
Investigating an Invariant Item Ordering for Polytomously Scored Items

Educational and Psychological
Measurement
XX(X) xx-xx
© 2009 Sage Publications
DOI: 10.1177/0013164409355697
<http://epm.sagepub.com>



Rudy Ligtvoet¹, L. Andries van der
Ark¹, Janneke M. te Marvelde¹,
and Klaas Sijtsma¹

Abstract

This article discusses the concept of an invariant item ordering (IIO) for polytomously scored items and proposes methods for investigating an IIO in real test data. Method manifest IIO is proposed for assessing whether item response functions intersect. Coefficient H^T is defined for polytomously scored items. Given that an IIO holds, coefficient H^T expresses the accuracy of the item ordering. Method manifest IIO and coefficient H^T are used together to analyze a real data set. Topics for future research are discussed.

Keywords

coefficient H^T , invariant item ordering, item response function for polytomous items, item step response function, polytomous item response theory models

In several measurement applications, it is convenient that the items have the same order with respect to difficulty or attractiveness for all respondents. Such an ordering facilitates the interpretation and the comparability of respondents' measurement results. An item ordering that is the same for all respondents is called an invariant item ordering (IIO; Sijtsma & Junker, 1996). Before we define an IIO, we first mention several measurement applications in which an IIO proves useful.

First, many intelligence tests present the items to children in the order according to ascending difficulty (Bleichrodt, Drenth, Zaal, & Resing, 1987; Wechsler, 1999). One reason for this presentation order is to comfort children and prevent them from

¹Tilburg University, Tilburg, The Netherlands

Corresponding Author:

Rudy Ligtvoet, Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE
Tilburg, The Netherlands
Email: r.ligtvoet@uvt.nl

panicking, which might result from starting with difficult items and which might negatively influence test performance. Another reason is that different age groups are administered different subsets of the items, and subsets are more difficult as age increases. For example, the youngest age group starts with the easiest items and a child stops when he or she fails, say, three consecutive items. The next age group always skips the five easiest items, because these items have been shown to be trivial to them, and starts at Item 6, and again a child stops when he or she fails, say, three consecutive items. And so on for the next age groups. Several intelligence tests use this administration mode, which assumes that the ordering of the items by difficulty is the same across age groups and persons. This assumption usually is ignored in the phase of test construction. In subsequent test use, test practitioners often are unaware that the assumption was never ascertained by means of empirical research, but they use the test as if it were.

Second, several developmental theories assume that abilities or skills go through different phases before they reach maturity (Bouwmeester & Sijtsma, 2007; Raijmakers, Jansen, & Van der Maas, 2004). A simple example is arithmetic ability, for which it may be assumed that development goes through mastering the operation of addition and then subtraction, multiplication, and, finally, division. An arithmetic test, which aims at measuring the degree to which these operations have been mastered, may be assembled and administered such that the hypothesized item ordering by difficulty reflects the assumed ordering of the operations or combinations of the operations. The hypothesized developmental ordering could be investigated using this test with either cross-sectional or, even better, longitudinal data from the population of interest. When the theory proves to be correct, this would lend credence to the diagnostic use of the test and the possibility to pinpoint children's problems with arithmetic as either normal developmental hurdles to be taken or signs of abnormal development.

Third, in attitude and personality testing, and also in the medical context researchers often assume their items to have a cumulative structure, reflecting a hierarchy of psychological or physical symptoms hypothesized to hold at the individual level (Van Schuur, 2003; Watson, Deary, & Shipley, 2008). For example, in measuring introversion it seems reasonable to expect a higher mean score on a rating scale statement like "I do not talk a lot in the company of other people" than on "I prefer not to see people and do things on my own," because the latter statement seems to refer to a more intense symptom of introversion. However, an ordering of these statements by group mean scores does not imply that this ordering also holds at the individual level. Indeed, several respondents may indicate a higher prevalence for doing things on their own, but the mixture of the two item orderings may be such that the first still has the highest mean score in the total group. Any set of items can be ordered by means of item mean scores, but whether such an ordering also holds for individuals has to be ascertained by means of empirical research. Only when the set of items has an IIO, can their cumulative structure be assumed to be valid at the lower aggregation level for individuals.

This study deals with the investigation of an IIO for a set of polytomously scored items and extends the previous work of Sijtsma and Meijer (1992) and Sijtsma and

Junker (1996) for dichotomously scored items. Very little work has been done in this area. Therefore, this study presents some first steps and has an exploratory character. An empirical data example shows that the results may be used for investigating whether an IIO holds in sets of polytomously scored items. Finally, directions for future research are discussed.

Definition of an Invariant Item Ordering

The context of this study is item response theory (IRT). Let a test contain k polytomously scored items, each of which is characterized by $m + 1$ ordered integer scores. These scores reflect the degree to which a person solved a complex problem (e.g., a physics problem or a text comprehension problem) or endorsed a statement (e.g., as in Likert-type items). For $m + 1 = 2$, items are dichotomous. Technically, the number of ordered item scores may vary across items but this hampers the comparison of expected item scores for different items. Hence, we follow Sijtsma and Hemker (1998) in only considering equal numbers of ordered item scores; equal numbers are common in many standard tests and questionnaires.

Let random variable X_i denote the score on item i , with realization $x_i \in \{0, \dots, m\}$. Let θ be the unidimensional latent variable from IRT on which the persons can be ordered. A test that consists of k items has an IIO (Sijtsma & Hemker, 1998) if the items can be ordered and numbered accordingly, such that for expected conditional item scores

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \dots \leq E(X_k|\theta), \text{ for all } \theta. \quad (1)$$

Equation (1) allows for the possibility of ties. The expected conditional item score $E(X_i|\theta)$ is called the item response function (IRF), and an IIO implies that the IRFs do not intersect. For dichotomously scored items, $E(X_i|\theta) = P(X_i = 1|\theta)$, which is the conditional probability that the answer was correct or the statement endorsed.

An IIO is a strong requirement in measurement practice. Researchers sometimes assume that a fitting IRT model implies that items have the same ordering by difficulty or popularity for all individuals, but this assumption requires modification. For dichotomous-item tests, Sijtsma and Junker (1996) showed that only IRT models that employ IRFs that cannot intersect, imply an IIO. Examples are the Rasch (1960) model and the Mokken (Mokken & Lewis, 1982) double monotonicity model, but the much-used two- and three-parameter logistic models (Birnbaum, 1968), which allow intersecting IRFs, do not imply the IIO property. For polytomous-item tests, Sijtsma and Hemker (1998) proved the surprising result that popular IRT models such as the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the graded response model (Samejima, 1969) do not imply an IIO. Thus, when any of these models gives an accurate description of the data, one cannot conclude that the items follow the same ordering by difficulty or popularity for each individual from the population of interest (Equation 1). Sijtsma and Hemker (1998) proved that only restrictive polytomous IRT models, such as the rating scale model

(Andrich, 1978), a rating scale version of Muraki's (1990) restricted graded response model, and the isotonic ordinal probabilistic model (Scheiblechner, 1995) imply an IIO.

Thus, there appears to be a mismatch between popular polytomous IRT models and the IIO property. This mismatch is due to an aggregation phenomenon, which we illustrate by means of the graded response model and a special case of this model. We assume a unidimensional latent variable θ , and item scores that are locally independent. Response functions of polytomous items are defined for separate item scores and given that an item has $m + 1$ different scores, for each item m such response functions are needed (Mellenbergh, 1995). An example of these response functions are the item step response functions (ISRFs) of the class of cumulative probability models, which are defined by the conditional probabilities $P(X_i \geq x|\theta)$, for $x = 1, \dots, m$; by definition, $P(X_i \geq 0|\theta) = 1$ and $P(X_i \geq m + 1|\theta) = 0$.

Given the definition of an IIO (Equation 1), one is interested in statistical information at the higher aggregation level of the item rather than the level of item scores. Hence, we consider the IRF, which is related to the m ISRFs by means of

$$E(X_i|\theta) = \sum_{x=1}^m P(X_i \geq x|\theta). \quad (2)$$

Sijtsma and Hemker (1998) used relationships like this one to prove that for many polytomous IRT models, combining the m ISRFs of items, $P(X_i \geq x|\theta)$, into IRFs, $E(X_i|\theta)$, does not result in an IIO as in Equation (1). These authors also showed that one needs restrictions on the mutual relationships between the ISRFs of different items in the test or the questionnaire to obtain an IIO. We give two examples of the relationships between ISRFs and IRFs, one resulting in failure of IIO and the other in an IIO; see Sijtsma and Hemker (1998) for mathematical proofs.

First, in Samejima's (1969) graded response model, each item has m threshold parameters such that $\beta_{i1} \leq \beta_{i2} \leq \dots \leq \beta_{im}$ (i.e., the m ISRFs have a fixed order), and one discrimination parameter α_i ; then, the ISRF for score x on item i is defined as

$$P(X_i \geq x|\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ix})]}{1 + \exp[\alpha_i(\theta - \beta_{ix})]}, \quad x = 1, \dots, m. \quad (3)$$

Summing the m ISRFs in Equation (3) across the m item scores yields IRF $E(X_i|\theta)$ (Equation 2). Figure 1a shows the ISRFs for two items with three different scores (solid ISRFs for one item, and dashed-dotted ISRFs for the other item) and Figure 1c shows their intersecting IRFs, which violate IIO.

Second, the restricted version of Muraki's (1990) rating scale version of the graded response model (Sijtsma & Hemker, 1998) places restrictions on the mutual relationships of the ISRFs of different items, which result in an IIO. Let α denote a general discrimination parameter, λ_i an item-dependent location parameter, and ε_x the distance of the x th ISRF to location λ_i , so that $\beta_{ix} = \lambda_i + \varepsilon_x$, and with the restriction that $\sum_x \varepsilon_x = 0$; then, the x th ISRF of item i is

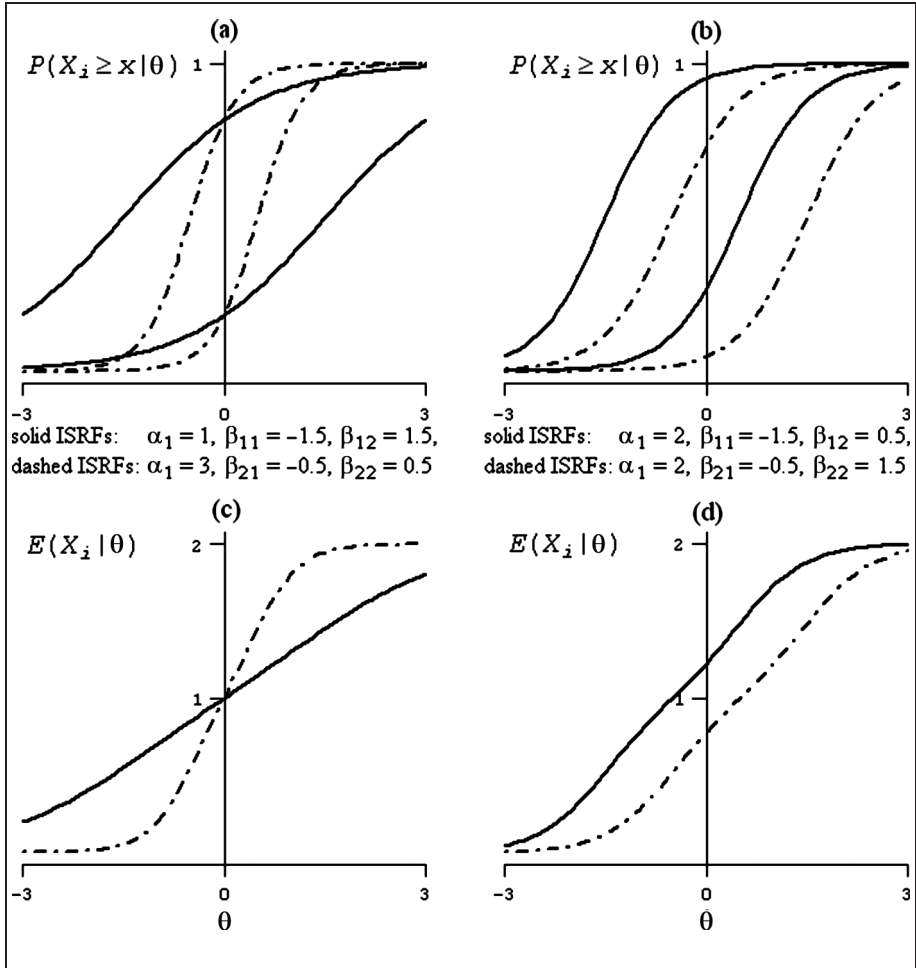


Figure 1. (a) Two items having two item step response functions (ISRFs) under the graded response model, (c) failing an invariant item ordering (IIO), and (b) two items having two ISRFs under the restricted rating scale version of the graded response model, (d) having an IIO

$$P(X_i \geq x | \theta) = \frac{\exp[\alpha(\theta - \lambda_i - \varepsilon_x)]}{1 + \exp[\alpha(\theta - \lambda_i - \varepsilon_x)]}. \quad (4)$$

All items show the same dispersion of the ISRFs around the location parameters λ_i . For two items satisfying Equation (4), Figures 1B and 1D show that they have an IIO.

Two sources of confusion seem to exist with respect to IIO. The first is that if an IRT model does not imply an IIO, the IIO property cannot be important. We emphasize that it is the measurement application, which determines whether an IIO is

important, not the psychometric model. If a particular IRT model does not give information about an IIO, other methods have to be used in data analysis for ascertaining whether an IIO is valid. The second source of confusion is that the IIO property applies to particular content areas but not to others and that it applies to rating scale items but not to constructed-response items. The examples given in the beginning of this article illustrated that an IIO may be important in different content areas. This is also true for different item types. For example, in intelligence tests many items require constructed responses, as in explaining to the test administrator the use of a particular object (e.g., a hammer, a car). If such items are administered in an ascending difficulty ordering, an IIO is assumed, which has to be supported by empirical research.

Investigating an Invariant Item Ordering

In IIO investigation for polytomous items, a distinction is made between sets of IRFs that are close together and sets of IRFs that are further apart. If IRFs are close together, respondents produce data that contain little information about the item ordering, resulting in an inaccurate ordering, and if IRFs are far apart, respondents produce data that contain much more information resulting in an accurate ordering. Thus, given an IIO, an index for the distance between the IRFs can be interpreted as an index of the accuracy of the ordering of the IRFs. In this study, we estimated the IRFs of k polytomous items, defined by $E(X_i|\theta)$, then we ascertained whether the items had an IIO and if they had, finally we used a generalization of coefficient H^T , proposed by Sijtsma and Meijer (1992) for dichotomous items, to polytomous items to express the degree to which an accurate item ordering was possible.

Sijtsma and Meijer (1992) demonstrated by means of a simulation study that for k invariantly ordered dichotomous items coefficient H^T increased as the mean distance between the item locations increased, or as the item discrimination increased (both manipulations have the effect that IRFs are further apart), whereas other properties of the IRFs and the distribution of θ were kept constant. They did not find convincing support for different values of H^T to distinguish failure of IIO from consistency with IIO (yet suggested tentative rules of thumb for making this distinction, to be discussed later), and in a pilot study, we found that this was even more difficult for polytomous items.

In what follows, we discuss a two-step procedure for investigating an IIO for polytomous items. First, we discuss the estimation of IRFs and propose *method manifest IIO*, which is based on the estimation of IRFs for dichotomous items (see Molenaar & Sijtsma, 2000, pp. 74-78) and which evaluates for each pair of IRF estimates whether or not they intersect. The sensitivity and specificity of this method was investigated by means of a simulation study. Second, we discuss coefficient H^T for polytomously scored items. We used a computational study to investigate how coefficient H^T reacts to different item and test properties, given an IIO. Finally, method manifest IIO and coefficient H^T were used to analyze a real data set.

Method Manifest Invariant Item Ordering

Theory: Estimation of IRFs, and Pairwise Inspection of Invariant Item Ordering

Method manifest IIO is available from the R package *mokken* (Van der Ark, 2007) as method `check.iio`. Let $R_{(i,j)} = X_+ - X_i - X_j$ be the rest score, defined as the total score on $k - 2$ items without the items i and j , and which has realization r , with $r = 0, \dots, (k - 2)m$. Let $E(X_i|R_{(i,j)})$ be the estimated IRF of item i . If population item means are ordered such that for pair (i, j) , $E(X_i) \leq E(X_j)$, then an IIO implies that

$$E(X_i|\theta) \leq E(X_j|\theta), \text{ for all } \theta. \quad (5)$$

Ligtvoet, Van der Ark, Bergsma, and Sijtsma (2009) showed that Equation (5) implies that

$$E(X_i|R_{(ij)} = r) \leq E(X_j|R_{(ij)} = r), \text{ for all } r. \quad (6)$$

Equation (6) is investigated for each pair of items using conditional sample means $\bar{X}_{i|r}$ and $\bar{X}_{j|r}$, for all r . If it is found that $\bar{X}_{i|r} > \bar{X}_{j|r}$, we use a one-sided one-sample t test for the null hypothesis that $E(X_i|R_{(ij)} = r) = E(X_j|R_{(ij)} = r)$ against the alternative that $E(X_i|R_{(ij)} = r) > E(X_j|R_{(ij)} = r)$, for all r . Rejection of the null hypothesis for at least one value of r leads to the conclusion that items i and j are not invariantly ordered. If the number of persons having a rest score r is too small for accurate estimation, adjacent rest score groups are combined until the group size exceeds a preset minimum (Molenaar & Sijtsma, 2000, p. 67; Van der Ark, 2007). A protection against taking very small violations seriously is to test sample reversals only when they exceed a minimum value denoted *minvi*. Molenaar and Sijtsma (2000, pp. 67-70) recommend for dichotomous items ($m = 1$) the default value $\text{minvi} = 0.03$. Polytomous items have a greater score range and a logical choice for *minvi* is $m \times 0.03$. Whether this is a reasonable choice was investigated in a simulation study (next section).

We used the following sequential procedure for method manifest IIO. First, for each of the k items the frequency is determined that the item is involved in significant violations that exceed *minvi*. If none of the items is involved in such violations, we conclude that an IIO holds for all k items; else, the item with the highest frequency is removed from the test. Second, the procedure is repeated for the remaining $(k - 1)(k - 2)/2$ item pairs, and if an item is removed, for the remaining $(k - 2)(k - 3)/2$ item pairs, and so on. When q items have the same number of significant violations, the $q - 1$ items having the smallest scalability coefficients (Sijtsma & Molenaar, 2002, p. 57) may be removed, but researchers may also consider other exclusion criteria, such as item content.

This procedure is suited for exploratory data analysis but for confirmatory purposes, when one wants to know whether all k items have an IIO, manifest IIO is checked for all item pairs but items are not removed. For the remaining item subset (exploration) or the complete item subset (confirmation), we compute the H^T value

to evaluate the degree to which an accurate item ordering is possible. Coefficient H^T is discussed in the next section.

Monte Carlo Study: Sensitivity and Specificity of Method Manifest Invariant Item Ordering

We used a Monte Carlo study to investigate the sensitivity (probability that IIO is correctly identified) and the specificity (probability that IIO is correctly rejected) of method manifest IIO.

Method

The design factors were defined as follows:

Failure of IIO and IIO. Samejima's (1969) graded response model (Equation 3), which does not imply IIO, was used to generate data for the design half in which an IIO did not hold. However, particular choices of item parameters may produce an IIO by coincidence and sampling fluctuations may have the same effect. A pilot study showed that IRFs almost always intersected in dense regions of the latent variable θ , so that it seemed safe to use the graded response model. The restricted version of Muraki's (1992) rating scale version of the graded response model (Equation 4) was used to generate data for the design half in which an IIO holds.

Minvi. We investigated 16 *minvi* values covering a wide range (0.00 to 0.45, using increments of 0.03, and including the suggestion that $minvi = m \times 0.03$). Value $minvi = 0.00$ implies that all violations, however small, were tested.

Item discrimination (α). Weak and normal levels were used. For weak discrimination, parameters α_i were sampled from $\log N(-0.5 \ln 20, \ln 5)$, corresponding with mean α_i equal to 0.5 and variance 1. For normal item discrimination, parameters were sampled from $\log N(-0.5 \ln 2, \ln 2)$, corresponding with mean α_i equal to 1 and variance 1. For data sets violating IIO, α_i s were sampled for each item separately. When an IIO held, one value $\alpha_i = \alpha$ was sampled for all items. Item locations β_{ix} and λ_i were sampled from $N(0, 1)$.

Sample size (N). We used $N = 200, 433, 800$ ($N = 433$ is the sample size in the real-data example discussed later); θ was sampled from $N(0, 1)$.

Number of items (k). We used $k = 5, 10$ (based on real-data example), 15.

Number of answer categories ($m + 1$). We used $m + 1 = 3, 5$ (based on real-data example), 7.

The design had size $2 \times 16 \times 2 \times 3 \times 3 \times 3$, thus resulting in 1,728 cells. For each of the 54 combinations of α , N , k , and $m + 1$, we generated 500 data sets violating IIO (Equation 3) and 500 data sets consistent with IIO (Equation 4). Each data set was

Table 1. Sensitivity and Specificity of Method Manifest Invariant Item Ordering for Different *minvi* Values for the Cases Corresponding to the Real-Data Example

<i>minvi</i>	Item Discrimination			
	Weak		Normal	
	Sensitivity	Specificity	Sensitivity	Specificity
0.00	.650	.924	.902	.998
0.03	.650	.924	.902	.998
0.06	.650	.924	.902	.998
0.09	.650	.924	.902	.998
0.12	.650	.924	.902	.998
0.15	.650	.924	.902	.998
0.18	.652	.924	.902	.998
0.21	.654	.918	.908	.998
0.24	.716	.892	.924	.998
0.27	.830	.850	.960	.996
0.30	.900	.776	.974	.986
0.33	.942	.718	.994	.980
0.36	.970	.654	.996	.962
0.39	.988	.596	1.000	.950
0.42	.998	.558	1.000	.924
0.45	1.000	.490	1.000	.902

analyzed by means of method manifest IIO for each of the 16 *minvi* values, and the sensitivity and the specificity were computed for each *minvi* value.

Results

The sensitivity of method manifest IIO ranged from .275 to 1.000 across all design cells ($M = 0.849$, $SD = 0.195$), and the specificity ranged from .013 to 1.000 ($M = 0.686$, $SD = 0.337$). Only significant main effects on sensitivity and specificity are discussed (Kruskal–Wallis test for several independent samples, nominal Type I error of .05).

Table 1 shows the sensitivity and specificity for the two levels of item discrimination, 16 levels of *minvi*, $N = 433$, $k = 10$, and $m + 1 = 5$ (choices corresponded to real-data example; results for N , k , and $m + 1$ are compared with results in Table 1). For $N = 433$, $k = 10$, and $m + 1 = 5$, sensitivity was lower for a low item discrimination and low levels of *minvi* and increased as *minvi* increased for both levels of item discrimination. Specificity decreased as *minvi* increased. Based on sensitivity and specificity, $minvi = m \times 0.03 = 0.12$ seemed suitable for the real-data example.

Across the design cells, an increase in *minvi* resulted in higher sensitivity (.760 for $minvi = 0.00$ and .970 for $minvi = 0.45$) and lower specificity (.790 for $minvi = 0.00$ and .490 for $minvi = 0.45$). Both sensitivity and specificity were higher for

Table 2. Summary of Main Effects on Sensitivity and Specificity

	Sensitivity	Specificity
<i>minvi</i>	+	–
Item discrimination	+	+
Sample size	+	
Number of items	–	+
Number of answer categories	–	+

normal discrimination (.908 and .758, respectively) than for low discrimination (.789 and .614, respectively). Greater sample size resulted in higher sensitivity: .769 ($N = 200$) and .915 ($N = 800$). Greater numbers of items resulted in lower sensitivity: .979 ($k = 5$) and .715 ($k = 15$), but higher specificity: .350 ($k = 5$) and .913 ($k = 15$). Finally, the number of answer categories negatively influenced sensitivity: .875 ($m + 1 = 3$) and .838 ($m + 1 = 7$), but positively influenced specificity: .486 ($m + 1 = 3$) and .827 ($m + 1 = 7$). Table 2 gives the significant positive (+) and negative (–) main effects.

Discussion

Higher *minvi* values result in a greater probability that IIO is correctly identified (i.e., higher sensitivity) but also to a greater probability that a violation of IIO is ignored (i.e., lower specificity). The choice of *minvi* thus depends on the specific application for which IIO is investigated. A high cost of incorrectly accepting IIO requires a lower *minvi* value, but in other cases, including our real-data example, $minvi = m \times 0.03$ may be appropriate. Method manifest IIO also benefits from higher discrimination, more item scores, and larger sample sizes. The sensitivity is worse for short tests, but the specificity is better.

Coefficient H^T for Polytomously Scored Items

Theory of Coefficient H^T

Let \mathbf{X} denote the data matrix of N respondents (rows) by k items (columns), with scores $x = 0, \dots, m$ in the cells. Coefficient H (Mokken & Lewis, 1982; Sijtsma & Molenaar, 2002, chap. 4) is a measure for the accuracy by which k items constituting a scale order respondents (Mokken, Lewis, & Sijtsma, 1986). Sijtsma and Meijer (1992) showed for dichotomous items that when H is computed on the transposed data matrix, the resulting coefficient H^T is a measure for the accuracy by which N respondents order k items. Here, we generalize coefficient H^T to polytomously scored items.

We index respondents by g and h , and let the vectors \mathbf{X}_g and \mathbf{X}_h ($g, h \in \{1, \dots, N\}$) contain the scores of respondents g and h on the k items in the test. We assume that the

k item scores show at least some variation, so that $\text{Var}(\mathbf{X}_g) > 0$, for $g \in \{1, \dots, N\}$. Let $\text{Cov}(\mathbf{X}_g, \mathbf{X}_h)$ be the covariance between the scores of respondents g and h , and $\text{Cov}_{\max}(\mathbf{X}_g, \mathbf{X}_h)$ the maximum possible covariance given the marginal distributions of the k item scores of respondents g and h . The total score on item i is denoted by $T_i = \sum_{g=1}^N X_{ig}$. Vector \mathbf{T} contains the k item totals and vector $\mathbf{T}_{(g)} = \mathbf{T} - \mathbf{X}_g$ contains the k item totals minus the contribution of respondent g . The person scalability coefficient H_g^T is defined as the weighted normalized covariance,

$$H_g^T = \frac{\sum_{h \neq g} \text{Cov}(\mathbf{X}_g, \mathbf{X}_h)}{\sum_{h \neq g} \text{Cov}_{\max}(\mathbf{X}_g, \mathbf{X}_h)} = \frac{\text{Cov}(\mathbf{X}_g, \mathbf{T}_{(g)})}{\text{Cov}_{\max}(\mathbf{X}_g, \mathbf{T}_{(g)})}. \tag{7}$$

Thus, coefficient H_g^T expresses the association between the k item scores of respondent g and the k item totals minus the scores of respondent g . Because even for small samples, $\mathbf{T} \approx \mathbf{T}_{(g)}$, coefficient H_g^T expresses the degree to which the scores of respondent g have the same ordering as the item totals.

When an IIO holds for the k items, theoretically we expect a perfect association between the ordering of the item scores in \mathbf{X}_g and the total scores $\mathbf{T}_{(g)}$. When IRFs are close together, we expect the ordering of the item scores to be unstable and the values of many coefficients H_g^T to be low. When IRFs are further apart, we expect the orderings of the item scores to be more stable and better in agreement with the ordering of the item totals, thus resulting in many higher H_g^T values. Coefficient H^T wraps up the N person coefficients as

$$H^T = \frac{\sum_g \text{Cov}(\mathbf{X}_g, \mathbf{T}_{(g)})}{\sum_g \text{Cov}_{\max}(\mathbf{X}_g, \mathbf{T}_{(g)})}. \tag{8}$$

When k items have an IIO, the value of coefficient H^T is higher the further the IRFs are apart.

For k invariantly ordered items, assuming local independence it follows that $0 \leq H_g^T \leq 1$ and $0 \leq H^T \leq 1$ (proof available from first author). The value of 0 is obtained if the k IRFs coincide and $\text{Cov}(\mathbf{X}_g, \mathbf{X}_h) = 0$ for all respondent pairs. Maximally, $H^T = 1$, and this value is obtained if the agreement between the respondents' ordering of item scores and the ordering of the corrected item totals is maximal. We used a computational study to investigate the influence of item and test properties on coefficient H^T for polytomously scored items.

Computational Study: Influence of Item Properties and Test Length on H^T

Figure 2 illustrates that for dichotomously scored items coefficient H^T cannot distinguish well between data generated under a model inconsistent with IIO (Figure 2a) and one consistent with IIO (Figure 2b); item locations and mean item discriminations are identical, and for $\theta \sim N(0, 1)$ one finds $H^T \approx .5$. Hence, Sijtsma and Meijer (1992)

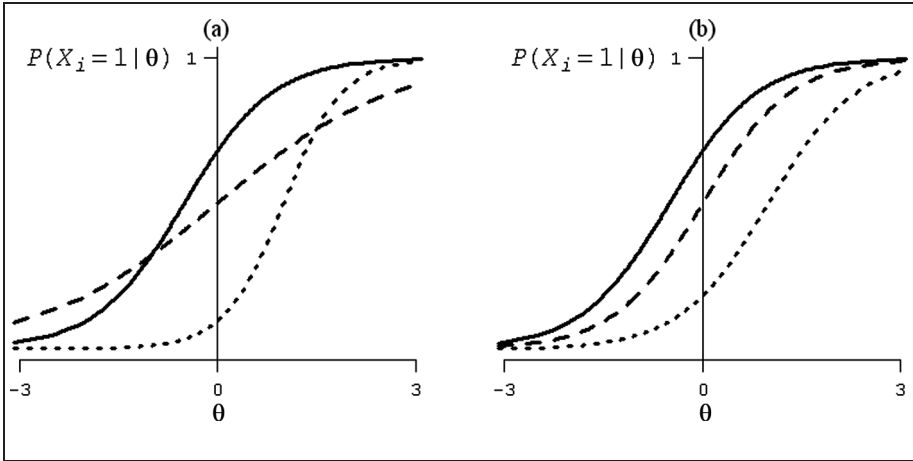


Figure 2. Failure of invariant item ordering (IIO; a) and IIO (b), both cases produce $H^T = .50$ and are consistent with the two-parameter logistic model: $\beta_i = 0.5, 0, 1$ (both cases); (a) $\alpha_i = 1.5, 0.75, 2.25$, and (b) $\alpha_i = 1.5, 1.5, 1.5$

recommended using Mokken Scale Analysis (e.g., Sijtsma & Molenaar, 2002) to first identify and remove items that have flat IRFs and tend to produce many intersections with other, often steeper IRFs. For the remaining items, they suggested concluding that an IIO held if $H^T \geq .3$, and the percentage of negative person scalability values (not discussed here) did not exceed 10; else, IIO was rejected.

We use method manifest IIO to select items, which have an IIO, and then compute coefficient H^T for the selected items. Instead of method manifest IIO, Sijtsma and Meijer (1992) suggested using Mokken Scale Analysis, but this method uses scalability coefficient H to assess the slopes of the IRFs but not whether different IRFs intersect. In their Monte Carlo study, these authors did not actually use Mokken Scale Analysis but a person scalability coefficient to have more power distinguishing failure of IIO from IIO. Because we used method manifest IIO to select an item set that is consistent with IIO, the use of coefficient H^T sufficed.

In their Monte Carlo study, for dichotomous items Sijtsma and Meijer (1992) found that coefficient H^T increases as distance between item locations increases or item discrimination increases. Sample size and test length hardly affected H^T values. We used a computational study for polytomous items involving parameter values for H^T (hence, sample size did not play a role) to investigate IIO conditions so as to learn how H^T may be used once an IIO has been ascertained by means of method manifest IIO. Based on Sijtsma and Meijer (1992), we included distance between item locations, item discrimination, and number of items, but with more variation in levels. We expected similar trends in H^T as for dichotomous items. The factors number of answer categories and distance between adjacent ISRFs were unique for polytomous items.

Method

Coefficient H^T was computed at the population level ($\theta \sim N(0, 1)$) for the restricted version of Muraki's (1990) rating scale version of the graded response model (Equation [3]), which implies IIO. The dependent variable was the expected value of coefficient H^T (computational details for coefficient H^T and its expectation under Equation [6] can be obtained from the first author). The five independent variables were

Number of items (k). Test length was: $k = 5, 10, 15$. Tests consisting of larger numbers of items were not investigated so as to facilitate interpretation of results.

Number of answer categories ($m + 1$). This number equaled $m + 1 = 2, 3, 5, 7$.

Item discrimination (α). Discrimination values were: $\alpha = 0.5, 1, 1.5, 2$.

Distance between adjacent item locations (λ_i). Item locations were symmetrical relative to the mean of the θ distribution ($\mu_\theta = 0$), and adjacent item locations were at a constant distance. The distance between the location of the most attractive item (λ_1) and the least attractive item (λ_k) is denoted as Δ ; $\Delta = 0, 2, 4$ (for $\Delta = 0$, all item locations coincide). The distance between adjacent items depended on Δ and test length k .

Distance between adjacent ISRFs (ε_{ν}). For dichotomously scored items, by definition $\varepsilon_1 = 0$ but for polytomously scored items, the parameters $\varepsilon_1, \dots, \varepsilon_m$ may vary. Two variations were considered. First, the extremes were fixed ($\varepsilon_1 = -1$ and $\varepsilon_m = 1$, for $m > 1$), and the other $m - 2$ ISRFs were located at equal distances between these extremes. Thus, for greater m , the ISRFs were more densely located around the item location, λ_i . Second, the distance between the locations of adjacent ISRFs was fixed at 0.5, which resulted in a greater dispersion of the ISRFs around the item location λ_i as m was greater.

The design had size $3 \times 4 \times 4 \times 3 \times 2$, thus resulting in 288 cells. Because dichotomously scored items only have one item step, the two cells in the design corresponding to the distance between adjacent ISRFs collapsed.

Results

For the design factors typical of polytomous items, which are number of answer categories and distance between adjacent ISRFs, we found little effect on coefficient H^T (no more than a few hundredths between corresponding design cells). This justifies discussing results for only the simplest case of $m + 1 = 2$. For the cells concerning coinciding IRFs ($\Delta = 0$), we found that $H^T = 0$ (consistent with mathematical proof obtainable from first author). Table 3 shows H^T values for all combinations of number of items, distance between item location (for $\Delta = 2, 4$) and item discrimination. Similar to results found by Sijtsma and Meijer (1992), distance between item location and item discrimination had positive effects on H^T . Unlike their results, however, where the number of items had no significant negative effect for $k = 9$ and 18, our results

Table 3. H^T Values for Varying Number of Items (k), Distance Between Item Locations (Δ), and Item Discrimination

k	Δ	Item Discrimination			
		0.5	1	1.5	2
5	2	0.038	0.145	0.279	0.413
	4	0.135	0.405	0.610	0.742
10	2	0.031	0.121	0.239	0.362
	4	0.112	0.355	0.557	0.698
15	2	0.029	0.114	0.227	0.346
	4	0.106	0.340	0.540	0.683

show a negative effect of the number of items. This discrepancy can be explained by the levels we used for the number of items ($k = 5, 10$, and 15), where we found the largest decrease in H^T between $k = 5$ and 10 . These results suggest that beyond approximately 10 items there is little to no effect of the number of items on the value of H^T .

Discussion

The computational results supported the expectation that when items are further apart, for a fixed θ the items' response probabilities show more variation and the ordering of a respondent's item scores better resembles the ordering of the items' total scores. Given IIO, coefficient H^T expresses the degree to which the ordering of the item totals is reflected by the individual vectors of item scores. The next section illustrates the practical use of method manifest IIO and the H^T coefficient.

A Real-Data Example

Method manifest IIO and coefficient H^T were used for investigating whether an IIO held in the two subscales for measuring deference ($k = 9$) and achievement ($k = 10$) from the Dutch version of the Adjective Checklist (Gough & Heilbrun, 1980). The subscales were not constructed with an IIO in mind, but are well suited for demonstrating the exploratory use of method manifest IIO. Items consist of an adjective and five ordered answer categories. Table 4 shows the item labels (negatively worded items were recoded). The respondents were 433 students, who were instructed to consider whether an adjective described their personality and rate the answer category that fitted best to this description. Vorst (1992) collected the data, which are available from the R package *Mokken* (Van der Ark, 2007).

Prior to investigating IIO, following Sijtsma and Meijer (1992) a Mokken Scale Analysis was done on both subscales. Inclusion of all items resulted in $H = .307$

Table 4. Number of Violations for the Deference Scale and the Achievement Scale

Deference		Achievement		
Items	Step	Items	Step	
	1		1	2
Impulsive	0	Quitting ^a	0	0
Demanding	0	Unambitious ^a	0	0
Forceful	0	Determined	0	0
Rebellious	0	Active	0	0
Uninhibited	0	Energetic	0	0
Bossy	0	Ambitious	1	0
Reckless	0	Alert	2	—
Boastful	0	Persevering	1	0
Conceited	0	Thorough	0	0
		Industrious	0	0
Coefficient H^T	0.320			0.116

a. Indicates negatively worded items.

for subscale Deference, and $H = .308$ for subscale Achievement. Following Mokken and Lewis (1982), $3 \leq H < .4$ stands for a weak scale.

For using method manifest IIO, the IRFs were estimated after adjacent rest scores were joined until each group contained at least $N/5 = 86$ respondents (Molenaar & Sijtsma, 2000, p. 67). Method manifest IIO was performed for $minvi$ values ranging from 0 to 0.45 using increments of 0.03 thus allowing how conclusions depended on different $minvi$ values. After method IIO had identified an item subset coefficient H^T was computed for this subset. The R package mokken (Van der Ark, 2007) was used for the computations.

Table 4 shows for $minvi = 0.03 \times m = 0.12$ that subscale Deference did not have significant violations of IIO, and that $H^T = 0.320$. Subscale Achievement had two significant violations, both involving item Alert. Removal of this item resulted in a subscale containing nine items for which an IIO held. Coefficient H^T cannot be computed for respondents that have the same scores on all items; hence, six respondents were excluded. For the remaining 427 respondents, we found $H^T = .116$. Support for IIO is stronger for Deference than for Achievement. Interpretation of H^T is discussed in the next section.

For subscale Deference, method manifest IIO produced the same results for varying $minvi$ values. For subscale Achievement, method manifest IIO produced the same results until $minvi = 0.21$ and resulted in 0 violations of IIO for higher $minvi$ values. Because $minvi$ values exceeding 0.24 are large in most applications, based on these results we concluded that method manifest IIO is robust for different $minvi$ values.

General Discussion

We used a top-down sequential procedure based on method manifest IIO for selecting a subset of items having nonintersecting IRFs. Thus, not all item subsets were investigated, and once removed, an item was not reevaluated for possible reselection in later steps of the procedure. Alternative selection procedures (e.g., genetic algorithms; Michalewicz, 1996), which assess all possible item subsets, may be investigated in future research so that possibly larger and different item subsets for which an IIO holds may be identified.

IIO research is new, and experience on how to interpret results has to accumulate as more applications become available. For the time being, we tentatively generalize the heuristic rules proposed by Mokken and Lewis (1982) for interpreting values of scalability coefficient H to the interpretation of H^T values, provided an IIO holds for an item set. Thus, we propose the following: $H^T < 0.3$ means that the item ordering is too inaccurate to be useful; $0.3 \leq H^T < 0.4$ means low accuracy; $0.4 \leq H^T < 0.5$ means medium accuracy; and $H^T \geq 0.5$ means high accuracy. Based on these rules, the nine items from the Deference subscale may be ordered with low accuracy ($H^T = 0.320$) and the remaining nine items from the Achievement scale do not have an IIO ($H^T = 0.116$).

The assumption of an IIO is both omnipresent and implicit in the application of many tests, questionnaires, and inventories. Test constructors and test users alike often assume that the same items are easy or attractive for each of the respondents to whom the items are administered but rarely put this strong assumption to the test of empirical evaluation. Yet an established IIO underpins and greatly facilitates the interpretation of the test results, for example, when the test administration procedure is based on the ordering of the items from easiest to most difficult, the items reflect a developmental sequence of cognitive steps assumed to be the same for everyone or when the set of items is assumed to reflect a hierarchical or cumulative structure. Invariant item ordering for polytomously scored items is an unexploited terrain. This study provides a first start for this interesting topic and shows directions for future explorations.

Declaration of Conflicting Interests

The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References

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

- Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1987). *Revisie Amsterdamse Kinder Intelligentie Test. Handleiding* [Revision Amsterdam Child Intelligence Test. Manual]. Lisse, The Netherlands: Swets & Zeitlinger.
- Bouwmeester, S., & Sijtsma, K. (2007). Latent class modeling of phase transition in the development of transitive reasoning. *Multivariate Behavioral Research*, 42, 457-480.
- Gough, H. G., & Heilbrun, A. B. (1980). *The Adjective Check List manual, 1980 Edition*. Palo Alto, CA: Consulting Psychologists Press.
- Ligtvoet, R., Van der Ark, L. A., Bergsma, W. P., & Sijtsma, K. (2009). *Polytomous latent scales for the investigation of the ordering of items*. Manuscript submitted for publication.
- 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.
- Michalewicz, Z. (1996). *Genetic algorithms + data structures = evolution programs*. Berlin, Germany: Springer.
- Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement*, 6, 417-430.
- Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "Mokken scale: A critical discussion." *Applied Psychological Measurement*, 10, 279-285.
- Molenaar, I. W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows*. Groningen, The Netherlands: iec ProGAMMA.
- 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: Applications for an EM algorithm. *Applied Psychological Measurement*, 16, 159-177.
- Raijmakers, M. E. J., Jansen, B. R. J., & Van der Maas, H. L. J. (2004). Rules in perceptual classification. *Developmental Review*, 24, 289-321.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.
- 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., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, 63, 183-200.
- 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.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20(11), 1-19.

- Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*, 139-163.
- Vorst, H. C. M. (1992). [Responses to the Adjective Checklist] Unpublished raw data.
- Watson, R., Deary, I., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine, 38*, 575-579.
- Wechsler, D. (1999). *WISC manual*. San Antonio, TX: Psychological Corporation.