - \beta_{ix})]}{1 + \exp[\alpha_{ix}(\theta - \beta_{ix})]}. \tag{3.8}$$

This model may be denoted the *two-parameter sequential model* (2p-SM).

3.2.3 Adjacent-category models and the np-PCM

In the class of adjacent category models an ISRF is defined by

$$A_{ix}(\theta) = \frac{\pi_{ix}(\theta)}{\pi_{i,x-1}(\theta) + \pi_{ix}(\theta)}. \tag{3.9}$$

By definition, $A_{i0}(\theta) = 1$ and $A_{i,m+1}(\theta) = 0$. Equation 3.9 implies that the x-th item step is passed by subjects that have an item score equal to x, but failed by subjects that have an item score equal to $x - 1$. None of the other categories contains information about item step x. The probability of obtaining a score x on item i in terms of Equation 3.9 is

$$\pi_{ix}(\theta) = \frac{\prod_{j=0}^{x} A_{ij}(\theta) \prod_{k=x+1}^{m} [1 - A_{ik}(\theta)]}{\sum_{y=0}^{m} \prod_{j=0}^{y} A_{ij}(\theta) \prod_{k=y+1}^{m} [1 - A_{ik}(\theta)]}. \tag{3.10}$$

The np-PCM assumes UD, LI, and ISRFs (Eq. 3.9) that are nondecreasing in θ for all i and all x (see also Hemker et al., 1996, 1997).

A well known parametric adjacent category model is the partial credit model (Masters, 1982), where the ISRF in Equation 3.9 is defined as a logistic function,

$$A_{ix}(\theta) = \frac{\exp(\theta - \delta_{ix})}{1 + \exp(\theta - \delta_{ix})}, \tag{3.11}$$

for all $x = 1, \ldots, m$, where δ_{ix} is the location parameter. The generalized partial credit model (Muraki, 1992) is a more flexible parametric model, which is obtained by adding a slope or discrimination parameter (cf. Eq. 3.4) denoted α_i that may vary across items.

3.3 Relationships Between Polytomous NIRT Models

The three NIRT models have been introduced as three separate models, but it can be shown that they are hierarchically related. Because the three models have UD and LI in common, the investigation of the relationship between the models is equivalent to the investigation of the relationships between the three definitions of the ISRFs (Eqs. 3.3, 3.5, and 3.9).

First, it may be noted that the ISRFs of the first item step in the np-SM and the np-GRM are equivalent; that is, $M_{i1} = C_{i1}$, and that the ISRFs of the last item step in the np-SM and the np-PCM are equivalent; that is, $M_{im} = A_{im}$. For dichotomous items there is only one item step and the first ISRF is also the last ISRF; therefore, $C_{i1}(\theta) = A_{i1}(\theta) = M_{i1}(\theta) = \pi_{i1}(\theta)$. This case is referred to as the *dichotomous NIRT model*.

Next, it is shown that the np-PCM implies the np-SM and that the np-SM implies the np-GRM, but that the reverse relationships are not true. As

a consequence, the np-PCM implies the np-GRM, which was already proved by Hemker et al. (1997).

THEOREM 1: *The np-PCM is a special case of the np-SM.*
PROOF: If the np-PCM holds, $A_{ix}(\theta)$ (Eq. 3.9) is nondecreasing in θ for all i and all x. This implies a monotone likelihood ratio of X_i in θ for all items (Hemker et al., 1997; Proposition); that is, for all items and all item scores c and k, with $0 \leq c < k \leq m$,

$$\frac{\pi_{ik}(\theta)}{\pi_{ic}(\theta)} \text{ is nondecreasing in } \theta. \tag{3.12}$$

Let $x \geq 1$, $c = x - 1$, and $k \geq x$, then Equation 3.12 implies that the ratio $\pi_{ik}(\theta)/\pi_{i,x-1}(\theta)$ is nondecreasing in θ, and also that $\sum_{k=x}^{m}[\pi_{ik}(\theta)/\pi_{i,x-1}(\theta)]$ is nondecreasing in θ. This is identical to

$$\frac{P(X_i \geq x|\theta)}{\pi_{i,x-1}(\theta)} \text{ nondecreasing in } \theta,$$

for all i and all x, and this implies that

$$\frac{\pi_{i,x-1}(\theta)}{P(X_i \geq x|\theta)} + \frac{P(X_i \geq x|\theta)}{P(X_i \geq x|\theta)} = \frac{P(X_i \geq x-1|\theta)}{P(X_i \geq x|\theta)} \tag{3.13}$$

is non*in*creasing in θ. The reverse of the right-hand side of Equation 3.13, $P(X_i \geq x-1|\theta)/P(X_i \geq x|\theta)$, which is identical to $M_{ix}(\theta)$ (Eq. 3.5), thus is non*de*creasing for all i and all x. This implies that all ISRFs of the np-SM $[M_{ix}(\theta)]$ are nondecreasing. Thus, it is shown that if the np-PCM holds, the np-SM also holds. The np-SM does not imply the np-PCM, however, because nondecreasingness of $\sum_{k=x}^{m}[\pi_{ik}(\theta)/\pi_{i,x-1}(\theta)]$ does not imply nondecreasingness of each of the ratios in this sum; thus, it does not imply Equation 3.12. Thus, the np-SM only restricts this sum, whereas the np-PCM also restricts the individual ratios.

THEOREM 2: *The np-SM is a special case of the np-GRM.*
PROOF: From the definition of the ISRF in the np-GRM, $C_{ix}(\theta)$ (Eq. 3.3), and the definition of the ISRF in the np-SM, $M_{ix}(\theta)$ (Eq. 3.5), it follows, by successive cancellation, that for all x

$$C_{ix}(\theta) = \prod_{j=1}^{x} M_{ij}(\theta). \tag{3.14}$$

From Equation 3.14 it follows that if all $M_{ij}(\theta)$ are nondecreasing, $C_{ix}(\theta)$ is nondecreasing in θ for all x. This implies that if the np-SM holds, the np-GRM also holds. The np-GRM does not imply the np-SM, however, because nondecreasingness of the product on the right-hand side of Equation 3.14 does not imply that each individual ratio $M_{ij}(\theta)$ is nondecreasing for all x.

To summarize, the np-PCM, the np-SM, and the np-GRM can be united into one hierarchical nonparametric framework, in which each model is defined by a subset of five assumptions:

1. UD;
2. LI;
3. $C_{ix}(\theta)$ nondecreasing in θ, for all i and all x;
4. $M_{ix}(\theta)$ nondecreasing in θ, for all i and all x;
5. $A_{ix}(\theta)$ nondecreasing in θ, for all i and all x.

Note that Theorem 1 and Theorem 2 imply that Assumption 3 follows from Assumption 4, and that Assumption 4 follows from Assumption 5. Assumptions 1, 2, and 3 define the np-GRM; Assumptions 1, 2, and 4 define the np-SM; and Assumptions 1, 2, and 5 define the np-PCM. This means that

$$\text{np-PCM} \Rightarrow \text{np-SM} \Rightarrow \text{np-GRM}.$$

Finally, parametric models can also be placed in this framework. A Venn-diagram depicting the relationships graphically is given in Hemker et al. (in press). Most important is that all well known parametric cumulative probability models and parametric adjacent category models are a special case of the np-PCM and, therefore, also of the np-SM and the np-GRM. All parametric continuation ratio models are a special case of the np-SM and, therefore, of the np-GRM, but not necessarily of the np-PCM. The proof that parametric continuation ratio models need not be a special case of the np-PCM had not been published thus far and is given here.

THEOREM 3: *The 2p-SM is a special case of the np-PCM only if* $\alpha_{ix} \geq \alpha_{i,x+1}$, *for all* i, x, *and* θ.
PROOF: Both the 2p-SM (Eq. 3.8) and the np-PCM (Eq. 3.9) assume UD and LI, thus it has to be shown that the ISRFs of the 2p-SM imply that $A_{ix}(\theta)$ (Eq. 3.9) is nondecreasing in θ only if $\alpha_{ix} \geq \alpha_{i,x+1}$, but not vice versa. First, $A_{ix}(\theta)$ is defined in terms of $M_{ix}(\theta)$. It can be shown, by applying Equation 3.6 to the right-hand side of Equation 3.9 and then doing some algebra, that

$$A_{ix}(\theta) = \frac{M_{ix}(\theta) - M_{ix}(\theta)M_{i,x+1}(\theta)}{1 - M_{ix}(\theta)M_{i,x+1}(\theta)}. \tag{3.15}$$

Next, applying Equation 3.8, the parametric definition of the ISRF of the 2p-SM, to Equation 3.15 and again doing some algebra, gives

$$A_{ix}(\theta) = \frac{\exp[\alpha_{ix}(\theta - \beta_{ix})]}{1 + \exp[\alpha_{ix}(\theta - \beta_{ix})] + \exp[\alpha_{i,x+1}(\theta - \beta_{i,x+1})]}. \tag{3.16}$$

If the np-PCM holds, the first derivative of $A_{ix}(\theta)$ with respect to θ is non-negative for all i, x and θ. Let for notational convenience $\exp[\alpha_{ix}(\theta - \beta_{ix})]$ be denoted $e_{ix}(\theta)$, and let $\exp[\alpha_{i,x+1}(\theta - \beta_{i,x+1})]$ be denoted $e_{i,x+1}(\theta)$. Let

the first derivative with respect to θ be denoted by a prime. Then for Equation 3.16 the np-PCM holds if

$$A_{ix}(\theta)' = \frac{e_{ix}(\theta)'[1 + e_{ix}(\theta) + e_{i,x+1}(\theta)] - [e_{ix}(\theta)' + e_{i,x+1}(\theta)']e_{ix}(\theta)}{[1 + e_{ix}(\theta) + e_{i,x+1}(\theta)]^2} \geq 0 \tag{3.17}$$

The denominator of the ratio in Equation 3.17 is positive. Note that $e_{ix}(\theta)' = \alpha_{ix}e_{ix}(\theta)$; and $e_{i,x+1}(\theta)' = \alpha_{i,x+1}e_{i,x+1}$. Thus, from Equation 3.17 it follows that the np-PCM holds if, for all θ,

$$\alpha_{ix} + (\alpha_{ix} - \alpha_{i,x+1})e_{i,x+1}(\theta) \geq 0. \tag{3.18}$$

Equation 3.18 holds if $\alpha_{ix} \geq \alpha_{i,x+1}$ because in that case α_{ix}, $(\alpha_{ix} - \alpha_{i,x+1})$, and $e_{i,x+1}$ are all nonnegative. However, if $\alpha_{ix} < \alpha_{i,x+1}$, it follows from Equation 3.18 that $A_{ix}(\theta)$ *decreases* in θ if

$$e_{i,x+1}(\theta) > \frac{\alpha_{ix}}{\alpha_{i,x+1} - \alpha_{ix}}.$$

Thus, if $\alpha_{ix} < \alpha_{i,x+1}$, $A_{ix}(\theta)$ *decreases* for

$$\theta > \beta_{i,x+1} + \left(\frac{\ln \alpha_{i,x+1} - \ln \alpha_{ix}}{\ln \alpha_{i,x+1}}\right).$$

This means that for $\alpha_{ix} < \alpha_{i,x+1}$, Equation 3.18 does not hold for all θ. Thus, the np-PCM need not hold if $\alpha_{ix} < \alpha_{i,x+1}$. Note that the reverse implication is not true because nondecreasingness of A_{ix} does not imply the 2p-SM (Eq. 3.8). For example, in the partial credit model (Eq. 3.11) A_{ix} is nondecreasing but the 2p-SM can not hold (Molenaar, 1983).

3.4 Ordering Properties of the Three NIRT models

The main objective of IRT models is to measure θ. NIRT models are solely defined by order restrictions, and only ordinal estimates of θ are available. Summary scores, such as X_+, may provide an ordering of the latent trait, and it is important to know whether the ordering of the summary score gives a stochastically correct ordering of the latent trait. Various ordering properties relate the ordering of the summary score to the latent trait. First, the ordering properties are introduced and, second, these properties for the NIRT models both on the theoretical and the practical level are discussed.

3.4.1 Ordering properties

Stochastic ordering properties in an IRT context relate the ordering of the examinees on a manifest variable, say Y, to the ordering of the examinees on the latent trait θ. Two manifest variables are considered, the item score,

X_i, and the unweighted total score, X_+. The ordering property of *monotone likelihood ratio* (MLR; see Hemker et al., 1996),

$$\frac{P(Y=K|\theta)}{P(Y=C|\theta)} \text{ nondecreasing in } \theta; \text{ for all } C,K;\ C<K, \tag{3.19}$$

is a technical property which is only interesting here because it implies other stochastic ordering properties (see Lehmann, 1986, p. 84). Two versions of MLR are distinguished: First, MLR of the item score (MLR-X_i) means that Equation 3.19 holds when $Y \equiv X_i$. Second, MLR of the total score (MLR-X_+) means that Equation 3.19 holds when $Y \equiv X_+$.

The first ordering property implied by MLR is *stochastic ordering of the manifest variable* (SOM; see Hemker et al., 1997). SOM means that the order of the examinees on the latent trait gives a stochastically correct ordering of the examinees on the manifest variable; that is,

$$P(Y \geq x|\theta_A) \leq P(Y \geq x|\theta_B), \text{ for all } x; \text{ for all } \theta_A < \theta_B. \tag{3.20}$$

Here, also two versions of SOM are distinguished: SOM of the item score (SOM-X_i) means that Equation 3.20 holds for $Y \equiv X_i$, and SOM of the total score (SOM-X_+) means that Equation 3.20 holds for $Y \equiv X_+$. It may be noted that SOM-X_i is equivalent to $P(X_i \geq x|\theta)$ (Eq. 3.3) nondecreasing in θ.

The second ordering property implied by MLR is *stochastic ordering of the latent trait* (SOL; see, e.g., Hemker et al., 1997). SOL means that the order of the examinees on the manifest variable gives a stochastically correct ordering of the examinees on the latent trait; that is,

$$P(\theta \geq s|Y=C) \leq P(\theta \geq s|Y=K), \text{ for all } s; \text{ for all } C,K;\ C<K. \tag{3.21}$$

SOL is more interesting than SOM because SOL allows to draw conclusions about the unknown latent trait. SOL of the item score (SOL-X_i) means that Equation 3.21 holds for $Y \equiv X_i$, and SOL of the total score (SOL-X_+) means that Equation 3.21 holds for $Y \equiv X_+$.

A less restrictive form of SOL, called *ordering of the expected latent trait* (OEL) was investigated by Sijtsma and Van der Ark (2001). OEL means that

$$E(\theta|Y=C) \leq E(\theta|Y=K), \text{ for all } C,K;\ C<K. \tag{3.22}$$

OEL has only been considered for $Y \equiv X_+$.

3.4.2 Ordering properties in theory

Table 3.1 gives an overview of the ordering properties implied by the np-GRM, the np-SM, the np-PCM, and the dichotomous NIRT model. A "+" indicates that the ordering property is implied by the model, and a "–" indicates that the ordering property is not implied by the model.

Table 3.1. Overview of Ordering Properties Implied by NIRT Models.

	Ordering properties						
Model	MLR-X_+	MLR-X_i	SOL-X_+	SOL-X_i	SOM-X_+	SOM-X_i	OEL
np-GRM	−	−	−	−	+	+	−
np-SM	−	−	−	−	+	+	−
np-PCM	−	+	−	+	+	+	−
Dich-NIRT	+	+	+	+	+	+	+

Note: The symbol "+" means "model implies property", and "−" means "model does not imply property". Dich-NIRT means dichotomous NIRT model.

Grayson (1988; see also Huynh, 1994) showed that the dichotomous NIRT model implies MLR-X_+, which implies that all other stochastic ordering properties also hold, both for the total score and the item score. For the np-GRM and the np-PCM the proofs with respect to MLR, SOL, and SOM are given by Hemker et al. (1996, 1997); and for the np-SM such proofs are given by Hemker et al. (in press). The proofs regarding OEL can be found in Sijtsma and Van der Ark (2001) and Van der Ark (2000). Overviews of relationships between polytomous IRT models and ordering properties are given in Sijtsma & Hemker (2000) and Van der Ark (2001).

3.4.3 Ordering properties in practice

In many practical testing situations X_+ is used to estimate θ. It would have been helpful if the NIRT models had implied the stochastic ordering properties, for then under the relatively mild conditions of UD, LI, and nondecreasing ISRFs, X_+ would give a correct stochastic ordering of the latent trait. The absence of MLR-X_+, SOL-X_+, and OEL for most polytomous IRT models, including all NIRT models, may reduce the usefulness of these models considerably. A legitimate question is whether or not the polytomous NIRT models give a correct stochastic ordering in the vast majority of cases, so that in practice under the polytomous NIRT models X_+ can safely be used to order respondents on θ.

After a pilot study by Sijtsma and Van der Ark (2001), Van der Ark (2000) conducted a large simulation study in which for six NIRT models (including the np-GRM, the np-SM, and the np-PCM) and six parametric IRT models the following two probabilities were investigated under various settings. First, the probability that a model violates a stochastic ordering property was investigated and, second, the probability that two randomly drawn respondents have an incorrect stochastic ordering was investigated. By investigating these probabilities under different circumstances (varying shapes of the ISRFs, test lengths, numbers of ordered answer categories, and distributions of θ) it was also possible to investigate which factors increased and decreased the probabilities.

The first result was that under many conditions the probability that MLR-X_+, SOL-X_+, and OEL are violated is typically large for all three NIRT models. Therefore, it not safe to assume that a particular fitted NIRT model will imply stochastic ordering given the estimated model parameters. Secondly, however, the probability that two respondents are incorrectly ordered, due to violations of OEL and SOL, is typically small. When tests of at least five items were used for ordering respondents, less than 2% of the sample was affected by violations of SOL or OEL. This means that, although the stochastic ordering properties are often violated, only a very small proportion of the sample is affected by this violation and, in general, this simulation study thus indicated that X_+ can be used safely to order respondents on θ.

Factors that increased the probability of a correct stochastic ordering were an increase of the number of items, a decrease of the number of answer categories, and a normal or uniform distribution of θ rather than a skewed distribution. Moreover, the np-PCM had a noticeable lower probability of an incorrect stochastic ordering than the np-SM and the np-GRM. The effect of the shape of the ISRFs was different for the three NIRT models. For the np-PCM and the np-SM similarly shaped ISRFs having lower asymptotes that were greater than 0 and upper asymptotes that were less than 1 yielded the best results. For the np-GRM the best results were obtained for ISRFs that differed in shape and had lower asymptotes equal to 0 and upper asymptotes equal to 1.

3.5 Three Approaches for Estimating Polytomous NIRT Models

Generally three approaches for the analysis of data with NIRT models have been proposed. The approaches are referred to as investigation of observable consequences, ordered latent class analysis, and kernel smoothing. The difference between the approaches lies in the assumptions about θ and the estimation of the ISRF. Each approach has its own software and uses its own diagnostics for the goodness of fit investigation. Not every model can be readily estimated with the available software. The software is discussed using two simulated data sets that consist of the responses of 500 simulees to 10 polytomous items with 4 ordered answer categories (these are reasonable numbers in practical psychological research).

Data Set 1 was simulated using an adjacent category model (Eq. 3.9) with ISRF

$$\frac{P(X_i = x|\theta)}{P(X_i = x|\theta) + P(X_i = x-1|\theta)} = \frac{\exp[\alpha_{ix}(\theta - \beta_{ix})]}{1 + \exp[\alpha_{ix}(\theta - \beta_{ix})]}. \tag{3.23}$$

In Equation 3.23 the parameters α_{ix} were the exponent of random draws from a normal distribution with mean 0.7 and variance 0.5; hence, $\alpha_{ix} > 0$. The θ values of the 500 simulees and the parameters β_{ix} both were random

draws from a standard normal distribution. Equation 3.23 is a special case of the np-PCM and, therefore, it is expected that all NIRT models will fit Data Set 1. An adjacent category model was chosen because continuation ratio models (Eq. 3.5) do not necessarily imply an np-PCM (see Theorem 3) and cumulative probability models (Eq. 3.3) are not very flexible because the ISRFs of the same item cannot intersect.

Data Set 2 was simulated using a two-dimensional adjacent category model with ISRF

$$\frac{P(X_i = x|\theta_1, \theta_2)}{P(X_i = x|\theta_1, \theta_2) + P(X_i = x-1|\theta_1, \theta_2)} = \frac{\exp[\sum_{d=1}^{2} \alpha_{ixd}(\theta_d - \beta_{ixd})]}{1 + \exp[\sum_{d=1}^{2} \alpha_{ixd}(\theta_d - \beta_{ixd})]} \tag{3.24}$$

In Equation 3.24, $\alpha_{ix2} = -0.1$ for $i = 1, \ldots, 5$, and $\alpha_{ix1} = -0.1$ for $i = 6, \ldots, 10$. The remaining α_{ix} parameters are the exponent of random draws from a normal distribution with mean 0.7 and variance 0.5 and, therefore, they are nonnegative. This means that the first five items have a small negative correlation with θ_2 and the last five items have a small negative correlation with θ_1. Equation 3.24 is not unidimensional and, due to the negative α_{ix}s, the ISRFs are decreasing in either θ_1 or θ_2. Therefore, it is expected that none of the models will fit Data Set 2. The θ values of the 500 simulees and the parameters β_{ix} both were random draws from a standard normal distribution, and θ_1 and θ_2 were uncorrelated.

3.5.1 Investigation of observable consequences

This approach was proposed by Mokken (1971) for nonparametric scaling of dichotomous items. The approach is primarily focused on model fitting by means of the investigation of observable consequences of a NIRT model. For polytomous items this approach was discussed by Molenaar (1997). The rationale of the method is as follows:

1. Define the model assumptions;
2. Derive properties of the manifest variables that are implied by the model assumptions (observable consequences);
3. Investigate whether or not these observable consequences hold in the data; and
4. Reject the model if the observable consequences do not hold; otherwise, accept the model.

Software. The computer program MSP (Molenaar, Van Schuur, Sijtsma, & Mokken, 2000; Molenaar & Sijtsma, 2000) is the only software encountered that tests observable consequences for polytomous items. MSP has two main purposes: The program can be used to test the observable consequences for

a fixed set of items (dichotomous or polytomous) and to select sets of correlating items from a multidimensional item pool. In the latter case, for each clustered item set the observable consequences are investigated separately. MSP can be used to investigate the following observable consequences:

- *Scalability coefficient* H_{ij}. Molenaar (1991) introduced a weighted polytomous version of the scalability coefficient H_{ij}, originally introduced by Mokken (1971) for dichotomous items. Coefficient H_{ij} is the ratio of the covariance of items i and j, and the maximum covariance given the marginals of the bivariate cross-classification table of the scores on items i and j; that is,

$$H_{ij} = \frac{Cov(X_i, X_j)}{Cov(X_i, X_j)_{\max}}.$$

  If the np-GRM holds, then $Cov(X_i, X_j) \geq 0$ and, as a result, $0 \leq H_{ij} \leq 1$ (see Hemker et al., 1995). MSP computes all H_{ij} s and tests whether values of H_{ij} are significantly greater than zero. The idea is that items with significant positive H_{ij} s measure the same θ, and MSP deletes items that have a non-positive or non-significant positive relationship with other items in the set.
- *Manifest monotonicity.* Junker (1993) showed that if dichotomous items are conditioned on a summary score that does not contain X_i, for example, the rest score

$$R_{(-i)} = X_+ - X_i, \tag{3.25}$$

  then the dichotomous NIRT model implies manifest monotonicity; that is,

$$P(X_i = 1|R_{(-i)}) \text{ nondecreasing in } R_{(-i)}. \tag{3.26}$$

  However, Hemker (cited by Junker & Sijtsma, 2000) showed that a similar manifest monotonicity property is not implied by polytomous NIRT models; that is, $P(X \geq x|R_{(-i)})$ need not be nondecreasing in $R_{(-i)}$. It is not yet known whether this is a real problem for data analysis. MSP computes $P(X \geq x|R_{(-i)})$ and reports violations of manifest monotonicity, although it is only an observable consequence of dichotomous items.

In search for sets of related items from a multidimensional item pool, MSP uses H_{ij} and the scalability coefficients H_i (a scalability coefficient for item i with respect to the other items) and H (a scalability coefficient for the entire test) as criteria. In general, for each scale found, $H_{ij} > 0$, for all $i \neq j$, and $H_i \geq c$ (which implies that $H \geq c$; see Hemker et al., 1995). The constant c is a user-specified criterion, that manipulates the strength of the relationship of an item with θ.

Example. It may be noted that the np-GRM implies $0 \leq H_{ij} \leq 1$, which can be checked by MSP. Because the np-GRM is implied by the np-SM and the np-PCM, MSP cannot distinguish these three models by only checking

the property that $H_{ij} > 0$, for all $i \neq j$. So, either all three NIRT models are rejected when at least one $H_{ij} < 0$, or none of the three NIRT models is rejected, when all $H_{ij} > 0$.

MSP can handle up to 255 items. Thus analyzing Data Set 1 and Data Set 2 was not a problem. For Data Set 1, which was simulated using a unidimensional adjacent category model (Eq. 3.23), the ten items had a scalability coefficient $H = .54$, which can be interpreted as a strong scale (see Hemker et al., 1995). None of the H_{ij} values were negative. Therefore, MSP correctly did not reject the np-GRM for Data Set 1. Although manifest monotonicity is not decisive for rejecting the np-GRM, violations may heuristically indicate non-increasing ISRFs. To investigate possible violations of manifest monotonicity in Data Set 1, MSP checked 113 sample inequalities of the type $P(X \geq x|R_{(-i)} = r) < P(X \geq x|R_{(-i)} = r - 1)$; four significant violations were found, which seems a small number given 113 possible violations.

For Data Set 2, which was simulated using a two-dimensional adjacent category model (Eq. 3.24), the ten items had a scalability coefficient of $H = .13$, and many negative H_{ij} values, so that the np-GRM was correctly rejected. If a model is rejected, MSP's search option may yield subsets of items for which the np-GRM is not rejected. For Data Set 2, the default search option yielded two scales: Scale 1 ($H = .53$) consisted of items 3, 4, and 5, and Scale 2 ($H = .64$) consisted of items 6, 7, 8, and 9. Thus, MSP correctly divided seven items of Data Set 2 into two subscales, and three items were excluded. For item 1 and item 2, the H_{ij} values with the remaining items of Scale 1 were positive but non-significant. Item 10 was not included because the scalability coefficient $H_{6,10} = -0.03$. It may be argued that a more conventional criterion for rejecting the np-GRM might be to test whether $H_{ij} < 0$, for all $i \neq j$. This is not possible in MSP, but if the minimum acceptable H is set to 0 and the significance level is set to 0.9999, then testing for $H_{ij} > 0$ becomes trivial. In this case, items 1 and 2 were also included in Scale 1.

3.5.2 Ordered latent class analysis

Croon (1990, 1991) proposed to use latent class analysis (Lazarsfeld & Henry, 1968) as a method for the nonparametric scaling of dichotomous items. The rationale is that the continuous latent trait θ is replaced by a discrete latent variable T with q ordered categories. It is assumed that the item score pattern is locally independent given the latent class, such that

$$P(X_1, \ldots, X_k) = \sum_{s=1}^{q} P(T = s) \times \prod_{i=1}^{k} P(X_i = x_i|T = s), \qquad (3.27)$$

with inequality restrictions

$$P(X_i = 1|T = s) \geq P(X_i = 1|T = s - 1), \text{ for } s = 2, \ldots, q, \qquad (3.28)$$

to satisfy the monotonicity assumptions. If $q = 1$, the independence model is obtained.

It may be noted that the monotonicity assumption of the dichotomous NIRT model [i.e., $P(X_i = 1|\theta)$ is nondecreasing in θ] implies Equation 3.28 for all discrete combinations of successive θ values collected in ordinal latent classes. As concerns LI, it can be shown that LI in the dichotomous NIRT model and LI in the ordinal latent class model (Eq. 3.28) are unrelated. This means that mathematically, the ordinal latent class model and the dichotomous NIRT model are unrelated. However, for a good fit to data an ordinal latent class model should detect as many latent classes as there are distinct θ values, and only θs that yield similar response patterns are combined into one latent class. Therefore, if LI holds in the dichotomous NIRT model, it holds by approximation in the ordinal latent class model with the appropriate number of latent classes.

Equation 3.28 was extended to the polytomous ordinal latent class model by Van Onna (2000), who used the Gibbs-sampler, and Vermunt (2001), who used maximum likelihood, to estimate the ordinal latent class probabilities. Vermunt (2001) estimated Equation 3.28 with inequality restrictions

$$P(X_i \geq x|T = s) \geq P(X_i \geq x|T = s - 1), \text{ for } s = 2, \ldots, q, \tag{3.29}$$

and

$$P(X_i \geq x|T = s) \geq P(X_i \geq x - 1|T = s), \text{ for } x = 2, \ldots, m. \tag{3.30}$$

Due to the restrictions in Equation 3.29, $P(X_i \geq x|T)$ is nondecreasing in T [cf. Eq. 3.5, where for the np-GRM probability $P(X_i \geq x|\theta)$ is nondecreasing in θ]. Due to the restrictions in Equation 3.30, $P(X_i \geq x|T)$ and $P(X_i \geq x - 1|T)$ are nonintersecting, which avoids negative response probabilities. The latent class model subject to Equation 3.29 and Equation 3.30, can be interpreted as an np-GRM with combined latent trait values. However, as for the dichotomous NIRT model, LI in the np-GRM with a continuous latent trait and LI in the np-GRM with combined latent trait values are mathematically unrelated.

Vermunt (2001) also extended the ordered latent class approach to the np-SM and the np-PCM, and estimated these models by means of maximum likelihood. For ordinal latent class versions of the np-PCM and the np-SM the restrictions in Equation 3.29 are changed into

$$\frac{P(X_i = x|T = s)}{P(X_i = x - 1 \vee x|T = s)} \geq \frac{P(X_i = x|T = s - 1)}{P(X_i = x - 1 \vee x|T = s - 1)}, \text{ for } s = 2, \ldots, q \tag{3.31}$$

and

$$\frac{P(X_i \geq x|T = s)}{P(X_i \geq x - 1|T = s)} \geq \frac{P(X_i \geq x|T = s - 1)}{P(X_i \geq x - 1|T = s - 1)}, \text{ for } d = 2, \ldots, q \tag{3.32}$$

respectively. For the np-PCM and the np-SM the ISRFs may intersect and, therefore, restrictions such as Equation 3.30 are no longer necessary.

Software. The computer program *ℓEM* (Vermunt, 1997) is available free of charge from the world wide web. The program was not especially designed to estimate ordered latent class models, but more generally to estimate various types of models for categorical data via maximum likelihood. The program syntax allows many different models to be specified rather compactly, which makes it a very flexible program, but considerable time must be spent studying the manual and the various examples provided along with the program. *ℓEM* can estimate the ordinal latent class versions of the np-PCM, the np-GRM, and the np-SM, although these options are not documented in the manual. Vermunt (personal communication) indicated that the command "or1" to specify ordinal latent classes should be changed into "or1(b)" for the np-PCM, and "or1(c)" for the np-SM. For the np-GRM the command "or1(a)" equals the original "or1", and "or1(d)" estimates the np-SM with a reversed scale (Agresti, 1990; Hemker, 2001; Vermunt, 2001). In addition to the NIRT models, *ℓEM* can also estimate various parametric IRT models. The program provides the estimates of $P(T = s)$ and $P(X_i = x|T = s)$ for all i, x, and s, global likelihood based fit statistics such as L^2, X^2, AIC, and BIC (for an overview, see Agresti, 1990), and for each item five pseudo R^2 measures, showing the percentage explained qualitative variance due to class membership.

Example. For Data Set 1 and Data Set 2, the np-GRM, the np-SM and the np-PCM with $q = 2$, 3, and 4 ordered latent classes we estimated. The independence model ($q = 1$) as a baseline model to compare the improvement of fit was also estimated. Latent class analysis of Data Set 1 and Data Set 2 means analyzing a contingency table with $4^{10} = 1,048,576$ cells, of which 99.96% are empty. It is well known that in such sparse tables likelihood-based fit statistics, such as X^2 and L^2, need not have a chi-squared distribution. It was found that the numerical values of X^2 and L^2 were not only very large (exceeding 10^6) but also highly different (sometimes $X^2 > 1000L^2$). Therefore, X^2 and L^2 could not be interpreted meaningfully, and instead the following fit statistics are given in Table 3.2: loglikelihood (L), the departure from independence (Dep. $= [L(1)-L(q)]/L(1)$) for the estimated models, and the difference in loglikelihood between the ordinal latent class model and the corresponding latent class model without order constraints (Δ). The latter two statistics are not available in *ℓEM* but can easily be computed. Often the estimation procedure yielded local optima, especially for the np-GRM (which was also estimated more slowly than the np-SM and the np-PCM). Therefore, each model was estimated ten times and the best solution was reported. For some models more than five different optima occurred; this is indicated by an asterisk in Table 3.2.

For all models the loglikelihood of Data Set 1 was greater than the loglikelihood of Data Set 2. Also the departure from independence was greater

Table 3.2. Goodness of Fit of the Estimated np-GRM, np-SM, and np-PCM With *ℓEM*.

		np-GRM			np-PCM			np-SM		
	Data q	L	Dep.	Δ	L	Dep.	Δ	L	Dep.	Δ
Data Set 1	1	-3576	.000	0	-3576	.000	0	-3576	.000	0
	2	-2949	.175	14	-2980	.167	45	-2950	.175	15
	3	-2853*	.202	34	-2872	.197	53	-2833	.208	24
	4	-2791*	.220	34	-2818	.212	61	-2778	.223	21
Data Set 2	1	-4110	.000	0	-4110	.000	0	-4110	.000	0
	2	-3868*	.058	1	-3917	.047	54	-3869	.059	6
	3	-3761*	.085	108	-3791	.078	138	-3767	.083	114
	4	-3745*	.089	51	-3775*	.092	181	-3763	.084	169

Note: L is the loglikelihood; Dep. is the departure of independence $\frac{L(q)-L(1)}{L(q)}$; Δ is the difference between the loglikelihood of the unconstrained latent class model with q classes and the ordinal latent class model with q classes.

for the models of Data Set 1 than for the models of Data Set 2, which suggests that modeling Data Set 1 by means of ordered latent class analysis was superior to modeling Data Set 2. The difference between the loglikelihood of the ordered latent class models and the unordered latent class models was greater for Data Set 2, which may indicate that the ordering of the latent classes was more natural for Data Set 1 than for Data Set 2. All these finding were expected beforehand. However, without any reference to the real model, it is hard to determine whether the NIRT models should be rejected for Data Set 1, for Data Set 2, or for both. It is even harder to distinguish the np-GRM, the np-SM, and the np-PCM. The fit statistics which are normally used to reject a model, L^2 or X^2, were not useful here. Based on the L^2 and X^2 statistics, only the independence model for Data Set 1 could have been rejected.

3.5.3 Kernel smoothing

Smoothing of item response functions of dichotomous items was proposed by Ramsay (1991) as an alternative to the Birnbaum (1968) three-parameter logistic model,

$$\pi_{i1}(\theta) = \gamma_i + (1 - \gamma_i)\frac{\exp[\alpha_i(\theta - \beta_i)]}{1 + \exp[\alpha_i(\theta - \beta_i)]}, \tag{3.33}$$

where γ_i is a guessing parameter, α_i a slope parameter, and β_i a location parameter. Ramsay (1991) argued that the three-parameter logistic model does not take nonmonotonic item response functions into account, that the sampling covariances of the parameters are usually large, and that estimation algorithms are slow and complex. Alternatively, in the monotone smoothing

approach, continuous nonparametric item response functions are estimated using kernel smoothing. The procedure is described as follows (see Ramsay, 2000, for more details):

1. *Estimation of* θ. A summary score (e.g., X_+) is computed for all respondents, and all respondents are ranked on the basis of this summary score; ranks within tied values are assigned randomly. The estimated θ value ($\hat{\theta}$) of the n-th respondent in rank is the n-th quantile of the standard normal distribution, such that the area under the standard normal density function to the left of this value is equal to $n/(N+1)$.
2. *Estimation of the CCC.* The CCC, $\pi_{ix}(\theta)$, is estimated by (kernel) smoothing the relationship between the item category responses and the $\hat{\theta}$s. If desired the estimates of θ can be refined after the smoothing. Douglas (1997) showed that under certain regularity conditions the joint estimates of θ and the CCCs are consistent as the numbers of respondents and items tend to infinity. Stout, Goodwin Froelich, and Gao (2001) argued that in practice the kernel smoothing procedure yields positively biased estimates at the low end of the θ scale and negatively biased estimates at the high end of the θ scale.

Software. The computer program TestGraf98 and a manual are available free of charge from the ftp site of the author (Ramsay, 2000). The program estimates θ as described above and estimates the CCCs for scales with either dichotomous or polytomous items. The estimates of θ may be expressed as standard normal scores or may be transformed monotonely to $E(R_{(-i)}|\hat{\theta})$ (see Equation 3.25) or $E(X_+|\hat{\theta})$. The program provides graphical rather than descriptive information about the estimated curves. For each item the estimated CCCs $[\pi_{ix}(\hat{\theta})]$ and the expected item score given $\hat{\theta}$ $[E(X_i|\hat{\theta})]$ can be depicted. For multiple-choice items with one correct alternative it is also possible to depict the estimated CCCs of the incorrect alternatives. Furthermore, the distribution of $\hat{\theta}$, the standard error of $\hat{\theta}$, the reliability of the unweighted total score, and the test information function are shown. For each respondent the probability of $\hat{\theta}$ given the response pattern, can be depicted.

Testing NIRT models with TestGraf98 is not straightforward because only graphical information is provided. However, if the np-GRM holds, which implies that $P(X_i \geq x|\theta)$ is nondecreasing in θ (Eq. 3.5), then $E(X_i|\theta)$ is also nondecreasing in θ, because

$$E(X_i|\theta) = \sum_{x=1}^{m} P(X_i \geq x|\theta).$$

If a plot in TestGraf98 shows for item i that $E(X_i|\hat{\theta})$ is not nondecreasing in $\hat{\theta}$, this may indicate a violation of the np-GRM and, by implication, a violation of the np-SM, and the np-PCM. Due to the lack of test statistics,

TestGraf98 appears to be a device for an eyeball diagnosis, rather than a method to test whether the NIRT models hold.

Example. For Data Set 1, visual inspection of the plots of $E(X_i|\hat{\theta})$ showed that all expected item scores where nondecreasing in $\hat{\theta}$. This means that no violations of the np-GRM were detected. For Data Set 2, for three items $E(X_i|\hat{\theta})$ was slightly decreasing in $\hat{\theta}$ over a narrow range of $\hat{\theta}$; $E(X_7|\hat{\theta})$ showed a severe decrease in $\hat{\theta}$. Moreover, three expected item score functions were rather flat, and two expected item score functions were extremely flat. This indicates that for Data Set 2, the np-GRM was (correctly) not supported by TestGraf98.

3.6 Discussion

In this chapter three polytomous NIRT models were discussed, the np-PCM, the np-SM, and the np-GRM. It was shown that the models are hierarchically related; that is, the np-PCM implies the np-SM, and the np-SM implies the np-GRM. It was also shown that the 2p-SM only implies the np-PCM if for all items and all item steps the slope parameter of category x is less or equal to the slope parameter of category $x + 1$. This final proof completes the relationships in a hierarchical framework which includes many popular polytomous IRT models (for overviews, see Hemker et al., in press).

NIRT models only assume order restrictions. Therefore, NIRT models impose less stringent demands on the data and usually fit better than parametric IRT models. NIRT models estimate the latent trait at an ordinal level rather than an interval level. Therefore, it is important that summary scores such as X_+ imply a stochastic ordering of θ. Although none of the polytomous NIRT models implies a stochastic ordering of the latent trait by X_+, this stochastic ordering will hold for many choices of ISRFs or CCCs in a specific model, and many distributions of θ. The np-PCM implies stochastic ordering of the latent trait by the item score. In the kernel smoothing approach an interval level score of the latent trait is obtained by mapping an ordinal summary statistic onto percentiles of the standard normal distribution. Alternatively, multidimensional latent variable models can be used if a unidimensional parametric IRT model or a NIRT model do not have an adequate fit. Multidimensional IRT models yield estimated latent trait values at an interval level (e.g., Moustaki, 2000). Multidimensional IRT models are, however, not very popular because parameter estimation is more complicated and persons cannot be assigned a single latent trait score (for a discussion of these arguments, see Van Abswoude, Van der Ark, & Sijtsma, 2001).

Three approaches for fitting and estimating NIRT models were discussed. The first approach, investigation of observable consequences, is the most formal approach in terms of fitting the NIRT models. For fitting a model based on UD, LI, and M, the latent trait is not estimated but the total score is

used as an ordinal proxy. The associated program MSP correctly found the structure of the simulated data sets.

In the ordinal latent class approach the NIRT model is approximated by an ordinal latent class model. The monotonicity assumption of the NIRT models is transferred to the ordinal latent class models, but the LI assumption is not. It is not known how this affects the relationship between NIRT models and ordinal latent class models. The latent trait is estimated by latent classes, and the modal class membership probability $P(T = t|X_1, \ldots, X_k)$ can be used to assign a latent trait score to persons. The associated software *ℓEM* is the only program that could estimate all NIRT models. *ℓEM* found differences between the two simulated data sets indicating that the NIRT models fitted Data Set 1 but not Data Set 2. It was difficult to make a formal decision.

The kernel smoothing approach estimates a continuous CCC and a latent trait score at the interval level. In this approach there are no formal tests for accepting or rejecting NIRT models. The associated software TestGraf98 gives graphical information. It is believed that the program is suited for a quick diagnosis of the items, but the lack of test statistics prevents the use for model fitting. Moreover, only a derivative of the np-GRM, $E(X_+|\hat{\theta})$, can be examined. However, the graphs displayed by TestGraf98 supported the correct decision about the fit of NIRT models.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397 – 424). Reading, MA: Addison-Wesley.

Croon, M. A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171-192.

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315-331.

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika, 62*, 7-28.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch Models: Foundations, recent developments and applications*. New York: Springer.

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

Hemker, B. T, (2001), Reversibility revisited and other comparisons of three types of polytomous IRT models. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds.), *Essays in item response theory* (pp. 275 – 296). New York: Springer.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional itembank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*, 337-352.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*, 679-693.

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

Hemker, B. T., Van der Ark, L. A., & Sijtsma, K. (in press). On measurement properties of continuation ratio models. *Psychometrika.*

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent variables. *Psychometrika, 59*, 77–79,

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

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lehmann, E. L. (1986). *Testing statistical hypotheses.* (2nd ed.). New York: Wiley.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140.*

Lord, F. M., & Novick M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses *Applied Psychological Measurement, 19*, 91-100.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton/Berlin: De Gruyter.

Molenaar, I. W. (1983). *Item steps* (Heymans Bulletin HB-83-630-EX). Groningen, The Netherlands: University of Groningen.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12(37)*, 97-117.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369 – 380). New York: Springer.

Molenaar, I. W., & Sijtsma, K. (2000). MSP for Windows [Software manual]. Groningen, The Netherlands: iec ProGAMMA.

Molenaar, I. W., Van Schuur, W. H., Sijtsma, K., & Mokken, R. J. (2000). MSP-WIN5.0; A program for Mokken scale analysis for polytomous items [Computer software]. Groningen, The Netherlands: iec ProGAMMA.

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement, 24*, 211-223.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-177.

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

Ramsay, J. O. (2000, September). TestGraf98 [Computer software and manual]. Retrieved March 1, 2001 from the World Wide Web: *ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf*

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17.*

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph, 18.*

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics, 25*, 391-415.

Sijtsma, K., & Van der Ark, L. A. (2001). Progress in NIRT analysis of polytomous item scores: Dilemmas and practical solutions. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders, (Eds.), *Essays on item response theory* (pp. 297–318). New York: Springer.

Stout, W., Goodwin Froelich, A., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders, (Eds.), *Essays on item response theory* (pp. 357–375). New York: Springer.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*, 39-55.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2001). *A comparative study on test dimensionality assessment procedures under nonparametric IRT models.* Manuscript submitted for publication.

Van der Ark, L. A. (2000). *Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models.* Manuscript submitted for publication.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273-282.

Van Onna, M. J. H. (2000). Gibbs sampling under order restrictions in a nonparametric IRT model. In W. Jansen & J. Bethlehem (Eds.) *Proceedings in Computational Statistics 2000; Short communications and posters* (pp. 117–118). Voorburg, The Netherlands: Statistics Netherlands.

Vermunt, J. K. (1997, September). *ℓEM*: A general program for the analysis of categorical data [Computer software and manual]. Retrieved September 19, 2001 from the World Wide Web: http://www.kub.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html

Vermunt, J. K. (2001). The use of latent class models for defining and testing non-parametric and parameteric item response theory models. *Applied Psychological Measurement, 25*, 283-294.