



Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Mokken scale analysis as time goes by: An update for scaling practitioners

Klaas Sijtsma^{a,*}, Rob R. Meijer^b, L. Andries van der Ark^a^a Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands^b Department of Psychometrics and Statistics, Faculty of Behavioral Sciences, University of Groningen, Groningen, The Netherlands

ARTICLE INFO

Article history:

Received 4 July 2010

Accepted 18 August 2010

Available online xxx

Keywords:

Hierarchical scale

Invariant item ordering

Item response theory

Non-cognitive measurement

Mokken scaling

Personality measurement

ABSTRACT

We explain why invariant item ordering (IIO) is an important property in non-cognitive measurement and we discuss that IIO cannot be easily generalized from dichotomous data to polytomous data, as some authors seem to suggest. Methods are discussed to investigate IIO for polytomous items and an empirical example shows how these methods can be used in practice.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Recently, Watson and colleagues (Stewart, Watson, Clark, Ebmeier, & Deary, 2010; Watson, Deary, & Austin, 2007; Watson, Roberts, Gow, & Deary, 2008) investigated for different personality inventories whether items measuring the same attribute formed a hierarchical scale. Items form a hierarchical scale when the ordering of the items according to their popularity (or mean score) is the same across different values of the latent trait. This property is also named invariant item ordering (IIO; Sijtsma & Junker, 1996). Meijer (2010) discussed that the way Watson and others investigated whether items form a hierarchical scale was not correct. In a reply to his article, Watson and Deary (2010) partly agreed with his criticism, but also referred to an article by Sijtsma, Debets, and Molenaar (1990) that allegedly used the P-matrix to investigate whether items form a hierarchical scale. However, in the Sijtsma et al. (1990) paper this matrix is not used to investigate whether items form a hierarchical scale, but whether *item step response functions* form a hierarchy. In the present article, we argue that a hierarchy of item step response functions need not imply a hierarchical scale for the items. Hence, the P-matrix is not an adequate tool for assessing whether a scale is hierarchical.

We applaud the use of more sophisticated techniques by Watson and colleagues, but apparently the literature on Mokken scale analysis (MSA) gives rise to some misunderstandings. Therefore, the aim of this paper is to discuss (1) why IIO is an important as-

pect of personality scales; (2) Mokken's model for the analysis of polytomous items; and (3) why the results for dichotomous items cannot be easily generalized to polytomous items. This takes us to a second source of confusion surrounding hierarchical scales, which is the assumption made by practitioners that high values of Mokken's scalability coefficient H support such a hierarchy. We argue that high values of H found in real-data analysis are not adequate for assessing whether a scale is hierarchical. Instead, high H values establish a person ordering, which is exactly what Hemker, Sijtsma, and Molenaar (1995, p. 340) claimed (however, see Watson & Deary, 2010). Hence, we discuss (4) why scalability coefficient H is not an index for IIO, and continue with (5) a method to investigate IIO for polytomous items; (6) an R program that can be used by practitioners to investigate IIO for polytomous items, and (7) an empirical example illustrating the use of the R program.

2. Why is invariant item ordering important in non-cognitive measurement?

The measurement of psychological traits often assumes, either implicitly or explicitly, that items used in inventories represent different levels of intensity with respect to the attribute of interest. For example, when measuring depression we assume that the item "thoughts of ending your life" represents a higher level of depression than the item "feeling no interest in things", and when measuring anxiety, the item "spells of terror or panic" has a higher intensity than the item "feeling tense" (Meijer & Baneke, 2004). Item intensity is often quantified as the mean item score in the group of interest. Suppose now that, after data collection, all items

* Corresponding author. Tel.: +31 13 4663222/4662544; fax: +31 13 4663002.

E-mail address: k.sijtsma@uvt.nl (K. Sijtsma).

have been recoded so that the higher the item scores, the higher the respondent's position on the attribute scale. If items are ordered by decreasing mean scores, then this is taken as an ordering by increasing intensity with respect to the measured attribute.

Now, we would make an aggregation error if we inferred from the item ordering that, without additional proof, this ordering not only holds in the complete group but also for each individual respondent. The error is that one cannot infer just like that from a property shown to hold at the group level that a similar property holds at the level of individuals. In general, it is not justified to go from a higher aggregation level to a lower level. In order to justify statements at the individual level, it needs to be established by means of empirical research whether the item ordering also holds at the lower, individual level. To put it most simply: If the group to which John belongs has a mean body height of 182 cm, we cannot infer that John's body height also is 182 cm. If we are interested in his individual body height, we need to measure John separately.

Suppose it had been empirically established that the ordering of items holds for individual respondents as well, then this would lend much credence to the constructed scale. First, if IIO has been established, we know for sure that the item ordering is the same for the individuals making up the population of interest, and also for interesting subgroups. If the item ordering holds for all individuals it also holds for subgroups of these individuals. This is true because one can always go the other way, from the lower to the higher aggregation level. Second, IIO gives a clear meaning to test scores. Let us conceive of items as symptoms; then, when IIO holds, compared to a person with a lower score, a person with a higher score has the same symptoms plus more symptoms representing higher intensity levels. This hierarchy of symptoms can be inferred from the total score and supports the useful interpretation of total

scores, not only as indicators of *attribute levels* but also as summaries of particular *sets of symptoms*. The higher the total score, the more the set of symptoms is extended with additional ones, and symptoms are always added in the same order, from low to high intensity.

3. Mokken's models for the analysis of dichotomous and polytomous items

Mokken (1971) proposed two models for dichotomous items, nowadays recognized as item response theory (IRT) models, one of which was meant for ordinal person measurement and the other both for ordinal person and item measurement. We discuss the polytomous-item versions as proposed by Molenaar (1982, 1986, 1991, 1997), of which Mokken's dichotomous-item models are special cases.

The first model is the monotone homogeneity model (MHM), which is based on the following assumptions

- (1) All J items in an inventory measure the same underlying attribute, which is represented by a latent variable θ (unidimensionality). A second assumption is local independence but unidimensionality implies local independence, and thus we do not further discuss the latter assumption. In the context of MSA, unidimensionality is investigated using an automated item selection procedure.
- (2) Items are monotone positively related to the underlying attribute (monotonicity). Let random variable X_j be the score on item j , which has values $x_j = 0, \dots, m$; for 5-point rating scales, this means $x_j = 0, \dots, 4$. We define the item step response function (ISRF) as the probability of obtaining an

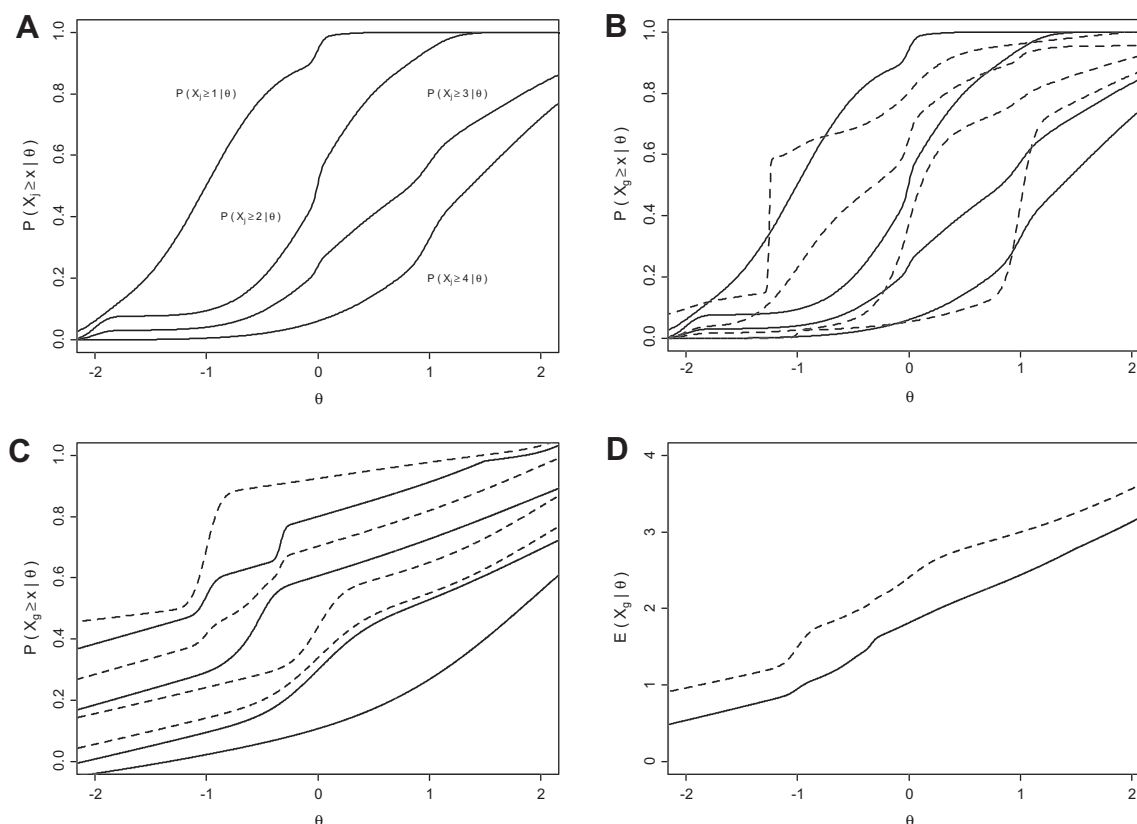


Fig. 1. (A) Four ISRFs of an item with five ordered item scores; (B) Two sets of four ISRFs each, each set for an item with five ordered scores (index g in y -axis label stands for either item j or k throughout); (C) Two sets of ISRFs consistent with the DMM; (D) Two IRFs with IIO, consistent with panel C.

item score of at least x_j as $P(X_j \geq x_j | \theta)$, for $x_j = 1, \dots, m$, thus ignoring $x_j = 0$ because the probability by definition equals $P(X_j \geq 0 | \theta) = 1$. This renders the probability not of interest.

Figure 1A shows the four unidimensional, monotone ISRFs for a 5-point rating scale item. For one item, the ISRFs are cumulative and by definition cannot intersect. For example,

$$P(X_j \geq 2 | \theta) = P(X_j = 2 | \theta) + P(X_j \geq 3 | \theta),$$

so that, due to probabilities being nonnegative, in Fig. 1A the ISRF on the left-hand side in the equation [$P(X_j \geq 2 | \theta)$] lies above the ISRF on the right-hand side [$P(X_j \geq 3 | \theta)$]. Fig. 1B shows the ISRFs of two items, j and k , and clarifies that ISRFs of different items are allowed to intersect (but this is not obligatory).

The MHM does not contain enough restrictions to estimate the latent variable θ . Molenaar (1982) proposed to use total score X_+ (which is the sum of the J item scores) for ordering persons on the scale of latent variable θ ; also, see Van der Ark (2005). Hence, the MHM is an ordinal measurement model for persons.

Several methods are available for investigating the goodness-of-fit of the MHM to rating-scale data. Program MSP5.0 (Molenaar & Sijtsma, 2000) contains a procedure, which selects items (whenever possible) in one or more unidimensional clusters, each of which is sensitive to another attribute. Nonparametric regression methods can be used to estimate the ISRFs from the data and investigate whether they are monotone. After the fit of the MHM to the data has been established, item-pair, item and total-scale quality can be assessed by means of scalability coefficients H_{jk} , H_j , and H , respectively. Coefficient H_j may be interpreted as a discrimination index and coefficient H is a weighted mean of the item coefficients, which expresses average discrimination power and thus may be interpreted as an index for the precision of ordering

persons by means of their total scores on the latent scale θ (Mokken, Lewis, & Sijtsma, 1986; Sijtsma & Meijer, 2007).

Molenaar (1982, 1986, 1997; Sijtsma et al., 1990) also proposed the DMM for polytomous items, which adds to the MHM the assumption that the ISRFs of different items do not intersect; see Fig. 1C for an example. Extending Mokken's theory for dichotomous items, Molenaar also proposed the use of the P-matrix and the P(0)-matrix for investigating whether nonintersection of the ISRFs of the different items in a particular inventory is plausible in real data. This method and other methods obtained for polytomous Mokken models were implemented in the program MSP. It is here that we suspect a misunderstanding has arisen among applied researchers who use this program, and in this article we intend to better explain the DMM and its consequences.

4. Why results for dichotomous items cannot be easily generalized to polytomous items

For dichotomous items (scores 0, 1), the ISRF equals $P(X_j \geq 1 | \theta) = P(X_j = 1 | \theta)$, so that one ISRF, now called item response function (IRF), suffices to describe the item scores. Figure 2A shows two IRFs for dichotomous items j (solid curve) and k (dashed curve). For each value of latent variable θ , the probability of obtaining a 1 score on item j is greater than for item k ; hence, they exhibit IIO. Sijtsma and Junker (1996) reviewed several methods for investigating the IIO property in dichotomous-item data.

Figure 2B shows for two items with three ordered answer categories the two sets of two ISRFs. None of the ISRFs intersect, which is consistent with the DMM, but does this imply we have IIO for these items? This can be investigated as follows. The IIO property refers to the ordering of items, not ISRFs. Hence, we must look at response functions that summarize the ISRFs for one item, and

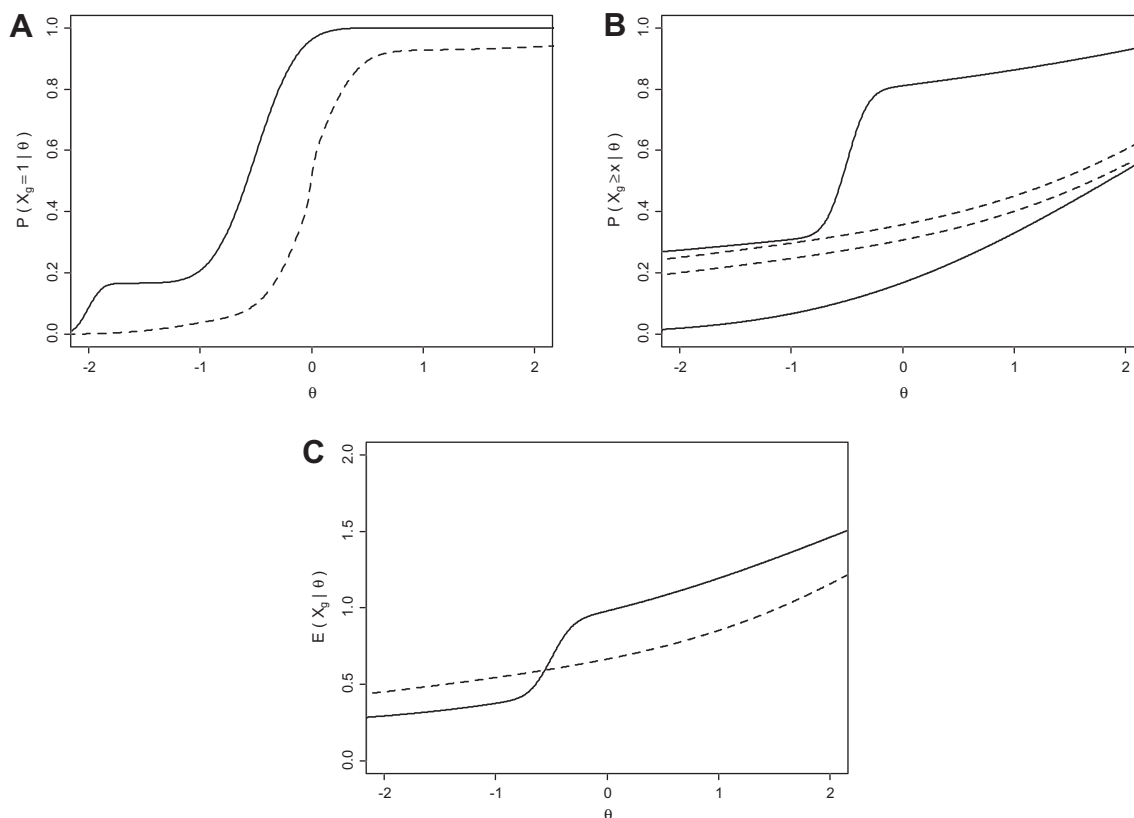


Fig. 2. (A) Two IRFs for two dichotomous items; (B) Two sets of two ISRFs, each set for an item with three ordered scores, consistent with the DMM; (C) Two IRFs violating IIO, consistent with panel B.

then check whether the summary functions of different items intersect. A good candidate for summarizing the ISRFs of one item is the sum of its ISRFs, which can be written as the conditional expectation of the item score, $E(X_j|\theta)$; see Sijtsma and Hemker (1998). Fig. 2C shows that these sum functions—polytomous-item IRFs for short or just IRFs—intersect for items j and k , even though their parent-ISRFs do not intersect (Fig. 2B). Hence, these items do not have the IIO property. Similar examples can be set up in which non-intersecting ISRFs do imply non-intersecting IRFs, hence implying IIO. For example, the sets of ISRFs in Fig. 1C do imply IIO, as Fig. 1D shows. The point is that the DMM may imply IIO for some item sets but not for others; hence, the DMM does not guarantee IIO. Thus, the DMM is not the appropriate model to investigate if one is interested in the IIO property.

We understand that it is tempting to infer from a fitting DMM that the items also have IIO, and that the result that this need not be the case is even somewhat counterintuitive. The researcher might now be tempted to resort to other polytomous IRT models, such as the popular partial credit model (Masters & Wright, 1997) or the graded response model (Samejima, 1997). However, this does not provide a solution, as Sijtsma and Hemker (1998) proved mathematically that, like the DMM, these and other well-known polytomous IRT models do not imply the IIO property. These authors also showed that only extremely restrictive polytomous IRT models do imply the IIO property (Meijer, 2010). However, use of these restrictive models is of little practical value since it is precisely their restrictiveness that makes these models unrealistic for many data sets.

We checked whether Sijtsma et al. (1990) erroneously suggested anywhere that the DMM implies IIO and found they did not. More interestingly, it may be noted that almost all polytomous

IRT models were defined at the level of individual item scores, thus aimed at estimating item parameters for each combination of an item and its answer categories instead of the item as a whole. Useful as this may be for some applications, we contend that this perspective seems to ignore the perspective of the practical researcher, who can only replace items from his inventory by other items, which are expected to function better, but not parts of items represented by estimated parameters by other parts. Hence, the researcher interested in IIO needs models and methods that have the whole item at the center of their interest.

5. Is coefficient H an index for IIO?

The second problem refers to coefficient H , which some authors mistakenly use as an index for IIO. The source of this confusion seems to reside with the deterministic Guttman model for dichotomous items, for which all H coefficients are equal to 1. Fig. 3 shows the typical step IRFs for four Guttman items, which do not intersect (although they mostly coincide) and hence exhibit IIO. Data consistent with the deterministic Guttman model are error-free but real data contain much error, and the Guttman model cannot describe such data. Instead, in practice researchers do not use the Guttman model but IRT models defining IRFs such as those in Fig. 2A, which describe real data much better as they do allow for random error. For real data, the H values usually are between, say, 0.3 and 0.6, and it has been erroneously inferred from such values that a “probabilistic version of the Guttman model”, with IRFs such as in Fig. 2A that do not intersect, is at the basis of these H values. Again, this is a misunderstanding as other sets of IRFs, including intersecting IRFs, may generate similar H values. This is illustrated by the sets of IRFs in Fig. 4A and B, which together with a standard normal latent variable θ both produce $H_{jk} = 0.40$ (computational details may be requested from the author). Thus, particular H values cannot distinguish sets of intersecting IRFs from sets of non-intersecting IRFs, and the conclusion must be that H is not an index of IIO.

Confusion may also have come from the result that the H coefficients can be written as decreasing functions of the number of so-called Guttman errors in a data matrix (Hemker et al. 1995; Sijtsma & Molenaar, 2002, p. 53). Thus, H can also be interpreted as expressing the distance of the data from the Guttman model: The higher H , the more the data resemble perfect data consistent with the Guttman model. Then, it may be tempting to think that, because the Guttman model has the IIO property, a value such as $H = 0.4$ or higher, which is common for real data, actually stands for a set of IRFs that are not Guttman IRFs but nevertheless non-intersecting as in Fig. 2A. But this would be the wrong inference, as we just saw, because any H value, not only the higher values,

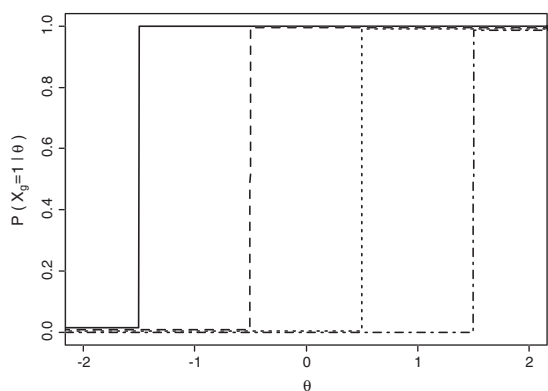


Fig. 3. Four IRFs for dichotomous items, consistent with the Guttman model.

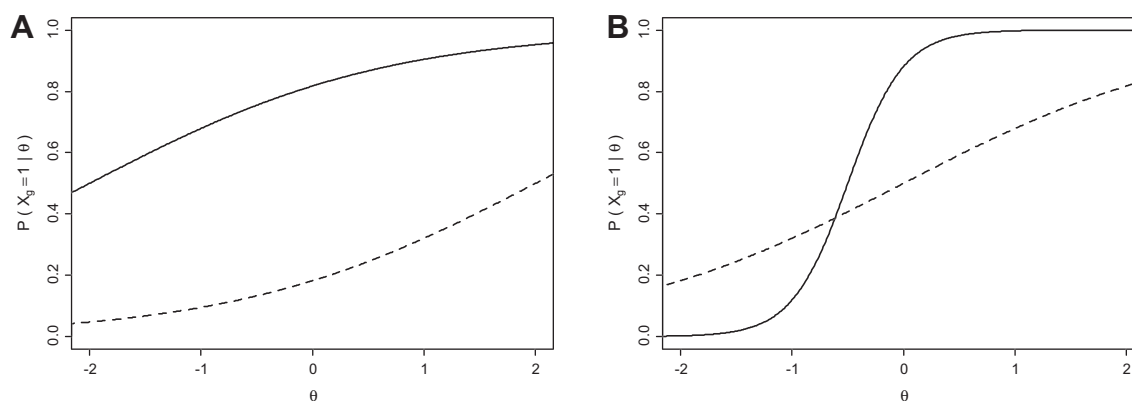


Fig. 4. Two IRFs for dichotomous items with IIO (panel A), and two IRFs for dichotomous items with IIO (panel B). For both, $H_{jk} = 0.40$ (θ standard normal).

might as well be consistent with a set of *intersecting* IRFs. To summarize, a set of items having IIO can have (but need not have) a high H value and, reversely, a high H value need not be (but can be) the product of a set of items having IIO. The idea that only high H values are consistent with IIO is incorrect.

Coefficient H must be used as an index for measurement precision on the ordinal person scale but not as an index for IIO. The same conclusion holds for polytomous items. We recommend first ascertaining that the IRFs or the ISRFs of all the items in the scale are monotone. If they are, we know that the ordering of persons by increasing total scores provides an ordinal scale. Positive H values alone do not suffice to draw this conclusion; the monotonicity of the IRFs has to be investigated and satisfied separately.

6. A method to investigate IIO for polytomous items

Ligtvoet, Van der Ark, Te Marvelde, and Sijtsma (2010) proposed a method to investigate IIO for polytomous items without the assumption of a particular IRT model. First, their *method manifest* IIO is checked for pairs of items. We define the rest score R as the total score on the $J - 2$ items excluding the scores on items j and k . For $J - 2$ items with $m + 1$ ordered scores each, rest score R theoretically runs from 0 to $(J - 2)m$. The method checks for each pair of items j, k with item means ordered such that $E(X_j) \leq E(X_k)$, whether the expected score for item j given the rest score is smaller or equal than the expected score for item k ; that is, whether $E(X_j|R) \leq E(X_k|R)$, for each value r of rest score R .

In a real-data analysis, the item means are estimated using sample mean scores, \bar{X}_j and \bar{X}_k , and items are numbered and ordered such that $\bar{X}_j \leq \bar{X}_k$. Then, if conditional sample means exhibit the reverse ordering for a particular value r of rest score R , such that $\bar{X}_j|R = r > \bar{X}_k|R = r$, a one-sided one-sample t -test is done to test the null hypothesis that the expected conditional item means are equal—the boundary of the permissible set of conditional item means given that $E(X_j) \leq E(X_k)$ —against the alternative that the expected conditional mean of item j exceeds that of item k , which is a violation of IIO. A protection against taking very small violations seriously is to test sample reversals only when they exceed a minimum value denoted *minvi* with a default value of $m \times 0.03$.

Ligtvoet et al. (2010) suggested the following data-analysis procedure. First, for each of the J items the frequency is determined that the item is involved in significant violations—reversals of the expected item ordering—that exceed *minvi* with any of the other $J - 1$ items. When there are items involved in violations of IIO, the item with the highest frequency is removed from the inventory and the procedure is repeated for the remaining items, and so on. This method is suited for exploratory data analysis. For confirmatory data analysis, Ligtvoet et al. (2010) suggested using a similar procedure for all item pairs, but items are not removed.

For the set of items for which IIO was found, the authors computed the H^T coefficient. For polytomous items, this coefficient is a generalization of the H^T coefficient for dichotomous items (Sijtsma & Meijer, 1992). The H^T coefficient has the same structure as the H coefficient but interchanges the roles of persons and items. Hence, it assesses the degree to which a sample of persons agrees on the ordering of the items. For J items having IIO, it can be shown that $0 \leq H^T \leq 1$. An important property of H^T is that when J items have IIO, the value of H^T is higher the further the IRFs are apart. We refer the reader to Ligtvoet et al. (2010), who provided tentative guidelines for interpreting numerical values.

7. A computer program to investigate IIO for polytomous items

Method manifest IIO is available in the R package *Mokken* as method *check.iio* (Van der Ark, 2007). Furthermore, this package

contains different functions (*coefH*, *aisp*, *check.monotonicity*, *check.pmatrix*, and *check.restscore*) to investigate different assumptions of the MHM and the DMM. Except for the graphics, the function names and the output in *Mokken* are similar to function names and output in the package *MSP5* for Windows (Molenaar & Sijtsma, 2000). An advantage of the R package *Mokken* over *MSP5.0* is that the latest developments with respect to checking model assumptions are incorporated and regularly updated. Readers unfamiliar with R packages may want to consult an introductory guide to *Mokken* (Van der Ark, 2010).

8. Example: analysis of SPPC data

We used R package *Mokken* to analyze the data from the six subscales of Harter's (1985) Self-Perception Profile for Children (SPPC) ($N = 268$, boys; see Meijer, Egberink, Emons, & Sijtsma, 2008). The SPPC measures how children between 8 and 12 years of age judge their own functioning in several specific domains and how they judge their global self-worth. Five of the six subscales represent specific domains of self-concept: Scholastic Competence (SC), Social Acceptance (SA), Athletic Competence (AC), Physical Appearance (PA), and Behavioral Conduct (BC). The sixth scale measures Global Self-worth (GS), which is a more general concept. Each subscale uses six 4-point rating scale items. Given the moderate sample size, of each IRF we estimated four discrete points so as to have enough precision in each combined group of adjacent restscores. Table 1 shows the results.

For five of the six subscales, we concluded that all six IRFs were monotone, and that the H values could be safely interpreted. Using widely accepted rules of thumb (Sijtsma, & Molenaar, 2002, p. 60), SC was a medium scale ($0.4 \leq H < 0.5$), PA and BC were weak scales ($0.3 \leq H < 0.4$), GS was a borderline case ($H < 0.3$) and AC clearly was unscalable. For subscale SA, we found one significant violation of IRF monotonicity, which means that five IRFs are monotone and one shows a local decrease. This raises the obvious practical question whether this should withhold the researcher from interpreting H (weak scale), whether he should remove the item with the violation, or whether he should consider one violation a borderline case and ignore it. This is a problem we cannot resolve here, as it requires judgment in combination with statistical robustness research, which is unavailable at the moment.

For five subscales, we conclude that IIO held. The corresponding H^T values were all smaller than 0.3, suggested by Ligtvoet et al. (2010) as the minimum values for the precision of an item ordering. Concretely, this means that the six IRFs are close together, so that respondents may find it difficult to distinguish one item from its neighbor in terms of intensity. For GS, we found two significant violations of IIO, both involving item 4. We recommend not interpreting the corresponding H^T value. After item 4 was removed, the remaining five items had IIO but $H^T = 0.12$ suggested their IRFs were close together.

Table 1

SPPC subscale results for IRF monotonicity and coefficient H , and invariant item ordering and coefficient H^T . "1(0)": 1 violation $>$ *minvi* found, 0 significant; etc.

Subscale	#Violations M	H	# Violations IIO	H^T
SC	1(0)	0.40	11(0)	0.08
SA	4(1)	0.35	14(0)	0.15
AC	5(0)	0.22	6(0)	0.07
PA	0(0)	0.38	8(0)	0.01
BC	2(0)	0.33	6(0)	0.09
GS (6 items)	2(0)	0.29	6(2)	0.11
GS (5 items)	0(0)	0.25	0(0)	0.12

Table 2
Relationships between models and coefficients for dichotomous and polytomous Mokken models.

Model	Assumptions	Methods	Measurement properties	H coefficient	H^T coefficient
<i>Dichotomous Items</i>					
MHM	UD&LI M	AISP IR-regr.	Ordinal Person Scale	0–1	NA
DMM	UD&LI M IIO	AISP IR-regr. S&M chap.6	Ordinal Person and Item Scales	0–1	0–1
<i>Polytomous Items</i>					
MHM	UD&LI M	AISP IR-regr.	Ordinal Person Scale	0–1	NA
DMM (Molenaar)	UD&LI, M, Non-intersecting ISRFs	AISP&IR-regr. S&M chap.7,8	Ordinal Person Scale + Invariantly Ordered ISRFs	0–1	NA
IIO-separate (Ligtvoet)	Non-Intersecting IRFs	<i>Method manifest</i> IIO	IIO	NA	0–1

UD & LI = Unidimensionality & Local independence; M = Monotonicity; NA = Not applicable; S&M = Sijtsma & Molenaar (2002).

9. Recommendations

We recommend that researchers first investigate whether the measurement model fits their data before they interpret the H or H^T coefficients. Table 2 summarizes MSA.

For dichotomous-item inventories and the MHM, we recommend first investigating unidimensionality by means of the automated item selection procedure (AISP) in MSP5.0 and Mokken. Monotonicity should be investigated by means of the item-rest-score regressions (IR-regr) in both programs. Coefficient H_j gives the strength of the relationship of item j with the latent variable as estimated by means of restscore R . It also expresses item discrimination. If item sets are unidimensional and IRFs monotone, coefficient H expresses the precision by which total score orders respondents; see Sijtsma and Molenaar (2002, chap. 4) for more details.

For dichotomous-item inventories and the DMM, in addition to the previous analyses it has to be investigated whether the IRFs intersect. Sijtsma and Molenaar (2002, chap. 6) provide several methods such as inspection of the $P(++)$ and $P(--)$ matrices (Mokken, 1971, p. 134, called these matrices the Π and Π^0 matrices, and others have called them P and $P(0)$ matrices), which are available in MSP5.0 and mokken. If IIO has been ascertained for a set of items, coefficient H^T expresses the precision of the item ordering; see Ligtvoet et al. (2010) for more details.

For polytomous-item inventories and the MHM, unidimensionality is investigated using the AISP and monotonicity using the IR-regr, and coefficients H_j and H fulfill the same role as with the dichotomous-item MHM. The DMM only adds the invariant ordering of ISRFs, which can be investigated using the same methods that are also available for dichotomous items (Sijtsma & Molenaar, 2002, chap. 7.8). However, knowing that the ISRFs are ordered may provide little practical merit for most applications, and instead we recommend investigating IIO using *method manifest* IIO; see Ligtvoet et al. (2010) for more details. When IIO has been established, next coefficient H^T expresses the precision of the item ordering.

References

- Harter, S. (1985). *Manual for the self-perception profile for children*. Denver: University of Denver.
- 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.
- Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, 40, 578–595.

- Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer.
- Meijer, R. R. (2010). A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale? *Personality and Individual Differences*, 48, 502–503.
- Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, 9, 354–368.
- Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, 90, 227–238.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.
- Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: A critical discussion'. *Applied Psychological Measurement*, 10, 279–285.
- Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitatieve Methoden*, 3(8), 145–164.
- Molenaar, I. W. (1986). Een vingeroefening in item response theorie voor drie geordende antwoordcategorieën (An exercise in item response theory for three ordered answer categories). In G. F. Pikkemaat & J. J. A. Moors (Eds.), *Liber amicorum Jaap Mulwijk* (pp. 39–57). Groningen: Econometrisch Instituut.
- Molenaar, I. W. (1991). A weighted Loevinger H -coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, 12(37), 97–117.
- Molenaar, I. W. (1997). Nonparametric models for polytomous items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.
- Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. A program for Mokken scale analysis for polytomous items*. Groningen: iecProGAMMA.
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.
- 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., & 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 intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16, 149–157.
- Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*, vol. 26: *Psychometrics* (pp. 719–746). Amsterdam: Elsevier, North Holland.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- Stewart, M. E., Watson, R., Clark, A., Ebmeier, K. P., & Deary, I. J. (2010). A hierarchy of happiness? Mokken scaling analysis of the Oxford Happiness Inventory. *Personality and Individual Differences*, 48, 845–848.
- Van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, 70, 283–304.
- Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20(11), 1–19.

- Van der Ark, L.A. (2010). Getting started with Mokken scale analysis in R. Unpublished manuscript, retrieved from <http://cran.r-project.org/web/packages/mokken/index.html>.
- Watson, R., & Deary, I. (2010). Reply to: A comment on Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale? *Personality and Individual Differences*, 48, 504–505.
- Watson, R., Deary, I., & Austin, E. (2007). Are personality trait items more or less “difficult”? Mokken scaling of the NEO-FFI. *Personality and Individual Differences*, 43, 1460–1469.
- Watson, R., Roberts, B., Gow, A., & Deary, I. (2008). A hierarchy of items within Eysenck’s EPI. *Personality and Individual Differences*, 45, 333–335.