Progress in NIRT analysis of polytomous item scores

Sijtsma, K.; van der Ark, L.A.

Published in:
Essays on item response theory

Publication date:
2001

16
Progress in NIRT Analysis of Polytomous Item Scores: Dilemmas and Practical Solutions

Klaas Sijtsma and L. Andries van der Ark[1]

ABSTRACT This chapter discusses several open problems in nonparametric polytomous item response theory: (1) theoretically, the latent trait θ is not stochastically ordered by the observed total score X_+; (2) the models do not imply an invariant item ordering; and (3) the regression of an item score on the total score X_+ or on the restscore R is not a monotone nondecreasing function and, as a result, it cannot be used for investigating the monotonicity of the item step response function. Tentative solutions for these problems are discussed. The computer program MSP for nonparametric IRT analysis is based on models that neither imply the stochastic ordering property nor an invariant item ordering. Also, MSP uses item restscore regression for investigating item step response functions. It is discussed whether computer programs may be based (temporarily) on models that lack desirable properties and use methods that are not (yet) supported by sound psychometric theory.

1 Mokken Scale Analysis for Polytomous Item Scores

Nonparametric item response theory (NIRT) for polytomous ordered item scores is characterized by the combination of practical methods for the analysis of empirical questionnaire data and complex theoretical developments, which have many pitfalls and several unsolved problems. The combination of unfinished and sometimes seemingly unsolvable theoretical problems and the need for practical solutions for data analysis problems poses an interesting challenge to the applied statistician. We focus on the nonparametric methods for scaling persons and items, initially proposed for dichotomous

[1]Both authors are at the Department MTO, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: *k.sijtsma@kub.nl* and *a.vdark@kub.nl*

item scores by Mokken (1971) and extended to polytomous ordered item scores by Molenaar (1982, 1986, 1997). In Section 1 we start by defining the two NIRT models that are central in the work of Mokken and Molenaar and then discuss the scaling procedures built into the computer program MSP5 for Windows (abbreviated throughout as MSP; Molenaar & Sijtsma, 2000). In Section 2 we discuss the state of the art for the theory underlying the two NIRT models and provide suggestions and results for open problems. In Section 3 we discuss the methodological issue of whether it is wise, desirable, or avoidable to implement as yet imperfect data analysis methods into a computer program.

1.1 Definition of the Models

We assume that a test or questionnaire consists of k items, each of which has $m+1$ ordered answer categories, reflecting the ordering on an underlying latent trait θ, such as an attitude or a personality trait. Scores are assigned to answer categories such that the ordering of the scores corresponds with the hypothesized ordering of the answer categories on θ. Let the items be indexed I_i, $i = 1, \ldots, k$, and let the random variable giving the score on item I_i be denoted X_i, with realizations $x_i = 0, \ldots, m$. We assume that the k items measure the same θ. An observable summary score based on the k item scores is used to describe the respondent's position on θ. Usually, this summary score is the unweighted sum of k item scores, denoted X_+, and is defined as

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

In addition to unidimensional measurement, we assume local independence of the item scores. Let $\mathbf{X} = (X_1, \ldots, X_k)'$ and $\mathbf{x} = (x_1, \ldots, x_k)'$. Furthermore, $P(X_i = x \,|\, \theta)$ denotes the conditional probability that a score of x has been obtained on item I_i. The assumption of local independence can be expressed as

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

Thus, θ alone explains the relationships between items. An implication of local independence is that the covariances between items conditional on θ equal 0. By integrating over the distribution of θ, denoted $G(\theta)$, the unconditional multivariate distribution of the k item scores is obtained,

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{i=1}^{k} P(X_i = x_i \,|\, \theta)\, dG(\theta) \ .$$

Suppes and Zanotti (1981) showed that, unless further restrictions are placed upon the conditional probabilities $P(X_i = x_i \mid \theta)$, the distribution of θ, or both, the multivariate distribution of the k item scores is not restricted. Usually, item response models further restrict the conditional probabilities by introducing additional assumptions.

With respect to the assumptions restricting the conditional probabilities, first we consider three possibilities (see also Agresti, 1990; Akkermans, 1998; Hemker & Sijtsma, 1998; Mellenbergh, 1995; Molenaar, 1983; Van Engelenburg, 1997) and then we concentrate on the model that is relevant in the present NIRT context. In our discussion, we use both $P(X_i = x_i \mid \theta)$ and $P(X_i \geq x_i \mid \theta)$. These probabilities are related by two simple equations:

$$P(X_i \geq x \mid \theta) = \sum_{n=x}^{m} P(X_i = n \mid \theta) \ ,$$

and

$$P(X_i = x \mid \theta) = P(X_i \geq x \mid \theta) - P(X_i \geq x + 1 \mid \theta) \ . \tag{16.1}$$

To further restrict the conditional response probabilities, the *cumulative probability* definition assumes that $P(X_i \geq x \mid \theta)$ is a monotone nondecreasing function of θ,

$$P(X_i \geq x \mid \theta) \text{ nondecreasing in } \theta, \text{ for all } x, i, \tag{16.2}$$

with $x = 0$ representing an uninteresting case; that is, $P(X_i \geq 0 \mid \theta) = 1$.

The *adjacent category* definition assumes that the conditional probability of having a score of x given that either a score of $x - 1$ or a score of x has been observed, and given θ, is a monotone nondecreasing function of θ,

$$P(X_i = x \mid X_i = x - 1 \vee x; \theta) \text{ nondecreasing in } \theta, \text{ for all } x, i, \tag{16.3}$$

where

$$P(X_i = x \mid X_i = x - 1 \vee x; \theta) = \frac{P(X_i = x \mid \theta)}{P(X_i = x - 1 \mid \theta) + P(X_i = x \mid \theta)} \ ,$$

with $x = 0$ producing a probability of 1.

Finally, the *continuation ratio* definition assumes that the conditional probability of having at least an item score of x, given that we know that at least a score of $x - 1$ was obtained, is a nondecreasing function of θ,

$$P(X_i \geq x \mid X_i \geq x - 1; \theta) \text{ nondecreasing in } \theta, \text{ for all } x, i, \tag{16.4}$$

where

$$P(X_i \geq x \mid X_i \geq x-1;\theta) = \frac{P(X_i \geq x \mid \theta)}{P(X_i \geq x-1 \mid \theta)} \ ,$$

with $x = 0$ again producing a probability of 1.

By choosing parametric functions for the conditional probabilities, parametric item response models for polytomous ordered item scores are obtained (Mellenbergh, 1995; Molenaar, 1983). For example, Samejima's (1969) graded response model is readily seen to be a parametric special case of (16.2),

$$P(X_i \geq x \mid \theta) = \frac{\exp[\alpha_i(\theta - \beta_{ix})]}{1 + \exp[\alpha_i(\theta - \beta_{ix})]} \ ,$$

where α_i denotes a nonnegative slope parameter and β_{ix} denotes a location parameter. Hemker and Sijtsma (1998) defined three NIRT models based on the assumptions of unidimensionality, local independence, and the order restrictions in (16.2), (16.3), and (16.4), respectively. The hierarchical relationships between the three classes of nonparametric and parametric item response models for polytomous items were investigated by Hemker, Sijtsma, Molenaar, and Junker (1996, 1997) and Hemker, Van der Ark, and Sijtsma (2000). Van Engelenburg (1997) investigated the psychological response processes underlying each type of model and Akkermans (1998) studied the item scoring rules for these models.

Molenaar (1986) considered the cumulative probability definition to be more plausible than the adjacent category definition, because the conditioning only on $x-1$ and x, thus isolating these two scores from their natural item context, is relatively unrealistic. Also, he dismissed the continuation ratio definition because conclusions based on it depend on the direction of the item scale whereas for most attitude items reversal of the scale direction should not lead to conclusions other than those based on the original scale direction (also, see Hemker, 2001; Hemker & Sijtsma, 1998).

Molenaar (1982, 1986, 1997) defined his model of *monotone homogeneity* (MH model) by assuming unidimensionality, local independence, and monotonicity in the sense of the cumulative probability definition, henceforth simply called monotonicity. In the MH model a higher θ implies a higher expected item score X_i, and also a higher expected unweighted total score X_+ (Hemker et al., 1997). In practice, however, the total score X_+ is used for ordering respondents on θ, so the inference uses X_+ to say something about θ. It has been shown (Hemker et al., 1997) that inference about the ordering on θ based on X_+ is not implied by the MH model for polytomous item scores. We thus run into the first theoretical problem of polytomous NIRT modeling, to which we return in Section 2.

In this chapter, the conditional probability $P(X_i \geq x \mid \theta)$ is called the item step response function (ISRF), because it describes the conditional

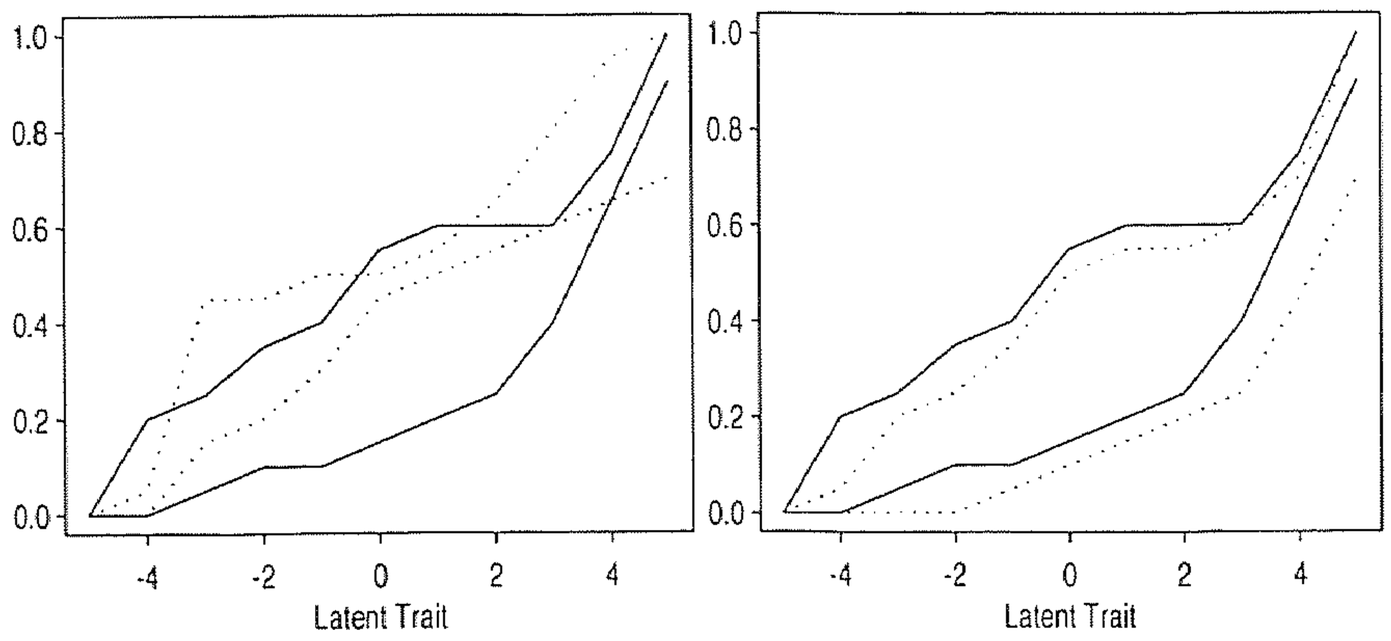

FIGURE 16.1. Two items with two ISRFs each (solid for one item, dotted for the other), satisfying monotonicity in the left panel, and double monotonicity in the right panel.

probability of taking the imaginary step x between the answer categories $x-1$ and x. Let Y_{xi} denote the binary variable indicating whether (score 1) or not (score 0) step x has been taken. Then we may write $P(X_i \geq x \,|\, \theta) = P(Y_{xi} = 1 \,|\, \theta)$. In the MH model the ISRFs are nondecreasing functions by definition.

Molenaar (1982) also defined the *double monotonicity* (DM) model which, like the MH model, is based on unidimensionality, local independence, and monotonicity, and adds the fourth assumption that the $k \times m$ ISRFs do not intersect. That is, for any pair of ISRFs, say, g of item I_i and h of item I_j, of which we know for one θ_0 that $P(X_i \geq g \,|\, \theta_0) < P(X_j \geq h \,|\, \theta_0)$, nonintersection means that

$$P(X_i \geq g \,|\, \theta) \leq P(X_j \geq h \,|\, \theta) \;, \quad \text{for all } \theta, g, h;\; i \neq j.$$

It may be noted that the ISRFs of the same item do not intersect by definition. Nonintersection of $k \times m$ ISRFs means that the ISRFs have the same ordering, with the exception of possible ties, for each value of θ. The DM model thus implies an ordering of item steps, which is valid for all θs and thus within all possible subpopulations from the population of interest. Monotonicity and double monotonicity are illustrated in Figure 16.1. The left panel of Figure 16.1 depicts the ISRFs of two items, indicated by a solid line for one item and a dotted line for the other, that satisfy monotonicity but not double monotonicity. The right panel of Figure 16.1 depicts the ISRFs of two items that satisfy both monotonicity and double monotonicity.

1.2 Scaling Procedure

Several researchers (e.g., Hemker, Sijtsma, & Molenaar, 1995; Molenaar, 1982, 1986, 1991; Molenaar & Sijtsma, 1988, 2000; Sijtsma, Debets, & Molenaar, 1990) discussed methods for determining the fit of the MH and DM models to polytomous item scores. These methods were implemented in three consecutive versions of the computer program MSP (Debets & Brouwer, 1989; Molenaar, Debets, Sijtsma, & Hemker, 1994; Molenaar & Sijtsma, 2000). The following methods have been implemented.

Scalability coefficients for pairs of items (denoted H_{ij}), for individual items with respect to the other $k-1$ items in the questionnaire (denoted H_i), and for a set of k items as a whole (denoted H) (Molenaar, 1991). These scalability coefficients can be written as normed (sums of) covariances between pairs of items. H_{ij} is defined as the ratio of the covariance of items I_i and I_j, $\text{Cov}(X_i, X_j)$, and the maximum covariance given the marginals of the bivariate cross-classification table of scores on items I_i and I_j, $\text{Cov}(X_i, X_j)_{\max}$; that is,

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

The item scalability coefficient H_i is defined as

$$H_i = \frac{\sum_{j \neq i} \text{Cov}(X_i, X_j)}{\sum_{j \neq i} \text{Cov}(X_i, X_j)_{\max}} ,$$

and the scalability coefficient H for k items is defined as

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{Cov}(X_i, X_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{Cov}(X_i, X_j)_{\max}} .$$

Hemker et al. (1995) have shown that given the MH model,

$$0 \leq H_{ij} \leq 1, \quad \text{for all } i, j;\ i \neq j ,$$

$$0 \leq H_i \leq 1, \quad \text{for all } i,$$

and

$$0 \leq H \leq 1 .$$

Under the MH model, H-values of 0 mean that for at least $k-1$ items the ISRFs are constant functions of θ (Hemker et al., 1995). Higher H-values mean that the slope of the ISRFs tends to be steeper, which implies

that the item steps and the items discriminate better among values of θ. Obviously, high H-values are desirable for a questionnaire and its items. When constructing questionnaires, we require both

$$H_{ij} > 0, \quad \text{for all } i, j;\ i \neq j\ , \tag{16.5}$$

and

$$H_i \geq c,\ c > 0, \quad \text{for all } i. \tag{16.6}$$

Because $H \geq \min(H_i)$, (16.6) implies that $H \geq c$. The usual choice is $c = 0.3$. Other values have been discussed by Hemker et al. (1995).

An **automated item selection procedure** of which the H-coefficients are the core. The item selection procedure is a bottom-up algorithm, which selects items from an item pool into subsets that satisfy (16.5) and (16.6). The item selection starts with a kernel of items, usually the two items with the highest H_{ij} value which is significantly greater than 0 (using a standard normal test; Molenaar & Sijtsma, 2000), and adds items from the remaining items one-by-one, in each selection step maximizing H, until no more items can be selected that satisfy (16.5) and (16.6). If there are items left after the first scale has been selected, the algorithm tries to select a second scale from the remaining items, a third, a fourth, and so on, until no more scales can be formed or no more items remain.

Descriptive methods, graphical displays, and statistical tests for investigating whether the data support the assumption that single ISRFs are nondecreasing functions and whether several ISRFs are nonintersecting. Among the methods for investigating intersection of the ISRFs are two that directly address the estimation of two ISRFs, say, g and h of the items I_i and I_j, respectively, through the regression of an item step score on the total score on the $k-2$ items excluding items I_i and I_j. Another method is an extension of the $P(++)$ and $P(--)$ matrices which Mokken (1971, pp. 132–133) introduced for investigating the nonintersection of the item response functions (IRFs) of dichotomously scored items. For dichotomous items, a method for investigating intersection of k ISRFs based on H-coefficients is available (Sijtsma & Meijer, 1992).

Comparative scalability results for several relevant subgroups from a group, for example, based on country, age, social-economic status, religion, or gender. This enables the comparison of groups on important scalability criteria, the assessment of the usefulness of a scale in different groups, and a first evaluation of differential item functioning (DIF) between groups.

A method for estimating the **reliability** of the total score X_+, which assumes that the ISRFs do not intersect (Molenaar & Sijtsma, 1988). Thus, before interpreting the reliability estimate it should be checked whether the DM model fits.

2 Three Open Theoretical Problems in NIRT

For each of the three problems discussed here, we first provide the theory, next we discuss current research aimed at solving the problem and, finally, we discuss the method for investigating the problem in real data.

2.1 Ordering Persons on the Latent Scale

Theory. Typical of NIRT is that ISRFs are subject to order restrictions but ISRFs are not parametrically defined. This means that the estimation of the latent person parameter θ by means of maximum likelihood methods, well known from parametric models, is not feasible here. Thus, NIRT necessarily resorts to observable person summary scores, such as the much used X_+. In an item response theory (IRT) context, one would expect the following order relationship between X_+ and θ: for higher values of X_+ we expect the conditional expected θ to increase or remain equal. That is, for two fixed values of X_+, say, s and t, we expect that

$$s < t \Rightarrow E(\theta \mid X_+ = s) \leq E(\theta \mid X_+ = t) \; . \tag{16.7}$$

We call this property *ordering of the expected latent trait* (OEL). OEL is implied by two better known properties. Hemker et al. (1996) investigated the monotone likelihood ratio (MLR) of X_+ given θ. MLR means that under the assumptions of unidimensionality, local independence, and monotonicity of the ISRFs, the ratio

$$\frac{P(X_+ = t \mid \theta)}{P(X_+ = s \mid \theta)}$$

is nondecreasing in θ whenever $s \leq t$. MLR implies OEL. They showed that none of the well-known polytomous IRT models from the class of cumulative probability models, including the MH and DM models, implies MLR. Of the popular models from the class of adjacent category models, only the partial credit model (PCM; Masters, 1982) and special cases of this model imply MLR. Hemker et al. (2000) showed that none of the known continuation ratio IRT models implies MLR.

Hemker et al. (1997) investigated whether θ was stochastically ordered by X_+. This property is called stochastic ordering of the latent trait (SOL). MLR implies SOL, but SOL does not imply MLR (Lehmann, 1959, p. 74). SOL means that, for any fixed value of θ, say, θ_0,

$$P(\theta \geq \theta_0 \mid X_+ = s) \leq P(\theta \geq \theta_0 \mid X_+ = t) \; , \quad \text{for all } s < t \; . \tag{16.8}$$

It may be noted that (16.8) expresses that with an increasing total score the cumulative distribution of θ is uniformly higher for s than for t (also, see Sijtsma & Hemker, in press). OEL (16.7) expresses the relationship between

the means of these distributions. Hemker et al. (1997) showed that none of the well-known cumulative probability models (including the MH and the DM models) implies SOL, and from the class of adjacent category models of the well-known models only the partial credit model (Masters, 1982) and its special cases imply SOL. For continuation ratio models, Hemker et al. (2000) showed that none of the known models implies SOL. Finally, it may be noted that for *dichotomous* items, the MH model and, by implication, the DM model imply MLR (see Grayson, 1988), and thus SOL and OEL.

Present research. Current research addresses two topics. The first is whether the MH and DM models *theoretically* imply the weaker OEL property (16.7). With an example we show that they do not, so the second topic is whether the MH model implies OEL in most *practical* situations.

The next example provides an answer to the first question. We use a numerical example with one item ($k = 1$) and three ordered answer categories ($m = 2$), and a discrete θ with two values θ_1 and θ_2, both with probability 0.5 and $\theta_1 < \theta_2$, to show that the DM model does not imply OEL (16.7); consequently, the more general MH model does not imply OEL either.

Example. *The DM model does not imply OEL.* Define $P(X \geq x \mid \theta)$ as

x	0	1	2
$P(X \geq x \mid \theta_1)$	1.0	.88	.40
$P(X \geq x \mid \theta_2)$	1.0	.94	.88

Next, $P(X = x \mid \theta)$ is computed using (16.1) as

x	0	1	2
$P(X = x \mid \theta_1)$	.12	.48	.40
$P(X = x \mid \theta_2)$	.06	.06	.88

$P(\theta \mid X = x)$ is obtained by $P(X = x \mid \theta)P(\theta)/P(X = x)$ and results in

x	0	1	2
$P(\theta_1 \mid X = x)$	.67	.89	.31
$P(\theta_2 \mid X = x)$	.33	.11	.69

If we let $\theta_1 = 0$ and $\theta_2 = 1$ then $E(\theta \mid X) = P(\theta_2 \mid X = x)$, and the last table shows that $E(\theta \mid X)$ is decreasing between $X = 0$ and $X = 1$. It may be readily checked that for all other choices of θ_1 and θ_2 with $\theta_1 < \theta_2$, $E(\theta \mid X)$ decreases between $X = 0$ and $X = 1$.

Next, we used a simulation study to investigate whether the MH model implies OEL in practical situations. The design of the simulation study was the following. We used a parametric definition of an ISRF within the context of the class of continuation ratio models (16.4), which was loosely inspired by Samejima's (1995) acceleration model, and which allowed us to simulate with great flexibility ISRFs within the context of the cumulative probability model (16.2), by using

$$P(X_i \geq x \mid \theta) = \prod_{n=1}^{x} \frac{P(X_i \geq n \mid \theta)}{P(X_i \geq n-1 \mid \theta)} \; . \tag{16.9}$$

Our definition of the continuation ratio ISRF uses five item parameters for describing the lower and upper asymptotes of the ISRFs (γ_L and γ_U, respectively), the location (β_{ix}) and the slope (α_{ix}) of the ISRF, and the acceleration parameter (ξ_{ix}), which pushes down ($\xi_{ix} > 1$) or lifts up ($0 < \xi_{ix} < 1$) the entire ISRF. The continuation ratio ISRF is now defined as

$$\frac{P(X_i \geq x \mid \theta)}{P(X_i \geq x-1 \mid \theta)} = \gamma_L + (\gamma_U - \gamma_L) \left\{ \frac{\exp[\alpha_{ix}(\theta - \beta_{ix})]}{1 + \exp[\alpha_{ix}(\theta - \beta_{ix})]} \right\}^{\xi_{ix}} . \tag{16.10}$$

Inserting (16.10) into (16.9) does not produce an ISRF with a parameter structure that has a sensible interpretation in the context of parametric IRT, but the result is a monotone increasing function of θ which, therefore, can be used for the investigation of OEL in the context of Molenaar's MH model for polytomous item scores.

In each cell of our simulation design we simulated 1000 tests, and evaluated for each test whether $E(\theta \mid X_+)$ is nondecreasing in X_+. If $E(\theta \mid X_+)$ decreased for some increasing values of X_+, we determined the probability that two randomly selected simulees are incorrectly ordered by X_+ given their ordering on θ. For example, a simulated test may yield

x_+	$P(X_+ = x_+)$	$E(\theta \mid X_+ = x_+)$	Nondecreasing
0	.08	$-.83$	
1	.21	.60	Yes
2	.42	**1.42**	Yes
3	.21	**1.41**	No
4	.08	1.74	Yes

For this test we say that $E(\theta \mid X_+)$ is not nondecreasing in X_+ and that the probability of incorrectly ordering two randomly selected simulees is $2 \times .42 \times .21 = .1764$.

In our simulation study, five items had $m + 1$ ordered categories whose ISRFs were determined by (16.9) and (16.10). Parameters α^*_{ix} were inde-

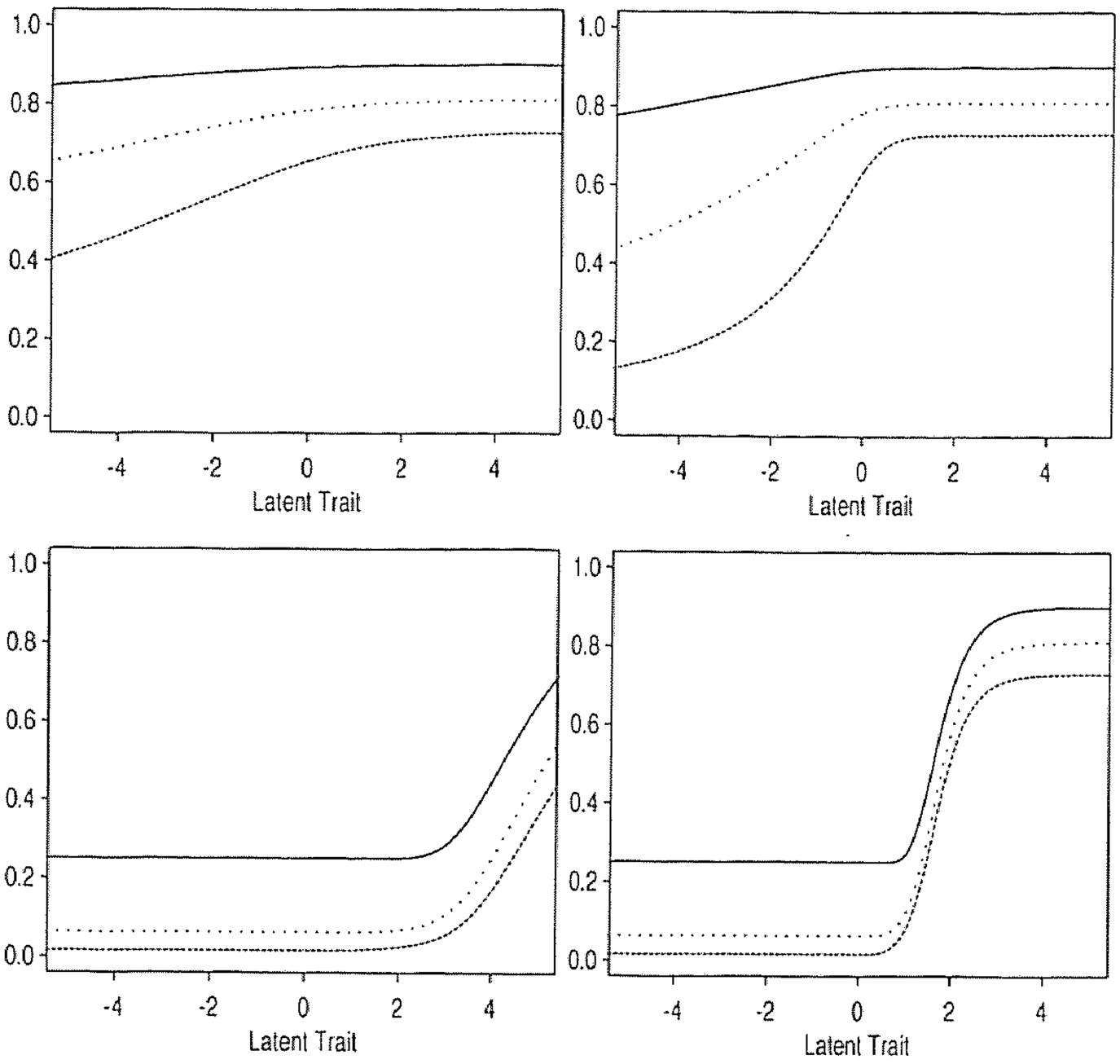

FIGURE 16.2. Typical ISRFs (simulation study, $m = 3$), each set based on $\gamma_L = .25$ and $\gamma_U = .9$, and also $\mu_{\alpha^*} = 0$, $\mu_{\xi^*} = -3$ (top left), $\mu_{\alpha^*} = 1$, $\mu_{\xi^*} = -3$ (top right), $\mu_{\alpha^*} = 0$, $\mu_{\xi^*} = 3$ (bottom left), $\mu_{\alpha^*} = 1$, $\mu_{\xi^*} = 3$ (bottom right).

pendent draws from $\mathcal{N}(\mu_{\alpha^*}, .2)$. To avoid negative values, the slope parameters in (16.10) were defined as $\alpha_{ix} = \exp(\alpha^*_{ix})$. The location parameters in (16.10), β_{ix}, were independent draws from $\mathcal{N}(0, .5)$. Parameters ξ^*_{ix} were independent draws from $\mathcal{N}(\mu_{\xi^*}, 1)$ and the acceleration parameters in (16.10) were defined as $\xi_{ix} = \exp(\xi^*_{ix})$. The latent trait θ was assumed to have a standard normal distribution. Different types of tests ($k = 5$) were generated by varying m, μ_{α^*}, μ_{ξ^*}, and two combinations of γ_L and γ_U. In the first combination $\gamma_L = 1/(m + 1)$ and γ_U was distributed uniformly, $\mathcal{U}(.9, 1)$; in the second combination $\gamma_L = 0$ and $\gamma_U = 1$. The complete simulation design had order 3 (number of answer categories, $m+1$) $\times 2$ (mean μ_{α^*} parameter) $\times 2$ (mean μ_{ξ^*} parameter) $\times 2$ (combination of γ_L and γ_U). It may be noted that within the class of continuation ratio models, α_{ix} can be interpreted as the slope parameter, and ξ_{ix} as the acceleration parameter, but the α^*_{ix} and ξ^*_{ix} parameters do not have a meaningful

TABLE 16.1. Number of times 1000 tests yield $E(\theta \mid X_+)$ nondecreasing in X_+.

γ_L/γ_U		$\frac{1}{m+1}/\sim \mathcal{U}(.9,1)$		0/1	
μ_{α^*}	$m+1$	$\mu_{\xi^*}=-3$	$\mu_{\xi^*}=3$	$\mu_{\xi^*}=-3$	$\mu_{\xi^*}=3$
0	3	530	791	629	977
	4	316	793	383	887
	5	245	702	209	770
1	3	698	784	747	835
	4	546	674	448	486
	5	423	514	228	231

TABLE 16.2. Probability of correctly ordering two randomly selected simulees.

γ_L/γ_U		$\frac{1}{m+1}/\sim \mathcal{U}(.9,1)$		0/1	
μ_{α^*}	$m+1$	$\mu_{\xi^*}=-3$	$\mu_{\xi^*}=3$	$\mu_{\xi^*}=-3$	$\mu_{\xi^*}=3$
0	3	.9792	.9875	.9884	1.0000
	4	.9765	.9683	.9807	1.0000
	5	.9843	.9596	.9767	1.0000
1	3	.9911	.9912	.9921	.9999
	4	.9908	.9833	.9886	.9998
	5	.9932	.9765	.9874	.9997

interpretation due to the logarithmic transformation. Within the class of cumulative probability models, of which the MH model is a special case, none of the parameters α_{ix}, ξ_{ix}, α^*_{ix}, and ξ^*_{ix} can be interpreted meaningfully due to the transformation expressed by (16.9). The only objective of different parameter choices is to obtain differently shaped ISRFs. Four examples of ISRFs resulting from inserting (16.10) into (16.9) are displayed in Figure 16.2.

The results of the simulation study are summarized in Tables 16.1 and 16.2. Table 16.1 shows for each cell in the design that often $E(\theta \mid X_+)$ was decreasing in some adjacent values of X_+. Models with $\mu_{\alpha^*} = 0$, $\mu_{\xi^*} = 3$, $\gamma_L = 0$, and $\gamma_U = 1$ produced the highest frequency of tests in which OEL held, probably because the slopes of the ISRFs were more similar than those in other design cells. Table 16.1 also shows that results became worse as $m + 1$ increased.

The results in Table 16.2 show a much brighter picture. The model with $\mu_{\alpha^*} = 0$, $m + 1 = 5$, $\mu_{\xi^*} = 3$, and $\gamma_L = 0.2$, and $\gamma_U \sim \mathcal{U}(.9, 1)$, yielded the lowest probability of ordering two randomly drawn simulees correctly,

which was still as high as .9596, which is the mean of 1000 probabilities, each based on two probabilities $P(X_+ = x_+)$ calculated using (16.10). Models with $\mu_{\xi\bullet} = 3$, and $\gamma_L = 0$ and $\gamma_U = 1$, produced the highest probabilities, which were all at least .999. Overall, no serious violations of OEL were found. Moreover, a visual inspection of the $E(\theta \mid X_+)$s suggested that if $E(\theta \mid X_+ = s) > E(\theta \mid X_+ = t)$ when $s < t$, the numerical difference between these expected values was very small. For example, a typical string of $E(\theta \mid X_+ = x_+)$ for $x_+ = 0, \ldots, 20$ was

$$\begin{array}{rrrrrrr} -.83, & .90, & 1.43, & 1.73, & 1.95, & 2.24, & 2.55, \\ 2.74, & 2.92, & 3.12, & 3.32, & 3.46, & 3.67, & \mathbf{3.93}, \\ \mathbf{3.88}, & 4.18, & 4.52, & 4.93, & 4.94, & 4.95, & 4.97, \end{array}$$

where the bold values indicate a reversal of the expected order. Results for $k = 10$ and results for slope parameters in (16.10), which are the exponential of independent draws from $\mathcal{N}(\mu_{\alpha\bullet}, \sigma^2_{\alpha\bullet} = 1)$, were also investigated for several cells of the design. These results were similar to the results for $k = 5$ and $\sigma^2_{\alpha\bullet} = .2$ and, therefore, are not presented here.

Based on these first results for the cumulative probability models, we tentatively conclude that the probability of incorrectly ordering persons on θ by means of X_+ is sufficiently small to continue the practice of measuring on the X_+ scale. Throughout this simulation study we assumed that θ had a standard normal distribution. It may be noted that OEL is not invariant for monotone transformations of θ (T.A.B. Snijders, personal communication, 2000) and ordering results may be different for nonlinear transformations of θ. This simulation study was only a first step in the investigation of OEL for polytomous IRT models. A more comprehensive study including investigation of SOL, for all three classes of polytomous IRT models, is presently being performed.

Method. MSP provides the user with methods for analyzing item response data (both dichotomous and polytomous) under the MH and DM models. When a model fits polytomous data the researcher is advised (Molenaar & Sijtsma, 2000, p. 18) to consider X_+ as a proxy for ordering respondents on θ. Our simulation results tentatively suggest that, in practice, the use of X_+ does not lead to serious errors when ordering respondents on θ. For dichotomous items, MLR and SOL hold; thus, X_+ can be used safely for ordering persons on θ.

2.2 Ordering Items by Their Difficulty

Theory. Parametric IRT models for polytomous item scores have location parameters on the θ scale for each answer category. Verhelst and Verstralen (1991) argued that these location parameters cannot be interpreted as difficulty parameters. For example, in the PCM (Masters, 1982) β_{ix} gives the

θ value at which the response curves of the adjacent answer categories $x-1$ and x intersect, but not the difficulty of answering in either one of these categories. Sijtsma and Junker (1996) gave a definition of item difficulty for parametric and nonparametric IRT models for dichotomous items, which was used also for polytomous items by Sijtsma and Hemker (1998; based on Chang & Mazzeo, 1994). The difficulty of an item with $m+1$ ordered answer categories is defined as $E(X_i \mid \theta)$, for $i = 1, \ldots, k$. By definition k items have an invariant item ordering (IIO) if they can be ordered (and numbered accordingly) such that

$$E(X_1 \mid \theta) \leq \ldots \leq E(X_k \mid \theta) \ , \quad \text{for all } \theta. \tag{16.11}$$

For dichotomous items, Rosenbaum (1987) said that items satisfying (16.11) have a *latent scale*.

An IIO can be important in applications that assume that respondents or subgroups of respondents have the same item ordering. For example, starting and stopping rules in intelligence testing assume that for all respondents the same items are easy and the same items are difficult. More specifically, the use of such rules assumes the same item ordering of the k items across θ, with the exception of possible ties; see (16.11). Another example is DIF which, for dichotomous items, assumes that the IRF $P(X_i = 1 \mid \theta)$ is identical for different subgroups from the population of interest. In a NIRT context, one could require an IIO, both for dichotomous and polytomous items (note that for dichotomous items with 0/1 scoring $E(X_i|\theta) = P(X_i = 1|\theta)$). In a particular population an IIO is defined at the θ level; therefore, an IIO implies the same item ordering by mean scores for subgroups, for example, boys and girls, different ethnic groups, and different age groups. One way to study DIF in a nonparametric context is by checking the item ordering at the subgroup level. Other applications of IIO were discussed by Sijtsma and Junker (1996); see also Hemker (2001).

Only a few polytomous IRT models imply an IIO. Sijtsma and Hemker (1998) found that of the popular IRT models from the classes of cumulative probability models and adjacent category models, only the rating scale model (Andrich, 1978) implies an IIO. The rating scale model is a special case of the PCM with linear restrictions on the location parameters. Hemker et al. (2000) found that within the class of continuation ratio models only the sequential rating scale model (Tutz, 1990) implies an IIO.

Sijtsma and Hemker (1998) gave a sufficient condition for IIO. First, unidimensionality, local independence, monotonicity of the ISRFs, and nonintersection of the ISRFs is assumed; that is, the DM model is assumed. Second, it is assumed that if for a given item score x_0,

$$P(X_1 \geq x_0 \mid \theta) < P(X_2 \geq x_0 \mid \theta) < \ldots < P(X_k \geq x_0 \mid \theta) \ , \text{ for all } \theta, \tag{16.12}$$

then for each item score x,

$$P(X_1 \geq x \mid \theta) \leq P(X_2 \geq x \mid \theta) \leq \ldots \leq P(X_k \geq x \mid \theta) \ , \quad \text{for all } \theta. \tag{16.13}$$

TABLE 16.3. Numerical example illustrating two methods of investigating IIO.

Restgroup Order	Subgroup Item Means	First Pooling	Second Pooling
1	0.40	0.40	0.40
2	**1.00**	0.80	0.80
3	**0.80**	0.80	0.80
4	**0.60**	0.80	0.80
5	1.20	1.20	1.20
6	2.00	2.00	2.00
7	2.20	**2.20**	2.07
8	**2.40**	**2.00**	2.07
9	**1.60**	**2.00**	2.07

Again, the item numbering follows the item ordering. In words, if we know the ordering of conditional probabilities for one item score x_0, then it follows that the ordering for the other item scores is the same with the exception of possible ties. Sijtsma and Hemker (1998) called the DM model with the additional restrictions (16.12) and (16.13) the strong DM model and noted that this is a highly restrictive model.

Present research. Here, we tentatively suggest a method for investigating whether a set of k polytomously scored items has an IIO, based on a first attempt published by Verweij, Sijtsma, and Koops (1999). First, the total group of interest is split into the relevant subgroups. Second, within each subgroup w, $w = 1, \ldots, W$, the item means are estimated using the sample means $\overline{X}_{iw}$, $i = 1, \ldots, k$. Third, for all w, the item means are also estimated within the total group with subgroup w excluded, say, the restgroup. The restgroup means are denoted $\overline{X}_{i(-w)}$, $i = 1, \ldots, k$. Fourth, for each subgroup the items are placed in the ordering suggested by the corresponding restgroup item means $\overline{X}_{i(-w)}$ and it is checked whether there are reversals in the subgroup item order compared to the corresponding restgroup item order, which thus serves as the criterion. It may be noted that the item ordering may vary across restgroups, and that other ordering criteria may be preferred. Fifth, when reversals of the criterion order are found within the subgroup, the string of items involved is marked as in Table 16.3, second column, where two strings are printed in boldface (the example involves nine items). We suggest certain procedures for evaluating reversals.

The first procedure investigates whether reversals within a string are significantly different from tied item means using the signed rank test. The second procedure searches for the nearest ordering that matches the rest-

group ordering according to the least-squares criterion. The principle is illustrated in the third and fourth columns of Table 16.3, and is known as the pool-adjacent-violators algorithm (Barlow, Bartholomew, Bremner, & Brunk, 1972, pp. 13–18). The idea is to replace strings of decreasing sample means (second column, boldface) by their mean (third column), and to repeat this routine until no more violations remain (fourth column). Sijtsma (1988, p. 46) suggested using a discrepancy measure based on the mean residual sum of squares as a descriptive statistic. These methods and variations are currently being subjected to more elaborate research.

Method. MSP provides the opportunity to investigate the scalability of a set of items within several groups, defined by means of a grouping variable. The program does not contain methods that explicitly investigate whether a set of items has an IIO. Because the MH and DM models do not imply an IIO, and because researchers often like to know whether item difficulty orderings are the same for different relevant subgroups, the inclusion of a method for investigating the item ordering in relevant subgroups would be an important new feature of MSP, or any other IRT program for models not implying an IIO.

2.3 Estimation of Item Step Response Functions

Theory. The monotonicity assumption says that $P(X_i \geq x \mid \theta)$ is nondecreasing in θ, for all x and all i. For dichotomous items, Junker (1993) showed that the MH model does *not* imply that the observable probability $P(X_i = 1 \mid X_+)$ is nondecreasing in X_+. This means that this probability cannot be used for investigating the monotonicity property of IRFs. Junker (1993) also showed that if we condition on a summary score that does not contain X_i, for example, the restscore

$$R = X_+ - X_i \ ,$$

the MH model implies manifest monotonicity (MM),

$$P(X_i = 1 \mid R) \text{ nondecreasing in } R.$$

Thus, MM can be used for investigating the monotonicity property of IRFs.

For polytomous items, Junker and Sijtsma (2000) discussed an example (due to B.T. Hemker, personal communication, 1997) which shows that the MH model does not imply MM; thus, in general,

$$P(X_i \geq x \mid R) \text{ NOT nondecreasing in } R.$$

Alternatively, Junker (1996) suggested conditioning on a restscore D based on the $k-1$ item scores all dichotomized at the same x, yielding dichotomized item scores D_j and $D = \sum_j D_j$ $(j \neq i)$, such that for consecutive fixed values $x = 1, \ldots, m$,

$$D_j = \begin{cases} 0\ , & \text{if } X_j < x\ , \\ 1\ , & \text{otherwise}\ , \end{cases} \tag{16.14}$$

and then investigating whether

$$P(X_i \geq x \mid D)\ , \quad \text{for all } x, \tag{16.15}$$

is nondecreasing in restscore D. Hemker (1996, Chap. 6) proved that if all k items are dichotomized as in (16.14)—thus including X_i (in 16.15), where $X_i \geq x$ is recoded as $X_i = 1$—the MH model still holds. This means that (16.15) has the property of MM and provides a valid means for investigating nondecreasingness of the ISRFs. Junker (1996) noted that different items can also be dichotomized at different values of x, but we refrain from this possibility. Finally, let $\mathbf{X}_{D_x}$ be the binary $N \times k$ data matrix resulting from dichotomization at item score x (16.14) and let $\mathbf{X}_P$ be the polytomous data matrix. Then, it is readily checked that

$$\mathbf{X}_P = \sum_{x=1}^{m} \mathbf{X}_{D_x}\ .$$

Thus, the m dichotomizations together contain all information from the polytomous data, which would not be the case when fewer dichotomizations were considered.

Future research. We suggest three lines of research. The first is to investigate the practical usefulness of the voluminous results produced by the method proposed by Junker (1996), based on the dichotomization of polytomous item response data. It may be noted that for each individual ISRF m estimates have to be investigated (one for each dichotomization of the $k-1$ items constituting the restscore D). The question then becomes how to combine all sample results into one conclusion, in particular, when for some but not for all ISRFs violations are found.

The second research topic is to take the polytomous data as the starting point and to investigate other conditioning variables based on observable data that are (1) substitutes for X_+ or R; (2) useful estimates of θ; and (3) when inserted in (16.15) produce MM. Candidate conditioning variables that are sum scores from the class of unweighted total scores X_+^* based on $1 \leq S \leq k-1$ items can be excluded beforehand, because the subset of S items can be defined to constitute a new test. We are currently investigating alternative and viable item summaries.

The third research topic is the investigation of the practical usefulness of $P(X_i \geq x \mid R)$ as an estimate of the ISRF by means of simulated polytomous item scores. This research is currently taking place, but no definitive results can be reported as yet.

Method. MSP makes ample use of $P(X_i \geq x|R)$ for estimating the ISRFs. These estimates are used for checking whether the ISRFs are monotone nondecreasing and whether ISRFs are nonintersecting. Future research has to (1) show whether for practical use these methods give valid information about the ISRFs in the sense that errors are negligible; and (2) indicate whether methods presently implemented should be replaced by theoretically sound methods as discussed above.

3 Discussion

3.1 Use Another Polytomous IRT Model?

Hemker et al. (1997) showed that the PCM (and special cases) is the only known polytomous IRT model that implies SOL, and Van der Ark (2000) showed that the PCM also is the only model that implies OEL. This means that when the researcher requires SOL or the weaker OEL as a measurement property, the PCM (or special cases) is the only option. Because the PCM is a restrictive model, assuming equal discrimination for all k items, its use in practical data analysis will often lead to the rejection of many items from the test. This is a sacrifice many researchers may not be prepared to make. Thus, the PCM is a limited option as a practical alternative for polytomous IRT models that do not imply SOL or OEL. This conclusion underlines the need for robustness studies that show to what degree SOL and OEL are violated under several models in practical testing situations. The first robustness results presented here tentatively support the use of X_+ for ordering respondents under the MH model for polytomous items.

Sijtsma and Hemker (1998) showed that only highly restrictive polytomous IRT models, such as their strong DM model and the parametric rating scale model (Andrich, 1978) imply an IIO. The application of any of these models to data probably leads to the rejection of many items. The restrictiveness of these models supports the need for methods that can be used for investigating IIO without the need for simultaneously fitting a particular IRT model. An example of such a method may be one that compares the item ordering by mean scores in each relevant subgroup with an overall item ordering found from the data or hypothesized on the basis of a priori knowledge.

Our simulation study gave some positive first results with respect to the practical robustness of models against failure of SOL or OEL. Of course, future research may show that there exist nontrivial situations in which SOL or OEL is not guaranteed. In those cases, new models have to be developed, for example, incorporating SOL as an assumption, and observable consequences from such models have to be studied, which can be used for investigating model-data fit. Based on our present knowledge we expect that methods for investigating IIO will be developed and applied success-

fully and will become part of future versions of MSP. Also, we expect that alternative methods for investigating properties of ISRFs as the one proposed by Junker (1996; see Equation 16.15) will be further developed.

3.2 Progress Through Success and Failure

The situation described in this chapter, in which several practical data analysis methods are supported by sound psychometric theory while others are only partly supported by theory, is typical of scientific research. New methods are developed and initially believed to be correct until someone shows that they suffer from flaws or someone else proposes a better method.

The availability of user-friendly software provides the opportunity for practical researchers to analyze their test and questionnaire data by means of advanced methods. For psychometricians, a user-friendly program is the only way in which they can effectively promote their newly developed methods among researchers. One could argue that programs be made available only after their theoretical basis has been established completely, but we have already seen that scientific knowledge develops rather through successes and apparent successes, which after a while may be found to suffer from flaws (or worse, turn out to be complete failures) and then are improved or replaced by better methods.

Much of the progress in psychometric theory is stimulated by the application of programs such as MSP to empirical data from various kinds of applications. For example, after a developmental psychologist had analyzed his polytomous data with MSP and OPLM (Verhelst, 1992), he asked whether the fit of the DM model and the generalized PCM implied that the items had equal ordering by mean scores for different age groups. This reasonable question could not be answered by most of the polytomous IRT models, and this fact then became the basis for the IIO research (Sijtsma & Junker, 1996; Sijtsma & Hemker, 1998). Although, looking back, their origins were less clear, without doubt the MLR, SOL, and item restscore regression research was also stimulated by practical data analysis problems and questions asked by researchers using MSP and other programs. It is this mutual cross-fertilization between the theory and practice of psychometrics that stimulates the progress in the development of polytomous IRT modeling.

References

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

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

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

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., & Brunk, H.D. (1972). *Statistical inference under order restrictions.* New York: Wiley.

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.

Debets, P., & Brouwer, E. (1989). MSP: A program for Mokken scale analysis for polytomous items [Software manual]. Groningen: ProGAMMA.

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

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. (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 on item response theory* (pp. 277–296). New York: Springer-Verlag.

Hemker, B.T., & Sijtsma, K. (1998). *A comparison of three general types of unidimensional IRT models for polytomous items.* Manuscript submitted for publication.

Hemker, B.T., Sijtsma, K., & Molenaar, I.W. (1995). Selection of unidimensional scales from a multidimensional item bank 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. (2000). *On measurement properties of sequential IRT models.* Manuscript submitted for publication.

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

Junker, B.W. (1996). *Exploring monotonicity in polytomous item response data.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

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

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

Masters, G.N. (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: With applications in political research.* The Hague: Mouton.

Molenaar, I.W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden, 8*, 145–164.

Molenaar, I.W. (1983). *Item steps* (Heymans Bulletin HB–83–630–EX). Groningen: University of Groningen, Vakgroep Statistiek en Meettheorie.

Molenaar, I.W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën [An exercise in item response theory for three ordered answer categories]. In G.F. Pikkemaat & J.J.A. Moors (Eds.), *Liber amicorum Jaap Muilwijk* (pp. 39–57). Groningen: Econometrisch Instituut.

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

Molenaar, I.W., Debets, P., Sijtsma, K., & Hemker, B.T. (1994). User's manual MSP [Software manual]. Groningen: ProGAMMA.

Molenaar, I.W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden, 28*, 115–126.

Molenaar, I.W., & Sijtsma, K. (2000). User's manual MSP5 for Windows: A program for Mokken scale analysis for polytomous items. Version 5.0 [Software manual]. Groningen: ProGAMMA.

Rosenbaum, P.R. (1987). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology, 40*, 157–168.

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

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*, 549–572.

Sijtsma, K. (1988). *Contributions to Mokken's nonparametric item response theory.* Amsterdam: Free University Press.

Sijtsma, K., Debets, P., & Molenaar, I.W. (1990). Mokken scale analysis for polychotomous items: Theory, a computer program and an application. *Quality & Quantity, 24*, 173–188.

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. (in press). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics.*

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. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149–157.

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

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

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

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.

Verhelst, N.D. (1992). *Het eenparameter logistisch model* (OPLM) [The one-parameter logistic model (OPLM)] (OPD Memorandum 92–3). Arnhem: CITO.

Verhelst, N.D., & Verstralen, H.H.F.M. (1991). *The partial credit model with non-sequential solution strategies.* Arnhem: CITO.

Verweij, A.C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23*, 241–264.