Behaviormetrika
Vol.28, No.1, 2001, 65-94

DEVELOPMENTS IN MEASUREMENT OF PERSONS AND ITEMS BY MEANS OF ITEM RESPONSE MODELS

Klaas Sijtsma*

This paper starts with a general introduction into measurement of hypothetical constructs typical of the social and behavioral sciences. After the stages ranging from theory through operationalization and item domain to preliminary test or questionnaire have been treated, the general assumptions of item response theory are discussed. The family of parametric item response models for dichotomous items (e.g., correct/incorrect scores) is introduced and it is explained how parameters for respondents and items are estimated from the scores collected from a sample of respondents who took the test or questionnaire. Next, the family of nonparametric item response models is explained, followed by the three classes of item response models for polytomous item scores (e.g., rating scale scores). Then, it is discussed to what degree the mean item score (the p-value for dichotomous items) and the unweighted sum of item scores for persons (the total test score) are useful for measuring items and persons in the context of item response theory. The concepts of invariant item ordering for items, and monotone likelihood ratio, stochastic ordering, and ordering of the expected latent trait for persons, are relevant here.

So far, the paper has concentrated on measurement of properties of persons and items, based on item response models. Such measurements make sense only when the item response model fits the data. Methods for fitting models to data are briefly discussed for parametric and nonparametric models, but also two recent hybrid methods are mentioned. Finally, the main applications of item response models are discussed, which include equating and item banking, computerized and adaptive testing, research into differential item functioning, person fit research, and cognitive modeling.

1. Introduction

In order to establish the empirical relationships between concepts, these concepts must be measured reliably and validly. For example, the relationship between analytical reasoning ability and attention span can only be investigated in a useful way if sound measurement instruments are available for both concepts. If reliability is low or even absent all that is measured is random noise, and the relationship of interest can only reveal itself weakly or not at all. The use of invalid measurements, even if reliable, may lead to the finding of a relationship that gives a misleading impression about the true relationship between the intended concepts. For example, due to a weak theoretical underpinning a measurement instrument intended to measure analytical reasoning may measure general intelligence instead.

Key Words and Phrases: applications of IRT, estimation of IRT models, invariant item ordering, item response theory, nonparametric item response theory, parametric item response theory, polytomous IRT models, stochastic ordering of persons

* Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. email: k.sijtsma@kub.nl. The author is grateful to Brian W. Junker, Ivo W. Molenaar and L. Andries van der Ark for their helpful comments on an earlier draft of this paper.

In psychology, psychometrics as a specialization of applied statistics emphasizes the need for reliable and valid person measurement. Sociological methodology traditionally focuses on the scaling of stimuli. These measurement traditions have made practical researchers in the social and behavioral sciences aware of the need for adequate measurement as the key to valid research conclusions. This paper explains the basic ideas behind measurement in the social and behavioral sciences, and summarizes the ways in which modern measurement models organized in item response theory (IRT) assign measurement values to persons and allow the assessment of the properties of the measurement instruments themselves.

2. The Basic Ideas of Measurement

2.1 Hypothetical constructs

Measurement in the social and behavioral sciences starts with defining hypothetical constructs. Examples of hypothetical constructs are:

- Abilities, such as spatial orientation and comprehensive reasoning;
- Achievements, for example, in school subjects such as arithmetics and grammar;
- Skills, for example, involving coordination when manipulating a panel controlling a complex technical process, such as the train traffic at a big railway station;
- Personality traits, such as introversion and extraversion, neuroticism and anxiety;
- Attitudes, for example, towards male and female role patterns, towards euthanasia, or towards the intervention of NATO in Kosovo;
- Preferences, for example, for brands of beer, automobiles, and political parties.

Different research areas tend to focus on different kinds of hypothetical constructs. For example, psychologists are interested, in particular, in the measurement of abilities, skills and personality traits, sociologists are more interested in attitudes and opinions, and marketing researchers more in motivations of consumers and their preferences for particular products. Clearly, in practical research for each of these hypothetical constructs measurement instruments have to be available.

In general, a hypothetical construct entails a theoretical definition of its relevant aspects and their relationships. For many important concepts theories are available that guide the measurement specialist when constructing a measurement instrument. For example, an intelligence theory can be helpful as a basis for intelligence test construction. In general, two problems may occur in this process.

One problem is the simultaneous existence of several competing theories, such as in the area of intelligence. These competing intelligence theories yield different definitions of intelligence and, thus, may pose a choice problem for the measurement of intelligence. For example, Spearman (1923) assumed one general intelligence faculty, Thurstone (1938) assumed seven general factors, and Guilford (1967) assumed three general dimensions subdivided into four, five, and six more specialized abilities, defining 120 meaningful combinations.

Another problem typical in some other areas of measurement may stem not so much from the variety in theories about a particular construct, but rather from the poverty of such theorizing and sometimes even the absence of it. For hypothetical constructs such as creativity, social and emotional intelligence, self-esteem, and leadership, the supporting theory may still be rather inarticulate and formulated at a highly abstract and general level, making it almost impossible to identify sets of behavior typical of the intended construct.

2.2 Operationalization, item domain, test

Operationalization entails the specification of the operations needed for the measurement of a hypothetical construct. First, the domain of behavior that is typical of the intended construct has to defined. When a hypothetical construct is supported by a well developed and tested theory, the definition of such a behavior domain may be rather straightforward. The presence of conflicting theories or, worse, the absence of a widely recognized theory may complicate the definition of a behavior domain.

Assuming a well-defined behavior domain, the next step is to define a domain of possible stimuli that can be presented to people from a population of interest, in order to elicit responses that are indicative of the relevant behaviors. Such stimuli are called *items*. Examples of items are:

- Statements, for example, about political or ethical issues (attitude measurement) or the respondents own behavior (personality trait measurement);
- Tasks, such as maze problems, building blocks to be used for copying a particular construction, and geometric figures to be rotated mentally to a prescribed position (intelligence and ability measurement);
- Questions, for example, about history or arithmetics, or about a text that has been read aloud to the respondents (achievement, ability measurement).

The measurement instrument consists of a representative sample from the item domain. From here on, I will generically use the term *test* to denote any measurement instrument consisting of a number of items. The number of items in a test typically ranges from 5 to 100, depending on the construct measured. The reason for having a larger number of items is that the statements about an individual's measurement value based on the test have to be free of measurement error to a

high degree (reliability) and sufficiently general so as to cover the construct well (validity).

2.3 Test construction, quantification, scores

A test administered to a representative sample of respondents elicits responses by each respondent to each of the items. These responses can be, for example,

- Solutions to problems, such as arithmetics or maze problems;
- Choices among alternatives from multiple-choice items or markings on a rating scale;
- Written or oral reports in response to passages read aloud to the respondents;
- Solution times on a test for attention and concentration.

Except for the solution times, these responses are qualitative; that is, they are not yet in the form of numbers that can be analyzed by means of statistical methods.

The quantification of the responses may be thought to consist of three stages. First, the responses are categorized into types that are assumed to be informative about the hypothetical construct. Second, the categories are ordered in the degree to which they reflect the measured construct. Third, scores are assigned to each of the ordered categories, reflecting this ordering. These scores are known as the *item scores.* Each respondent obtains an item score on each item (s)he has answered. The adequacy of the assumption that a higher item score reflects a higher standing on the measured construct, should be investigated by means of the statistical analysis of the empirical item scores.

Everything that goes wrong during the stages of test construction discussed thus far, be it the incompleteness of the underlying theory, an inappropriate operationalization where important aspects of the theory are overlooked, an unfortunate choice of the item format, or an inadequate quantification of the item responses, cannot be, or can only partly be, repaired by statistical modeling once the data have been collected. On the other hand, a well-constructed test based on sound theory and a fine-grained operationalization will yield data that will more or less speak for themselves. That is, the role of the statistical model here is to simply show the general data structure without the need for extensive further manipulation and exploration of the data.

2.4 Results of data analysis, practical use of tests

The statistical analysis of item scores obtained by N respondents on J items yields one of three general results. First, the measurement instrument; that is, the item set retained after a careful statistical analysis of the quality of each of the separate items and the item set as a whole. Second, quality measures of

the measurement instrument, such as reliability and validity estimates. Third, the measurement values or *test scores*, which locate individuals on a continuum representing the yard stick on which we measure the construct of interest.

The two main applications of test scores are scientific research and individual diagnosis. In scientific research test scores are used to compare groups or one group with itself at a later point in time, as in longitudinal research. Here, the focus is on *group statistics* such as mean test scores, for example, when boys and girls are compared with respect to verbal ability, and correlations of a test score with another variable of interest, for example, when an intelligence test score is correlated with school performance after one year. Use of test scores in scientific research can be found in psychology, education (achievement testing), sociology, political science (e.g., opinion polls), medical research (e.g., quality of life), demographic research (attitudes toward moral issues, e.g., evaluated at the national level), and marketing research.

Test use in an individual diagnosis context focuses on *individual test scores* and uses these scores for decision making about individuals. Examples are assigning a patient to a particular therapy on the basis of his personality profile, hiring an applicant on the basis of intelligence and achievement test results, and advising a pupil or his/her parents to continue education at another school type. Because individual scores are more subject to measurement error than group means, the quality requirements of tests intended for individual diagnosis are much higher than of tests used in scientific research.

3. Item Response Theory and Models

Many statistical models have been proposed for the analysis of the item scores. Nowadays, the most important are the family of item response models, collectively defining *item response theory* (IRT; e.g., Van der Linden & Hambleton, 1997; Junker & Sijtsma, 2001a). Before we explain IRT, we introduce some notation that we will use throughout. Let X_j denote the random variable for the ordered score on item j, with $j = 1, \ldots, J$; and let x_j be a realization (often simply denoted x). When items have two ordered scores, they are called dichotomous items ($x_j = 0, 1$) and when they have three or more ordered scores, they are called polytomous items ($x_j = 0, \ldots, m$). It may be noted that we assume that all J polytomous items have a fixed number of m ordered answer categories because, in general, not all theoretical results hold when the number of answer categories is free to vary across items. Also, in practice many researchers choose their items to have the same number of answer categories. Further, we need a random vector $\mathbf{X} = (X_1, \ldots, X_J)$ and its realization $\mathbf{X} = \mathbf{x}$. Finally, the latent trait is denoted θ, and has a density $f(\theta)$ and a cumulative distribution function $F(\theta)$. The latent trait does not coincide with the hypothetical construct of which the test is an operationalization; instead, it is a variable representing person performance on the item domain of which the test is a sample.

For polytomous items, several conditional probabilities may describe the relationship between X_j and θ. Here, we mention two of these probabilities, but later on we will also encounter others. The two conditional probabilities are:

$$P(X_j = x_j|\theta), \text{ known as the } \textit{category characteristic curve}; \tag{1}$$

and

$$P(X_j \geq x_j|\theta), \text{ known as the } \textit{item step response function} \text{ (ISRF)}. \tag{2}$$

The relationships between these response functions are

$$P(X_j = x|\theta) = P(X_j \geq x|\theta) - P(X_j \geq x + 1|\theta), \tag{3}$$

and

$$P(X_j \geq x|\theta) = \sum_{y=x}^{m} P(X_j = y|\theta). \tag{4}$$

For dichotomous items we have one relevant response function, which is

$$P_j(\theta) \equiv P(X_j = 1|\theta), \text{ this is the } \textit{item response function} \text{ (IRF)}, \tag{5}$$

which is also known as the item characteristic curve or the trace line. In general, we call category characteristic curves (1), ISRFs (2), and IRFs (5) *response functions*. IRT models the relationship between the probability of an item score and the latent trait by making several assumptions about the response process. We discuss the general assumptions, and then discuss how specific models are specializations of these general assumptions.

4. General Assumptions of IRT

4.1 Dimensionality of measurement and relationships between items

The first kind of assumptions is about the dimensionality of measurement. Some models assume that a response to an item is governed by several latent traits, collected in a multidimensional $\boldsymbol{\theta}$. In this case, the response probabilities (1), (2), and (5) condition on $\boldsymbol{\theta}$ and represent response surfaces rather than curves. A multidimensional IRT model (e.g., Kelderman & Rijkes, 1994; Reckase, 1997) may be appropriate when, for example, some of the items from an arithmetics test are solved with greater probability for higher levels of verbal ability and a few others are solved with greater probability for higher levels of spatial orientation. The "psychological" assumption underlying most tests, however, is that one latent trait is measured, so that we may use a unidimensional IRT model for describing the data, and a scalar θ. The majority of the IRT models assume unidimensionality.

The second kind of assumptions describe the relationships between the items from a test, given that we have complete knowledge of the dimensionality. The most common assumption is local independence, defined as

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{J} P(X_j = x_j|\boldsymbol{\theta}). \tag{6}$$

Local independence means that during test administration the probability of getting an item correct is determined only by that item and θ, and is not influenced by success or failure on any other items. Processes such as learning through practice while being tested are assumed not to influence test results, that is, the measurement procedure does not affect the measurement results. Such processes may be hard to control in a testing situation, so that learning and development may influence test results irrespective of the efforts of the experimenter to suppress or eliminate them. IRT models incorporating such effects have been developed (e.g., Jannerone, 1997). Also, in a dynamic testing context where pupils are trained and their development is monitored by subsequent tests, Embretson (1991) has formalized development by adding latent traits as the training and testing proceeds. Despite such efforts, most psychometric models assume that the test is locally independent and unidimensional.

One may ask whether the common assumptions of unidimensionality and local independence constitute a model in the sense that the joint conditional distribution of the items scores in (6) restricts the J-variate data distribution,

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{j=1}^{J} P(X_j = x_j|\theta) dF(\theta). \tag{7}$$

Suppes and Zanotti (1981) and Holland and Rosenbaum (1986) showed that $P(\mathbf{X} = \mathbf{x})$ is not restricted, meaning that unidimensionality and local independence do not constitute a falsifiable model, unless the response functions are restricted, the distribution of θ is restricted, or both are restricted. Junker (1993) discussed that restrictions are always necessary on the dimensionality, the conditional relationships between items, and the response functions, and that dropping either one of the three kinds of restrictions leads to an unrestricted distribution $P(\mathbf{X} = \mathbf{x})$.

Taking (7) as a starting point, IRT modeling always places restrictions on the response functions, and often on the θ distribution when estimating the model parameters (for each item one or more parameters to be discussed shortly; and depending on the dimensionality, one or more θs for each person). Examples are marginal maximum likelihood estimation (e.g., Baker, 1992), where the density $f(\theta)$ often is assumed to be normal, and Markov Chain Monte Carlo estimation (e.g., Hoijtink & Molenaar, 1997; Patz & Junker, 1999), where model parameters of interest are assumed to originate from distributions. When the model structure

is sufficiently simple (Rasch, 1960; Masters, 1982; Verhelst & Glas, 1995), conditional maximum likelihood estimation which avoids making assumptions about $f(\theta)$ may be used, but this is not realistic for most models as we will see in the next section.

4.2 Restrictions on response functions

The restrictions on IRFs (5) and ISRFs (2) have the same common purpose, expressing that the higher θ, the higher the response probability; for example, the higher someone's intelligence, the higher the probability that (s)he will correctly solve items from an intelligence test. Restrictions are also possible on category characteristic curves (1), and by restricting ISRFs by implication the former curves are also restricted; see (3). In this paper, we mostly discuss ISRFs (and IRFs), as will be clear throughout. Restrictions on response functions can be parametric or nonparametric. First, we will make this distinction for models for dichotomous item scores, and then for models for polytomous items.

4.2.1 Parametric IRT models for dichotomous items, estimation

An example of a parametric IRT model is the 3-parameter logistic model (e.g., Birnbaum, 1968, pp. 395-479). Let δ_j denote the location of item j on the θ scale, also interpreted as the item difficulty; $\alpha_j (\alpha_j > 0)$ the slope parameter at δ_j, which indicates the degree of separation of low θs and high θs by means of item j, and also known as the discrimination parameter; and γ_j the lower asymptote for $\theta \rightarrow -\infty$, also known as the guessing parameter, but more generally capturing the idea that low θ respondents may have response probabilities greater than 0, not only due to guessing correct, but also because the item may contain cues for its solution, which are not dependent on θ; then the 3-parameter logistic model is defined as

$$P_j(\theta) = \gamma_j + \frac{(1-\gamma_j)\exp[\alpha_j(\theta-\delta_j)]}{1+\exp[\alpha_j(\theta-\delta_j)]}. \tag{8}$$

The function in (8) is strictly increasing in θ. Other well known parametric IRT models also have logistic IRFs; for example, the 2-parameter logistic model (Birnbaum, 1968), for which $\gamma_j = 0$ for all j, and the 1-parameter logistic model or Rasch (1960) model, for which $\gamma_j = 0$ and $\alpha_j = 1$, for all j. Characteristic of parametric IRT is that the person parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ and the item parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_J)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)$ can be estimated from the likelihood functions of the models. Based on the scores of N respondents on J items, collected in the data matrix $\mathbf{X}_{N\times J}$, the general expression for the likelihood, $L(\text{model})$, is

$$L(\text{model}) = P[\mathbf{X}_{N\times J}|\text{model}] = \prod_{i=1}^{N}\prod_{j=1}^{J} P_j(\theta)^{x_{ij}}[1-P_j(\theta)]^{1-x_{ij}}. \tag{9}$$

For example, for the Rasch model, defined as

$$P_j(\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}, \tag{10}$$

after substituting (10) into (9), followed by some algebra, the likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\delta}) = P[\mathbf{X}_{N\times J}|\boldsymbol{\theta}, \boldsymbol{\delta}] &= \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{[\exp(\theta_i - \delta_j)]^{x_{ij}}}{1 + \exp(\theta_i - \delta_j)} \\ &= C_{N,J}(\boldsymbol{\theta}, \boldsymbol{\delta}) \exp\left[\sum_{i=1}^{N} \theta_i r_i - \sum_{j=1}^{J} \delta_j s_j\right], \end{aligned} \tag{11}$$

where

$$C_{N,J}(\boldsymbol{\theta}, \boldsymbol{\delta}) = \left\{ \prod_{i=1}^{N}\prod_{j=1}^{J} [1 + \exp(\theta_i - \delta_j)] \right\}^{-1}, \tag{12}$$

which does not depend on the data; and

$$r_i = \sum_{j=1}^{J} x_{ij}, \tag{13}$$

which is the number of items correctly answered by respondent i; and

$$s_j = \sum_{i=1}^{N} x_{ij}, \tag{14}$$

which is the number of persons in the sample who gave a correct answer to item j. In (11), which is a likelihood function from the exponential family (Molenaar, 1995), the observable sum scores r_i and s_j are sufficient statistics for the parameters θ_i and δ_j, respectively. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ can be estimated jointly but inconsistently, or using the *conditional maximum likelihood* method separately but consistently (Andersen, 1970; Baker, 1992, chap. 5). In practice, this method is used only to estimate the δs, not only consistently but also without bias, and next maximum likelihood is used to estimate the θs assuming that the estimated δs are the true parameters [see Hoijtink & Boomsma (1995) for a justification of this two-step procedure]. Lord (1983) has shown that, unfortunately, the estimates of the extreme θs in particular are biased, and Warm (1989) has suggested corrections for most of this bias.

For the 3-parameter logistic model and the 2-parameter logistic model, conditional maximum likelihood estimation is not feasible. For example, in the two-parameter logistic model the likelihood is an expression similar to (11), but with

$$C_{N,J}(\boldsymbol{\theta}, \boldsymbol{\delta}, \boldsymbol{\alpha}) = \left\{ \prod_{i=1}^{N}\prod_{j=1}^{J} \{1 + \exp[\alpha_j(\theta_i - \delta_j)]\} \right\}^{-1}, \tag{15}$$

which also is independent of the data, and statistics

$$r_i^* = \sum_{j=1}^{J} \alpha_j x_{ij}, \tag{16}$$

which is the sufficient statistic for θ provided $\boldsymbol{\alpha}$ is known; and

$$s_j^* = s_j = \sum_{i=1}^{N} x_{ij}. \tag{17}$$

which is the sufficient statistic for the product $\delta_j^* \equiv \alpha_j \delta_j$. For conditional maximum likelihood statistics are needed that only depend on the data. However, r_i^* also depends on the unknown parameters $\boldsymbol{\alpha}$; and s_j^* estimates a re-scaled location parameter δ_j^*, but no information is retrieved about the slope parameter α_j, necessary for estimating θ_i.

Alternatively, *marginal maximum likelihood* may be used to estimate the item parameters. Here, a distribution $f(\theta|\boldsymbol{\omega})$ with parameters $\boldsymbol{\omega}$ is assumed, and the likelihood is integrated over this distribution; for the two-parameter logistic model,

$$L(\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\delta}) = P[\mathbf{X}_{N\times J}|\boldsymbol{\omega}, \boldsymbol{\alpha}, \boldsymbol{\delta}] = \int_{\theta} \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{\exp[\alpha_j(\theta_i - \delta_j)]^{x_{ij}}}{1 + \exp[\alpha_j(\theta_i - \delta_j)]} f(\theta|\boldsymbol{\omega}) d\theta. \tag{18}$$

The resulting marginal likelihood can be maximized with respect to the item parameters, and $\boldsymbol{\omega}$, the parameters of $f(\theta|\boldsymbol{\omega})$. This yields consistent estimates, assuming that $f(\theta)$ was correctly specified (other versions exist, in which $f(\theta)$ is first estimated from the data, and then used in a likelihood as (18)). The person parameters are estimated next by maximum likelihood, assuming that the item parameter estimates are the true values (e.g., Baker, 1992, chap. 6; Warm, 1989).

Verhelst & Glas (1995) used conditional maximum likelihood for estimating a hybrid model with a location parameter δ_j and a user-specified slope *index* A_j, meaning that the slopes can be treated as known. This way, full advantage was taken of the simplicity of the Rasch model, because only location parameters δ had to be estimated, but no αs. More complex models can be estimated using *Markov Chain Monte Carlo estimation* (e.g., Albert, 1992). For example, Béguin (2000) used this method to estimate the parameters from unidimensional and multidimensional IRT models that assume normal ogive curves instead of logistic functions. Patz and Junker (1999) used Markov Chain Monte Carlo estimation for 2-parameter and 3-parameter logistic IRT models for tests consisting of both multiple-choice items and constructed-response items (e.g., open-ended questions), with missing item responses; and also extensions of the Rasch model with linear restrictions on the location parameters to accommodate rater effects (e.g., when the answers to questions have been assessed and scored by several raters).

4.2.2 Nonparametric IRT models for dichotomous items

The second class of restrictions on the response functions is found with nonparametric IRT models, which define *order restrictions* on the response functions, without the parametric restrictions as in (8) and (10). An example is the monotone homogeneity model (Mokken, 1971, chap. 4; Mokken & Lewis, 1982).

DEFINITION 1. The monotone homogeneity model assumes a unidimensional θ, local independence (Eq. (6)), and monotonicity; that is, for fixed values θ_a and θ_b,

$$P_j(\theta_a) \leq P_j(\theta_b) \text{ whenever } \theta_a < \theta_b. \tag{19}$$

Another nonparametric model is the model of double monotonicity (Mokken, 1971, chap. 4; Mokken & Lewis, 1982), which is a special case of the monotone homogeneity model.

DEFINITION 2. The double monotonicity models assumes unidimensionality, local independence (Eq. (6)), monotonicity (Eq. (19)), and nonintersecting IRFs. Nonintersection means that if we know that for a fixed value θ_0 the response probabilities for items j and k are ordered $P_j(\theta_0) < P_k(\theta_0)$, then we know that

$$P_j(\theta) \leq P_k(\theta), \text{ for all } \theta. \tag{20}$$

In general, nonparametric IRT models are less restrictive than their parametric counterparts and will fit more often to data. For example, the monotone homogeneity model may be seen as a liberalization of the 3-parameter logistic model, because in addition to variation in lower asymptotes γ, slopes α, and locations δ, IRFs are not restricted to the logistic function, as long as they are nondecreasing. Thus, they may have an irregular shape with several inflection points and need not be symmetric. Likewise, the double monotonicity model may be seen as a liberalization of the Rasch model, into IRFs that are nondecreasing and nonintersecting, but not necessarily logistic. Likelihood equations for nonparametric models do not allow for the estimation of item parameters such as δ, α, and γ, simply because they are not contained in the likelihood (Eq.(9)). We will see later on, however, that in a nonparametric context it is possible to order persons on θ by means of observable test scores, and that information about item properties can be obtained through other parameters. Measurement by means of nonparametric IRT models constitutes an important topic of this paper.

The parametric and nonparametric models we have discussed so far were all *strictly unidimensional* IRT models (Junker, 1993; Sijtsma & Junker, 1996), as opposed to *essentially unidimensional* models (Stout, 1987, 1990). Essential unidimensionality means that the items measure one dominant latent trait, but that several items may also measure nuisance traits. This implies that local independence does not hold for all J items from the test when the conditioning is only on

the dominant trait. The defining characteristic of nuisance traits, however, is that for $J \rightarrow \infty$ their influence vanishes, which is defined by essential independence rather than local independence. Define a vector $\boldsymbol{\theta} = (\theta, \theta_1, \theta_2, \ldots)$ containing the dominant latent trait θ and all relevant nuisance traits. Essential independence is defined (Stout, 1990) using conditional covariances as

$$\binom{J}{2}^{-1} \sum_{1 \leq j < k \leq J} |\text{Cov}(X_j, X_k|\theta)| \rightarrow 0 \text{ if } J \rightarrow \infty. \tag{21}$$

Just as essential independence relaxes local independence, Stout's (1990) weak monotonicity assumption relaxes IRF monotonicity (Eq. (19)). Weak monotonicity says that the test characteristic curve, which may be defined as the mean of J IRFs, is an increasing function of $\boldsymbol{\theta}$ (Stout, 1987). More specifically, if $\boldsymbol{\theta}_a \leq \boldsymbol{\theta}_b$ in each coordinate, then weak monotonicity is defined as

$$J^{-1} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}_a) \leq J^{-1} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}_b), \text{ all } \boldsymbol{\theta}_a \leq \boldsymbol{\theta}_b, \text{ coordinatewise.} \tag{22}$$

Under weak monotonicity not all individual IRFs have to be increasing, as long as their mean is increasing. Equations (21) and (22) together define essential unidimensionality.

A further restriction on weak monotonicity, which we need later on when we discuss properties of estimates of the latent trait θ, is *local asymptotic discrimination*. This restriction says that the mean IRF from the test is sufficiently discriminating locally for each θ; that is, for every latent trait from $\boldsymbol{\theta}$ and for every value θ_0 of a latent trait θ there exists $\epsilon_{\theta_0} > 0$, such that for every $\theta_0{}^*$ close to θ_0

$$J^{-1} \sum_{j=1}^{J} \frac{P_j(\theta_0{}^*) - P_j(\theta_0)}{\theta_0{}^* - \theta_0} \geq \epsilon_{\theta_0} > 0, \text{ all } J. \tag{23}$$

Local asymptotic discrimination is both a relaxation of monotonicity because it pertains to the sum or the mean of the IRFs rather than individual IRFs, and a strengthening, because for the mean IRF strict increasingness rather than nondecreasingness at the individual IRF level is assumed.

4.2.3 Polytomous IRT models, estimation

The discussion of parametric polytomous IRT models is more lucid when it is embedded in a general nonparametric IRT context. We discuss three general classes of polytomous IRT models, and within each class give both parametric and nonparametric examples of polytomous IRT models. Hemker, Van der Ark, and Sijtsma (in press) also discuss these three general classes, which have also been distinguished at the level of parametric polytomous IRT models (Agresti, 1990; Akkermans, 1998; Mellenbergh, 1995; Molenaar, 1983a; Van Engelenburg, 1997).

Adjacent Category Models. The first is the class of adjacent category models (ACMs). In general, the ISRF (different from (2)) is defined as

$$f_{jx}^{ACM}(\theta) \equiv P(X_j = x|\theta; X_j = x-1 \vee X_j = x) = \frac{P(X_j = x|\theta)}{P(X_j = x-1|\theta) + P(X_j = x|\theta)}, \tag{24}$$

for $x = 1, \ldots, m$. It is assumed that this is a nondecreasing function of θ. A well known parametric model from this class is the partial credit model (Masters, 1982), which can be seen as a polytomous version of the Rasch model (Eq. (10)). The partial credit model defines the ISRF as

$$P(X_j = x|\theta; X_j = x-1 \vee X_j = x) = \frac{\exp(\theta - \delta_{jx})}{1 + \exp(\theta - \delta_{jx})}, \text{ all } x = 1, \ldots, m. \tag{25}$$

Here, given that there are two possibilities, either a score of $x-1$ or a score of x, the Rasch model governs the probability that a score of x is obtained. Thus, Eq. (25) models the response process for two adjacent response categories that are isolated from the other $m-1$ answer categories. As a result, within an item the ordering of the δ_{jx}s is not fixed, and thus may vary over items. This may suggest that the partial credit model fits best to an item type that consists of m subtasks that may be solved in an arbitrary order. For example, a text comprehension item may consist of three separate questions about the content of a text that can each be solved without considering the other two. Each question may yield 0 or 1 points and the item is scored $0, 1, 2, 3$. Van Engelenburg (1997) argued, however, that the partial credit model and this item type (or other item types) do not logically imply one another.

When Eq. (25) is defined for m pairs of adjacent item score pairs, the category characteristic curve of the partial credit model is

$$P(X_j = x|\theta) = \frac{\exp\left[\sum_{s=1}^{x}(\theta - \delta_{js})\right]}{\sum_{q=0}^{m}\exp\left[\sum_{s=1}^{q}(\theta - \delta_{js})\right]}, \tag{26}$$

with $\sum_{s=1}^{0}(\theta - \delta_{js}) \equiv 0$. The parameters of this model can be estimated using conditional maximum likelihood (Masters, 1982). Muraki (1992) generalized the partial credit model using $\alpha_j(\theta - \delta_{jx})$ in the exponents rather than sums of terms $(\theta - \delta_{jx})$, thus allowing category characteristic curves to have different slopes across but not within items. Patz and Junker (1999) used Markov Chain Monte Carlo for estimating this model from data with missing item scores.

Cumulative Probability Models. The second class of polytomous IRT models is the class of cumulative probability models (CPMs), with nondecreasing ISRFs defined as

$$f_{jx}^{CPM}(\theta) \equiv P(X_j \geq x|\theta), \tag{27}$$

for $j = 0, \ldots, m$; and with $P(X_j \geq 0|\theta) = 1$. The homogeneous case of the graded response model (Samejima. 1969) is a well-known parametric IRT model from this class. The graded response model defines the ISRF as a logistic function with a slope (α_j) and a location (λ_{jx}) parameter. The slope parameter is the same for each of the m ISRFs,

$$P(X_j \geq x|\theta) = \frac{\exp[\alpha_j(\theta - \lambda_{jx})]}{1 + \exp[\alpha_j(\theta - \lambda_{jx})]}. \tag{28}$$

The parameters of this model can be estimated, for example, using joint maximum likelihood (see Baker, 1992, chap. 8)

Van Engelenburg (1997) argued that the graded response model is best suited for modeling item scores that result from a global assessment task, for example, the rating of a response on a Likert-type item measuring an attitude or a personality trait. Here, the respondent forms a general impression of his/her position on the scale relative to the item. For example, the respondent is asked to determine on a Likert scale the degree to which the main character in a text expressed a hostile attitude toward the other characters.

Continuation Ratio Models. The third class of polytomous IRT models is the class of continuation ratio models (CRMs), that define the nondecreasing ISRF as

$$f_{jx}^{CRM}(\theta) \equiv P(X_j \geq x|\theta; X_j \geq x - 1) = \frac{P(X_j \geq x|\theta)}{P(X_j \geq x - 1|\theta)}, \tag{29}$$

for $x = 1, \ldots, m$ (note that (29) is different from (2)). An example of a parametric CRM is the sequential model (Tutz, 1990), which is defined by the ISRF,

$$P(X_j \geq x|\theta; X_j \geq x - 1) = \frac{\exp(\theta - \beta_{jx})}{1 + \exp(\theta - \beta_{jx})}. \tag{30}$$

Tutz (1997) uses joint maximum likelihood and marginal maximum likelihood for estimating the parameters of the model.

Here, the typical item consists of a fixed sequence of m subtasks, and failure on the $(x+1)st$ subtask implies an item score of x. This means that the subtasks of the item have to be executed in a fixed order, and failure on one subtask means failure on the subsequent subtasks. For example, in a text comprehension item it may first be checked whether the respondent has understood the topic of the text (if not, $x = 0$), then whether (s)he has grasped a particular fact about an event explicitly described in the text (if not, $x = 1$) and, finally, whether (s)he has understood the implicitly mentioned intention of the main character portrayed (if not, $x = 2$; otherwise, $x = 3$). Samejima (1972, chap. 4) showed that for CRMs, the category characteristic curve can be expressed as (assuming $f_{jx}^{CRM} = 1$ for $x < 1$; and $f_{jx}^{CRM} = 0$ for $x > m$)

$$P(X_j = x|\theta) = \prod_{y=0}^{x} f_{jy}^{CRM}(\theta)[1 - f_{j,x+1}^{CRM}(\theta)]. \tag{31}$$

That is, the probability of having a score of x is the product of x ISRFs for the first x subtasks that were answered correctly, and one probability of failing the $(x+1)st$ subtask.

General Results for Polytomous IRT Models. Each of the three classes of polytomous IRT models contains several models [see Hemker, Sijtsma, Molenaar, & Junker (1997); Hemker et al. (in press); and Sijtsma & Hemker (2000) for overviews]. The three definitions of the three classes represent the most important differences. Within classes different parametric models have different parameterizations. For example, Muraki's (1992) generalized partial credit model allows varying discrimination between items, whereas the partial credit model (Eq. (26)) assumes constant discrimination.

Hemker et al. (1997) investigated the hierarchical relationships between the well known parametric and nonparametric models from the classes of ACMs and CPMs. Hemker et al. (in press) investigated the hierarchical relationships within the class of CRMs, and related their results to the results found by Hemker et al. (1997) for the other two classes. For the purpose of this paper, we summarize the main results as follows.

1. Definitions of general nonparametric models from each of the three classes:

DEFINITION 3. ACM class: The *nonparametric Partial Credit Model* (np-PCM) assumes unidimensionality, local independence (Eq. (6)), and $f_{jx}^{ACM}(\theta)$ (Eq. (24)) nondecreasing in θ.

DEFINITION 4. CPM class: The *nonparametric Graded Response Model* (np-GRM) assumes unidimensionality, local independence (Eq. (6)), and $f_{jx}^{CPM}(\theta)$ (Eq. (27)) nondecreasing in θ.

DEFINITION 5. CRM class: The *nonparametric Sequential Model* (np-SM) assumes unidimensionality, local independence (Eq. (6)), and $f_{jx}^{CRM}(\theta)$ (Eq. (29)) nondecreasing in θ.

2. Within each of the three classes of ACMs, CPMs, and CRMs, the np-PCM (Definition 3), the np-GRM (Definition 4), and the np-SM (Definition 5), respectively, are the most general models. Also, each of these nonparametric models contains all other *parametric* and *nonparametric* models from its class as special cases. That is, when we represent each model as a set, a Venn-diagram displaying the relationships between the models from the class of ACMs would show the np-PCM as the outer set encompassing all other ACMs as subsets, for example, the partial credit model (Eq. (26)); and likewise for Venn-diagrams for CPMs and CRMs.

3. Hemker (1996, chap. 6) proved that the following relationships hold for the np-PCM (Definition 3), the np-SM (Definition 5), and the np-GRM (Definition 4):

$$\text{np-PCM} \Rightarrow \text{np-SM} \Rightarrow \text{np-GRM}. \tag{32}$$

That is, of the well known polytomous IRT models the np-GRM is the most general model, which has all other models from the other three classes as special cases [also see Hemker et al. (in press); and Van der Ark (2001)]. This is an important conclusion that we will use later on.

5. Measuring Persons and Items

Properties of items, such as their difficulty (δ or a related parameter) or their discrimination power (α), are relevant in the phase of instrument construction, whereas person properties (latent traits θ) are relevant when the test is put to practical use. Here, we will summarize results that relate classical observable statistics, such as the number-correct and the item mean, to IRT models.

5.1 Classical person and item summaries

Classical test theory (CTT; Nunnally, 1978; Lord & Novick, 1968) uses simple observable statistics for measuring persons and items. For person i ($i = 1, \ldots, N$), the sum of J item scores, X_{+i}, is used, and is defined as

$$X_{+i} = \sum_{j=1}^{J} X_{ij}; \quad X_{ij} = 0, 1, \ldots, m; \; X_{+i} = 0, 1, \ldots, mJ. \tag{33}$$

It may be noted that $X_{+i} = r_i$ (Eq. (13)), which is the sufficient statistic for θ_i in the Rasch model (Eq. (10)). Total score X_{+i} can be used to estimate the true proportion-correct score,

$$\overline{T}_i \equiv J^{-1} E(X_{+i}), \; i = 1, \ldots, N, \tag{34}$$

where the expectation is over hypothetical independent replications of the test for person i. For the difficulty of an item, the item mean,

$$\overline{X}_j = N^{-1} \sum_{i=1}^{N} X_{ij}, \; j = 1, \ldots, J, \tag{35}$$

is used, which estimates the population mean μ_j. For binary scores, the item mean is the p value.

5.2 Ordering items using the item mean

Several applications assume that the same items are difficult or easy for all respondents, that is, for all θs. [It may be noted, by the way, that a distinction can be made between respondents and θs and, related to this, different definitions of the response probability; see Holland (1990) and Ellis and Van den Wollenberg (1993). We will ignore this distinction here for practical purposes.] For example, some intelligence tests assume the same starting rules and stopping rules, based

on the ordering of the items according to difficulty, to hold for each respondent. Based on this assumption, for a particular age group, say, the first ten easiest items may be skipped, because they are assumed to be too easy, and each individual stops when (s)he has failed on three consecutive items, assuming that the next items are too difficult. These generally applied rules only make sense when for all θs the item ordering is the same.

We will consider the item ordering by mean item score, $E(X_j)$, instead of latent location parameters, such as δ_j from the 3-parameter logistic model (Eq. (8)), δ_{jx} from the partial credit model (Eq. (26)), λ_{jx} from the graded response model (Eq. (28)), and β_{jx} from the sequential model (Eq. (30)). Sijtsma and Hemker (2000) have argued that these parameters cannot be interpreted meaningfully as item difficulties. For dichotomous items, when IRFs intersect, as in the 3-parameter logistic model, the ordering of item difficulties expressed as response probabilities, $P_j(\theta)$, depends on θ as well as δ. For polytomous items location parameters give information, for example, on intersection points of pairs of category characteristic curves, as in the partial credit model (Eq. (26)). Also, it is not clear how m of these δ_{jx}s for each item should be combined into one difficulty index.

A more familiar and simpler item difficulty parameter is the item mean. We will consider the item mean conditional on θ, that is, $E(X_j|\theta)$, which Chang and Mazzeo (1994) defined as the IRF for polytomous items. For J items an invariant item ordering (IIO; Sijtsma & Junker, 1996) is defined when the items can be ordered and numbered accordingly, such that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_J|\theta), \text{ for all } \theta. \tag{36}$$

That is, given θ the item means have the same ordering with the exception of possible ties for some θs. For dichotomous items, it is easily checked that $E(X_j|\theta) = P_j(\theta)$, which is the IRF, so that IIO is identical to

$$P_1(\theta) \leq P_2(\theta) \leq \cdots \leq P_J(\theta), \text{ for all } \theta. \tag{37}$$

From (37) it is easily seen that IRT models with nonintersecting IRFs imply an IIO. Examples are the Rasch model (Eq. (10)) and the double monotonicity model (Definition 2). For polytomous items,

$$E(X_j|\theta) = \sum_{x=1}^{m} P(X_j \geq x|\theta), \tag{38}$$

and for the difference between two conditional expected item scores we have

$$E(X_j|\theta) - E(X_k|\theta) = \sum_{x=1}^{m} [P(X_j \geq x|\theta) - P(X_k \geq x|\theta)]. \tag{39}$$

From (39) it is clear that for the items j and k to have an IIO, for consecutive θs the sum of the m differences between ISRFs for $x = 1, \ldots, m$ must not change

sign. Sijtsma and Hemker (1998) have shown that of the well known polytomous IRT models from the ACM class only the rating scale model (Andrich, 1978), and from the CPM class only the isotonic ordinal probabilistic model (Scheiblechner, 1995) imply an IIO (Eq. (36)). Hemker et al. (in press) have shown that from the CRM class only the sequential rating scale model (Tutz, 1990) implies an IIO.

When an IIO holds for J items, the items also have the same ordering in any subpopulation g from the population of interest, with distribution $F_g(\theta)$. Arbitrarily assume that in (39) the sign of the difference on the left-hand side is nonnegative for all θs, then given an IIO it follows that

$$E(X_j) - E(X_k) = \int_\theta \left\{ \sum_{x=1}^{m} [P(X_j \geq x|\theta) - P(X_k \geq x|\theta)] \right\} dF_g(\theta) \geq 0. \quad (40)$$

Because in (40) the sum in the integrand has nonnegative sign for all θs, the difference in item means on the left also has nonnegative sign for any selection from $F(\theta)$, that is, for any subpopulation $F_g(\theta)$. Research aimed at investigating IIO in real data uses this result, and checks for relevant subgroups whether the item ordering according to $E(X_j)$ within subgroups is invariant between subgroups. That is, let $g = 1, \ldots, G$ index subgroups, then an invariant item ordering at the level of subgroups implies

$$E(X_1|g) \leq E(X_2|g) \leq \cdots \leq E(X_J|g), \; g = 1, \ldots, G. \quad (41)$$

See Sijtsma and Van der Ark (2001) for more information.

5.3 Ordering persons on θ using X_+

An interesting question is whether X_+ can be used for ordering persons on θ. One reason is that in a nonparametric IRT context we do not have numerical estimates of θ, so that the use of the observable X_+ as a proxy for θ would be highly convenient. Also, in the general IRT context the practical use of θ is hampered by its complicated metric in terms of logits (e.g., Mellenbergh, 1995; Sijtsma & Hemker, 2000; also see Ramsay, 1995), which renders it quite troublesome when communicating test results to laymen (clients, pupils, parents, teachers) and test practitioners. The advantage of X_+ is that most people are familiar with the number-correct (dichotomous items) or number-of-points earned (polytomous items) interpretation. Our central question for person ordering is: When we have X_+, and the test or questionnaire was constructed using an IRT model, does the ordering of respondents on X_+ reflect, in a statistical sense (i.e., allowing random error), the ordering on θ? To answer this question, we consider person ordering in a more technical sense. We distinguish results for dichotomous IRT models and results for polytomous IRT models.

5.3.1 Person ordering using dichotomous item scores

Grayson (1988; also see Huynh, 1994) proved an extremely important result that relates X_+ to θ in a stochastic way. In particular, Grayson (1988) showed for IRT models for dichotomous items, and assuming unidimensionality, local independence, and nondecreasing IRFs, that for any pair of test scores, such that $0 \leq x_{+a} < x_{+b} \leq J$,

$$g(x_{+a}, x_{+b}; \theta) = \frac{P(X_+ = x_{+b}|\theta)}{P(X_+ = x_{+a}|\theta)} \text{ is nondecreasing in } \theta. \tag{42}$$

This property is known as monotone likelihood ratio (MLR) of X_+ in θ. MLR is important because of two implications (Lehmann, 1959, 1986). The first is stochastic ordering of θ by X_+ (SOL), which is defined for any pair $x_{+a} < x_{+b}$, and any $\theta = t$, as

$$P(\theta \geq t|X_+ = x_{+a}) \leq P(\theta \geq t|X_+ = x_{+b}). \tag{43}$$

It may be noted that SOL takes the observable test score X_+ as the starting point for inferences about the unobservable θ. This means that any IRT model that implies SOL allows the stochastic ordering of respondents on θ by means of X_+. There may be *random error* when ordering respondents on θ using X_+, but Eq. (43) says that there is *no systematic distortion.* An implication of SOL is that the expectations of the conditional distributions of θ are ordered, such that

$$E(\theta|X_+ = x_{+a}) \leq E(\theta|X_+ = x_{+b}). \tag{44}$$

This ordering property is called ordering of the expected latent trait (OEL; Sijtsma & Van der Ark, 2001). The OEL property will be studied more closely in the next section on person ordering based on polytomous items.

The second implication of MLR (Eq. (42)) is stochastic ordering of the manifest score X_+ by the latent trait θ (SOM), defined for any pair $\theta_a < \theta_b$ and any x_+ as

$$P(X_+ \geq x_+|\theta_a) \leq P(X_+ \geq x_+|\theta_b). \tag{45}$$

It may be noted that SOM takes the latent trait as known, which is certainly not realistic in nonparametric IRT. Thus, SOL is more important from a practical point of view and we will further concentrate on SOL.

What are the important implications of Grayson's MLR result for IRT person measurement using dichotomous items? The first is that for any model based on unidimensionality, local independence, and nondecreasing IRFs, ordering on θ using X_+ can be done without systematic distortion. This is even true for nonparametric IRT models, such as the monotone homogeneity model (Definition 1), that do not allow the numerical estimation of θ, but that imply the SOL property, meaning that persons can be ordered on θ by means of X_+. The second implication is that SOL is not dependent on X_+ being a sufficient statistic for θ, as

in the Rasch model (Eq. (10)). Even when models have other sufficient statistics, such as r^* (given that $\boldsymbol{\alpha}$ is known) in the 2-parameter logistic model (Eq. (16)), or when models have no sufficient statistics for θ at all, SOL is still implied by such models. The third implication is that if a dichotomous IRT model implies SOL, then SOL holds for any subset from the J items from the test. This follows simply because the proof of MLR (Grayson, 1988) holds for any J, and any new item subset defines another J. This can be an important property when items are removed from an itemset for which SOL holds. For the remaining subset SOL still holds.

5.3.2 Person ordering using polytomous item scores

A Consistency Result. Junker (1991) showed for the np-GRM (Definition 4), which is the most general of all polytomous IRT models (Hemker, 1996; Hemker et al., in press), that X_+ is a consistent estimator of θ. The proof of this consistency result uses Chang and Mazzeo's (1994) "polytomous" IRF,

$$A_j(\theta) \equiv E(X_j|\theta), \tag{46}$$

the mean $\overline{X}_J$ of J item scores X_j $(j = 1, \ldots, J)$ and, taking the mean across J items, the test response function or test characteristic curve, defined as

$$\overline{A}_J(\theta) = E(\overline{X}_J|\theta) = J^{-1} \sum_{j=1}^{J} A_j(\theta). \tag{47}$$

$\overline{A}_J(\theta)$ is the mean conditional item score, which also equals the true mean item score given θ; that is, based on (34), and conditioning on θ, we might define

$$\overline{T}(\theta_i) \equiv J^{-1} E(X_+|\theta_i) = J^{-1} \sum_{j=1}^{J} E(X_j|\theta_i) = \overline{A}_J(\theta_i). \tag{48}$$

We will assume that $\overline{T}(\theta_i) = \overline{T}_i$ (Eq. (34)) (here, the distinction between a fixed θ and a single examinee is important but will be ignored, as we said earlier; see Holland, 1990). Next, defining the inverse function of the test response function as $\overline{A}_J^{-1}(u)$, which maps test scores u onto latent trait values θ, we have that

$$\widetilde{\theta} = \overline{A}_J^{-1}(\overline{X}_J), \tag{49}$$

and the question now is under which conditions

$$\widetilde{\theta} = \theta. \tag{50}$$

Junker (1991) generalized essential independence (Eq. (21)) and local asymptotic discrimination (Eq. (23)) to polytomous items, with ISRFs $f_j^{CPM}(\theta) = P(X_j \geq x|\theta)$ as defined under the np-GRM (Definition 4), and showed, under fairly general

conditions, that for each θ and each $\epsilon > 0$, and given several technical conditions that we will not go into, that

$$\lim_{J\to\infty} P\{|\overline{A}_J^{-1}(\overline{X}_J) - \theta| > \epsilon|\theta\} = 0, \tag{51}$$

[based on Stout (1990), who established the same result for dichotomous items]. We give this important result, because it says that for infinitely many items, the true mean item score as defined by (48), and assumed to be equal to the true proportion-correct score defined by (34), and which can be estimated from the observable count of the number correct, X_+, contains all the information about θ. Also, for infinite J we know that X_+, which then coincides with the true score from (34), gives the exact ordering of respondents on θ. Moreover, because these results were obtained for the np-GRM (Definition 4) and this is the most general of all known polytomous IRT models (Eq. (32)), by implication we have that for infinite J, total score X_+ gives the exact ordering of respondents on θ *for all polytomous IRT models* from the three classes of ACMs, CPMs, and CRMs. Also, by implication the consistency result holds for the monotone homogeneity model (Definition 1) and all special cases (Stout, 1990). Junker's (1991) consistency result is an asymptotic result, however, and we also need to know whether SOL is implied by polytomous IRT models for any *finite* J.

Stochastic Ordering Results. For polytomous IRT models, Hemker et al. (1996) showed that MLR is implied by Masters' (1982) partial credit model (Eq. ((26)) and a special case of this model with linear restrictions on the δ_{jx} parameter, such that $\delta_{ix} = \delta_i + \tau_x$, known as the rating scale model (Andrich, 1978), but by none of the other well known models from the classes of ACMs and CPMs. Hemker et al. (in press) showed, in addition, that none of the CRMs implies MLR. Hence, because MLR implies SOL and OEL, we know by implication that for the partial credit model and its special cases also the SOL and OEL ordering properties hold. In addition, Hemker et al. (1997) showed that from the class of ACMs no other well known models imply the SOL property. None of the CPMs imply SOL, and Hemker et al. (in press) showed that none of the CRMs imply SOL. Sijtsma and Van der Ark (2001) and Van der Ark (2000) showed that only the partial credit model and its special cases imply the OEL property, but no other well known polytomous IRT model implies OEL.

The conclusion is that at the theoretical level, only the partial credit model and its special cases imply MLR and its implications, SOL and OEL. For these models, X_+ also is the sufficient statistic for θ (Masters, 1982). The implication for all other polytomous IRT models is that if we use X_+ for ordering respondents, this does not fully represent their ordering on θ under one of these models. Not only is there random error present, which is common to all psychological measurement by means of tests, but the ordering that uses X_+ may also systematically *distort* the ordering on θ. The practical question then becomes how often OEL is distorted for consecutive X_+ values, and how large these systematic distortions are.

For this purpose, Sijtsma and Van der Ark (2001) did a small simulation study in the context of the np-GRM (Definition 4), in which items were not extremely easy or difficult, and discrimination ranged from weak to strong; that is, items could be considered representative for the practical use of tests. The number of answer categories varied over design cells ($m + 1 = 3, 4, 5$). The distribution of θ was standard normal, and the number of items was 5. In each design cell, 1000 tests were drawn (i.e., given the IRT model and the θ distribution, item and person parameters were sampled from specified distributions), and it was counted how often $E(\theta|X_+)$ was nondecreasing in X_+ (OEL), which was evaluated for all adjacent values of X_+.

The general conclusions from the first results were:

- When the slopes of the ISRFs were more similar, and the response functions had minimum and maximum asymptotes of 0 and 1, respectively, the percentage of tests showing *no* violations of OEL was relatively large; in particular, this percentage ranged from 77 to 98 percent;
- For the whole simulation study, the number of violations increased with the number of answer categories; for example, from 2 percent ($m + 1 = 3$) to 23 percent ($m + 1 = 5$);
- The proportion of times that two randomly drawn simulees were ordered correctly for the whole study ranged from .96 to over .999; and
- When the expected ordering did not appear, the typical result for ordering $E(\theta|X_+)$ was (e.g., for $X_+ = 0, \ldots, 20$),

−0.83	0.90	1.43	1.73	1.95	2.24	2.55	2.74	2.92	3.12
3.32	3.46	3.67	**3.93**	**3.88**	4.18	4.52	4.93	4.94	4.95	4.97

The tentative conclusions were that at the individual level there were not many violations, and when violations appeared they usually were small. This means that for practical purposes OEL seems to hold well, with the exception of mostly small violations. More comprehensive results for models from the ACM, the CPM, and the CRM classes are discussed by Van der Ark (2000). He found that the probability that two randomly drawn simulees were incorrectly ordered decreased as J increased.

6. Model Fit Assessment and Scale Construction Using IRT Data

Measures of persons and items may be of little practical use if the IRT measurement model does not fit the data. A model-data fit investigation may begin with exploring the properties of the IRFs. In a nonparametric context, this means using nonparametric regression methods, such as *kernel smoothing* (e.g., Fox, 2000), for estimating a smooth function from the data [see Ramsay (1991), and Douglas

(1997)], or *binning* (Fox, 2000) where each bin is a group of respondents with the same summary score and within each group the proportion answering the item j of interest correct (for dichotomous items) estimates a discrete point of the IRF. Junker and Sijtsma (2000) discussed this approach, known as the item-restscore regression, in much detail; also see Rosenbaum (1984). For investigating local independence, Stout, Habing, Douglas, Kim, Roussos, and Zhang (1996) discussed methods based on conditional inter-item covariances, $\mathrm{Cov}(X_j, X_k|\theta)$, and averaged over θ, that can be used to determine the dimensionality of a data set; and Douglas, Kim, Habing, and Gao (1998; also, see Habing, 2001) discussed the conditional covariance as a diagnostic tool for investigating, for example, speededness as a function of θ.

Typical of the nonparametric IRT context is the existence of several automated item selection procedures, intended for clustering unidimensional item subsets from a larger item pool. The most popular procedures are contained in the programs MSP (Molenaar & Sijtsma, 2000), using scalability coefficients, and DETECT (Zhang & Stout, 1999a, 1999b), using averaged conditional covariance. Also, see Bolt (2001) for a geometric representation of multidimensional test structure.

We have seen that nonparametric IRT corroborates the use of X_+ for measuring persons on θ when items are dichotomous and that for polytomous items the use of this statistic leads to useful person ordering, without much danger of systematically ordering persons incorrectly. For parametric models, where the researcher pursues estimates of latent parameters such as θ, goodness-of-fit methods have been extensively studied and proposed for the Rasch model, and also, but to a lesser extent for several other models. For the Rasch model, Glas and Verhelst (1995a) discuss several useful statistical tests for local independence at the level of the test (i.e., all J items are evaluated simultaneously) and for pairs of items (also, see Molenaar, 1983b); and tests for the simultaneous evaluation of the logistic shape of the J IRFs, and for individual IRFs (also, see Molenaar, 1983b). Tests for other parametric IRT models, both for dichotomous and polytomous items, and for unidimensional and multidimensional data, are surveyed in Van der Linden and Hambleton (1997). Examples are tests for investigating the IRFs of the 2- and 3-parameter logistic models (Orlando & Thissen, 2000); tests for the slopes of the IRFs of the one parameter logistic model (Verhelst & Glas, 1995), which is a hybrid model with imputed slope indices, in between the Rasch model and the 2-parameter logistic model; and tests for the fit of the partial credit model and related polytomous Rasch models [see Glas & Verhelst (1995b) for an overview].

Some recent results come from Douglas and Cohen (2001), who use kernel smoothing for nonparametrically estimating the IRFs, and also estimate the item parameters under a hypothesized parametric model. Then they use resampling from the estimated parametric model to find for each item the parametric IRF that is nearest to the nonparametric IRF. When the nearest parametric IRF is far from

the nonparametric IRF, the latter estimate serves as a diagnostic for interpreting misfit. Vermunt (2001) uses ordered latent classes instead of the continuous θ, and order restrictions on response probabilities for polytomous items from each of the three classes discussed earlier, to estimate models using maximum likelihood, and tests their fit by means of likelihood-ratio statistics.

7. Practical Applications of IRT models

This paper started with the basic ideas of measurement, such as hypothetical constructs, operationalization, definition of an item domain, and the construction of a test or questionnaire. We will end with listing some of the practical applications of the IRT tool kit, after the model-data fit has been established and person and item measures have been estimated. Model-data fit research and parameter estimation lead to the final composition of the *test*. We have already made clear that tests are important in any area of the social and behavioral sciences but also, for example, in medical research.

When an IRT model fits the data, the measurement scale of the parameters, which is *implied* by the model, is assumed to hold for the particular test. For example, for the Rasch model (Eq. (10)), θ is measured on a difference scale, $\theta^* = \theta + c$ (c is a fixed real), but other monotone transformations of θ and the IRFs are possible, such as $\xi = \exp(\theta)$ and $\eta_j = \exp(-\delta_j)$, yielding $P_j(\xi) = (\xi\eta_j)/(1 + \xi\eta_j)$ and a ratio measurement level. The θ scale may constitute the basis for further scientific research, but for practical purposes the complicated logit θ metric may be transformed, for example, to the more convenient true score scale, which is justified by the SOL property (Hemker et al., 1997). Thus, the *calibration* of the scale as implied by the IRT model, may be conveniently transformed by the researcher to the well-known X_+ scale.

The θ scale seems to be useful in particular for such applications as the *equating* of scales based on different tests for the same latent trait, with the purpose of making the measurements of pupils who took these different tests directly comparable. This may eventually lead to the formation of an *item bank*, consisting of hundreds of items which measure the same latent trait, but with varying difficulty and other item properties. From such a bank a computer can assemble new tests fit to a particular application (Van der Linden, 1998). Also, for individual examinees tests can be assembled by first presenting a few items of average difficulty to an examinee with an unknown θ value and then, on the basis of a preliminary estimate of θ, improving θ estimation stepwise by selecting items in each step that are tailored to the estimated θ. This procedure, known as *adaptive testing* (e.g., Van der Linden & Glas, 2000), uses fewer items than the conventional standard (paper-and-pencil) tests for estimating θ with adequate accuracy, and is mostly convenient in large scale testing programs in education and job selection and placement.

In a multicultural society people with different backgrounds are often assessed

with the same measurement instruments, and an important issue is whether persons having the same θ level, but differing on relevant covariates such as gender, and social-economic and ethnic background, have the same response probabilities on the items from the test. If not, the test is said to exhibit *differential item functioning*. This can be investigated with parametric IRT methods (for an overview, see Holland & Wainer, 1993) and nonparametric IRT methods (Shealy & Stout, 1993). Items functioning differently between groups may be replaced by items which function identically between groups. For example, differential item functioning may occur when people from two groups are tested with an arithmetics test that also requires verbal ability, and one group has a systematically lower verbal ability level because the language of the test is not their native tongue. Then it can be expected that a representative of this group will have a lower response probability on the items than someone from the other group who has the same arithmetic level.

Respondents may be confused by the item format; they may be afraid of situations, including a test, in which they are evaluated; they may underestimate the level of the test and miss the depth of several of the questions; they may have learned an incorrect solution strategy; they may cheat by copying answers from an able respondent sitting next to them or from notes hidden in their lap; or they may guess to the answers on most of the items. Each of these mechanisms, as well as several others, may produce a pattern of J item scores that is unexpected give the predictions from IRT models. For example, confusion by the item format may lead to many incorrect answers on the first few items from the test, which may also be the easiest items. Likewise, cheating may lead to a few correct answers to the most difficult items, while many much easier items are answered incorrectly. Nonparametric *person-fit methods* (e.g., Meijer, 1994; Sijtsma & Meijer, in press) and parametric person-fit methods (e.g., Molenaar & Hoijtink, 1990; Drasgow, Levine, & Zickar, 1996) have been proposed to identify nonfitting item score patterns (for an overview, see Meijer & Sijtsma, in press). The identification of such patterns may contribute to the diagnosis of the problem behind the pattern and, depending on the cause, a θ estimate may be corrected (e.g., in case of cheating), an individual may be given a second chance (e.g., in case of test anxiety), or a pupil may be subjected to remedial teaching (e.g., in case of an incorrect solution strategy).

Finally, the recent development of *cognitive modeling* has taken measurement one or two steps beyond assigning scores to people, in that the cognitive process or the solution strategy that produced these scores is part of the IRT model and measurement is connected to a psychological explanation. Several possibilities have been explored: modeling the item difficulty δ of the Rasch model as a linear combination of subtask parameters (Fischer, 1995); assuming multiple latent traits for subabilities connected to parts of the cognitive process (Kelderman & Rijkes, 1994); and modeling the steps of a solution scheme as Rasch models that are combined multiplicatively, resulting in a noncompensatory model (Embret-

son, 1997). A nonparametric approach has been proposed by Junker and Sijtsma (2001b). Each of these approaches models the correct/incorrect scores (Kelderman & Rijkes' model is also suited for polytomous scores) that are the outcome of processes or strategies, but it is easy to see that also collecting data on the cognitive processes and solution strategies themselves and incorporating these data into psychometric models, will probably lead to new models with better explanatory power for the outcomes. Such models may also be envisaged in the area of attitude and personality measurement, where it can easily be imagined that, for example, different motivations led to the same rating on a particular item. Incorporating variables for these motivational factors into a model again might improve explanatory power and understanding of measurement outcomes.

REFERENCES

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

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, **17**, 251-269.

Akkermans, L. M. W. (1998). *Studies on statistical models for polytomously scored test items.* Unpublished doctoral dissertation, University of Twente, the Netherlands.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, **32**, 283-301.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, **43**, 561-573.

Baker, F. B. (1992). *Item response theory. Parameter estimation techniques.* New York: Marcel Dekker.

Béguin, A. A. (2000). *Robustness of equating high-stakes tests.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bolt, D. M. (2001). Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement.*

Chang, H. & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika*, **59**, 391-404.

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

Douglas, J. & Cohen, A. (2001). Nonparametric ICC estimation to assess fit of parametric models. *Applied Psychological Measurement.*

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, **23**, 129-151.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, **9**, 47-64.

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

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, **56**, 495-515.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-321). New York: Springer.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 131-155). New York: Springer.

Fox, J. (2000). *Nonparametric simple regression.* Thousand Oaks, CA: Sage.

Glas, C. A. W. & Verhelst, N. D. (1995a). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 69-95). New York: Springer.

Glas, C. A. W. & Verhelst, N. D. (1995b). Tests of fit for polytomous Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 325-352). New York: Springer.

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

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Habing, B. (2001). A survey of nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement.*

Hemker, B. T. (1996). *Unidimensional IRT models for polytomous items, with results for Mokken scale analysis.* Unpublished doctoral dissertation, Utrecht University, The Netherlands.

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.*

Hoijtink, H. & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 53-68). New York: Springer.

Hoijtink, H. & Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika*, **62**, 171-189.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, **55**, 577-601.

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

Holland. P. W. & Wainer, H. (Eds.). *Differential item functioning.* Hillsdale NJ: Erlbaum.

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

Jannerone, R. J. (1997). Models for locally dependent responses: conjunctive item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 465-479). New York: Springer.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, **56**, 255-278.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *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.

Junker, B. W. & Sijtsma, K. (2001a). Nonparametric IRT in action: An overview of the special issue. *Applied Psychological Measurement.*

Junker, B. W. & Sijtsma, K. (2001b). Cognitive assessment models with few assumptions, and connections with nonparametric IRT. *Applied Psychological Measurement.*

Kelderman, H. & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, **59**, 149-176.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* New York: Wiley.

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, **48**, 233-245.

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

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

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person fit statistic. *Applied Psychological Measurement*, **18**, 311-314.

Meijer, R. R. & Sijtsma, K. (in press). Methodology review: evaluating person fit. *Applied Psychological Measurement.*

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.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, **6**, 417-430.

Molenaar, I. W. (1983a). *Item steps.* (Heymans Bulletin 83-630-EX). Groningen, The Netherlands: University of Groningen, Department of Statistics and Measurement Theory.

Molenaar, I. W. (1983b). Some improved diagnostics for failure of the Rasch model. *Psychometrika*, **48**, 49-72.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 39-51). New York: Springer.

Molenaar, I. W. & Hoijtink. H. (1990). The many null distributions of person fit indices. *Psychometrika*, **55**, 75-106.

Molenaar, I. W. & Sijtsma, K. (2000). *MSP5 for Windows. User's manual.* Groningen, The Netherlands: iec ProGAMMA.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, **16**, 159-176.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, **24**, 50-64.

Patz, R. J. & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, **24**, 342-366.

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

Ramsay, J. O. (1995). A geometrical approach to item response theory. *Behaviormetrika*, **23**, 3-16.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, **49**, 425-435.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, No.17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No.18.*

Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika*, **60**, 281-304.

Shealy, R. T. & Stout, W. F. (1993). An item response theory model for test bias and differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale NJ: Erlbaum.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

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

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. & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, **49**, 79-105.

Sijtsma, K. & Meijer, R. R. (in press). The person response function as a tool in person-fit research. *Psychometrika.*

Sijtsma, K. & Van der Ark, L. A. (2001). Progress in IRT analysis for polytomous item scores: dilemma's 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. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, **52**, 589-617.

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

Stout, W. F., Habing, B., Douglas, J., Kim, H., Roussos. L., & Zhang. J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, **20**. 331-354.

Suppes, P. & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, **48**, 191-199.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometrika monograph, No.1.*

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

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139-152). New York: Springer.

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). An overview of relationships in polytomous IRT, and some applications. *Applied Psychological Measurement.*

Van der Linden, W. L. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, **22**, 195-211.

Van der Linden, W. L. & Glas, C. A. W. (Eds.) (2000). *Computer adaptive testing. Theory and practice.* Boston, MA: Kluwer.

Van der Linden, W. J. & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory.* Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.

Verhelst, N. D. & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 215-237). New York: Springer.

Vermunt, J. K. (2001). On the use of (order-)restricted latent class models for defining and testing (non-)parametric IRT models. *Applied Psychological Measurement.*

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika*, **54**, 427-450.

Zhang, J. & Stout, W. F. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, **64**, 129-152.

Zhang, J. & Stout, W. F. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, **64**, 213-249.

Received January 2001. Final version accepted March 2001.