PSYCHOMETRIKA—VOL. 63, NO. 2, 183–200
JUNE 1998

NONPARAMETRIC POLYTOMOUS IRT MODELS FOR INVARIANT ITEM ORDERING, WITH RESULTS FOR PARAMETRIC MODELS

KLAAS SIJTSMA

TILBURG UNIVERSITY

BAS T. HEMKER

CITO NATIONAL INSTITUTE FOR EDUCATIONAL MEASUREMENT

It is often considered desirable to have the same ordering of the items by difficulty across different levels of the trait or ability. Such an ordering is an invariant item ordering (IIO). An IIO facilitates the interpretation of test results. For dichotomously scored items, earlier research surveyed the theory and methods of an invariant ordering in a nonparametric IRT context. Here the focus is on polytomously scored items, and both nonparametric and parametric IRT models are considered.

The absence of the IIO property in two *nonparametric* polytomous IRT models is discussed, and two nonparametric models are discussed that imply an IIO. A method is proposed that can be used to investigate whether empirical data imply an IIO. Furthermore, only two *parametric* polytomous IRT models are found to imply an IIO. These are the rating scale model (Andrich, 1978) and a restricted rating scale version of the graded response model (Muraki, 1990). Well-known models, such as the partial credit model (Masters, 1982) and the graded response model (Samejima, 1969), do no imply an IIO.

Keywords: invariant item ordering, item response theory, nonparametric polytomous IRT models, parametric polytomous IRT models.

Many researchers use tests and questionnaires consisting of polytomously scored items with ordered answer categories to measure personality traits, attitudes, opinions and, occasionally, knowledge and abilities. One of the topics researchers are often interested in is the ordering of the items according to difficulty level. Usually this item ordering is required to be the same in different subgroups that are relevant to the investigation. For example, it is often considered a sign of differential item functioning or item bias if such orderings are different for different gender, ethnic or social groups. Furthermore, within one homogeneous group some respondents may produce unexpected patterns of item scores given the results of the vast majority or given the statistical model used to construct the test. The person-fit results of such aberrant individuals are better understood if one standard item ordering has been found to hold for the total group. In general, an ordering of items that is the same, except for possible ties, in all possible subgroups from the population of interest, to be denoted an *invariant item ordering* (IIO; Sijtsma & Junker, 1996), facilitates the interpretation of test results.

This paper has the following structure. First, the property of IIO is formally defined, followed by some preliminary results. For *nonparametric* polytomous item response theory (IRT) models it is investigated which models imply an IIO, and a new nonparametric IRT model is defined that implies an IIO. Next, the property of IIO is investigated for *parametric* polytomous IRT models, and it is demonstrated that two restrictive models imply an IIO. The nonparametric model that was introduced here is more general than the two

Requests for reprints should be sent to Klaas Sijtsma, Department of Research Methodology, Faculty of Social Sciences, Tilburg University, PO Box 90153, 5000 LE Tilburg, THE NETHERLANDS. E-mail: k.sijtsma@kub.nl

0033-3123/98/0600-96455$00.75/0

existing parametric polytomous IRT models which imply an IIO. Finally, a nonparametric method is proposed to investigate the IIO property in empirical data, and this method is illustrated with an empirical data analysis. The large number of acronyms used in this paper is summarized in the Appendix.

Definition of Invariant Item Ordering, and Preliminary Results

Let each of the k polytomously scored items in the test or questionnaire have $m + 1$ *ordered* answer categories. A fixed number of identically formatted answer categories across items is typical of attitude measurement where response categories usually are the same across the items; for example, "strongly disagree", "disagree", "neutral", "agree", and "strongly agree". For achievement measurement this number sometimes varies across items, for example, because different problems can require solution paths of varying complexity and for one item a finer grading of ordered responses may be more feasible than for another item. The important results for IIO discussed here are only valid for *equal* numbers of answer categories.

IRT models for polytomous ordered item scores typically assume the existence of a unidimensional scalar latent trait θ for person measurement (see Rosenbaum, 1987a, for treatment of multidimensional θ), and m latent parameters that characterize the thresholds between the answer categories. In different IRT models these threshold parameters can have different interpretations (Andrich, 1995; Masters, 1982; Mellenbergh, 1995). Andrich discusses two classes of polytomous IRT models, which he coines Thurstone models and Rasch models, and concludes that these classes are incompatible both algebraically and with respect to the underlying process that leads to the item response. More specifically, he notes that in order to arrive at a response, Rasch models assume a process characterized by the simultaneous consideration of all thresholds whereas Thurstone models assume that the choice of a particular answer category only depends on the thresholds bounding that category (Andrich, p. 115). The incompatibility of these classes of models, however, has no consequences for our investigation of IIO within these classes.

We assume that for each item a score is recorded which is the count of the number of ordered thresholds passed by the respondent starting from the lowest category upwards and arriving at the category of his/her response. Let the random variable X_i denote the observable count on item i ($i = 1, \ldots, k$). We define the conditional expectation of the item score, $\mathscr{E}(X_i|\theta)$ ($i = 1, \ldots, k$), also known as the item response function (IRF; Chang & Mazzeo, 1994) within polytomous IRT models. $\mathscr{E}(X_i|\theta)$ was used by Sijtsma and Junker (1996) to study the IIO property for IRT for binary item scores. Note that for $x = 0, 1$, $\mathscr{E}(X_i|\theta) = P(X_i = 1|\theta)$, which is the IRF for dichotomous items. Thus, for polytomous items $\mathscr{E}(X_i|\theta)$ seems to be an excellent choice to order items.

Definition. A set of k items with $m + 1$ ordered answer categories per item and thus m thresholds per item have an invariant item ordering (IIO) if the items can be ordered and numbered accordingly such that

$$\mathscr{E}(X_1|\theta) \leq \mathscr{E}(X_2|\theta) \leq \cdots \leq \mathscr{E}(X_k|\theta); \text{ all } \theta. \quad (1)$$

Equation 1 allows for the possibility that for certain θs the ordering contains ties.

Figure 1 shows IRFs for three items with five ordered answer categories each, scored $x = 0, \ldots, 4$. The two items with the steepest curves have an IIO, but all three items taken together do not have an IIO.

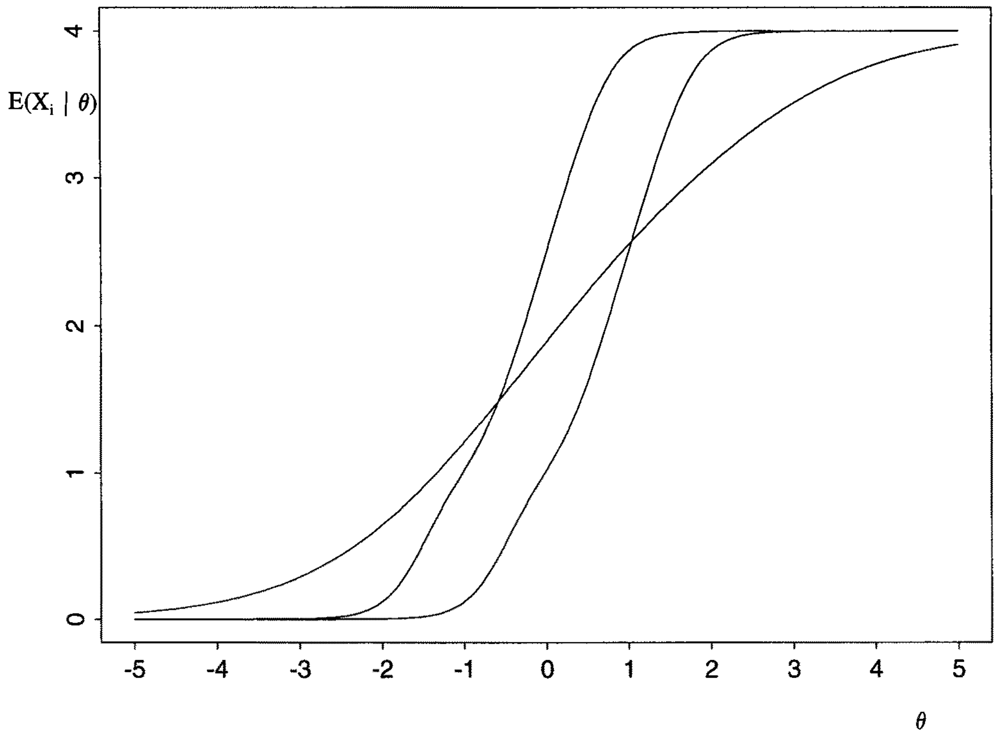

FIGURE 1
Item response functions for three items with five ordered answer categories each, scored $x = 0, \ldots, 4$.

To investigate whether a particular polytomous IRT model implies an IIO, for items i and j an expression is derived for the difference $\mathscr{E}(X_i|\theta) - \mathscr{E}(X_j|\theta)$:

$$\begin{aligned}\mathscr{E}(X_i|\theta) - \mathscr{E}(X_j|\theta) &= \sum_x x.P(X_i = x|\theta) - \sum_x x.P(X_j = x|\theta) \\ &= \sum_x [P(X_i \geq x|\theta) - P(X_j \geq x|\theta)].\end{aligned} \tag{2}$$

Note that $P(X_i \geq 0|\theta) = 1$ *by definition*. Further, we assume that $P(X_i \geq m + 1|\theta) = 0$. Items i and j have an IIO if the sign of the difference on the left does not change across θ from plus to minus, or vice versa.

The conditional probabilities $P(X_i = x|\theta)$ (Equation (2)) are of central interest in the class of adjacent-category models (Hemker, 1996, chap. 6, p. 6; Mellenbergh, 1995), which is closely related to the classes of divide-by-total models (Thissen & Steinberg, 1986) and Rasch models (Andrich, 1995). Andrich's (1978) rating scale model (RSM), Master's (1982) partial credit model (PCM), Verhelst & Glas' (1995) one parameter logistic model (OPLM), and Muraki's (1992) generalized PCM are adjacent-category models. The conditional probabilities $P(X_i \geq x|\theta)$ (Equation (2)) are central in the class of cumulative probability models (Hemker, 1996, chap. 6, p. 5; Mellenbergh, 1995). This class is closely related to the class of difference models (Thissen & Steinberg, 1986) or Thurstone models (Andrich, 1995). Examples of cumulative probability models are Samejima's (1969) homogeneous case of the graded response model (GRM) and Muraki's (1990) rating scale version of the GRM (RS-GRM).

For adjacent-category models, $P(X_i \geq x|\theta)$ can be obtained from the definition of $P(X_i = x|\theta)$ by using

$$P(X_i \geq x|\theta) = \sum_{s=x}^{m} P(X_i = s|\theta). \tag{3}$$

For cumulative probability models, $P(X_i = x|\theta)$ can be obtained from the definition of $P(X_i \geq x|\theta)$ by using

$$P(X_i = x|\theta) = P(X_i \geq x|\theta) - P(X_i \geq x + 1|\theta). \tag{4}$$

Note that converting one kind of probability into the other within the context of one particular IRT model can yield a result that is difficult to interpret in relation to the latent threshold parameters (e.g., Andrich, 1995; Mellenbergh, 1995). For our purposes, however, it suffices to note that we can use $P(X_i \geq x|\theta)$ under any model to check Equation (2) for sign changes across θ. We call this probability the item step response function (ISRF; e.g., Hemker, Sijtsma, & Molenaar, 1995).

The most general polytomous IRT model investigated here is the monotone homogeneity model (MHM; Hemker, Sijtsma, & Molenaar, 1995; Molenaar, 1997). The MHM assumes (1) unidimensionality (UD) of measurement; (2) local independence (LI) of the item scores; and (3) ISRFs $P(X_i \geq x|\theta)$ (all x, all i) that are nondecreasing in θ without parametric restrictions (monotonicity assumption, denoted M). Because of the absense of parametric restrictions on the ISRFs, the MHM is a *nonparametric* IRT model. The MHM is related to models studied by Holland and Rosenbaum (1986) and Junker (1991). The next lemma concerning M will provide useful here:

Lemma. The MHM *assumes* that the ISRF $P(X_i \geq x|\theta)$ is nondecreasing in θ (M). By implication, polytomous IRT models that are special cases of the MHM have this property: the double monotonicity model (DMM; Molenaar, 1997), Scheiblechner's (1995) isotonic ordinal probabilistic (ISOP) model, the generalized PCM (Muraki, 1992), the OPLM (Verhelst & Glas, 1995), the PCM (Masters, 1982), the rating scale model (RSM; Andrich, 1978), the GRM (Samejima, 1969), and the RS-GRM (Muraki, 1990).

Proof. One of the assumptions of the DMM is M; the DMM further assumes UD and LI, and adds a fourth assumption that restricts the ordering of the ISRFs (Molenaar, 1997; also see (5), to be discussed later on). The ISOP model is based on UD, LI, and M, and adds a fourth assumption, different from the fourth assumption of the DMM, that restricts the ordering of the ISRFs (see (6); to be discussed later on). Hemker, Sijtsma, Molenaar, and Junker (1997) showed that the generalized PCM, and its special cases the PCM and the RSM, and the GRM are special cases of the MHM. This also holds for the OPLM, which may be characterized as a generalized PCM with imputed slopes. The RS-GRM is a special case of the GRM by definition (Muraki, 1990). Hence, for each of these models $P(X_i \geq x|\theta)$ in nondecreasing in θ (M). □

In the sequel we will not only investigate the IIO property for several polytomous IRT models, but also the property of an invariant ordering of the ISRFs.

Nonparametric Polytomous IRT Models and IIO

The Monotone Homogeneity Model for Polytomous Items

From (3) it follows that the ISRFs of the same item can not intersect. ISRFs of *different* items are allowed to intersect. Thus the MHM does not imply an invariant ordering of the ISRFs. From this result it follows that the sign of the difference between the x-th ISRF of item i and the x-th ISRF of item j can change across θ (Equation (2), last expression). Since this is true for all $x = 1, \ldots, k$, the sign of (2) can change across θ. Thus the MHM does not imply an IIO.

The Double Monotonicity Model for Polytomous Items

The DMM (Molenaar, 1997) is based on the assumptions of UD, LI, and M, plus the fourth assumption that the ISRFs of different items have an *invariant ordering* across θ; that is, they *do not intersect*. This means that, for any pair of ISRFs of different items, say the s-th ISRF of item i and the r-th ISRF of item j, if for some θ the first ISRF is smaller than the second, then for all θ

$$P(X_i \geq s|\theta) \leq P(X_j \geq r|\theta). \tag{5}$$

Equation (5) implies that for fixed x ($s = r$) the difference of the ISRFs of items i and j can not show a sign change across θ. Different signs can occur, however, for different values of x: for example, it is possible to have $P(X_1 \geq 1|\theta) < P(X_2 \geq 1|\theta)$ for all θ, and $P(X_1 \geq 2|\theta) > P(X_2 \geq 2|\theta)$ for all θ, and so on for higher x values. Thus, the *sum* of the differences of the ISRFs over x (last expression of Equation (2)) can show sign changes across θ. It follows that the DMM does not imply an IIO.

A New Double Monotonicity Model that Implies an IIO

The last expression of (2) suggests how to specialize the DMM into a new model that implies an invariant ordering of the ISRFs and an IIO. We consider two items, i and j; arbitrarily assume that $i < j$ (see Definition); and require that

$$P(X_i \geq x|\theta) \leq P(X_j \geq x|\theta), \text{ all } \theta, \text{ all } x. \tag{6}$$

Then from (2) we find that $\mathscr{E}(X_i|\theta) \leq \mathscr{E}(X_j|\theta)$ for all θ. Thus an IIO is obtained for these two items. Equation (6) provides a restriction for pairs of corresponding ISRFs of *different* items. Equation (6) can be refuted by the data, in contrast with the structural restriction on the ordering of the ISRFs from the *same* item which holds by definition; see (4). The inequality relations between ISRFs defined in Equation (6) are a special case of a more general assumption discussed by Scheiblechner (1995, p. 285, Definition) which he called *weak item independence*, abbreviated W2.

The new DMM version which implies an IIO is defined by UD, LI, and M; and further by an invariant ordering of the $k \times m$ ISRFs (Equation (5)); and for each x the same ordering, except for possible ties, of the x-th ISRFs of the k items (Equation (6)). This new model will be denoted the *strong* DMM in this paper. The original DMM, which is characterized by an invariant ordering of the ISRFs (Equation (5)), will be denoted the *weak* DMM.

The strong DMM is *sufficient* for an IIO, but *not necessary*. This is apparent from the next counterexample, in which it is shown that if Equation (6) is *not* true for all x, then it is still possible to construct examples such that $\mathscr{E}(X_i|\theta) \leq \mathscr{E}(X_j|\theta)$, all θ.

Example. Let $m = 2$; thus $x = 0, 1, 2$. Convenient choices for the ISRFs are:

$$P(X_i \geq 1|\theta) = \exp(\theta)/[2 + 2\exp(\theta)];$$

$$P(X_i \geq 2|\theta) = \exp(\theta)/[3 + 3\exp(\theta)];$$

$$P(X_j \geq 1|\theta) = \exp(\theta)/[1 + \exp(\theta)]; \text{ and}$$

$$P(X_j \geq 2|\theta) = \exp(\theta)/[6 + 6\exp(\theta)].$$

Table 1

Assumptions and Ordering Properties of ISRFs and Items of Four Nonparametric, Polytomous IRT Models; a '+' Means that an Assumption or Property is Present, a '-' that It is Absent.

Assumption	MHM	Weak DMM	Schei.DMM	StrongDMM
Unidimensionality	+	+	+	+
Local Independence	+	+	+	+
Monotonicity in θ	+	+	+	+
Nonintersection ISRFs	-	+	-	+
Ordering ISRFs (Eq.6)	-	-	+	+
Ordering Property				
Invariant ISRF Ordering	-	+	-	+
Invariant Item Ordering	-	-	+	+

With these definitions it can easily be checked that $P(X_i \geq 1|\theta) < P(X_j \geq 1|\theta)$; however, $P(X_i \geq 2|\theta) > P(X_j \geq 2|\theta)$. Together these inequalities contradict Equation (6); therefore, the strong DMM does *not* hold. For this parameter setup, however,

$$\mathscr{E}(X_i|\theta) = P(X_i \geq 1|\theta) + P(X_i \geq 2|\theta) = \frac{5}{6} \times \frac{\exp(\theta)}{1 + \exp(\theta)};$$

$$\mathscr{E}(X_j|\theta) = P(X_j \geq 1|\theta) + P(X_j \geq 2|\theta) = \frac{7}{6} \times \frac{\exp(\theta)}{1 + \exp(\theta)}.$$

This result readily shows that $\mathscr{E}(X_i|\theta) < \mathscr{E}(X_j|\theta)$, all θ.

If the assumption of an invariant ordering of the ISRFs (Equation (5)) is dropped, the resulting model based on UD, LI, M, and (6) still implies an IIO. This is exactly Scheiblechner's (1995) ISOP model specialized to polytomous items, to be denoted here as *Scheiblechner's DMM.*

Table 1 summarizes the assumptions of the MHM, the weak DMM, Scheiblechner's DMM, and the strong DMM. In addition, the results with respect to the invariant ordering of the ISRFs and the items have been included.

Parametric Polytomous IRT Models and IIO

The Partial Credit Model and a Special Case

The Partial Credit Model. The PCM (Masters, 1982) is based on UD and LI and, further, parametrically defines the probability $P(X_i = x|\theta)$, also denoted the category characteristic curve (CCC). Each item is characterized by m transition parameters (Masters, 1982) or threshold parameters (Andrich, 1995) denoted δ_{ix} ($x = 1, \ldots, m$). In the PCM

$$P(X_i = x|\theta) = \frac{\exp\,[\sum_{s=1}^{x} (\theta - \delta_{is})]}{\sum_{q=0}^{m} \exp\,[\sum_{s=1}^{q} (\theta - \delta_{is})]}. \tag{7}$$

There are no restrictions on the distances between the locations of the CCCs of one item. The next numerical example shows that the CCCs of *different* items can be chosen such that an IIO is *not* implied. Thus, the PCM does not imply an IIO. By implication it also follows that the OPLM (Verhelst & Glas, 1995), and the generalized PCM (Muraki, 1992) do not imply an IIO.

Example. Let us assume that the PCM holds, and that $k = m = 2$. For item i, $\delta_{i1} = -1$ and $\delta_{i2} = 1$, and for item j, $\delta_{j1} = -2$ and $\delta_{j2} = 2$. Substitution of these values in (7) yields

$$P(X_i = 1|\theta) = \exp\,(\theta + 1)/\Psi_i;$$

$$P(X_i = 2|\theta) = \exp\,(2\theta)/\Psi_i;$$

$$P(X_j = 1|\theta) = \exp\,(\theta + 2)/\Psi_j; \text{ and}$$

$$P(X_j = 2|\theta) = \exp\,(2\theta)/\Psi_j;$$

with $\Psi_i = 1 + \exp\,(\theta + 1) + \exp\,(2\theta)$, and $\Psi_j = 1 + \exp\,(\theta + 2) + \exp\,(2\theta)$. By means of these probabilities the cumulative probabilities needed in (2), $P(X_i \geq x|\theta)$ and $P(X_j \geq x|\theta)$, can be obtained:

$$P(X_i \geq 1|\theta) = [\exp\,(\theta + 1) + \exp\,(2\theta)]/\Psi_i;$$

$$P(X_i \geq 2|\theta) = \exp\,(2\theta)/\Psi_i;$$

$$P(X_j \geq 1|\theta) = [\exp\,(\theta + 2) + \exp\,(2\theta)]/\Psi_j; \text{ and}$$

$$P(X_j \geq 2|\theta) = \exp\,(2\theta)/\Psi_j.$$

Summation of the first two probabilities yields $\mathscr{E}(X_i|\theta)$, and of the last two $\mathscr{E}(X_j|\theta)$:

$$\mathscr{E}(X_i|\theta) = [\exp\,(\theta + 1) + 2\exp\,(2\theta)]/\Psi_i; \text{ and}$$

$$\mathscr{E}(X_j|\theta) = [\exp\,(\theta + 2) + 2\exp\,(2\theta)]/\Psi_j.$$

After some algebra it follows that $\mathscr{E}(X_i|\theta) \geq \mathscr{E}(X_j|\theta)$ if

$$\exp\,(\theta)[\exp\,(2\theta) - 1][\exp\,(2) - \exp\,(1)] \geq 0,$$

thus, if $\exp(2\theta) \geq 1$. Therefore, if $\theta \geq 0$, then $\mathscr{E}(X_i|\theta) \geq \mathscr{E}(X_j|\theta)$, otherwise $\mathscr{E}(X_i|\theta) < \mathscr{E}(X_j|\theta)$.

The Rating Scale Model. The RSM (Andrich, 1978) is a special case of the PCM in that it is assumed that $\delta_{ix} = \delta_i + \tau_x$; δ_i is a location parameter, and the thresholds are characterized by m parameters τ_x ($x = 1, \ldots, m$). The total number of item parameters thus is reduced from $k \times m$ in the PCM to $k + m$ in the RSM. The item parameter δ_i is defined as the mean of the δ_{ix}s across x. The CCC is defined as

$$P(X_i = x|\theta) = \frac{\exp\left[\sum_{s=1}^{x} (\theta - \delta_i - \tau_s)\right]}{\sum_{q=0}^{m} \exp\left[\sum_{s=1}^{q} (\theta - \delta_i - \tau_s)\right]}. \tag{8}$$

Patterns of corresponding τ's of different items i and j can be obtained through translations equal to $\delta_i - \delta_j$. We will show that the RSM implies an IIO.

Theorem 1. The RSM implies an IIO.

Proof. We will use the notational convention that $i < j$ implies $\delta_j \leq \delta_i$. Let $\Delta_{ij} = \delta_i - \delta_j \geq 0$, then in the RSM

$$P(X_j = x|\theta) = P(X_i = x|\theta + \Delta_{ij}), \text{ all } x.$$

To show that $\mathscr{E}(X_i|\theta) - \mathscr{E}(X_j|\theta)$ (see Equation (2)) has no sign changes across θ, we take the step from CCCs to ISRFs:

$$\begin{aligned} P(X_j \geq x|\theta) &= \sum_{s=x}^{m} P(X_j = s|\theta) \\ &= \sum_{s=x}^{m} P(X_i = s|\theta + \Delta_{ij}) \\ &= P(X_i \geq x|\theta + \Delta_{ij}); \text{ all } x. \end{aligned} \tag{9}$$

Under the RSM, $P(X_i \geq x|\theta)$ is nondecreasing in θ (Lemma). Combination of this knowledge and (9) yields

$$P(X_i \geq x|\theta) \leq P(X_i \geq x|\theta + \Delta_{ij}) = P(X_j \geq x|\theta). \tag{10}$$

Equation (10) is equivalent to Equation (6); therefore, the RSM implies an IIO. Together with UD, LI, and M, this result further demonstrates that the RSM is a special, parametric case of Scheiblechner's DMM. □

The Graded Response Model and a Special Case

The Graded Response Model. The GRM (Samejima, 1969) is based on UD and LI and, further, has a parametric definition of the ISRF, $P(X_i \geq x|\theta)$. Within the same item, the ISRFs have a fixed order, parameterized by m threshold parameters with $\lambda_{i1} \leq \lambda_{i2} \leq \ldots \leq \lambda_{im}$, but the distances between adjacent ISRFs of the same item are free to vary. Furthermore, each item is characterized by a positive discrimination parameter, α_i. The ISRF is defined as

$$P(X_i \geq x|\theta) = \frac{\exp\left[\alpha_i(\theta - \lambda_{ix})\right]}{1 + \exp\left[\alpha_i(\theta - \lambda_{ix})\right]}. \tag{11}$$

The relative position of the ISRFs of different items is not restricted. In addition, the ISRFs of different items can have different slopes which causes these ISRFs to cross. Both characteristics separately imply that patterns of ISRFs of different items can be constructed so as to create violations of an IIO. This is shown for the positioning of ISRFs, even if they have equal slopes, by means of the next numerical example.

Example. Let us assume that $\alpha_i = \alpha_j = 1$, and that $k = m = 2$. For item i, $\lambda_{i1} = 0$ and $\lambda_{i2} = \ln(20)$, and for item j, $\lambda_{j1} = \ln(2)$ and $\lambda_{j2} = \ln(10)$. Substitution of these values in (11) yields

$P(X_i \geq 1|\theta) = \exp(\theta)/[1 + \exp(\theta)]$;

$P(X_i \geq 2|\theta) = \exp[\theta - \ln(20)]/\{1 + \exp[\theta - \ln(20)]\} = \exp(\theta)/[20 + \exp(\theta)]$;

$P(X_j \geq 1|\theta) = \exp[\theta - \ln(2)]/\{1 + \exp[\theta - \ln(2)]\} = \exp(\theta)/[2 + \exp(\theta)]$; and

$P(X_j \geq 2|\theta) = \exp[\theta - \ln(10)]/\{1 + \exp[\theta - \ln(10)]\} = \exp(\theta)/[10 + \exp(\theta)]$.

The first two probabilities determine $\mathscr{E}(X_i|\theta)$, and the last two $\mathscr{E}(X_j|\theta)$. Thus,

$$\mathscr{E}(X_i|\theta) = [21.\exp(\theta) + 2.\exp(2\theta)]/\{[1 + \exp(\theta)][20 + \exp(\theta)]\}; \text{ and}$$

$$\mathscr{E}(X_j|\theta) = [12.\exp(\theta) + 2.\exp(2\theta)]/\{[2 + \exp(\theta)][10 + \exp(\theta)]\}.$$

After some algebra, it follows that $\mathscr{E}(X_i|\theta) \geq \mathscr{E}(X_j|\theta)$ if

$$9\exp(\theta) \times [20 - \exp(2\theta)] \geq 0,$$

thus, if $\exp(2\theta) \leq 20$. Therefore, if $\theta \leq [\ln(20)]/2$, then $\mathscr{E}(X_i|\theta) \geq \mathscr{E}(X_j|\theta)$, otherwise $\mathscr{E}(X_i|\theta) < \mathscr{E}(X_j|\theta)$. Thus, the GRM does not imply an IIO.

A Rating Scale Version of the GRM. The RS-GRM (Muraki, 1990) is a special case of the GRM in that it restricts the location parameter. Let λ_i denote the location parameter of item i, and β_x the location parameter of the x-th ISRF. By assuming that $\lambda_{ix} = \lambda_i + \beta_x$, the ISRF of the RS-GRM is defined as

$$P(X_i \geq x|\theta) = \frac{\exp[D\alpha_i(\theta - \lambda_i - \beta_x]}{1 + \exp[D\alpha_i(\theta - \lambda_i - \beta_x]}, \tag{12}$$

where D is a scaling constant that puts the θ-scale in the same metric as the normal ogive model, and α_i is a positive discrimination parameter that varies over items.

Unlike the RSM, the RS-GRM does not allow an IIO. This follows immediately from the model property that ISRFs of different items can have different slopes. By introducing the restriction that the slope is equal for all ISRFs ($\alpha_i = \alpha$; all i), an IIO can, however, be obtained. We will call this model the *Restricted RS-GRM.*

Theorem 2. The Restricted RS-GRM implies an IIO.

Proof. Assuming that $\lambda_j \leq \lambda_i$; defining $\Lambda_{ij} = \lambda_i - \lambda_j \geq 0$; and maintaining the notational conventions used thus far, in the Restricted RS-GRM

$$P(X_j \geq x|\theta) = P(X_i \geq x|\theta + \Lambda_{ij}), \text{ all } x;$$

from (9). From the Lemma we have that $P(X_i \geq x|\theta)$ is nondecreasing in θ. Combination of this knowledge and Equation (9) for the RSM yields (10) (with Δ_{ij} replaced by Λ_{ij}). Therefore, the Restricted RS-GRM implies an IIO. Because all ISRFs have the same slope, it also readily follows that ISRFs of different items cannot intersect. Therefore, all five assumptions listed in Table 1 hold for the Restricted RS-GRM, which thus is a special, parametric case of the strong DMM. □

Four Models that Imply an IIO

The strong DMM, the RSM, and the Restricted RS-GRM are special cases of Scheiblechner's DMM, and the Restricted RS-GRM is a special case of the strong DMM. The relation between the RSM and the strong DMM, and between the RSM and the Restricted RS-GRM is studied next.

The RSM and the strong DMM. The RSM shares four assumptions with the strong DMM, but it is unknown for the RSM whether the ordering of the ISRFs is invariant across θ. To investigate this, the difference of the ISRFs, $P(X_i \geq x|\theta)$ and $P(X_j \geq x + 1|\theta)$, of the RSM is rewritten using (4) and (10):

$$P(X_i \geq x|\theta) - P(X_j \geq x + 1|\theta) = P(X_i = x|\theta) + [P(X_i \geq x + 1|\theta) - P(X_i \geq x + 1|\theta + \Delta_{ij})].$$

The difference between brackets is always negative because in the RSM the ISRF is increasing in θ (see Lemma). Thus the sign of the total sum on the right-hand side depends on $P(X_i = x|\theta)$; equivalently, the sign of the difference on the left-hand side can vary across θ. Let us consider an example for $k = m = 2$; $\delta_i = 3$; and $\tau_1 = -1$ and $\tau_2 = 1$. We will assess $P(X_i \geq 1|\theta) - P(X_j \geq 2|\theta)$. For $\theta = 0$ and $\Delta_{ij} = 3$ (meaning that $\delta_j = 0$) the difference between the ISRFs equals $-.091$; for $\theta = 5$ (same Δ_{ij}) the difference equals .005. Under the RSM the ISRFs thus are not invariantly ordered; therefore the RSM is not a special case of the strong DMM. Because the RSM has a parametric CCC (Equation (8)) and the strong DMM a nonparametric CCC (defined as in Equation (4)), the latter model neither is a special case of the former.

The RSM and the Restricted RS-GRM. Thissen and Steinberg (1986) showed that the more general PCM and GRM do not have a hierarchical relation. A proof for the special cases RSM and Restricted RS-GRM follows the same line of reasoning, and is therefore omitted here. Table 2 shows the relations between the four models that imply an IIO.

Methods to Investigate Invariant Ordering of ISRFs and an IIO

Investigating Invariant Ordering of ISRFs Across θ

In this section, methods are proposed for investigating in empirical data whether a set of ISRFs have an invariant ordering across θ (Equation (5)), and whether an IIO holds (Equation (6)). The methods do not assume a particular parametric definition of the ISRF, for example, as is done in the RSM and the Restricted RS-GRM. A model-data fit investigation of, e.g., the RSM would also provide evidence of IIO, but is not pursued here.

First, univariate and bivariate proportions are defined that are relevant for the investigation whether ISRFs intersect, and whether items can be invariantly ordered. Let $G(\theta)$ be a probability distribution function. Then the univariate proportions π_{ix}^{+} are equal to

$$\pi_{ix}^{+} = \int_{\theta} P(X_i \geq x|\theta)\, dG(\theta) = \sum_{s=x}^{m} P(X_i = s). \tag{13}$$

Table 2

Assumptions and Ordering Properties of Four Polytomous IRT Models that Imply an Invariant Item Ordering; a '+' Means that an Assumption or Property is Present, a '-' that It is Absent.

Assumption	Schei.DMM	Strong DMM	RSM	ResRS-GRM
Unidimensionality	+	+	+	+
Local Independence	+	+	+	+
Monotone ISRF in θ	+	+	+	+
Parametric ISRF/CCC	-	-	+	+
Nonintersection ISRFs	-	+	-	+
Ordering ISRFs (Eq.6)	+	+	+	+
Ordering Property				
Invariant ISRF Ordering	-	+	-	+
Invariant Item Ordering	+	+	+	+

The population proportion π_{ix}^{+} can be estimated by summation of the appropriate sample frequencies n_{ix} in answer category x of item i and dividing by the sample size n (Sijtsma, Debets, & Molenaar, 1990).

Besides x and s, item score indices g, h, r and t are used. The bivariate population proportion that has at least score s on item i and at least score r on item j is

$$\pi_{is,jr}(++) = \int_{\theta} P(X_i \geq s|\theta)P(X_j \geq r|\theta)\, dG(\theta)$$
$$= \sum_{g=s}^{m} \sum_{h=r}^{m} P(X_i = g, X_j = h). \tag{14}$$

Let $n_{is,jr}(++)$ denote the joint sample frequency with a score of at least s on item i and a score of at least r on item j, then $\pi_{is,jr}(++)$ can be estimated by summation of appropriate sample frequencies and dividing by n (Sijtsma et al., 1990).

Finally, we need the joint proportions that have at most a score $s - 1$ on item i, and at most a score $r - 1$ on item j:

$$\pi_{is,jr}(--) = \int_{\theta} P(X_i < s|\theta)P(X_j < r|\theta)\, dG(\theta)$$
$$= \sum_{g=0}^{s-1} \sum_{h=0}^{r-1} P(X_i = g, X_j = h). \tag{15}$$

Assume that the joint sample frequency with a score lower than s on item i and a score lower than r on item j is denoted by $n_{is,jr}(--)$. Then $\pi_{is,jr}(--)$ can be estimated by summation of appropriate sample frequencies and dividing by n (Sijtsma et al., 1990).

Next, a method is discussed for investigating whether the ISRFs have an invariant ordering across θ. An adaptation of the method can be used to investigate whether an IIO holds. Assume that the r-th ISRF of item j_1 and the t-th ISRF of item j_2 are ordered such that

$$P(X_{j_1} \geq r|\theta) \leq P(X_{j_2} \geq t|\theta), \text{ all } \theta. \tag{16}$$

Given LI and the invariant ordering of the ISRFs, it can be shown (Molenaar, 1997) that

$$\pi_{is,j_1r}(++) \leq \pi_{is,j_2t}(++). \tag{17}$$

The symmetric $P(++)$ matrix of order $km \times km$ with elements $\pi_{is,jr}(++)$ ($i, j = 1, \ldots, k$; $i \neq j$; $s, r = 1, \ldots, m$) is defined. Rows and columns are ordered corresponding to the increasing ordering along the marginals of the proportions π_{ix}^+ (Equation (13)). Given this ordering, the rows and columns must be monotonely *nondecreasing* if the ISRFs are invariantly ordered across θ; see (17). Proportions referring to the *same* item, $\pi_{is,ir}(++)$ (s, $r = 1, \ldots, m$), can not be observed through sample fractions because this would require independent replications of the same item with the same subjects.

Analogously to the $P(++)$ matrix, the symmetric km $\times$ km $P(--)$ matrix is defined. This matrix contains the joint proportions $\pi_{is,jr}(--)$. We assume the ordering of the r-th ISRF of item j_1 and the t-th ISRF of item j_2 to be the same as in Equation (16). Then it can be shown along similar lines as with (17) that

$$\pi_{is,j_1r}(--) \geq \pi_{is,j_2t}(--). \tag{18}$$

The $P(--)$ matrix can thus be arranged such that the orderings of rows and columns correspond with the decreasing ordering along the marginals of the proportions $1 - \pi_{ix}^+$. Given this arrangement, rows and columns must be monotonely *nonincreasing* if the ISRFs have an invariant ordering across θ; see (18).

Rewriting (17) and (18) in the form of conditional probabilities yields the following results, respectively (see Sijtsma & Junker, 1996, for dichotomous items):

$$P(X_{j_1} \geq r|X_i \geq s) \leq P(X_{j_2} \geq t|X_i \geq s); \tag{19a}$$

$$P(X_{j_1} \geq r|X_i < s) \leq P(X_{j_2} \geq t|X_i < s). \tag{19b}$$

From these equations it can be concluded that the $P(++)$ and $P(--)$ matrices provide independent sources of information about the invariant ordering of ISRFs. The sample fractions corresponding to the probabilities in (19a) and (19b) can be used to investigate the invariant ordering of the ISRFs in groups that are located at relatively low and high regions of the scale. This is done for all pairs of ISRFs that belong to different items, and per pair the conditioning is on a large number of different splits of the sample.

Investigating IIO

First, an example is discussed using data from three items which contain many violations of the assumption of invariantly ordered ISRFs and which do not support an IIO. Next, results for four other items are discussed which support invariant ordering of ISRFs, and also an IIO.

Example 1. This example pertains to a subscale ($k = 3$; $x = 0, 1, 2, 3$) of a questionnaire (Cavalini, 1992) on annoyance due to industrial malodour ($n = 828$). In addition to invariance of the ordering of ISRFs, nondecreasingness of ISRFs in θ is also inspected because violations of M can be an important source of information about invariant ordering of ISRFs.

Table 3 shows the sample $P(++)$ and $P(--)$ matrices under the *weak* DMM. These matrices contain several violations of the expected orderings in rows and columns (Equations (17) and (18)). A detailed analysis with the computer program MSP (Molenaar, Debets, Sijtsma, & Hemker, 1994), not displayed here for reasons of space, showed that the two matrices together contained 27 violations, of which 14 were significant (5% level; test by Molenaar, 1970, chap. 3, Formula 5.5). The significant violations each involved intersections of the ISRFs of Item #1 with the ISRFs of Item #2 and Item #4.

The nondecreasingness of the ISRFs of item #1 was investigated by means of the empirical regression of the *proportion* of respondents with at least a score x on item i on the *total score*, denoted R, on the other two items ($R = 0, 1, \ldots, 6$). Hemker, Sijtsma, Molenaar, and Junker, (1996) argued that in testing, the highly frequent use of the unweighted total score as a proxy for θ has a long history. Further motivation for interest in the unweighted total score comes from its usually high correlation with many statistics that may be more appropriate to estimate θ, or an ordering on θ (Hemker et al., 1997), and from the ordinal consistency results of e.g. Junker (1991). We thus recognize potential weaknesses of the unweighted total score, and use the regressions, denoted $\pi_{ix|R}$ (π denotes a sample fraction), only as *proxies* of the ISRFs.

Figure 2 shows that $\pi_{12|R}$ and $\pi_{13|R}$ are decreasing at the lower end of the scale, and relatively flat in the middle and at the higher end. Statistical testing (Molenaar, 1970, chap. 4, Formula 2.37) revealed that $\pi_{12|R}$ and $\pi_{13|R}$ had two and three significant decreases, respectively. For Item #2, $\pi_{21|R}$, and for Item #4, $\pi_{41|R}$, both had one significant decrease. These results provide an explanation for the results on intersection spotted by the $P(++)$ $- P(--)$ methodology.

An appropriate permutation of the rows and columns of the $P(++)$ and $P(--)$ matrices renders them suitable for the investigator of (6). The item means are 1.33 (#4), 1.38 (#2), and 1.86 (#1). The item ordering by mean score suggests the same ordering of the x-th ($x = 1, 2, 3$) ISRF across the items (Equation (6)). In Table 3 it can be seen that the third ISRF of Items #4 and #2 has an ordering reverse to the ordering based on item means. This implies that the first two rows and columns of the $P(++)$ and $P(--)$ matrices must be interchanged so as to create matrices denoted $P(++)^s$ and $P(--)^s$. Theoretically, in $P(++)^s$ rows and columns must be monotonely nondecreasing and in $P(--)^s$ monotonely nonincreasing.

A visual inspection of Table 3 reveals that the permutation would create two *additional* violations in the $P(++)$ matrix, and one in the $P(--)$ matrix. These results suggest that the permutation of only the ISRFs "23" and "43" is insufficient to satisfy Equation (6) and that these data do not support an IIO. The researcher could be adviced to look for meaningful subgroups for which different IIOs hold, or to inspect item contents. Indeed, Items #2 ("no laundry outside") and #4 ("no blankets outside") seem more strongly related with each other than with Item #1 ("keep windows closed"), which seems to be a more general reaction to industrial malodour in the vicinity of one's home.

Example 2. Detailed results, not reported here, for another subscale ($k = 4$) from the same questionnaire showed that the $P(++)$ and $P(--)$ matrices had the correct orderings under the *strong* DMM (based on the sample item means: 0.54 (#6), 0.65 (#13), 0.78 (#15), and 0.98 (#14): These matrices thus were equivalent with the $P(++)^s$ and $P(--)^s$ matrices. The matrices did not contain significant violations from the expected orderings. Figure 3 shows the four triples of regressions, $\pi_{ix|R}$.

None of the local decreases was significant. Apart from small fluctuations, the order-

Table 3

P(+ +) and P(--) Matrices for 3 Items with 4 Ordered Answer Categories Each. Original Item Numbering is Maintained. Rows and Columns to be Interchanged for Strong DMM in Italics.

P(++) matrix											
Item			2	4	1	4	2	1	4	2	1
≥			3	3	3	2	2	2	1	1	1
π_{is}^{+}			.28	.30	.38	.43	.43	.60	.60	.66	.89
Item	≥	π_{is}^{+}									
2	3	.28		*.24*	*.20*	*.25*		*.24*	*.26*		*.27*
4	3	.30	*.24*		*.20*		*.28*	*.26*		*.29*	*.29*
1	3	.38	*.20*	*.20*		.23	.24		.25	.27	
4	2	.43	*.25*		.23		.37	.35		.41	.41
2	2	.43		*.28*	.24	.37		.36	.40		.42
1	2	.60	*.24*	*.26*		.35	.36		.41	.43	
4	1	.60	*.26*		.25		.40	.41		.55	.57
2	1	.66		*.29*	.27	.41		.43	.55		.62
1	1	.89	*.27*	*.29*		.41	.42		.57	.62	
P(--) matrix											
Item			2	4	1	4	2	1	4	2	1
≥			3	3	3	2	2	2	1	1	1
$1-\pi_{is}^{+}$			.72	.70	.62	.57	.57	.40	.40	.34	.11
Item	≥	$1-\pi_{is}^{+}$									
2	3	.72		*.66*	*.54*	*.54*		*.36*	*.38*		*.10*
4	3	.70	*.66*		*.52*		*.54*	*.36*		*.33*	*.10*
1	3	.62	*.54*	*.52*		.42	.42		.27	.23	
4	2	.57	*.54*		.42		.51	.32		.32	.10
2	2	.57		*.54*	.42	.51		.32	.36		.10
1	2	.40	*.36*	*.36*		.32	.32		.20	.17	
4	1	.40	*.38*		.27		.36	.20		.29	.08
2	1	.34		*.33*	.23	.32		.17	.29		.08
1	1	.11	*.10*	*.10*		.10	.10		.08	.08	

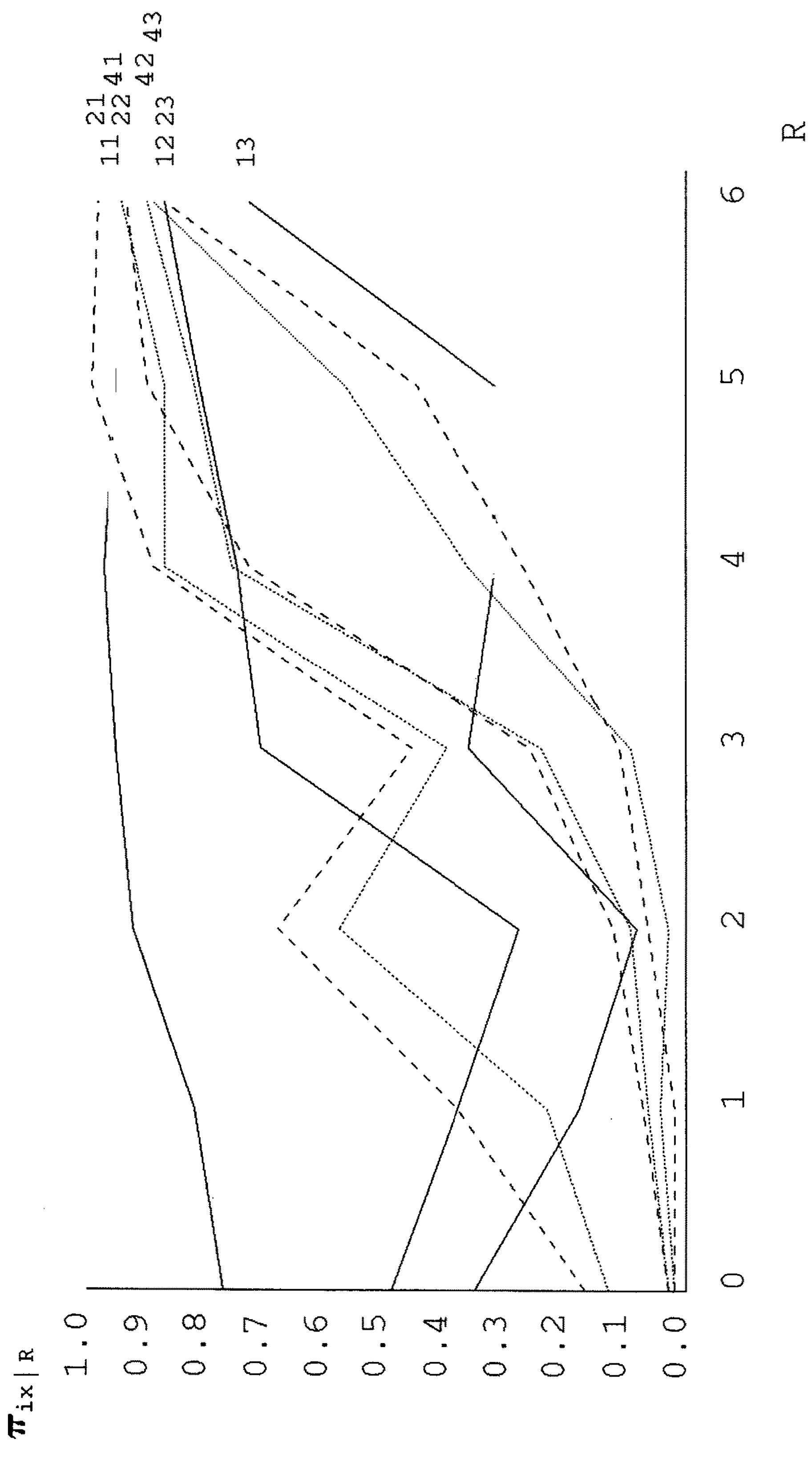

FIGURE 2
Estimated proportion with at least score x on item i as a function of rest score R; Item #1 (solid lines), Item #2 (dashed lines), and Item #4 (dotted lines).

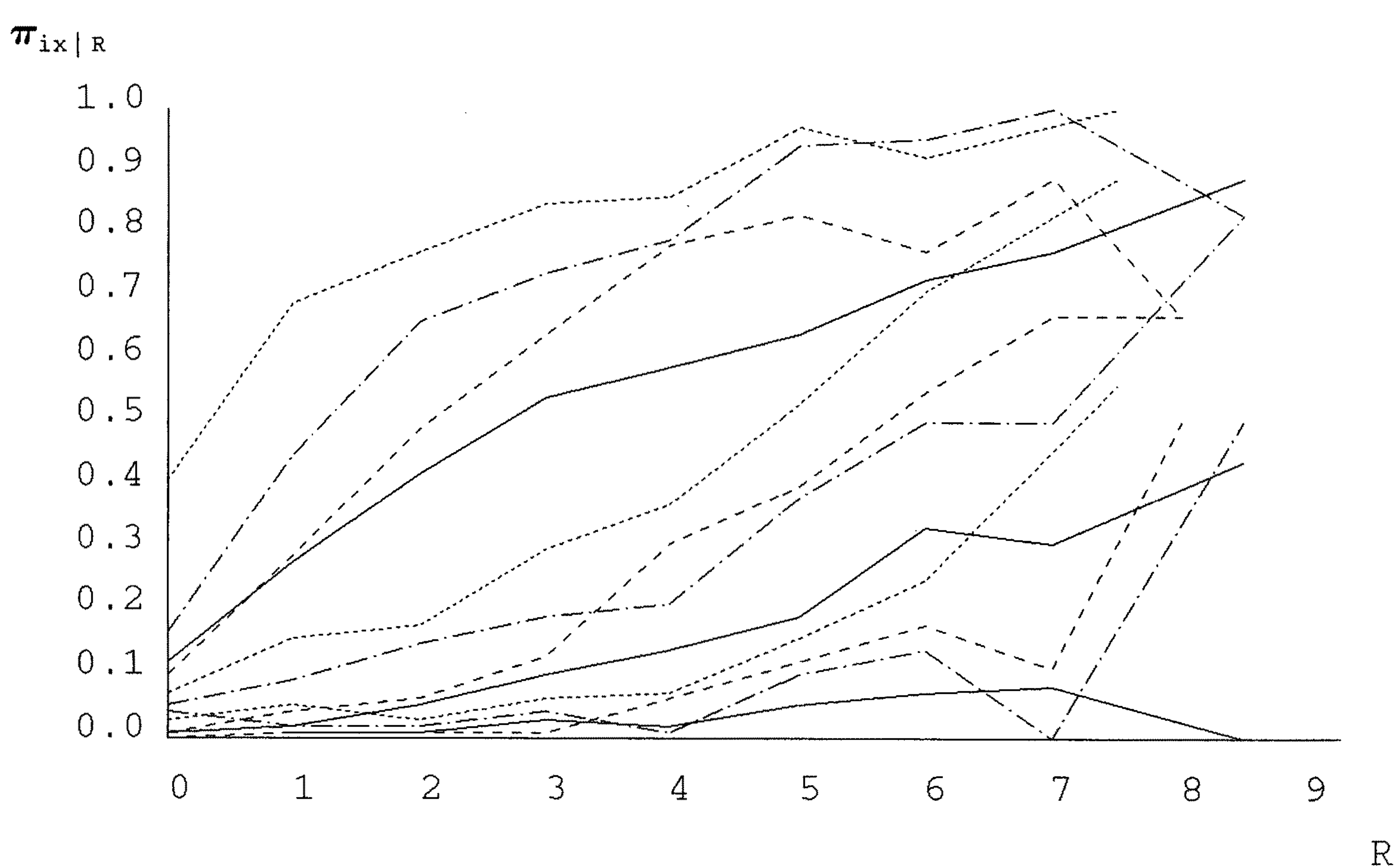

FIGURE 3

Estimated proportion with at least score x on item i as a function of rest score R; Item #6 (solid lines), Item #13 (dashed lines), Item #14 (dotted lines), and Item #15 (dashed/dotted lines). (Note: The highest restscore groups contained few observations and were sometimes merged. E.g., the solid regression lines run to $R = 8.5$ (groups 8 and 9 merged), etcetera.)

ing of the regressions for fixed x is invariant to a considerable degree across the four items. Most intersections appear at the higher end of the scale where the fractions were based on small restscore groups ($R = 7, 8, 9$; group size varied from 6 to 24). A detailed analysis revealed 18 reversals within pairs from the expected ordering, of which only one reached significance (test discussed by Molenaar, 1970, chap. 3, Formula 5.5). The second item set thus supports an IIO.

Discussion

An IIO can prevent cumbersome problems of interpretation that might arise if item orderings are different in different relevant subgroups. Different item orderings for different measurement levels would call at least for additional research to reveal the cause of these differences. This is not to say that psychometric models that do not imply an IIO are not useful. Indeed, many of such IRT models have proven themselves to be very useful in test construction.

The weak DMM (Molenaar, 1997), the PCM (Masters, 1982), and the GRM (Samejima, 1969) do not imply an IIO. By implication, this is also true for generalizations of these models, such as Muraki's (1992) *generalized* PCM, the OPLM (Verhelst & Glas, 1995), and the MHM. The RSM (Andrich, 1978), the Restricted RS-GRM (a special case of a model proposed by Muraki, 1990), the strong DMM, and Scheiblechner's (1995) DMM do imply an IIO.

The usefulness of the $P(++) - P(--)$ methodology was investigated for checking the fifth assumption (see also (6)) of the strong DMM with respect to the ordering of the ISRFs, which is also an assumption of Scheiblechner's DMM, and which secures an IIO. This methodology may be seen as a first attempt to check this crucial assumption. Much is unknown so far, and future research might address issues of power, Type I error, and chance capitalization. Other methods to investigate the IIO property may be derived from methods for dichotomous items proposed and surveyed by Sijtsma and Junker (1996). Such methods include the use of ordering properties based on joint proportions of item score patterns on n items ($2 \leq n < k$) as a generalization of the $P(++) - P(--)$ methodology, and the pairwise comparison of the IRFs, mainly based on work of Rosenbaum (1987a, 1987b).

Appendix

List of acronyms

Technical terms:

CCC : category characteristic curve
IIO : invariant item ordering
IRF : item response function
IRT : item response theory
ISRF : item step response function
LI : local independence
M : monotonicity
UD : unidimensionality

Item response models:

DMM : double monotonicity model
GRM : graded response model
ISOP : isotonic ordinal probabilistic model

MHM	: monotone homogeneity model
OPLM	: one parameter logistic model
PCM	: partial credit model
RS-GRM	: rating scale version of the graded response model
RSM	: rating scale model

References

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

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19*, 101–119.

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

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.

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. (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.

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

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

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.

Molenaar, W. (1970). *Approximations to the Poisson, binomial, and hypergeometric distribution functions*. Amsterdam: Mathematical Centre Tracts 31.

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

Molenaar, I. W., Debets, P., Sijtsma, K., & Hemker, B. T. (1994). *User's manual MSP*. Groningen, The Netherlands: iecProGAMMA.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59–71.

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

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

Rosenbaum, P. R. (1987b). Comparing item characteristic curves. *Psychometrika, 52*, 217–233.

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

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

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

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.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567–577.

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.

Manuscript received 1/30/96
Final version received 6/11/97