The Use of Nonparametric Item Response Theory to Explore Data Quality

Meijer, Rob R.; Tendeiro, Jorge N.; Wanders, Rob K.

Published in:
Handbook of Item Response Theory Modeling

Publication date:
2014

Citation for published version (APA):
Meijer, R. R., Tendeiro, J. N., & Wanders, R. K. (2014). The Use of Nonparametric Item Response Theory to Explore Data Quality. In S. P. Reise, & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (1st ed., pp. 85-110). Routledge.

5 The Use of Nonparametric Item Response Theory to Explore Data Quality

Rob R. Meijer, Jorge N. Tendeiro,
and Rob B. K. Wanders

Introduction

The aim of this chapter is to provide insight into a number of commonly used nonparametric item response theory (NIRT) methods and to show how these methods can be used to describe and *explore* the psychometric quality of questionnaires used in patient-reported outcome measurement and, more in general, typical performance measurement (personality, mood, health-related constructs). NIRT is an extremely valuable tool for preliminary data analysis and for evaluating whether item response data are acceptable for parametric IRT modeling. This is in particular useful in the field of typical performance measurement where the construct being measured is often very different than in maximum performance measurement (education, intelligence; see Chapter 1 of this handbook). Our basic premise is that there are no "best tools" or "best models" and that the usefulness of psychometric modeling depends on the specific aims of the instrument (questionnaire, test) that is being used. Most important is, however, that it should be clear for a researcher how sensitive a specific method (for example, DETECT, or Mokken scaling) is to the assumptions that are being investigated. The NIRT literature is not always clear about this, and in this chapter we try to clarify some of these ambiguities.

NIRT may serve two aims (Sijtsma & Meijer, 2007): (1) providing a theoretical framework that may serve as a basis for IRT modeling (e.g., Hemker, Sijtsma, Molenaar, & Junker, 1997) and (2) providing methods and procedures that can be used to analyze test and questionnaire data mainly to explore data quality. In this chapter we are primarily interested in how to apply NIRT as an exploratory model to obtain information about the data structure. Software for nonparametric IRT analyses can be found in the R-packages *irtoys* (Partchev, 2013), *KernSmoothIRT* (Mazza, Punzo, & McGuire, 2012), *mokken* (van der Ark, 2007, 2012), and *PerFit* (Meijer, Niessen, & Tendeiro, 2014; Tendeiro, 2014).

There are several good introductions written on NIRT at varying levels of technical sophistication. For example, Santor and Ramsay (1998) and Meijer and Baneke (2004) provided an introduction and empirical examples of the usefulness of NIRT in the context of personality and psychopathology measurement. Sijtsma and Molenaar (2002) wrote a monograph on NIRT and Sijtsma and Meijer (2007) provided an extensive psychometric and statistical overview of the foundations of NIRT. Also Stout and colleagues (e.g., Stout, 1987; Stout, Habing, Douglas, & Kim, 1996) wrote many theoretical and empirical papers on NIRT.

In general, the advantage of NIRT approaches is that through inspecting the simple covariance structure between items and using tools like nonparametric regression (a) results are obtained that can be interpreted very easily by practitioners, (b) data are not

forced into a structure they do not have, and (c) analyses can be run through user-friendly software programs. Finally, NIRT models may serve as alternatives to model data to parametric models to answer the question of "the robustness of conclusions under variations in model choice" (Molenaar, 2004).

Interesting in this respect was a study by Chernyshenko, Stark, Chan, Drasgow, and Williams (2001), who found that nonparametric maximum-likelihood formula scoring models to noncognitive dichotomous and polytomous data provided a better description of these data than the two- and three-parameter logistic model and the graded response model. Meijer and Baneke (2004) also argued that for data for which the response process is not well understood, such as noncognitive and clinical data (see Reise & Waller, 2009), NIRT may provide an interesting basis to investigate data quality. This is not to argue for an overall replacement of parametric by nonparametric models. Parametric IRT models lead to point estimates of the latent trait. Such scales can be very convenient, for example, for comparing the results from different questionnaires selected from the same item bank or for the study of change. However, we do think that NIRT approaches may serve as handy tools to describe and explore item response data. As Molenaar (2004) stated, "The question whether misfit is significant or not is in my view less relevant than the exploration, detection, and description." Exploring data quality may help researchers to uncover the structure of the data so that, when needed, appropriate confirmatory parametric models can be chosen.

Although NIRT models are based on weaker assumptions than parametric IRT models, NIRT models often do make several assumptions about the data. Sijtsma and Meijer (2007) stated that "the reason for calling one class of IRT models nonparametric and another parametric, is that the former only puts order restrictions on response functions and the latter assumes a specific parametric function, such as the logistic or the normal ogive." Thus in this sense the parametric assumptions are weakened. For example, NIRT typically assumes that there is a monotone relationship between the probability of endorsing an item and the latent variable. On the other hand programs like TESTGRAF do not specify this relationship.

In this chapter, we present and explore the usefulness of methods to investigate nonparametric assumptions in typical performance data and we apply these methods to data from an Aggression Questionnaire and a Physical Functioning scale. We present a critical evaluation of a number of often-used NIRT models and, in particular, we explore some recently proposed methods to investigate invariant item ordering (IIO). A set of items possesses the IIO property when the ordering of the items according to their mean score is the same for persons with different latent variable scores. There are three reasons why we consider this an interesting property to investigate for typical performance data: (1) many applied researchers and practitioners seem unaware of the fact that the item ordering according to the mean score in the population does not necessarily hold at the individual level, (2) the few studies that investigate IIO in clinical practice often use the wrong methods, and experience on how to interpret results has to accumulate as more applications become available, and (3) investigating IIO gives us a good idea about the general quality of the data. In particular it may provide us information about the psychometric quality of items in a particular population.

Research Methods

Both parametric and nonparametric IRT models try to explain the structure in the manifest item and test responses by assuming the existence of a latent scale (θ) on which persons and items have a position. Because an IRT model places restrictions on the data,

a researcher can check whether the data fit the model. Models have been formulated for both dichotomous (e.g., true-false) and polytomous (e.g., disagree-neutral-agree) items. We start with the models for dichotomous items. For dichotomous items, most IRT models assume unidimensionality (UD assumption) and a specified form for the so-called item response function (IRF). The IRF denotes the probability that an item i is answered correctly or is endorsed in the keyed direction as a function of the latent trait θ and is denoted $P_i(\theta)$. We refer to Chapter 2 of this handbook for a discussion of unidimensionality and "essential unidimensionality." From this chapter it is clear that dimensionality depends on the population of persons. Especially in clinical and outcome measurement, this is a constant source of worry because in clinical assessment it is often unclear what the exact population is. Unidimensionality implies the assumption of *local independence* (LI assumption), which states that the responses in a test are statistically independent conditional on θ. Furthermore, it is assumed that the probability of endorsing an item is *monotonically* non-decreasing in θ (M assumption).

Nonparametric IRT models differ from parametric models in that the form of the IRFs is not parametrically defined. In parametric IRT models, like the well-known two- and three-parameter logistic models, the IRF is specified through the logistic function and through item parameters like the discrimination parameter (*a* parameter) and the item difficulty or threshold parameter (*b* parameter). In nonparametric models there are no restrictions with regard to the form of the IRFs, except that they should be non-decreasing. Furthermore, in parametric IRT often specific distributions are required for the trait scores; this is not the case in nonparametric IRT.

Because NIRT models do not parameterize the response function, there is no θ metric available as in parametric models. However, dichotomous NIRT models imply an ordinal scale in the sense that the latent variable θ is stochastically ordered by the total score X_+ (Sijtsma & Molenaar, 2002). Thus, in practice any IRT model that is based on UD, LI, and M implies the ordering of respondents on the latent variable θ by means of the ordering of respondents on the observable sum score X_+, except for error.

Mokken Models

Two models that are based on the three assumptions of UD, LI, and M are Mokken's (Mokken, 1971) monotone homogeneity model (MHM) and Mokken's double monotonicity model (DMM). These models are probably the most researched and applied nonparametric IRT models, and chapters have been devoted to these models in several psychometric handbooks (e.g., Mellenbergh, 2011; Rao & Sinharay, 2007; van der Linden & Hambleton, 1997). Furthermore, there is a stand-alone computer program MSP5.0 (Molenaar & Sijtsma, 2000) and an R program package *mokken* (van der Ark, 2007, 2012) that can be used to analyze data through these models. In this chapter, we therefore use these Mokken models as the basis to discuss nonparametric data analyses. This is not to say that we are not critical about some of the methods suggested in the literature to check the fit of these models. However, we believe that these models provide a good basis to discuss useful nonparametric data analyses for typical performance measurement, as we show later in this chapter.

As mentioned earlier, the MHM assumes UD, LI, and increasing IRFs. The DMM also assumes UD, LI, and increasing IRFs, and an additional assumption is that the IRFs do not intersect. This makes the DMM a special case of the MHM, which means that when the DMM holds the weaker MHM also holds, but the reverse is not true. The assumption of nonintersecting IRFs implies IIO. More formally, when IIO holds for a set of k items and

the items are ordered in decreasing popularity (or decreasing proportion-correct score), it applies that:

$$P_1(\theta) \geq P_2(\theta) \geq \ldots \geq P_k(\theta), \text{ for all } \theta. \tag{5.1}$$

Molenaar (1997) discussed polytomous versions of Mokken's original dichotomous models, which are based on the same set of assumptions as the dichotomous models. Central in his approach is the item step response function (ISRF). Let X_i be the score on item i, with values $x_i = 0, \ldots, m$; thus for five-point rating scales, this means $x_i = 0, \ldots, 4$. The ISRF is the probability of obtaining an item score of at least x_i and is denoted $P(X_i \geq x_i | \theta)$ for $x_i = 1, \ldots, m$, thus ignoring $x_i = 0$ because this probability by definition equals one. Molenaar (1997) also discussed the DMM for polytomous items, which adds to the MHM the assumption that the ISRFs of different items do not intersect.

The polytomous DMM, however, does *not* imply that items can be invariantly ordered, it only implies that the ISRFs do not intersect across items. This was extensively discussed in Sijtsma, Meijer, and van der Ark (2011) and Meijer (2010), but it has been a source of confusion in a number of empirical papers. For example, Watson, Deary, and Austin (2007) claimed to investigate whether sets of the items have IIO. In that study, however, methods were used that were sensitive to checking whether sets of ISRFs do not intersect, *not* whether items have IIO.

Methods to Check Nonparametric IRT

As discussed earlier, NIRT models do not offer estimates of item parameters, neither do they allow for point estimates of θ. However, several methods have been proposed to check the assumptions of the NIRT models without using model parameters. In the following paragraphs we discuss methods for each model assumption.

Monotonicity Assumption

To investigate whether IRFs or ISRFs are non-decreasing in the latent variable, several nonparametric regression methods are available. An approach that obtained some popularity especially in applied clinical testing is TESTGRAF (Ramsay, 1991, 2000). For example, Santor, Ramsay, and Zuroff (1994) used TESTGRAF to inspect the IRFs for a depression questionnaire. Through TESTGRAF, category response functions (CRFs) can be plotted. A CRF gives the probability of endorsing each response category as a function of the trait score (here approximated by the total score). In general, good items should have CRFs with steep trace lines that exceed all other response functions in one interval of the total score. When the categories are ordered in increasing order, the total score values for which a response function is optimal should reflect the order in which the response choices of an item are presented. If this is not the case, two or more response functions may be combined to one response function. Items with flat or unsatisfactory response functions may be reformulated or may be removed from a questionnaire.

TESTGRAF uses so-called kernel smoothing to estimate IRFs/CRFs. This technique takes weighted averages at each trait score value; the weights are determined by the kernel function. The formula is given by:

$$\hat{P}_{im}(\theta_q) = \sum_{a=1}^{N} \frac{K\left[(\theta_a - \theta_q)/h\right]}{\sum_{b=1}^{N} K\left[(\theta_b - \theta_q)/h\right]} y_{ima}, \tag{5.2}$$

where *i*, *m*, index the item and the response category, respectively, and *a*, *b*, and *q* index persons (ranked by total score), respectively, *K* is the kernel function, *h* is a smoothing parameter, and y_{ima} is an indicator variable (1 if person *a* chose option *m* of item *i*; 0 otherwise). The trait score θ_q of person *q* is estimated as a quantile of the standard normal distribution, taking the ranking of the persons by the total sum score into account. The kernel function $K()$ is chosen so that larger weights are given to trait values in the neighborhood of θ_q; common options are the uniform, quadratic, and Gaussian functions (Ramsay, 1991). The user-specified bandwidth value *h* controls the trade-off between bias and sampling variation. Low values of *h* yield estimated functions with large variance and small bias, and high values of *h* yield estimated functions with small variance but large bias. Generally, the bottom line is to choose a bandwidth minimizing the mean-square error, which is the sum of the variance and the squared bias (i.e., $MSE = Bias^2 + Variance$).

It should be realized that smoothing can be affected by the bandwidth and that one should be careful in choosing a bandwidth, especially when there are not many observations. TESTGRAF uses the value $1.1N^{-1/5}$ as default. Note that kernel smoothing does not enforce monotonicity in θ, and as such it may be a handy tool to investigate local deviations from monotone increasing IRFs.

The *KernSmoothIRT* (Mazza et al., 2012) R package provides another possibility for performing nonparametric regression smoothing to estimate IRFs. As an alternative, both the MSP5.0 and the R package *mokken* provide a simpler method to investigate monotonicity. Here the discrete estimates of IRFs and ISRFs based on binning are constructed. That is, groups of persons who have similar total scores (or rest scores, that is the total score minus the score on the researched item) are formed for each item *i*, and for each group the proportion that answers item *i* correct is calculated. These proportions are plotted as a function of the total scores. This approach yields a limited number of plots (at most the number of items in the test). For two sample proportions it can be tested whether the item proportions are non-decreasing. This graph then forms the IRF or in the case of polytomous items the ISRF. The reason why sometimes a rest score is used instead of total score is that, theoretically, it can be shown that for polytomous item scores the total score does not need to have monotonicity in θ (Sijtsma & Meijer, 2007, p. 732).

The assumption of monotonicity can be fairly easily investigated using graphical methods, eyeball inspection, and simple statistical significance testing. In general, however, there seems to be a great reluctance by especially trained psychometricians to use graphs. We often see fit statistics and large tables full of numbers that certainly do not provide more information than graphs (see also Wainer, 2005, for a more thorough presentation of this topic). As we discuss later in this chapter, graphs are also excellent tools to inspect the assumption of invariant item ordering.

Unidimensionality Assumption

When reviewing the NIRT literature, it is not always clear how sensitive the different methods are to investigating the unidimensionality assumption. Later in this chapter, we discuss a number of techniques that are all presented under the label of unidimensionality assessment tools but that are sensitive to sometimes different characteristics of the data.

As mentioned in Sijtsma and Meijer (2007), nonparametric unidimensionality analysis is based on *conditional association* (Holland & Rosenbaum, 1986). Let **X** denote a vector of ordered item scores. Split **X** in two disjoint vectors, **X** = (**Y,Z**). Let f_1 and f_2 denote

non-decreasing functions in the ordered item scores from $\mathbf{Y}$, and let g be some function of the item scores in $\mathbf{Z}$. Conditional association (CA) means that:

$$\text{Cov}[f_1(\mathbf{Y}), f_2(\mathbf{Y}) \mid g(\mathbf{Z})=z] \geq 0, \tag{5.3}$$

for all z. CA is a necessary condition for UD, LI, and M. Therefore, necessary conditions for NIRT model fit based on CA are available after proper specification of $\mathbf{Y}$, $\mathbf{Z}$, f_1, f_2, and g. One such condition that can be used to check unidimensionality for the MHM states that all $k(k$-$1)/2$ inter-item covariances within a test should be non-negative in the sample. Strictly speaking, one negative covariance indicates misfit of the MHM. However, important is that if one finds only positive signs in the data this does not mean that the MHM fits. Hence, having nonnegative inter-item covariances is a necessary, but not sufficient, condition for unidimensionality to hold.

To investigate the unidimensionality assumption in the context of the Mokken models, several methods have been proposed. A popular method, but sensitive to specific item characteristics as we discuss later in this chapter, is the automated item selection algorithm (AISP). This algorithm uses the scalability coefficient H. H is defined at the item(step)-pair level (H_{ij}) and item level (H_i). All coefficients can be expressed as ratios of (sums of) observed covariances and maximum possible covariances. For the interpretation of H, Sijtsma and Molenaar (2002, p. 60) give the following guidelines. The scale H coefficient should be above 0.3 for the items to form a scale. When $0.3 \leq H < 0.4$ the scale is considered weak, when $0.4 \leq H < 0.5$ the scale is considered medium, and when $H \geq 0.5$ the scale is considered strong. There is, however, a lack of literature with respect to the exact meaning of these benchmarks.

The AISP is based primarily on the inter-item covariances and the strengths of the relationship between items and the rest score as expressed by the item H_i coefficients. Based on such information, clusters of related items may be identified. The AISP "bottom-up" procedure starts by selecting the pair of items for which (a) H_{ij} is significantly larger than 0 and (b) H_{ij} is the largest among the coefficients for all possible item pairs. Then a third item l is selected that (c) correlates positively with the items already selected, (d) has an H_l coefficient that is larger than 0, and (e) has an H_l coefficient that is larger than a user-specified value c. The procedure continues to select items as long as items are available that satisfy conditions c, d, and e. The end result may be one or more item clusters such that each taps a different latent trait or latent trait composite. The substantive interpretation of the clusters is done on the basis of the content of the clustered items and the substantive knowledge one has about the data structure. For Mokken scale analysis the AISP plays a crucial role. Sijtsma and Molenaar defined a Mokken scale as a set of items that "a) measure a common trait . . . with b) reasonable discriminative power determined by lowerbound c" (Sijtsma & Molenaar, 2002, p. 68).

The AISP is a bottom-up selection algorithm that selects an item only once, without the possibility of taking back the assignment later on. Thus, AISP does not consider all possible item partitionings. Straat, van der Ark, and Sijtsma (2013) suggested a new selection algorithm that uses a genetic search (GA) algorithm that avoids this problem. In contrast to the AISP algorithm, this GA is probabilistic and approximates all possible subdivisions. As a result it may produce different item partitionings than the AISP procedure.

Because for a fixed distribution of θ both H and H_i are increasing functions of the slopes of logistic IRFs, this selection procedure selects items that have (together) high discrimination parameters. Thus, what is very important to understand is that *the monotonicity assumption* forms the basis of the dimensionality analysis in Mokken procedures. This may have important consequences for the selection of items in the noncognitive area.

Theoretically, selecting items on the basis of monotonicity may result in selecting items within a cluster that are locally dependent. As several authors have shown for parametric IRT models (e.g., Chen & Thissen, 1997), item discrimination parameters may be inflated when the items are locally dependent. How strong this effect is in the AISP procedure is not clear from the literature. Recently, Straat (2012, chapter 6) suggested a number of statistics to detect violations of local independence, but there is not yet much experience with these statistics.

Local Independence

Chen and Thissen (1997) showed that, for parametric IRT models, item discrimination parameters may be inflated when the items are locally dependent. To investigate what the effect of local dependence on the H_i scalability coefficient is we replicated and extended the study by Chen and Thissen (1997, p. 266) as follows. A data set with scores of 1,000 subjects on six items was generated using the 2PL model; the true a and b parameters that were used are shown in Table 5.1 (second and third columns). Two cases were considered: Local independence (LI) and local dependence (LD). The LI case consisted of using the generated data set without any further changes. In the second data set LD was simulated using the same item scores for both item 3 and item 6 (like in the Chen & Thissen, 1997 study). Item parameters a's, b's and scalability coefficients H_i were then estimated for each case (LI, LD) using IRTPRO (using default program options) and the R *mokken* package, respectively. Results are depicted in Table 5.1. It is clear that the estimation of the parameters of items 3 and 6 in the LD case was greatly affected by the fact that their scores are dependent (equal in this case). Moreover, the associated H_i values seem to be larger, on average, than the H_i values for items that do not contribute to the LD problem. This effect was confirmed by replicating the experiment 100 times, as shown in Table 5.2 (second column).

The study was further extended by increasing the number of items from 6 to 12. Only item scalability H_i values were computed. Table 5.3 shows the results for one data set only; Table 5.2 (third column) shows the average results of more than 100 replications. It can be seen that the effect of LD on the H_i values was somewhat reduced, implying that test length is an important factor. More specifically, the scalability of locally dependent

Table 5.1 True Item Parameters, Estimated Item Parameters, and Item Scalability Coefficients H_i Under Local Independence and Local Dependence of Items 3 and 6

	True parameters		First case: LI			Second case: LD (item 3 = item 6)		
Item	a	b	a (SE)	b (SE)	H_i (SE)	a (SE)	b (SE)	H_i (SE)
1	1.0	−1.0	1.07 (.14)	−0.88 (.11)	0.27 (.03)	0.62 (.08)	−1.42 (.18)	0.27 (.03)
2	1.0	−0.5	1.11 (.14)	−0.45 (.08)	0.25 (.02)	0.54 (.07)	−0.85 (.15)	0.24 (.03)
3	1.0	0.0	0.98 (.13)	−0.05 (.08)	0.22 (.02)	207.75 (*)	−0.11 (**)	0.44 (.02)
4	1.0	0.5	1.10 (.14)	0.53 (.09)	0.25 (.02)	0.52 (.07)	0.89 (.17)	0.24 (.03)
5	1.0	1.0	1.03 (.14)	0.98 (.12)	0.27 (.03)	0.66 (.08)	1.31 (.17)	0.27 (.03)
6	1.0	0.0	1.06 (.14)	0.02 (.07)	0.23 (.02)	207.75 (*)	−0.11 (**)	0.44 (.02)

Note: * = 51964.62. ** = 37.64. LI = local independence. LD = local dependence.

Table 5.2 Mean (SD) H_i Values Across 100 Replications for Data Sets With 6 and 12 Items

	$k = 6$	$k = 12$
LI	0.24 (.03)	0.23 (.03)
LD	0.43 (.02)	0.32 (.02)

Note: k = number of items. LI = local independence. LD = local dependence.
All H_i values in each replicated data set are used in the LI cases.
Only H_i values concerning items 3 and 6 ($k = 6$) and items 6 and 12 ($k = 12$) in each replicated data set are used in the LD cases.

Table 5.3 True Item Parameters and Item Scalability Coefficients H_i Under Local Independence and Local Dependence of Items 6 and 12

	True parameters		First case: LI	Second case: LD (item 3 = item 6)
Item	a	b	H_i (SE)	H_i (SE)
1	1.0	−1.25	0.26 (.03)	0.24 (.03)
2	1.0	−1.00	0.26 (.02)	0.26 (.02)
3	1.0	−0.75	0.24 (.02)	0.25 (.02)
4	1.0	−0.50	0.25 (.02)	0.25 (.02)
5	1.0	−0.25	0.24 (.02)	0.25 (.02)
6	1.0	0.00	0.21 (.02)	0.32 (.02)
7	1.0	0.25	0.23 (.02)	0.23 (.02)
8	1.0	0.50	0.22 (.02)	0.21 (.02)
9	1.0	0.75	0.19 (.02)	0.20 (.02)
10	1.0	1.00	0.29 (.02)	0.29 (.02)
11	1.0	1.25	0.23 (.03)	0.24 (.03)
12	1.0	0.00	0.21 (.02)	0.32 (.02)

Note: LI = local independence. LD = local dependence.

items may become less problematic if the number of items violating the LI assumptions is relatively small with respect to the total number of items. More studies are needed to clarify which proportions of locally dependent items are admissible in a data set to still allow for good scaling of individual differences.

Maximizing H *Values: Some Drawbacks*

In the literature there is a strong emphasis on selecting items with H_i values larger than some lower bound as, say, $H_i = 0.3$. Egberink and Meijer (2011) observed, however, that a researcher should also be careful when H_i values are very high. They argued that repeating items with a similar content will result in scales with high H_i values but, sometimes, scales that measure extremely narrow-band constructs. That is, strong Mokken scales may very reliably differentiate persons, but at the cost of construct validity. High H_i values may also point at items that define the construct ("*I am often depressed*" in a depression questionnaire).

Following this line of reasoning, Meijer, de Vries, and van Bruggen (2011) also argued to consider the 18 items of the BSI-18 scale (Derogatis, 1983) as one scale instead of forming subscales with each higher H_i value. Often subscales of clinical questionnaires contain items that are semantically similar, like "*I get a sort of frightened feeling as if something awful is about to happen*" and "*I get a sort of frightened feeling like butterflies in the stomach*" (both items from the Anxiety scale in the HADS questionnaire, Zigmond & Snaith, 1983), or "*I still enjoy the things I used to enjoy*" and "*I look forward with enjoyment to things*" (Depression items from the HADS). These items will result in high H_{ij} values, but a researcher may question whether it is strictly necessary to have both items in the same scale. A similar observation was reported by Reise (2009). He reported that fatigue and sleep disturbance are conceptually narrow and have a limited pool of item indicators. The Mokken scalability coefficient for the fatigue item bank equaled $H = 0.71$. As Reise (2009) remarked, there is little conceptual distance between the symptoms (items) and the construct. In this case one may conclude that it is not necessary to ask several questions, but that asking a few questions may suffice to estimate a person's standing on the latent trait scale.

Because the AISP focuses on the monotonicity assumption, several studies showed that this procedure performs worse in recovering the correct dimensionality structure than other methods (Mroch & Bolt, 2006; Smits, Timmerman, & Meijer, 2012; van Abswoude, van der Ark, & Sijtsma, 2004). In particular, those studies revealed that Mokken scale analysis does not function well in conditions in which the traits correlate (Mroch & Bolt, 2006; van Abswoude et al., 2004), or in which the items load on more than one trait (Smits et al., 2012; van Abswoude et al., 2004).

DETECT

Several alternative procedures have been proposed to assess unidimensionality in non-parametric contexts. In this chapter we restrict ourselves to DETECT because it is the most often used alternative to Mokken scale analyses and alternatives such as DIMTEST are based on assumptions that assume the knowledge of existing unidimensional subtests. DETECT uses conditional covariances between items to assess dimensionality and it is based on the early work by Stout (1987, 1990). He defined essential unidimensionality (Stout, 1987), essential independence, and weak monotonicity. Essential unidimensionality states that there is one dominant latent variable and several nuisance parameters. In this respect, it resembles common (general) factors and unique (group) factors as in factor analysis. Essential independence states that for infinitely many items the mean inter-item covariance conditional on θ equals 0. Also weak monotonicity assumes that the mean of the k IRFs is monotone on θ. This means that assumption M is dropped as a requirement for individual IRFs but that it is active for the test response function under the name of weak monotonicity. A drawback of these procedures is that the properties implied by UD, LI, and M with respect to stochastically ordering persons on θ by means of the total score are lost. Therefore, we restrict ourselves here to dimensionality analyses.

To check the dimensionality assumption, Stout and coworkers based their method on a special case of conditional association, namely that the covariance between items i and j must be nonnegative for subgroups that have the same rest score. Assuming that the items measure Q latent variables to a different degree, we may assume that θ_q is a linear combination of these variables. The performance on the Q latent variables is estimated by means of total score or rest scores that both summarize test performance but ignore multidimensionality. Zhang and Stout (1999), however, showed that the sign of $\text{cov}(X_i, X_j \mid \theta_q)$ provides useful information about the dimensionality of the data. It is positive when two items measure the same latent variable and negative when they clearly measure different

latent variables. This observation forms the basis of DETECT that divides a data set of items into clusters that together approach essential independence as well as possible given all possible item clusters.

Assume that several of the latent variables measured by the items are dominant and that others are nuisance variables. Consider an arbitrary partitioning, denoted P, of the set of items into disjoint clusters; and let $d_{ij}(P) = 1$ if items i and j are in the same cluster and $d_{ij}(P) = -1$ if they are not. DETECT is then defined as:

$$d(P) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d_{ij}(P) E[cov(X_i, X_j \mid \theta)]. \tag{5.4}$$

The DETECT procedure tries to find the partitioning of the items, denoted P*, that maximizes Equation (5.4). This is the item clustering that best approximates weak LI between item sets, and it is taken as the best description of the dimensionality of the data. Thus, important here is that DETECT is sensitive to violations of LI, whereas Mokken scaling is sensitive to the monotonicity assumption. Several studies suggested rules of thumb that can be used to decide whether a data set is unidimensional or multidimensional. Stout and colleagues (1996) considered DETECT values smaller than 0.1 indicating essential unidimensionality and DETECT values larger than 1 as indicating multidimensionality. Roussos and Ozbek (2006) suggested the following rules of thumb; *DETECT < 0.2*: weak multidimensionality/approximate unidimensionality; *0.2 < DETECT < 0.4*: weak to moderate multidimensionality; *0.4 < DETECT < 1.0* = moderate to large multidimensionality, and *DETECT > 1.0*: strong multidimensionality. Recently, however, Bonifay, Reise, Scheines, and Meijer (2014) discussed that these values are sensitive to the factor structure of the data set and the relation between general and group factors. Thus, these values should be used with care. We will return to this in the data analyses we discuss at the end of this chapter.

Invariant Item Ordering

The measurement of patient-reported outcomes through questionnaires sometimes assumes that the items used in these questionnaires represent different levels of intensity with respect to the attribute of interest. For example, when we consider two items of the Aggression Questionnaire that we analyze later, then it is assumed that an item "Sometimes I fly off the handle for no good reason" represents a much higher level of aggression than the item "I tell my friends openly when I disagree with them." On the other hand, we often encounter questionnaires that consist of items that are repetitions of similar questions and with conditional item means that cluster together, suggesting that there are no differences between the intensity of the items. The question then is whether the item ordering according to severity (or mean score) established at the *group* level can be used for persons at different *individual* trait levels, that is, whether IIO holds (Ligtvoet, van der Ark, te Marvelde, & Sijtsma, 2010; Sijtsma et al., 2011). Thus the IIO assumption holds when the ordering of the items according to their severity is the same across different values of the latent variable.

Methods to Investigate IIO

Several methods have been developed to establish IIO for dichotomously scored items (see Sijtsma & Junker, 1996, for an overview). Meijer and Egberink (2012) suggested,

however, to first inspect plots of item-rest score regression before checking possible violations of IIO. Then, to further investigate IIO for dichotomous items, the rest-score method, the P-matrix method, and the coefficient H^T can be used. These methods are implemented in the R package *mokken*.

It should be emphasized that two different kinds of methods can be used together. One type of method checks whether IRFs intersect (e.g., the rest-score method and the P-matrix methods, see later in this chapter). The other method (H^T) provides information about the spread of IRFs, that is, how far the IRFs are apart. The idea is that when the IRFs are further apart, we may have more confidence in IIO.

Rest-Score Method

Under the null hypothesis of nonintersecting IRFs Equation (5.1) applies. In the rest-score method the IRFs for each pair of items are estimated and compared using the item rest-score functions. For the dichotomous case, item rest-score functions relate the rest score (summed score minus the item score) on the *x*-axis with the observed response proportions for each rest-score group on the *y*-axis. Rest scores are grouped together when there are not enough persons at each possible rest score to compute reliable item proportions. For each pair of items *i* and *j*, with item *j* more popular than item *i* in the total group, it is checked whether:

$$P\left(X_j = 1 \mid R_{(ij)} = r\right) \geq P\left(X_i = 1 \mid R_{(ij)} = r\right), \text{ for } r = 0, \ldots, k-2, \tag{5.5}$$

with $R_{(ij)}$ denoting the summary score that contains neither X_i nor X_j. Thus, when two item rest-score functions are compared, the conditioning is based on the remaining $k-2$ items. When IIO holds, the order of the item proportions for the different rest-score groups is the same as the ordering of the item proportions for the total group. Rest-score functions for all item pairs are compared to investigate whether IIO holds for a set of *k* items. A violation is reported when the response proportion is reversed in a rest-score group as compared to the response proportion order for the total group. The null hypothesis of equal item proportions can be tested against the one-sided alternative that the proportion of item *i* is larger than the proportion of item *j*.

Because significant violations sometimes have low impact, Molenaar and Sijtsma (2000) discuss an effect size measure named *Crit* that consists of a weighted number of different indicators of violations for which the following guidelines have been suggested: *Crit* values smaller than 40 indicate no serious violations; *Crit* values between 40 and 80 indicate minor violations, and *Crit* values larger than 80 indicate serious violations. The numerical values are based on several practical data analyses but should be interpreted with care. For example, there are no simulation studies that endorse these values. We will use these *Crit* values to get an idea about the seriousness of model violations in our data analyses.

The P-Matrix Method

To investigate nonintersection of the IRFs, the P-matrix method uses two square symmetric *k* x *k* matrices in which the items are ordered from difficult to easy (based on the probabilities in the overall group). The cells in the first matrix, P(+ +), contain all joint proportions $P_{ij}(1,1)$ of persons "passing" both items *i* and *j*. The cells in the second matrix, P(− −), contain all joint probabilities $P_{ij}(0,0)$ of persons "failing" both items *i* and *j*. Nonintersection of the IRFs means that the rows and columns of

the P(+ +) matrix are non-decreasing and of the P(– –) matrix non-increasing. A violation is reported when there is a decrease in one of the rows or columns of the P(+ +) matrix and when there is an increase in one of the rows or columns of the P(– –) matrix. Like the rest-score method, significance testing is done using the McNemar's test.

Coefficient H^T

Coefficient H^T (Sijtsma & Meijer, 1992) can be used as a measure for the accuracy of the item ordering. A low H^T value suggests that the IRFs are close together, whereas a high value of H^T suggests that the IRFs are further apart. When IIO holds for k items, it can be shown that $0 \le H^T \le 1$. For practical purposes, Sijtsma and Meijer (1992) suggested using $H^T \ge 0.3$ as a lower bound. It is important to emphasize that H^T is only related to all k items together, and cannot be used to assess which items cause intersections. Therefore, Sijtsma and Meijer (1992) suggested combining information from H^T with the results from other methods such as the rest-score method and the P-matrix method.

Polytomous Data: Method Manifest IIO

Ligtvoet and colleagues (2010) developed a method to investigate IIO for polytomous items, which is named method manifest IIO. Method manifest IIO (MIIO) compares the ordering of the item means for all item pairs for different rest-score groups. This is investigated by numbering and ordering the items in increasing order of difficulty (i.e., from easy to difficult) for the entire test. IIO holds when:

$$E\left(X_i \mid R_{(ij)} = r\right) \ge E\left(X_j \mid R_{(ij)} = r\right), \text{ for all } r \text{ and all } i < j, \tag{5.6}$$

and with $R_{(ij)}$ denoting, once more, the total test score excluding the scores on items i and j. Then, one-sided one-sample t-tests are conducted to test the significance of each sample violation of the IIO condition stated earlier. The null hypothesis is that the expected conditional item means are equal, and it is tested against the alternative that the expected conditional mean of item j exceeds that of item i, which is a violation of IIO. A violation is reported when there is a reverse ordering of the conditional sample means for a particular rest score. To prevent taking very small violations seriously, these reverse orderings are only reported when they exceed a minimum value, denoted *minvi*.

Reliability in Nonparametric IRT

In *parametric* IRT a latent trait variable is estimated and the variance of this estimate across (hypothetical) repeated test administrations of the same test is the variance of the trait score. When this variance is small the estimate is precise, when it is large it is not precise. Because this variance is defined for a particular person it is within-person precision. Often the reciprocal of the within-person variance of the trait estimate is used, which is the test information at the given trait estimate.

In parametric IRT, reliability is concerned with the differentiation between trait scores of different persons. When the trait score can be precisely predicted from the test score reliability is satisfactory and it is unsatisfactory when it cannot.

Because in *nonparametric* IRT there are no θ estimates, Mokken (1971) proposed a method in the context of the DMM model that is based on the same definition as classical reliability and that results in a reliability estimate. TESTGRAF gives plots for a reliability estimate conditional on the estimated score.

Fit of Item Score Patterns

Thus far, we have discussed model data fit from the perspective of the items and set of items. Once a researcher concluded that a set of items can be described by an NIRT model, it can be checked whether there are persons who do not comply to the model assumptions.

Psychologists and clinicians have become increasingly aware of the need to evaluate carefully and routinely the veracity of information obtained from testing, clinical interviews, and other methods that are based on self-report. Deliberate exaggeration, or fabrication of problems to obtain a desired external goal, may invalidate test scores. Thus, feedback provided by simple statistics about the consistency of answering behavior can be useful, especially in clinical treatment and outcome measurement settings.

In personality and clinical assessments a tradition exists to detect invalid test scores using different types of validity scales. The utility of validity scales to detect "faking bad" or exaggerating symptoms has generated quite some debate. For example, Pinsoneault (2007) found that different MMPI validity scales had enough power to be used in practice. However, other authors have questioned the usefulness of validity scales to detect "faking good" or social desirability. One of the problems of validity scales is that they may be confounded with valid personality trait variance and show a relationship with other content scales. For example, Ones, Viswesvaran, and Reiss (1996) found that social desirability measures correlate with Emotional Stability ($r = 0.37$), Conscientiousness ($r = 0.20$), and Agreeableness ($r = 0.14$) in normal (honest) assessment situations. Because of this, a high social desirability score can indicate either legitimate favorable trait elevations or dissimulation. Also, many clinical scales do not have separate validity scales.

As an alternative, some authors have suggested using consistency measures (Tellegen, 1988). Since the 1980s with the increased use of item response theory modeling, statistics have been proposed to identify item score patterns that are unlikely under an IRT model. These unlikely patterns can be identified and further inspected. Although many statistics can be used to identify invalid score patterns (see Meijer & Sijtsma, 2001 for an overview), only a few studies show their practical usefulness (e.g., Meijer, Egberink, Emons, & Sijtsma, 2008).

Despite this lack of empirical evidence, we think that although a particular questionnaire can be a good measure of a psychological construct for a group of persons, it may be a poor measure of the construct for a particular individual, and that checking the consistency of answering behavior on questionnaires may be of help in outcome measurement. Perhaps it is the lack of easy-to-use software and the lack of the presence in often-used software programs (such as Multilog or IRTPRO) that make this technique relatively unpopular. However, the recently developed R packages *irtoys*, *PerFit*, and *mirt* (Chalmers, 2012) contain parametric person-fit statistics, which may increase the use of these statistics.

Application

To illustrate the different nonparametric IRT methods described so far we used two different data sets we obtained from the PROMIS® Wave 1 data (www.nihpromis.org). First, we analyzed data from two subscales of the Aggression Questionnaire (Buss & Perry, 1992): the Verbal Aggression scale and the Anger scale. We analyzed the 12 items of the combined subscales as one scale (denoted as AQ) because this enabled us to illustrate the performance of the different nonparametric IRT methods. Items were scored on a five-point Likert scale from 0 (*extremely uncharacteristic of me*) through 4 (*extremely characteristic of me*). Second, we analyzed data from the Physical Functioning (PF) scale of the SF-36

(Ware & Sherbourne, 1992) consisting of 10 items scored from 0 through 2 (0 = *limited a lot*; 1 = *limited a little*; 2 = no, not *limited at all*).

Aggression Scale

The sample consisted of 819 persons of the normal population (M_{age} = 51.3, SD = 18.21); 50.7 percent were men. Sijtsma and colleagues (2011) suggested a procedure for analyzing data according to the DMM. Because the DMM implies the MHM this procedure nicely fits the aim of the present study. For dichotomous items they distinguished the following steps: (1) investigate scalability/unidimensionality through an automated item selection procedure (AISP), (2) investigate monotonicity through inspecting item rest-score regressions, (3) investigate IIO through inspecting methods like rest score and P-matrix, and finally (4) investigate the precision of the item ordering through the H^T coefficient.

For polytomous data, a similar methodology was proposed by Sijtsma and colleagues (2011), with the exception that the method manifest IIO proposed by Ligtvoet and colleagues (2010) should be used to investigate IIO. We will loosely follow this methodology. That is, our aim is to illustrate how the data can be checked with respect to the MHM and the DMM. Analyses were performed using the *R* package *mokken*.

For the methodology as proposed by Sijtsma and colleagues (2011) it is not strictly necessary to use $c = 0.3$ as a lower bound in the AISP procedure. An alternative is to check whether all inter-item covariances are positive, thus using $H_{ij} = 0$ as a lower bound.

For both data sets, we started with the AISP procedure with $c = 0$. Although this lower bound is seldom used, it provides interesting information about which items comply to the minimum requirements of the MHM. Then we checked monotonicity by inspecting the IRFs.

Second, we used a lower bound of $c = 0.30$ for the AISP, *minvi* = 0.03 to investigate monotonicity, and *minvi* = 0.03 times the number of item step response functions (m) to investigate IIO. Ligtvoet and colleagues (2010) investigated the sensitivity and specificity of method manifest IIO. They used different *minvi* values and their simulation study showed that a *minvi* of 0.03 times m is an appropriate choice for investigating IIO with polytomous items. Furthermore, we used the following rules of thumb for the H^T coefficient: $H^T < 0.3$ implies that the item ordering is too inaccurate to be useful; $0.3 \leq H^T < 0.4$ implies low accuracy of item ordering; $0.4 \leq H^T < 0.5$ implies medium accuracy; and $H^T \geq 0.5$ implies high accuracy.

Ligtvoet and colleagues (2010) suggested the following sequential data-analysis procedure for method manifest IIO. First, for each of the k items the number of significant violations (i.e., that exceed *minvi*) is determined and the item with the highest number of violations is removed. When different items have the same number of significant violations, the item with the smallest H_i coefficient may be removed, but other criteria might also be considered, for example the item content. Second, this procedure is repeated for the remaining items until none of the remaining items have significant violations, which means that IIO holds for all items. When IIO holds for the (remaining) items, the H^T coefficient for polytomous items can be computed. This coefficient is a generalization of the original H^T coefficient to polytomous data and it allows checking the accuracy of the item ordering.

Results

Descriptive Statistics

Table 5.4 shows the content of the items, the mean item scores, and the item-total score correlation. A first observation is that the distribution of the item scores is skewed to the right; that is, most persons chose options 0, 1, or 2. This is as expected because most persons in

Table 5.4 Descriptive Statistics for the Aggression Scale

Item	Subscale	Item content	Mean	SD	Item-total correlation
1	Anger	Some of my friends think I am a hothead	0.66	1.04	0.70
2	Verbal	I tell my friends openly when I disagree with them	2.54	1.12	0.39
3	Verbal	I can't help getting into arguments when people disagree with me	1.06	1.13	0.67
4	Anger	I am an even-tempered person (reverse scored)	1.01	1.08	0.55
5	Anger	I flare up quickly but get over it quickly	1.69	1.28	0.54
6	Verbal	When people annoy me, I may tell them what I think of them	1.51	1.28	0.56
7	Anger	I have trouble controlling my temper	0.74	1.09	0.72
8	Anger	When frustrated, I let my irritation show	1.90	1.21	0.64
9	Verbal	I often find myself disagreeing with people	1.48	1.09	0.64
10	Anger	I sometimes feel like a powder keg ready to explode	0.55	1.02	0.63
11	Verbal	My friends say that I'm somewhat argumentative	1.12	1.22	0.71
12	Anger	Sometimes I fly off the handle for no good reason	0.53	0.97	0.66

the general population will not endorse statements that indicate (very) aggressive behavior. In Figure 5.1 we depicted the conditional reliability as calculated through TESTGRAF. As can be seen, reliability was highest for scores between 5 and 15 (to be further discussed later in this chapter). To obtain a first impression about the relative first-factor strength to assess unidimensionality we determined the ratio of the first to second eigenvalues—the larger this ratio, the more appropriate the item response data are thought to be for unidimensional IRT modeling. The ratio between first and second order factors was $4.7/1.5 = 3.1$ for the Aggression scale and $6.3/1.0 = 6.0$ for the Physical Functioning scale, so there is a stronger general factor running through the PF data than the AQ data.

Checking Monotonicity and Scalability

We start with inspecting the mean ISRF plots to obtain a first impression about the functioning of the items across different rest-score groups (see Figure 5.2). Note that we grouped many rest scores at the higher end of the total scores because there are relatively few persons with a high rest score. This plot already teaches us some interesting things about the item quality. It tells us (1) which items are most related to the underlying latent variable, aggression (steepest functions) and (2) in what range of the rest scores the items discriminate best. From Figure 5.2 it is clear that the item rest-score functions for all items are non-decreasing, although the mean ISRF of item 2 "*I tell my friends openly when I disagree with them*" is rather flat (i.e., not discriminating well between different rest scores). The rest-score function of item 11 "*My friends say that I'm somewhat argumentative*" has one of the steepest item rest-score functions. Note that one should be careful in interpreting the steepness of the IRFs because high rest scores are pooled together, thus the *x*-axis is not interval scaled.

Also note that items 10 and 12 discriminate well in the high rest-score range, but do not discriminate in the low rest-score range. Another observation is that for low rest-score levels the mean scores on items 2, 5, 6, and 8 are far from zero and for high rest-score
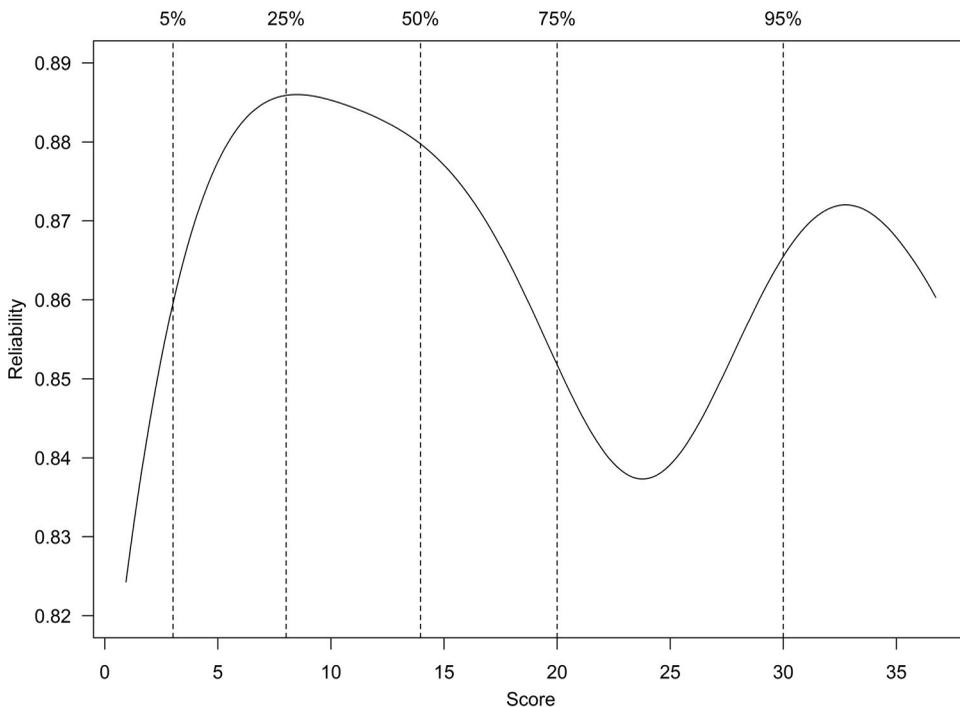
Figure 5.1 Conditional reliability for the Aggression scale scores.

levels this mean score is below three out of a maximum of four, thus indicating that the upper asymptote is far from the maximum category. This would suggest that for these data a parametric IRT model with an upper asymptote parameter lower than one is needed. These plots and similar results can also be obtained using TESTGRAF. To further explore item quality we return to the descriptive statistics in Table 5.4; these statistics confirm some of the observations obtained from inspecting the graphs in Figure 5.2. For example, item 2 does have a relatively low item-test correlation and the scores are skewed to the right.

Next, we ran the AISP using $c = 0$. Interesting is that 11 out of 12 items were selected; only item 2 was not selected because of a negative correlation with item 4. From these first analyses, it is clear that item 2 is a candidate for removal from the scale. When we ran the AISP with $c = 0.3$ as a lower bound item 6 was also removed from the scale because of an H_i value smaller than 0.3. When using the new AISP approach proposed by Straat and colleagues (2013) we found the same results; that is, both items 2 and 6 were not selected in the final scale.

Checking the monotonicity assumption for the remaining 10 items, we observed that the most problematic case concerned item 4. In Figure 5.3 we depicted the ISRFs of item 4. Note that there are five options (scores from 0 through 4) and thus four ISRFs. The figure shows that ISRF 1 is (almost) non-decreasing, but that the remaining ISRFs decrease between rest-score groups (3–4) and (5–6). ISRF 3 is also decreasing between rest-score groups 9–11 and 12–14 and ISRF 4 is also decreasing between rest-score groups 5–6 and 7–8. This results in a decreasing mean item step (bold line) and thus a violation of the MHM model.

To explain how monotonicity is checked consider Figure 5.3. There are eight rest-score groups with a varying number of persons (with a minimum of 81 persons per group in the present analysis). For each item step we now compare the proportions across the
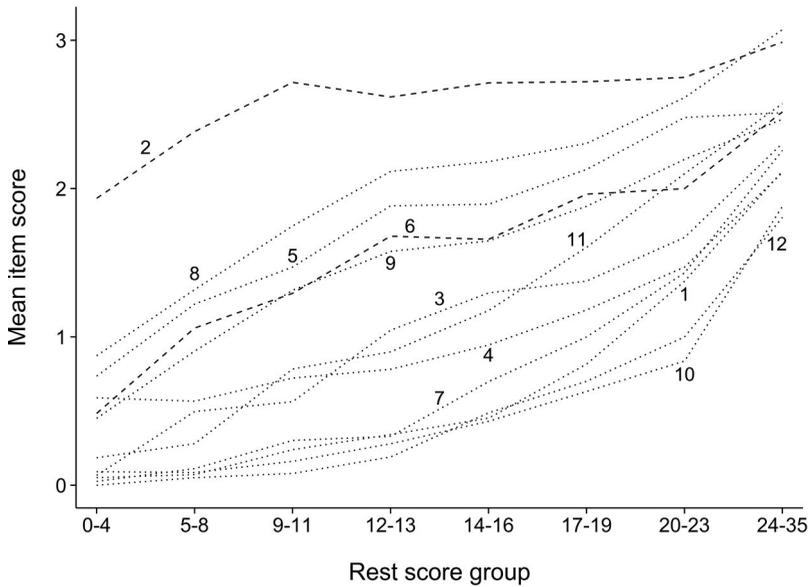
Figure 5.2 Mean ISRFs for the 12 items of the Aggression Questionnaire.

rest-score groups and a decrease larger than 0.03 is indicated as a violation against mono-tonicity. For example, for item step 2 there were violations for rest-score groups 1 and 3, 2 and 3, and 2 and 4.

To further explore this and to investigate how serious these violations are, in Table 5.5 we depicted some relevant statistics for all 10 selected items, but we concentrate on item 4 (in bold). In Table 5.5 we give the H_i values, the so-called number of *active pairs*, number of violations, and *Crit* values. The number of *active pairs* is the number of pairs that are taken into account when comparing the proportion-correct scores for each ISRF.

As can be seen in Table 5.5, item 4 had 15 violations out of 108 active pairs.[1] Between each two rest-score groups an observed decrease in proportion could be due to sampling variation. For example, for item step 2, the maximum violation was found when comparing the proportions of rest-score groups 3–4 and 5–6; the conditional probability $P(X_{(4)} \geq 2|R_{(4)})$ decreased 0.05. This decrease could be due to random fluctuation and not necessarily due to a large difference between both rest-score groups. This can be tested in a 2*2 table containing the number of persons in the rest-score groups 2 and 3 and the number of persons with item step < 2 and item step ≥ 2. Using a normal approximation for this hypergeometric distribution resulted in $z = 0.91$, thus we conclude that this specific violation is not significant. The same line of reasoning allows us to confirm that no significant violations exist for item steps 1 and 2. For item step 3 there were 4 violations (all significant) and for item step 4 there were 8 violations (6 significant). Thus in total there were 10 significant violations that resulted in a *Crit* value of 92. The researcher may consider removing this item from the item set.

---

1 There are 108 active pairs for item 4 because there are eight rest-score groups (see Figure 5.2) and there are four ISRFs per item. For each item there are thus (8*7/2)*4 = 112 possible comparisons. However, because there were zero results for the fourth rest group, the comparisons with rest-score groups 5, 6, 7, and 8 did not exist. Hence the number of active pairs for item 4 is 112 − 4 = 108.
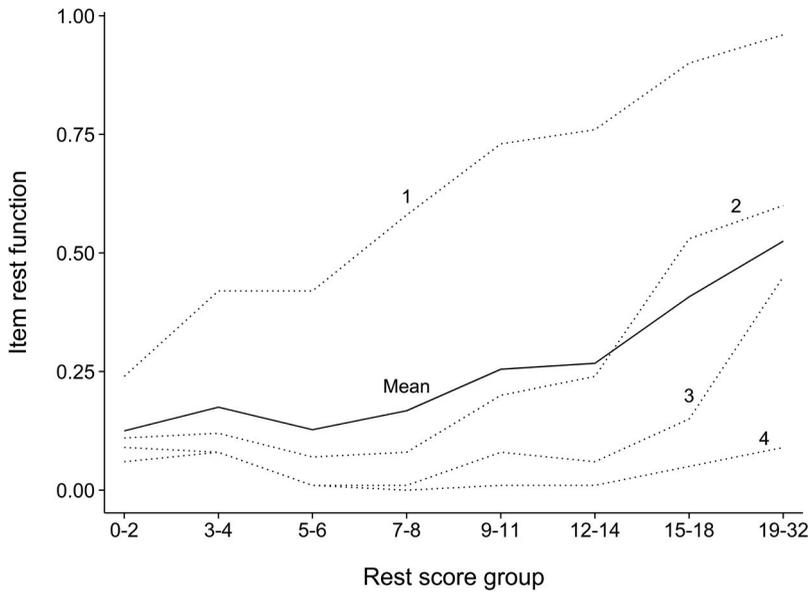
Figure 5.3 ISRFs for item 4, Aggression scale.

Table 5.5 Item H_i Values and Monotonicity Checks for 10 Items of the Aggression Questionnaire

Item	H_i	# Ac pairs	# violations	# sign Z-test	crit
1	.47	72	2	0	12
3	.41	112	1	1	33
4	**.37**	**108**	**15**	**10**	**92**
5	.32	112	3	0	21
7	.51	92	1	0	3
8	.43	105	3	0	19
9	.42	97	2	0	12
10	.47	92	0	—	—
11	.45	103	0	—	—
12	.49	64	0	—	—

Note: # Ac pairs = number of active pairs. # violations = number of violations. # sign Z-test = number of significant z tests. crit = *Crit* values.

Sijtsma and Molenaar (2002) suggested investigating multidimensionality by increasing the lower bound c values in the AISP. When this increase clearly resulted in the creation of different scales that are easy to interpret, multidimensionality may exist. If not, then the scale may be considered unidimensional. When we used $c = 0.4$ we found two scales: a first scale with items (10, 12, 7, 8, 1, 11, 9, 3) and a second scale with items (2, 6); items 4 and 5 were unscalable. With a lower bound of $c = 0.5$ there were two scales (10, 12, 7, 8, 1) and (11, 9, 3) and items 2, 4, and 6 are unscalable. There was no clear distinction

between anger items and verbal aggression items. As we discussed earlier, maximizing *H* values leads to clusters of items that are very narrow in content. For example, item cluster (3, 9, 11) only consists of verbal disagreeing items and these items are thus very similar in content. On the basis of this analysis we conclude that there is no clear multidimensionality present in the data, and that items 2, 4, and 6 are candidates for removal from the scale.

Results Using DETECT

The DETECT procedure was used to further study the dimensionality of the data. We used the stand-alone program polyDETECT described in Zhang (2007) and obtained from the author. For the AQ items, DETECT selected three subscales: one consisting of items from the Anger subscale (items 1, 4, 7, 10, 12), another subscale consisted of Verbal Aggression items (items 2, 3, 6, 9, 11), and a cluster of two items (5, 8) formed a third subscale. The maximum DETECT value equaled 6.1, indicating multidimensionality. When we forced the data to have two dimensions, both the Anger and the Verbal Aggression subscales were identified. We will further elaborate on these results in relation to the Mokken analyses results when we discuss the results for the Physical Functioning scale.

Invariant Item Ordering

We checked for violations of IIO without items 2, 4, and 6. There was only one violation for items 10 and 12 and this violation was not significant according to MIIO, therefore these items were kept in the scale. For the subscale consisting of the remaining nine items we found $H^T = 0.38$, which following the rules of thumb tentatively proposed by Ligtvoet and colleagues (2010) indicates that the accuracy of the ordering of the items is low. This is also reflected in the plots of the mean ISRFs (see Figure 5.2). The ISRFs of items 1, 7, 10, and 12 cluster together and even cross (although not resulting in a significant different ordering of the items, as discussed earlier). Note that MIIO only investigates a reversed ordering and that item response functions may cluster together without violating the assumption of IIO. Important now is that it is clear that many mean ISRFs cluster together as is typical for many clinical scales (Meijer & Egberink, 2012).

How can we apply this information for further analyses? Based on these analyses one may choose a parametric model with a fixed *a*-parameter to describe the data. Another implication is that there is not a strong item ordering according to the *p*-values (i.e., the items' proportion-correct). Note that item 2 is not a good item because it does not differentiate between respondents with a low and high total score, although the item is responsible for a larger spread in the mean ISRFs; removing this item results in a decrease of the H^T coefficient from 0.41 to 0.31.

Physical Functioning Scale

The sample consisted of 714 persons ($M_{age} = 51.1$, $SD = 18.94$); 48.9 percent were men. An interesting conceptual difference between the AQ and the PF scale is that the AQ measures a psychological construct, whereas the PF scale measures physical activities and clearly not a psychological construct. As our results show, this has a strong influence on the psychometric quality of the scale. Because the questions of the PF scale refer to concrete activities a person can perform and because there seems to be a clear ordering of those activities from easy to difficult, the scale seems to be much more like a Guttman (1950) scale than
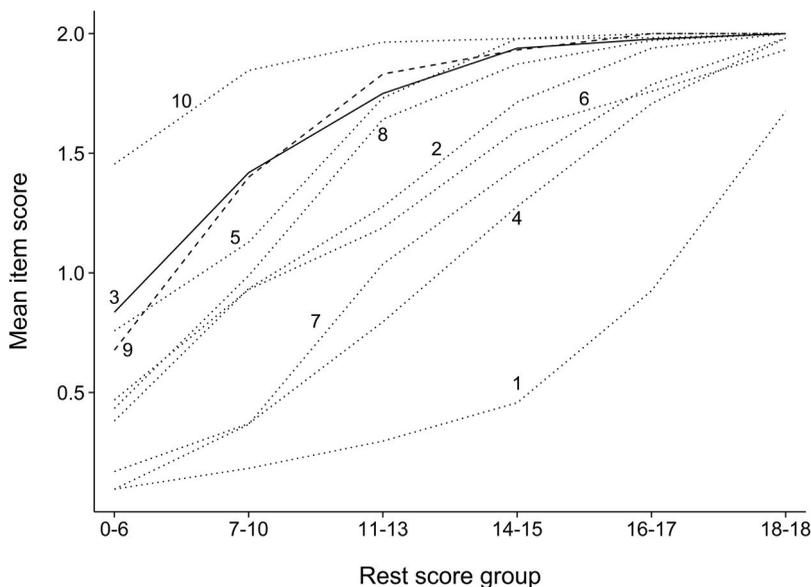
Figure 5.4 Mean ISRFs for the 10 items of the Physical Functioning scale.

the AQ. Thus, for the PF scale there is much similarity between the construct and the items that measure the construct. In contrast, for the AQ items the difference between the construct and the indicators (items) is larger than for the PF items and as a result scalability is lower. This is indeed reflected in the mean ISRF for the PF scale as compared to the AQ scale: Compared to the PF scale, the mean ISRFs of the AQ scale are less steep.

Table 5.6 gives the descriptive statistics for the 10 items of the PF scale. As expected, item-test correlations and H_i values are very high. H_i values varied between 0.69 and 0.80, resulting in an overall H value of 0.77 and thus indicating a very strong scale. As a consequence all items were selected by the AISP procedure using $c = 0.3$ (and thus also for $c = 0$). Inspecting the mean ISRFs of the PF items it is clear that most items discriminate much better across groups with different rest scores than the mean ISRFs of the Aggression scale (see Figure 5.4).

This is also reflected in the result from the IIO analyses. There were no violations against IIO with $H^T = 0.65$ indicating a high precision of the item ordering according to the rules proposed by Ligtvoet and colleagues (2010). However, one should be careful when interpreting this value. As Meijer and Egberink (2012) showed, H^T is sensitive to outliers. The mean ISRFs of items 1 and 10 are far apart from the other IRFs and these mean ISRFs are partly responsible for the high H^T value. When we removed these items, H^T dropped from 0.65 to 0.40. Thus, it is important to inspect the IRFs to get an idea about the spread of the IRFs.

Furthermore, it is clear that the IRFs of item 3 (*Lifting or carrying groceries*) and item 9 (*Walking one block*) almost coincide, indicating that these items do not reflect a different severity of physical functioning, at least in this population (note that the mean value of both items is 1.78). Given the fact that these items are so highly related ($H_{ij} = 0.69$), one may consider removing one of these items from the scale because each item does not overly contribute to the item ordering in the presence of the other. Also the mean ISRFs of items 2 and 6 are close together.

Table 5.6 Descriptive Statistics of the PF Scale

Item	Item content	Mean	SD	Item-total correlation	H_i
PF1	Vigorous activities, running, lifting	1.00	0.83	.74	.79
PF2	Moderate activities, moving table	1.60	0.63	.87	.79
PF3	Lifting or carrying groceries	1.78	0.50	.77	.74
PF4	Climbing several flights of stairs	1.38	0.78	.85	.78
PF5	Climbing one flight of stairs	1.75	0.53	.82	.78
PF6	Bending, kneeling, or stooping	1.52	0.66	.77	.69
PF7	Walking more than a mile	1.43	0.75	.88	.80
PF8	Walking several blocks	1.68	0.62	.84	.78
PF9	Walking one block	1.78	0.52	.80	.78
PF10	Bathing or dressing yourself	1.91	0.35	.55	.72

Results Using DETECT

Concerning data dimensionality, for the PF scale DETECT identified three subscales: (1,4,5), (2,3,6,10), and (7,8,9) with a maximum DETECT value of 0.89, implying some multidimensionality (Roussos & Ozbek, 2006). Note that cluster (7,8,9) consists of the "walking items" and two out of the three items from the cluster (1,4,5) are about "climbing stairs."

What should we make of these results? It is clear that the much lower DETECT value for the PF data as compared to the AQ data is the result of the stronger general factor in the PF data than the AQ data. As we mentioned earlier, there are several rules of thumb (e.g., Roussos & Ozbek, 2006) that would classify the AQ scale as multidimensional and even the PF scale would be considered moderately to largely multidimensional. However, Bonifay and colleagues (2014) discussed that "the flaw of DETECT, in terms of predicting parameter bias, is that its values are based on the size of the conditional covariances. These conditional covariances are determined by the size of the loadings on the group factor, but they are, by definition, independent of the general factor." Bonifay and colleagues (2014) in the context of bifactor modeling suggested therefore to consider DETECT values in the context of the general factor strength. In a nonparametric context it is interesting to consider the Mokken results in relation to the DETECT results. Remember that the ratio between the first and second order factors equaled six for the Physical Function scale and that the Mokken analysis showed that the H_{ij} values were very high (roughly between 0.7–0.9). These results point at a strong general factor. Thus, because DETECT is aimed at maximizing the differences between the item clusters and because it does not take the general factor into account, it is very sensitive to classifying items into different clusters. And perhaps most importantly, content-wise the DETECT results also lead to very narrow-band subscales, namely a three-item "walking scale," a three-item "climbing flights of stairs scale," and a subscale with the other items. Thus, it seems rather absurd to consider the PF data as multidimensional.

These conclusions are in line with the results reported in Bonifay and colleagues (2014), who found in a simulation study that using the rules of thumb suggested by Roussos and Ozbek (2006) for DETECT leads to many classifications of "multidimensionality"

whereas a strong common factor was present in the data. Like in the present study, this was due to strong group factors.

Person-Fit Results

To illustrate the usefulness of person-fit analysis, we investigated the score patterns of the PF scale in more detail. As shown earlier, the scale consists of 10 items that allow for IIO.

Given that we order the PF items from most popular to least popular, a simple and powerful person-fit statistic is the number of Guttman errors (Meijer, 1994). For dichotomous items, the number of Guttman errors equals the number of zero scores preceding a one score in a score pattern, thus, the pattern (10110010) contains five Guttman errors. A drawback of this statistic, however, is that it is confounded with the total score (Meijer, 1994). For polytomous items Emons (2008), therefore, proposed a normed version of the number of Guttman errors:

$$G_N^p = \frac{G^p}{\max\left(G^p | X_+\right)}.$$

In this statistic the number of Guttman errors (G^p) is weighted by its maximum value given the sum score (for details see Emons, 2008). G_N^p values ranged from zero (i.e., no misfit) through one (i.e., maximum misfit); for perfect response patterns the statistic is undefined.

We used the *R*-package *PerFit* (Tendeiro, 2014) to calculate G_N^p. The mean G_N^p equaled M = 0.05 (SD = 0.09). Most persons obtained low G_N^p values, indicating model conform behavior. However, there were outliers. One of the most extreme patterns was the pattern of person 535 with $G_N^p = 0.66$. It is instructive to compare the score pattern of person 535 with the score pattern of a randomly chosen person (person 165) with a low G_N^p value and a similar total score of $X_+ = 8$.

Person 165: [2 1 2 1 0 1 0 1 0 0] total score 8: $G_N^p = 0.06$.

Person 535: [0 0 1 2 0 1 0 0 2 2] total score 8: $G_N^p = 0.66$.

The items are ordered so that item 1 reflects the easiest tasks (*Bathing or dressing yourself*) and item 10 the most difficult tasks (*Vigorous activities*). Intriguing is that person 535 reports problems in bathing and dressing (item score 0), whereas there are no problems with "*Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports*" (item score 2). This is rather unexpected and certainly deserves closer attention. It is also clear that the total score patterns of both persons do indicate different things about physical functioning.

Summary

Reise and Waller (2009) wrote: "the clinical measures we examined are characterized by relatively small samples of poorly defined mixtures of patient groups of convenience, highly skewed score distributions, poorly articulated content domains, constructs with a

limited number of potential indicators, narrow band constructs (e.g., fatigue), quasi-traits, and high correlations among scales measuring different traits (e.g., due to the omnipresent negative affectivity dimension)." In this context, we hope to have shown that nonparametric IRT methods that explore data quality can be very useful before using more complex models or models that are based on more restrictive assumptions than nonparametric approaches. In contrast to the statistical tradition of data exploration, plotting data, looking for outliers, and carefully considering different models and approaches, the psychometric literature sometimes seems obsessed with all kinds of "model wars" and to formulate "do's" and "don'ts" to support a "Holy Grail" model. We do not endorse this. We see psychometric modeling more as a way to explore data quality than as a model fit contest (so researchers need to stop being fitidiots). Nonparametric approaches are excellent tools to decide whether parametric models are justified. Moreover, given the often not-so-easy-to-interpret fit statistics for parametric models, nonparametric tools provide a nice extension of the parametric toolkit to IRT modeling.

This is not to say that in some situations parametric models and even restrictive parametric models may be preferred over nonparametric models. For example, using θ scoring instead of total scores can reduce ceiling effects. Also the use of computer adaptive testing is much easier using parametric models than nonparametric approaches.

Finally and returning to the content of many outcome measures, our analyses once again showed that many clinical scales are simple instruments that require little or no understanding of psychological theory. This also makes it difficult, for example, to investigate IIO.[2] We seldom find detailed descriptions that made clear that the content of one item reflected a less severe type of the construct that is being measured than the content of another item. For example, we do not encounter many clinical or personality scales that a priori describe symptoms in order of severity. Perhaps this reflects the fact that many symptoms go together and that there is no real ordering in item severity. On the other hand we have the impression that not much thought is given to the theoretical underpinnings of many constructs.

References

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (In press). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling*.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, 63(3), 452–459.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289.

Chernyshenko, O. S., Stark, S., Chan, K., Dragow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36(4), 523–562.

2 Ligtvoet (2010) showed that very large samples (almost 3,000 persons) are needed to distinguish realistic numbers of ISRFs. Ligtvoet (2010) suggested that alternative approaches are needed to establish IIO in practice. One suggested approach was to form clusters of items that have similar adjacent IRFs and to establish IIO across different item clusters. A drawback of this approach may be that item clusters may consist of items that are redundant.

Derogatis, L. R. (1983). Brief Symptom Inventory: Administration, scoring, and procedures manual. Minneapolis, MN: National Computer Systems.

Egberink, I. L., & Meijer, R. R. (2011). An item response theory analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. *Assessment*, *18*(2), 201–212.

Emons, W. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224–247.

Guttman, L. (1950). *The basis for scalogram analysis*. In: S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazersfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.

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*(3), 331–347.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, *14*(4), 1523–1543.

Ligtvoet, R. (2010). *Essays on invariant item ordering*. Unpublished doctoral dissertation, University of Tilburg.

Ligtvoet, R., van der Ark, L., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*(4), 578–595.

Mazza, A., Punzo, A., & McGuire, B. (2012). KernSmoothIRT: An R package for kernel smoothing in item response theory. Retrieved from http://arxiv.org/pdf/1211.1183v1.pdf.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311–314.

Meijer, R. R. (2010). Watson, Deary, and Austin (2007) and Watson, Roberts, Gow, and Deary (2008): How to investigate whether personality items form a hierarchical scale?: Comment. *Personality and Individual Differences*, *48*(4), 502–503.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, *9*(3), 354–368.

Meijer, R. R., de Vries, R. M., & van Bruggen, V. (2011). An evaluation of the Brief Symptom Inventory–18 using item response theory: Which items are most strongly related to psychological distress? *Psychological Assessment*, *23*(1), 193–202.

Meijer, R. R., & Egberink, I. L. (2012). Investigating invariant item ordering in personality and clinical scales: Some empirical findings and a discussion. *Educational and Psychological Measurement*, *72*(4), 589–607.

Meijer, R. R., Egberink, I. L., Emons, W. 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*(3), 227–238.

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2014). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests*. The Hague: Eleven International.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague: De Gruyter.

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

Molenaar, I. W. (2004). About handy, handmade and handsome models. *Statistica Neerlandica*, *58*(1), 1–20.

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

Mroch, A. A., & Bolt, D. M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education*, *19*(1), 67–91.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679.

Partchev, I. (2013). Package "irtoys": Simple interface to the estimation and plotting of IRT models. Retrieved from http://cran.r-project.org/web/packages/irtoys/irtoys.pdf.

Pinsoneault, T. B. (2007). Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory-2 protocols. *Psychological Assessment*, *19*(1), 159–164.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611–630.

Ramsay, J. O. (2000). TestGraf: A program for the graphical analysis of multiple choice test and questionnaire data. Retrieved from www.psych.mcgill.ca/faculty/ramsay/TestGraf.html.

Rao, C. R., & Sinharay, S. (2007). *Handbook of Statistics 26. Psychometrics*. Amsterdam: Elsevier.

Reise, S. P. (2009). The emergence of item response theory models and the patient reported outcomes measurement information systems. *Austrian Journal of Statistics*, *38*(4), 211–220.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Roussos, L. A., & Ozbek, O. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*(3), 215–243.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, *10*(4), 345–359.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment*, *6*(3), 255–270.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*(1), 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*(2), 149–157.

Sijtsma, K., & Meijer, R. R. (2007). *Nonparametric item response theory and special topics*. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics 26. Psychometrics* (pp. 719–746). Amsterdam: Elsevier.

Sijtsma, K., Meijer, R. R., & Andries van der Ark, L. L. (2011). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences*, *50*(1), 31–37.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Smits, I. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory Mokken scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement*, *36*(6), 516–539.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293–325.

Stout, W., Habing, B., Douglas, J., & Kim, H. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*(4), 331–354.

Straat, J. H. (2012). *Using scalability coefficients and conditional association to assess monotone homogeneity* (Doctoral dissertation). Retrieved from http://arno.uvt.nl/show.cgi?fid=128094.

Straat, J. H., Van der Ark, L. A., & Sijtsma K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification*, *30*, 75–99.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*(3), 621–663.

Tendeiro, J. N. (2014). PerFit (version 1.2) [Computer software]. University of Groningen. Available from http://cran.r-project.org/web/packages/PerFit/.

van Abswoude, A. H., van der Ark, L., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3–24.

van der Ark, L.A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20, 1–19.

van der Ark, L.A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48, 1–27.

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

Wainer, H. (2005). *Graphic discovery. A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.

Ware, J.E., Jr., & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36): Conceptual framework and item selection. *Medical Care*, 30, 473–483.

Watson, R., Deary, I., & Austin, E. (2007). Are personality trait items reliably more or less "difficult"? Mokken scaling of the NEO-FFI. *Personality and Individual Differences*, 43(6), 1460–1469.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, 72(1), 69–91.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64(2), 213–249.

Zigmond, A.S., & Snaith, R.P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, 67(6), 361–370.