15

Unidimensional Item Response Theory

Rob R. Meijer and Jorge N. Tendeiro

Unidimensional item response theory (IRT) models have become important tools to evaluate the quality of psychological and educational measurement instruments. Strictly unidimensional data are unlikely to be observed in practice because data often originate from complex multifaceted psychological traits. Still, unidimensional models may provide a reasonable description of these data in many cases. In large-scale educational testing IRT is now the standard. Also for the construction and evaluation of psychological measurement instruments, IRT is starting to replace classical test theory (CTT). To illustrate this: When we recently obtained reviews of a paper from *Psychological Assessment*, one of the leading journals with respect to measurement and empirical evaluation of clinical instruments, it was stated that we did not have to explain in detail our IRT models because those models "are well-known to the audience of the journal." We would not have received this message, say, 10 years ago.

In this chapter, we distinguish parametric and nonparametric IRT models, and IRT models for dichotomous and polytomous item scores. We describe model assumptions, and we discuss model-data fit procedures and model choice.

Standard unidimensional IRT models do not take test content into account, that is, IRT models are formulated without specific reference to maximum performance testing (intelligence, achievement) and typical performance testing (personality, mood, vocational interest). Yet, when these models are applied to different types of data, there are interesting differences that will be discussed in this chapter and that may guide the use of these models in different areas of psychology.

Item Response Theory

Although CTT contributed to test and questionnaire construction for many years, in the papers by Lord (1952, 1953) and Birnbaum (1968) the foundation of modern test theory, or what was later called item response theory, was formulated. In these models the responses to items are explicitly modeled as the result of the interaction between characteristics of the items (e.g., difficulty, discrimination) and a person's latent variable (often denoted by the Greek letter θ). This variable may be intelligence, a personality

The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, First Edition. Edited by Paul Irwing, Tom Booth, and David J. Hughes.

trait, mood disorder, or any other variable of interest. Another important contribution in the development of, in particular, nonparametric IRT (NIRT) was made by Guttman (1944, 1950). His deterministic approach was based on the idea that, in the case of maximum performance testing, when a person p knows more than person q, then p responds positively to the same items as q plus one or more additional items. Furthermore, the items answered positively by a larger proportion of respondents are always the easiest or most popular items in the test. Because empirical data almost never satisfied these very strong model assumptions, stochastic nonparametric IRT versions of his deterministic model were formulated that were more suited to describe both typical and maximum performance data.

Researchers started with formulating models for dichotomous data, which were later extended for polytomous data. Because conceptually it is also easier to first explain the principles of dichotomous IRT models we first describe these types of models.

Dichotomous parametric item response models

All unidimensional IRT models (dichotomous and polytomous) are based on a number of assumptions with respect to the data. The data in this chapter are the answers of k persons to n items. In the case of dichotomous items these answers are almost always scored as 0 (incorrect, disagree) and 1 (correct, agree). In the case of polytomous items there are more than two categories. For example, in maximum performance testing these scores may be 0 (incorrect), 1 (partly correct), or 2 (correct), or in typical performance testing the scores may be 0 (agree), 1 (do not agree nor disagree), or 2 (disagree).

Assumptions and basic ideas The assumption of unidimensionality (UD) states that between-persons' differences in item responses are mainly caused by differences in one variable. Although all tests and questionnaires require more than one variable (or trait) to explain response behavior, some of these variables do not cause important differences in response behavior of respondents of a given population. Because items may generate different response behavior in different populations, dimensionality also depends on the population of persons. Instead of total (sum) scores as in CTT, scores are expressed on a θ scale (representing the assumed unique dimension of interest). This scale has a mean of zero and a standard deviation of 1, and can be interpreted as a z-score scale. Thus, someone with $\theta = 1$ has a θ-score that is 1SD above the mean score in the population of interest.

Another important assumption in IRT modeling is local independence (LI), which states that the responses in a test are statistically independent conditional on θ. Finally, it is assumed that the probability of giving a positive or correct response to an item is *monotonically* nondecreasing in θ (M assumption). This conditional probability is also called the item response function (IRF) and is denoted $P_i(\theta)$, where i indexes the item. The UD, LI, and M assumptions form the basis of the most widely used nonparametric and parametric IRT models in practice. All NIRT and IRT models presented in this chapter are based on these assumptions.

Parametric dichotomous item response models are further constrained by imposing well-defined mathematical models on the IRF. These models typically differ with respect to the number of parameters used. In the one-parameter logistic model (1PLM) or the Rasch (1960) model, only an item location parameter (denoted b_i) is

used to define an IRF, in the two-parameter Birnbaum model (2PLM) a discrimination parameter is added (denoted a_i), and in the three-parameter model (3PLM) an additional guessing parameter (denoted c_i) is used to describe the data. In Figure 15.1 we depict IRFs that comply with the 1, 2, and 3 PLM. Note that for the Rasch model the IRFs do not intersect because it is assumed that all IRFs have the same discrimination parameter (this parameter is not in the equation and thus it does not vary between items), whereas for the 2PLM different items may have different discrimination parameters and as a result the IRFs can cross; for the 3PLM the additional guessing parameter may result in IRFs that also have different lower asymptotes. Some authors also explore the use of a four-parameter logistic model with an additional parameter for the upper asymptote, but there are few published research examples of this model and we will not discuss it any further.

The IRF of the 3PLM for item i is given by

$$P_i(\theta) = P(X_i = 1|\theta) = c_i + (1 - c_i)\frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))},$$

where X_i is the random variable representing the score of item i. The 2PLM can be obtained from the 3PLM by setting the guessing parameter (c_i) equal to zero and the Rasch model can be obtained from the 2PLM by setting the discrimination parameter (a_i) to 1. As an example, consider the IRF of item 3 displayed in Figure 15.1(c). The probability that a person 1SD below the mean ($\theta = -1$) gives a positive answer to this item is equal to $.25 + (1-.25) \times \exp(.5(-1-1))/(1+\exp(.5(-1-1))) = .45$, whereas for a person 1SD above the mean ($\theta = 1$) the probability is $.25 + (1-.25)/2 = .63$.

The item location parameter, b_i, is defined as the point at the θ scale where the probability of giving a positive answer to an item equals $(1 + c_i)/2$ (i.e., halfway between c_i and 1). When $c_i = 0$ (in the 1PLM and 2PLM) the item location is defined as the point at the θ scale where the probability of endorsing this item equals .5. Thus, when we would move the IRF to the right side of the scale, the IRF would pertain to a more difficult item in the case of maximum performance testing; when we would move the IRF to the left side of the scale it would pertain to an easier item. For this reason, parameter b_i is also known as the *difficulty* parameter. Item location parameters usually range from −2.5 through +2.5. Furthermore, in parametric IRT models the item difficulties and the θ values are placed on the same scale. This is not the case in CTT where a total score has a different metric than the item difficulty, which in CTT is the proportion-correct score. The advantage of a common scale for the item difficulty and θ is that they can be very easily interpreted in relation to each other.

The steepness of the IRF is expressed in the discrimination parameter a_i. This parameter is a function of the tangent to the IRF at the point $\theta = b_i$. For most questionnaires and tests a_i parameters fluctuate between $a_i = .5$ and $a_i = 2.5$. The M assumption prevents negative values for this parameter. Moreover, values close to zero are related to items that discriminate poorly between persons close together in the θ scale (i.e., the associated IRFs are "flat"). The magnitude of the discrimination parameters depends on the type of questionnaire or test. Our experience is that for typical performance questionnaires (especially for clinical scales) as are in general somewhat higher than for maximum performance questionnaires. This has to do with the broadness of the construct. Many clinical scales consist of relatively homogeneous constructs, where questions are very similar, whereas maximum performance measures tap into broader constructs. When scales consist of items that are similar, all items have a strong relation

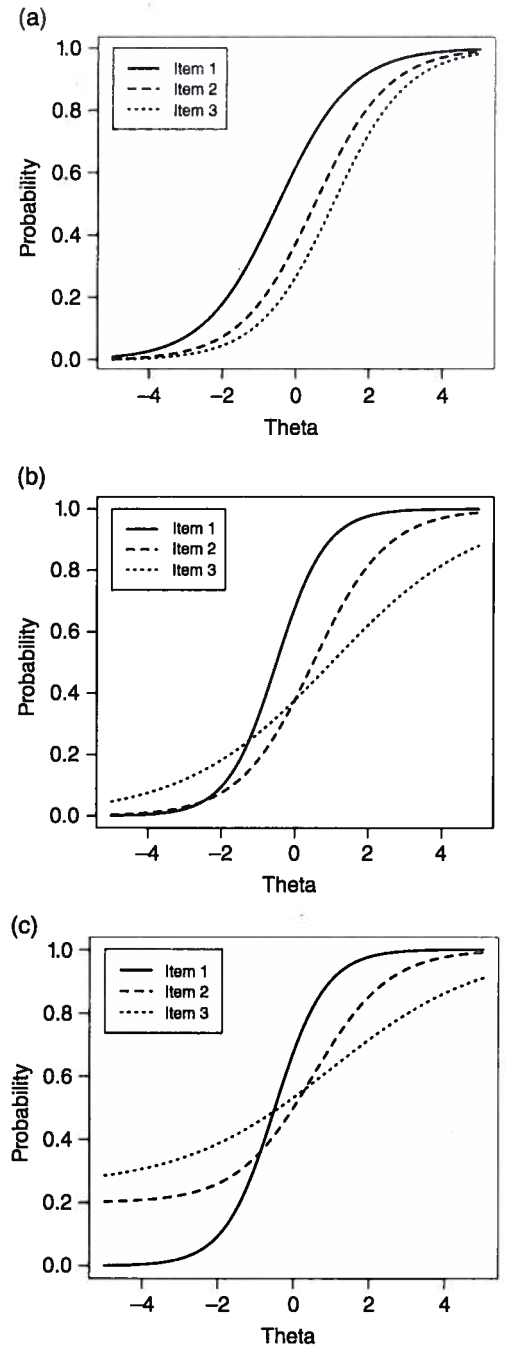

Figure 15.1 (a) Three IRFs from the 1PLM (Item 1: b_1=−.5; Item 2: b_2=.5; Item 3: b_3=1.0). (b) Three IRFs from the 2PLM (Item 1: a_1=1.5, b_1=−.5; Item 2: a_2=1.0, b_2=.5; Item 3: a_3=.5, b_3=1.0). (c) Three IRFs from the 3PLM (Item 1: a_1=1.5, b_1=−.5, c_1=0; Item 2: a_2=1.0, b_2=.5, c_2=.2; Item 3: a_3=.5, b_3=1.0, c_3=.25).

to the underlying trait and as a result the IRFs will be relatively steep. When the trait being measured is more heterogeneous in content IRFs will be, in general, less steep.

Although test constructors will, in general, strive for tests with items that have high discrimination parameters, there is a trade-off between tests measuring relatively narrow constructs with high discrimination parameters and tests measuring relatively broad constructs with lower discrimination parameters.

The guessing parameter of the 3PLM, c_i, specifies the lower asymptote of the IRF. For example, a value of $c_i = .20$ (Figure 15.1(c), item 2) implies that any person, regardless of his or her ability, has at least a 20% probability of answering the item correctly. This assumption is adequate for a multiple-choice item with five possible answer options, because a person may just try to *guess* the correct answer. The guessing parameter of item 3 in Figure 15.1(c) is .25, which is adequate for a multiple-choice item with four possible answer options. Of the three models presented, the Rasch model is most restrictive to the data because it has only one parameter, whereas the 3PLM is the least restrictive (more flexible).

Figure 15.2 further illustrates the use of IRFs. Two IRFs are shown from a Social Inhibition (SI) Scale with answer categories true/false (see Meijer & Tendeiro, 2012). We used the 2PLM to describe these data. Note that the probability of giving a positive answer is an increasing function of θ. First consider item SI23, "I find it difficult to meet strangers" ($a = 2.21$, $b = 0.22$). It is clear that someone with a trait value $\theta = 0$ has a probability of about .4 to endorse this item, whereas someone with, for example, $\theta = 1$ has a probability of about .8. The IRF of item SI23 is steep between $\theta = -1$ and $\theta = +1$, which means that this item discriminates well between persons that are relatively close on this region of the θ scale. Furthermore, persons with θ values smaller than $\theta = -1$ have a probability of endorsing this item of almost 0, whereas persons with θ at or above +1.5 have a probability of almost 1. Now consider item SI105, "I find it difficult to make new friends" ($a = 1.5$, $b = 1.65$). This item is less popular than item SI23: The difficulty parameter is larger, so the IRF is more to the right than the IRF of

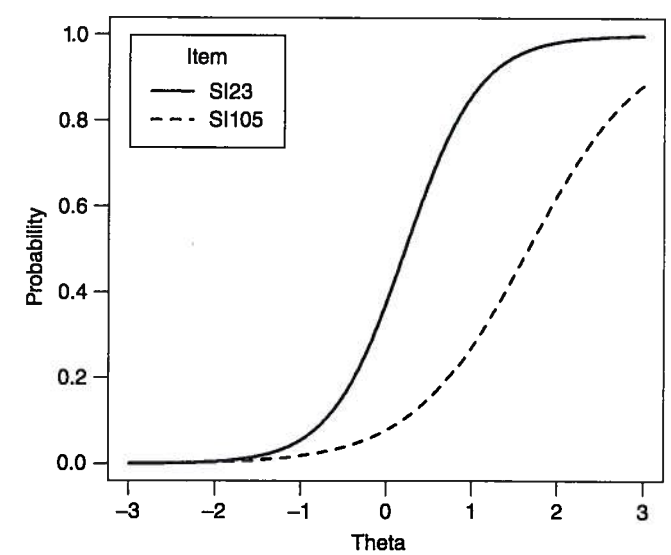

Figure 15.2 Two IRFs from the SI scale.

item SI23. Thus, a person should have a higher level of social inhibition to endorse item SI105 than to endorse item SI23. Moreover, item SI105 is less steep (smaller discrimination parameter).

Polytomous models

There are different types of polytomous IRT models and the theoretical foundations of the models are sometimes different (Embretson & Reise, 2000). However, the practical implications of the different models are often negligible. For example, Dumenci and Achenbach (2008) showed that differences between trait estimates obtained under the partial credit model and the graded response model were trivial. We will not discuss the different theoretical foundations of the models, see for example Embretson and Reise (2000) for more details, but instead emphasize their practical usefulness.

Polytomous item response models can be used to describe answers to items with more than two categories. In psychological assessment polytomous item scores are mostly used in combination with typical performance data like personality and mood questionnaires. Often five-point Likert scales are used where the score categories are ordered from "not indicative" to "indicative." An example is the question "I like to go to parties" from an Extraversion scale. Answer categories may be "Agree strongly" (scored 0), "Agree" (1), "Do not agree or disagree" (2), "Disagree" (3), and "Disagree strongly" (4). To model the response behavior for these types of items several models have been proposed.

In contrast to dichotomous IRT models, the unit of analysis is not the item but the answer categories. Each answer category has an associated response function (CRF). In polytomous IRT various models have been formulated to describe these CRFs. In van der Linden and Hambleton (1997), Embretson and Reise (2000), and Nering and Ostini (2010) a detailed overview is given of the nature and statistical foundations of the different polytomous IRT models. Next, we discuss the most often-used polytomous models for which easy-to-use software is available. We discuss the nominal response model, the partial credit model, the generalized partial credit model, and the graded response model.

Nominal response model The most general and most flexible polytomous model is the nominal response model (NRM) proposed by Bock (1972; see Thissen, Cai, & Bock, 2010 for a recent discussion). For example IRTPRO code, please see Appendix Code 1. Originally the NRM was proposed to model item responses to nominal data, such as the responses to multiple-choice items. Hence, and in contrast to other polytomous IRT models discussed next, in the NRM it is not assumed that the responses are ordered along the θ continuum. Assume that item i has $m+1$ response categories $k = 0, 1, \ldots, m$. The CRF $P_{ik}(\theta) = P(X_i = k|\theta)$ is the probability that a person with latent variable θ responds in category k on item i. Thus, an item has as many CRFs as response categories. In the NRM the probability of answering in category k depends on slope (a_{ik}) and intercept (c_{ik}) parameters, one pair per category response $k = 0, 1, \ldots, m$. The CRF for category k on item i is given by

$$P_{ik}(\theta) = \frac{\exp(a_{ik}\theta + c_{ik})}{\Sigma_{j=0}^{m} \exp(a_{ij}\theta + c_{ij})} \tag{15.1}$$

Parameter a_{ik} is related to the slope and parameter c_{ik} to the intercept of the k-th CRF. Because the model is not identified, Bock (1972) used the following constraint:

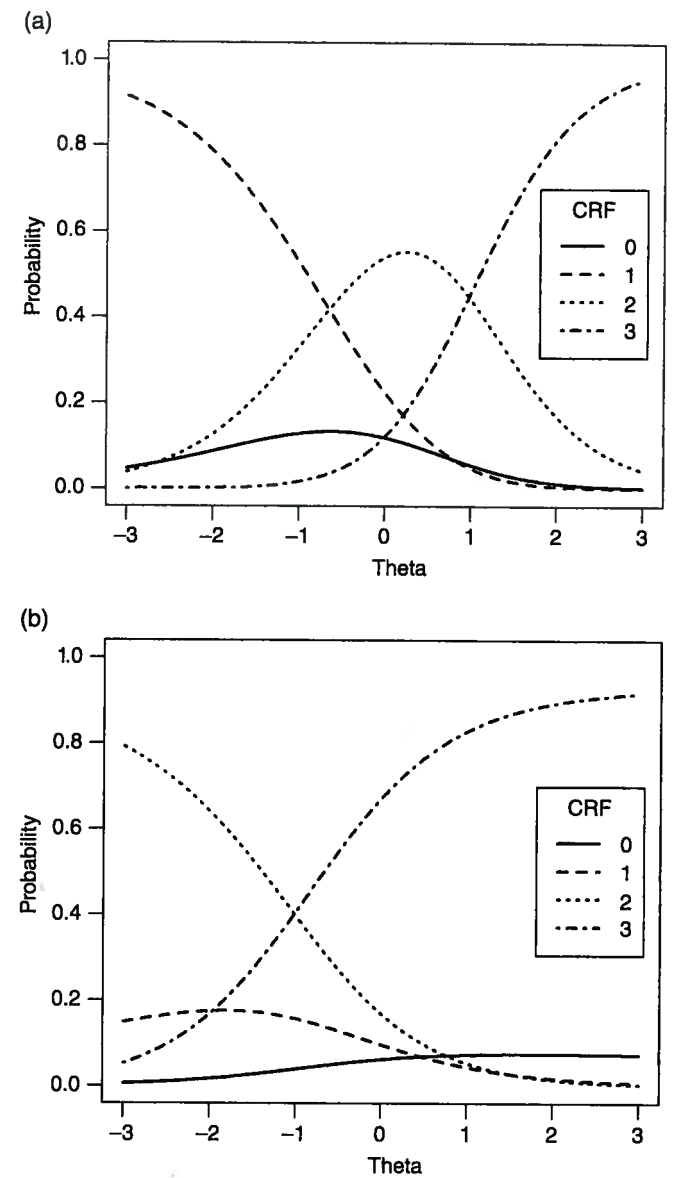

Figure 15.3 (a) CRFs for item 4 of the SPPC estimated using the NRM. (b) CRFs for item 5 of the SPPC estimated using the NRM.

$\Sigma a_{ik} = \Sigma c_{ik} = 0$. Alternatively, the parameters of the lowest CRF can be constrained to zero (e.g., the default in IRTPRO): $a_{i0} = c_{i0} = 0$.

In contrast to the graded response model and the generalized partial credit model (both discussed next), the NRM allows for different discrimination parameters within one item. This makes it a very interesting model to explore the quality of individual items. For example, Preston, Reise, Cai, and Hays (2011) argued that the NRM is very useful to check that presumed ordered responses indeed elicit ordered response behavior. Furthermore, the NRM can be used to check whether all item categories discriminate equally well between different θ values.

Table 15.1 Estimated item parameters of items 4 and 5 from the subscale Athletic Competence of Harter's SPPC.

Item (SPPC)	*Parameter*	*CRF 0*	*CRF 1*	*CRF 2*	*CRF 3*
Item 4	a_{ik}	0.00	−.78	.57	2.11
	c_{ik}	0.00	.64	1.52	.02
Item 5	a_{ik}	0.00	−.94	−1.31	.05
	c_{ik}	0.00	.44	1.00	2.38

Note. a_{ik} is the slope parameter and c_{ik} the intercept parameter; CRF is the category response function.

In Figure 15.3 we depicted the CRFs of the NRM for two items of the subscale Athletic Competence of Harter's Self Perception Profile for Children (SPPC, see Meijer et al. 2008 for a further description of this questionnaire and data). When a child fills out the SPPC, first he or she has to choose which of two statements applies to him or her and then indicates if the chosen statement is "sort of true for me" or "really true for me." Parameters were estimated using the IRTPRO software; estimates are shown in Table 15.1 In this example the actual ordinal nature of the answer categories of both items was disregarded by the NRM. Figure 15.3(a) shows an item that performs relatively well ("*Some children think they are better in sports than other children*"). It can be seen that category 1 is the most popular for low-θ children (for θ less than about −1.0). Note that θ here represents the amount of self-perceived Athletic Competence. Category 2 is preferred for children with θ between about −1.0 and +1.0, and for children with θ larger than about 1.0 category 3 is preferred. Category 0 was relatively unpopular across the entire θ scale. Now consider Figure 15.3(b). For this item ("*I am usually joining other children while playing in the schoolyard*") most children (with θ larger than about −1.0) chose category 3 independently of their position on the Athletic Competence scale. Category 2 was the most preferred category for children with ability below about −1.0. Furthermore, two out of the four category response functions are relatively flat. This item might be a badly functioning item: Half of its answer categories are seldom chosen. This item might need to be rephrased, or some answer options might need to be dropped.

Partial credit model, generalized partial credit model The partial credit model (PCM; Masters 1982) is suitable for items that involve a multistep procedure to find the item's correct answer. Partial credit is assigned to each step. Hence the item's score reflects the extent to which a person approached the correct answer. The PCM defining the CRF for category $k(k=0,\ldots,m)$ of item i involves parameters $b_{ij}(j=1,\ldots,m)$, which are often described as item-step difficulties. Item-step difficulties are the imaginary thresholds to take the step from one item score to the next. So, for a three-category item there are two item steps. b_{ij} is the point on the θ-axis where two consecutive CRFs intersect (more precisely, b_{ij} is the value of θ for which the probability of endorsing category j is the same as endorsing category $(j-1)$, $P_{ij}(\theta)=P_{i,j-1}(\theta)$) with $j=1,\ldots,m$. The CRF for category k on item i is given by

$$P_{ik}(\theta)=\frac{\exp\Sigma_{j=0}^{k}(\theta-b_{ij})}{\Sigma_{h=0}^{m}\exp\Sigma_{j=0}^{h}(\theta-b_{ij})},\ \text{with}\ \Sigma_{j=0}^{0}(\theta-b_{ij})\equiv 0 \qquad (15.2)$$

As an example, consider an item "I am good at sports" with three score categories, scored 0 (*not characteristic for me*), 1 (*a bit characteristic for me*), and 2 (*very characteristic for me*). In this case, we have two item-step difficulties, say $b_1=-.5$ and $b_2=1.5$. What is the probability that a person that is very sport-minded and performs at a national level in soccer, say with $\theta=2$, will choose the answer category 2? To obtain this probability we fill out the numerator in Equation 15.2 noting that in this case $k=2$, and thus: $\exp(2.5+.5)=\exp(3)=20.09$. The denominator in Equation 15.2 equals $\exp(0)+\exp(2.5)+\exp(2.5+.5)=33.27$, and thus the required probability equals $20.09/33.27=0.60$. Thus, there is a probability of 60% that this good athlete will choose option 2. Figure 15.4 displays the three CRFs for this item. Observe how b_1 and b_2 correspond to the intersection points of consecutive CRFs, as previously explained. It can be seen that persons with θ below −.5 have a high probability of not passing the first step (i.e., not collecting any credit for the item), persons with θ between −.5 and 1.5 have high probability of passing the first step, and persons with θ above 1.5 have a high probability of passing the second step.

An important observation is that in the PCM there is no discrimination parameter specified, so that the probability of endorsing a category only depends on the item-step locations and the person parameter. Like for the dichotomous Rasch model this may be a strong assumption, too strong for many data. Therefore, in the *generalized* partial credit model (Muraki, 1992) a slope parameter is added to the model. The CRF for category k on item i under the generalized PCM is given by

$$P_{ik}(\theta)=\frac{\exp\Sigma_{j=0}^{k}a_i(\theta-b_{ij})}{\Sigma_{h=0}^{m}\exp\Sigma_{j=0}^{h}a_i(\theta-b_{ij})},\ \text{with}\ \Sigma_{j=0}^{0}a_i(\theta-b_{ij})\equiv 0 \qquad (15.3)$$

Important is that the item discrimination depends on a combination of the slope parameter and the category intersections. Large slope parameters indicate steep category response functions and low slope parameters indicate flat response functions. The rating scale model (RSM; Andrich, 1978a, 1978b) can be derived from the PCM, but in the RSM each item has its own location parameter and the item-step difficulties are the same across items.

Graded response model The graded response model (GRM; Samejima, 1969, see Appendix Code 2 for example IRTPRO code) is suitable when answer categories are ordered (e.g., in Likert scales). Each item i is defined by a slope parameter, a_i, and by several threshold parameters, $b_{ij}(j=1,\ldots,m)$. To define the CRFs, we first define the item-step response function (ISRF) given by

$$P_{ik}^{*}(\theta)=P(X_i\geq k|\theta)=\frac{\exp[a_i(\theta-b_{ik})]}{1+\exp[a_i(\theta-b_{ik})]} \qquad (15.4)$$

that is, the probability of responding in category k or higher $(k=1,\ldots,m)$ computed using the 2PLM. Because the probability of responding in or above the lowest category equals one and because responding above the highest category equals 0, the CRF for category k is given by $P_{ik}(\theta)=P_{ik}^{*}(\theta)-P_{i(k+1)}^{*}(\theta)$, with $P_{i0}^{*}(\theta)=1$ and $P_{i(m+1)}^{*}(\theta)=0$. More specifically, for an item with three item score categories $(k=0,1,2)$, the item's CRFs are given by $P_{i0}(\theta)=1.0-P_{i1}^{*}(\theta)$, $P_{i1}(\theta)=P_{i1}^{*}(\theta)-P_{i2}^{*}(\theta)$, and $P_{i2}(\theta)=P_{i2}^{*}(\theta)-0$. In Figure 15.5 we depicted the CRFs for the two items of the SPPC Athletic Competence subscale discussed previously, parameters were estimated using IRTPRO. For

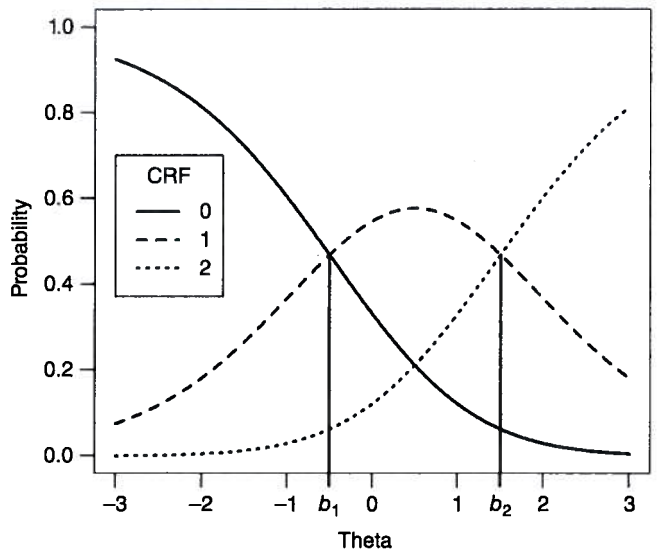

Figure 15.4 CRF of a polytomous item (three answer categories) using the PCM ($b_1 = -.5$, $b_2 = 1.5$).

good performing items, CRFs should be relatively steep (reflecting larger discrimination values) and separate (reflecting spread of the threshold parameters). The model shown in Figure 15.5a performs relatively well. It can be seen that category 0 is the most popular for low-ability children (for θ less than about −2.0). For children with θ between about −2.0 and −0.5 category 1 is preferred, for children with θ between about −0.5 and 1.0 category 2 is preferred, and for children with θ larger than about 1.0 category 3 is preferred. On the other hand, for the item shown in Figure 15.5b most children (with θ larger than about −1.0) chose category 3 independently of their position on this part of the θ scale. Furthermore, three out of the four category response functions are relatively flat. This item should be reviewed, for instance perhaps response categories 0, 1, 2 may be collapsed.

Item parameters estimation

Because parameter estimation is a relatively technical topic, we restrict ourselves here to some basic ideas and refer the reader to Baker and Kim (2004) and van der Linden and Hambleton (1997) for further details. There are essentially two types of methods to estimate the parameters of an IRT model: Maximum likelihood estimation (MLE) and Bayesian estimation. There are three types of MLE procedures: Joint maximum likelihood estimation (JML; Birnbaum, 1968), conditional maximum likelihood (CML; Rasch, 1960, Andersen, 1972), and marginal maximum likelihood (MML; Bock & Lieberman, 1970). Although JML allows estimating both item and person parameters jointly, an important drawback is that item parameter estimates are not necessarily consistent. CML solves this problem for the 1PLM by using a sufficiency property of this model that states that the likelihood function (a function with both item and person parameters as variables) of a person's response vector, conditional on his/her total score, does not depend on θ. This property allows estimating the 1PLM's item

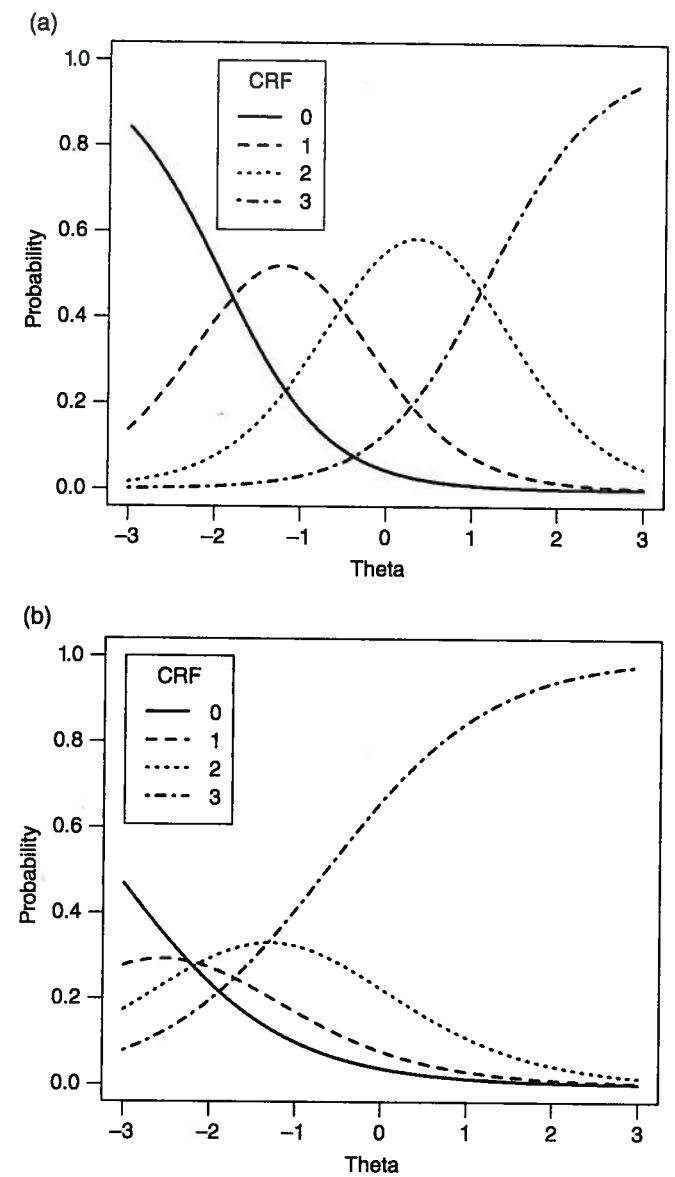

Figure 15.5 (a) CRFs for item 4 of the SPPC estimated using the GRM ($a_i = 1.59$, $b_1 = -1.94$, $b_2 = -.49$, $b_3 = 1.19$). (b) CRFs for item 5 of the SPPC estimated using the GRM ($a_i = 1.04$, $b_1 = -3.12$, $b_2 = -1.96$, $b_3 = -0.64$).

difficulty parameters independently from θ. Unfortunately, the CML only applies to the 1PLM, since there are no sufficient estimators for θ under the 2PLM or 3PLM. As an alternative, MML can be used. For MML it is assumed that the θ values have some known distribution (the normal distribution is typically used). This allows integrating the likelihood function over the ability distribution, thus estimation of item parameters can be freed from the person parameters. In conclusion, for the Rasch model both CML and MML can be chosen, whereas for the 2PLM and the 3PLM only the MML applies.

As an alternative to these procedures, one may use a Bayesian approach. Bayesian approaches have the advantage that they can be used in cases for which MLE methods lead to unreasonable estimated values or even fail to provide parameter estimates (e.g., for all 0's or all 1's response vectors). Bayesian methods based on marginal distributions (Mislevy, 1986) are currently the most widely used.

Bayesian methods are also used in Markov chain Monte Carlo (MCMC) methods that are applied in more advanced IRT models to solve complex dependency structures.

Test scoring and information

Once item parameters have been estimated using any of the methods explained in the previous section, it is possible to estimate the person parameters. A person parameter describes the person's position on the latent trait variable (θ). Both MLE and Bayesian estimation approaches are available. In MLE, the value of θ that maximizes the likelihood function for a particular response pattern is used as the estimate for θ. Advantages of MLE are that they tend to be consistent and efficient. The main disadvantages of MLE are that the peak of the likelihood function does not exist for perfect score patterns and patterns with all items incorrect. As a consequence, MLE can over- or under-estimate θ for nearly perfect response vectors. Warm (1989) proposed a weighted maximum likelihood estimation procedure (WML) that takes this problem into consideration.

As an alternative to MLE, two Bayesian approaches can be used. Both the expected a posteriori (EAP) and the modal or maximum a posteriori (MAP) methods rely on the person's response vector and on a prior distribution for θ. The likelihood (estimated from the response vector) is combined with the prior distribution for θ, which results in a posterior distribution for θ. The EAP estimate consists of the expected value of the posterior distribution, whereas the MAP consists of the mode of the same distribution. An advantage of Bayesian estimation is that the extra information obtained using the prior can improve the estimation of θ. A limitation of this type of procedure is that if the distance between a parameter and the mean of the prior distribution is large, the resulting estimated parameter will tend to regress to the mean of the prior (shrinkage).

Model-data fit

Item and model fit Several statistical methods are available to check whether an IRT model is in agreement with the data. There are global methods that can be used to investigate the fit of the IRT model to the complete test and there are methods to investigate item fit. For fit tests for the Rasch model we refer to Suárez-Falcón and Glas (2003) and Maydeu-Olivares and Montaño (2013). Next, we concentrate on a number of fit statistics that can be obtained when running the program IRTPRO and that are relatively easy to understand. Traditional approaches concern Pearson (Bock, 1972; Yen, 1981) and likelihood ratio (McKinley & Mills, 1985) χ^2 procedures. We will focus on dichotomous items unless stated otherwise because the procedures underlying fit for polytomous items do not fundamentally differ from the fit procedures for dichotomous items.

The Pearson approach is based on a statistic which assesses the distance between observed and expected scores. Large differences between observed and expected scores indicate misfit. Originally it was required to divide the latent scale in a number of disjoint intervals (say, u) such that roughly the same number of persons was placed in each

interval, according to their estimated ability. Yen's (1981) Q_1 statistic, for example, prespecified $u = 10$ such intervals. Next, observed and predicted scores were computed for each ability interval and each item score. Bock (1972) suggested using the median of the ability estimates in each interval to compute the predicted scores, whereas Yen (1981) suggested using the sample ability mean in each interval (Yen's statistic). The test statistic is given by

$$X_i^2 = \Sigma_{v=1}^{u} N_v \frac{(O_{iv} - E_{iv})^2}{E_{iv}(1 - E_{iv})}, \tag{15.5}$$

where i indexes the item, v indexes the group defined on the ability latent scale (Bock, 1972; Yen, 1981) u is the number of groups, N_v is the number of persons in group v, and O_{iv} and E_{iv} are the observed and expected proportion-correct responses for item i in group v, respectively. This test statistic follows approximately a χ^2 distribution with $u - g$ degrees of freedom, where g is the number of item parameters estimated by the IRT model. However, because groupings are based on an estimate of θ, which is both sample- and model-based and violates the assumption of the χ^2 statistic, Orlando and Thissen (2000) proposed instead to use NC scores on the test to create the groups of persons; their item fit statistic is denoted by $S-X_i^2$. For dichotomous items the summation in Equation 15.5 runs through NC scores 1 and $n - 1$ (n = number of items), since the proportion of persons answering item i correctly when NC = 0 is always 0 and the proportion of persons answering item i correctly when NC = n is always 1. The $S-X_i^2$ statistic is approximately χ^2 distributed with $(n - 1 - g)$ degrees of freedom. An extension of the $S-X_i^2$ statistic to polytomous items is readily available (Kang & Chen, 2008). The $S-X_i^2$ statistic is available in the IRTPRO software.

The likelihood ratio approach (McKinley & Mills, 1985) uses a different test statistic denoted G_i,

$$G_i^2 = 2\Sigma_{v=1}^{u} N_v \left[O_{iv} \log \frac{O_{iv}}{E_{iv}} + (1 - O_{iv}) \log \frac{1 - O_{iv}}{1 - E_{iv}} \right], \tag{15.6}$$

with the same notation as Equation 15.5. This statistic is also based on groups defined on the θ scale and follows approximately a χ^2 distribution with $(u - g)$ degrees of freedom. Orlando and Thissen (2000; see also Orlando & Thissen, 2003) proposed $S-G_i^2$, which is based on NC-groups.

Model-fit tests other than the χ^2 procedures just discussed have been proposed. Limited information fit tests (Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005) use observed and expected frequencies based on classifications of all possible response patterns. Specifically, low-order margins of contingency tables are used. Such approaches arose because it was verified that the traditional χ^2 (i.e., full information) statistics, when applied to contingency tables, led to empirical type I errors larger than the nominal errors of their asymptotic distributions (due to the sparseness of the tables for even realistic test lengths and/or number of response categories). An example of the limited information approach is the M_2 (Maydeu-Olivares & Joe, 2006) statistic, which is also available in the program IRTPRO.

There are item fit approaches that evaluate violations of LI. Yen (1984) proposed a statistic, Q_3, which was one of the first statistics used to investigate LI between item

responses conditional on θ. Q_3 statistic is the correlation between the scores of a pair of items from which the model's expected score has been partialled out. There are two problems associated to Q_3. On the one hand, Q_3 relies on estimated θ values for each response pattern. Such estimates are not always available because the likelihood function is sometimes not well defined, as explained previously. On the other hand, the reference distribution of Q_3 suggested by Yen (a normal distribution after a Fisher's transformation) does not seem to work well (Chen & Thissen, 1997). Chen and Thissen (1997) proposed instead a Pearson's χ^2 statistic to test LI between any pair of items. This statistic is given in IRTPRO as the LD *X2* statistic. According to the software manual these statistics are standardized χ^2 scores that are approximately z-scores. However, the LD *X2* LI statistics given in IRTPRO are difficult to interpret. As discussed in the IRTPRO manual, because these statistics are only approximately standardized, values of 2 or 3 should not be considered large. Instead, only values of 10 or larger should be taken as a serious violation. Our own experience with using these statistics to identify locally dependent item pairs is that it is often advisable to take item content into account. Moreover, very high a parameters (say, larger than $a = 3$) are sometimes a better indication of redundant items than local independence statistics like the LD *X2*.

Another model-fit approach is based on the Lagrange multiplier (LM) test (Glas, 1999). The idea of a LM test is to consider a model where an additional parameter is added to the IRT model of interest. Under the null hypothesis of LI this additional parameter equals zero. The LM test statistics are asymptotically χ^2 distributed with a number of degrees of freedom equal to the number of parameters fixed under the null hypothesis. This approach can also be used to test deviations between theoretical and empirical IRFs. Glas and Suárez-Falcón (2003) compared the detection performances between LM and other tests and concluded that the LM tests work relatively well. Extensions to polytomous models exist (Glas, 1999).

Recently, Ranger and Kuhn (2012) proposed fit statistics based on the information matrix and compared these statistics with other fit statistics. More details, and comparisons to other methods, can be found in their article.

Person fit Although an IRT model may give a reasonable description of the data, the item score patterns of some persons may be very unlikely under the assumed IRT model. For these persons, it is questionable whether the estimated θ score gives an adequate description of θ. Several statistical methods have been proposed to investigate whether an item score pattern is unlikely given the assumed IRT model. Meijer and Sijtsma (2001) give an overview of the different approaches and statistics that are available. The most often-used statistic is the standardized log-likelihood statistic l_z (Drasgow, Levine, & Williams, 1985). This statistic is based on the likelihood of a score pattern given the estimated trait value. To classify an item score pattern as fitting or misfitting a researcher needs a distribution of person-fit scores. One major problem with the l_z statistic is that its asymptotic standard normal distribution is only valid when true (not estimated) θs are used. This is a severe limitation in practice, since true abilities are typically unknown. Snijders (2001) proposed an extension of l_z, denoted l_z^*, which takes this problem into account. Magis, Raîche, and Béland (2012; see also Meijer & Tendeiro, 2012 for some important additional remarks) wrote a very readable tutorial and also provided *R* code to calculate the l_z^*.

Alternative approaches to calculating likelihood statistics were proposed by van Krimpen-Stoop and Meijer (2001) and recently by Tendeiro, Meijer, Schakel, and Maij-de Meij (2013). They used the so-called cumulative sum statistics that are sensitive to strings of item scores that may indicate aberrant behavior, like cheating behavior or random responding.

How serious is misfit and what does it mean? As some authors have mentioned, fit research is not unproblematic. Because IRT models are simple stochastic models that will never perfectly describe the data, fit always is a matter of degree. Also, for large datasets a model will always be rejected because of high power, even if model violations are small and have no practical consequences.

Furthermore, as we discussed before, the numerical values of many fit indices are sensitive to particular characteristics of the data. For example, the LD *X2* local independence statistics given in IRTPRO are difficult to interpret because the associated standardization has limitations. Thus, there is always an important subjective element in deciding when an item or item score pattern does not fit the model. Therefore, some authors argue for more research that investigates the effects of model misfit on the estimation of structural parameters.

When practically applying IRT models, it is often difficult to decide on the basis of fit research which items to remove from a scale. Some researchers only use some general indicators of misfit, others conclude after some detailed fit research that "despite the model misfit for the scale, we used the full scale, because the effects on the outcome measures were small." Perhaps removing items from a scale because of flat IRFs or violations of monotonicity is easiest because it is clear that such items do not contribute to any meaningful measurement. For example, Meijer, Tendeiro, and Wanders (2015) showed that an item from the aggression scale "I tell my friends openly when I disagree with them" did not discriminate between different trait levels and as such does not contribute to meaningful measurement.

With respect to person-fit research a sometimes-heard criticism is that although it is technically possible to identify misfitting item score patterns, the practical usefulness has not yet been shown. That is, it is often unclear what the misfit of an item score pattern really means in psychological terms. Is misfit due to random response behavior because of unmotivated test behavior? Or is it due to misunderstanding the questions? One of the few studies that tried to explain person misfit is Meijer et al. (2008). They combined person-fit results with qualitative information from interviews with teachers and other background variables to obtain information why children produced unlikely response patterns on a self-evaluation scale. Another interesting application was given in Conijn (2013) who conducted several studies to explain person misfit. For example, Conijn (2013) found that patients were more likely to show misfit on clinical scales when they experienced higher levels of psychological distress. What is clearly needed here are studies that address the psychological meaning of misfitting response patterns: We are very curious to see more empirical studies that explain *why* a score pattern is misfitting.

Nonparametric IRT

Nonparametric IRT (NIRT) models are based on the same set of assumptions as parametric IRT models (UD, LI, and M). However, unlike parametric IRT models, in NIRT the IRFs (dichotomous case) or CRFs (polytomous case) do not need to have

a logistic or any other functional form. In other words, no parameterized models of θ involving item and person parameters are defined. As a consequence, it is not possible to estimate person parameters even though a θ-latent scale is still assumed to exist. Instead of estimating θ, in NIRT the ordering of respondents on the observable sum score (total score) X_+ is used to stochastically order persons on the latent θ scale (Sijtsma & Molenaar, 2002). Hence, only the *ordinal* nature of the latent scale is of interest in NIRT. Under the UD, LI, and M assumptions the stochastic ordering of persons on the θ scale by X_+ holds for dichotomous items. Although for polytomous items this stochastic ordering does not hold in all cases (in theory; Hemker, Sijtsma, Molenaar, & Junker, 1997), van der Ark (2005) showed that this is not problematic in practical settings. However, this ordering holds for the *rest score* (total score minus score on an item) and therefore the rest-score is used instead of the total score. Also, item difficulty parameters are not estimated in NIRT. Instead, item proportion-correct scores similar to the ones in CTT are used. However, unlike CTT, in NIRT explicit models have been formulated and methods have been proposed to check these models.

The most popular nonparametric models are the Mokken (1971) models. Sijtsma and Molenaar (2002) devoted a complete monograph to these models and there are many papers that discuss measurement properties of these models and/or show how these models can be used to investigate the psychometric quality of tests and questionnaires (e.g., Meijer & Baneke, 2004).

Mokken (1971) proposed two models: The monotone homogeneity model (MHM) and the double monotonicity model (DMM). Both models have been formulated for dichotomous and polytomous item scores.

Monotone homogeneity model The MHM applies to both dichotomously and polytomously scored items. Both the dichotomous and polytomous MHMs are based on the UD and LI assumptions. Furthermore, monotonicity is assumed for the nonparametric IRFs (in the dichotomous case) or ISRFs (in the polytomous case). To check these assumptions several methods have been proposed that are incorporated in the *R* package ***mokken*** (van der Ark, 2007, 2012). We will discuss some of these methods in this chapter.

The MHM can be considered a nonparametric version of the GRM (for model selection, see next). In Figure 15.6 we plotted the nonparametric ISRFs for the SPPC Athletic Competence items 4 and 5. In Figure 15.5 we already showed the associated CRFs for the GRM; it is now interesting to compare the corresponding Figures 15.5 and 15.6. Persons were grouped according to their rest score; proportions of positive response per item step were then computed for each rest-score group of persons. Focusing first on item 5, it can be observed that item steps $P(X_i > 1)$ and $P(X_i > 2)$ are close together (Figure 15.6b). This shows that there is little difference between the first two answer options: Persons passing the first item step had a high probability of also passing the second step. In other words, item 5 does not discriminate well among persons, which confirms what we found via the GRM's CRFs (see Figure 15.5b). Figure 15.6a, on the other hand, shows ISRFs that are well separated from each other, highlighting item 4 as a good, discriminating item (supporting our previous findings using the GRM, see Figure 15.5a).

Double monotonicity model In his book in 1971 Mokken proposed the double monotonicity model for dichotomous items, later adapted for polytomous items (Molenaar, 1997). The DMM also assumes UD, LI, and M. Moreover, for ***dichotomous*** items, this

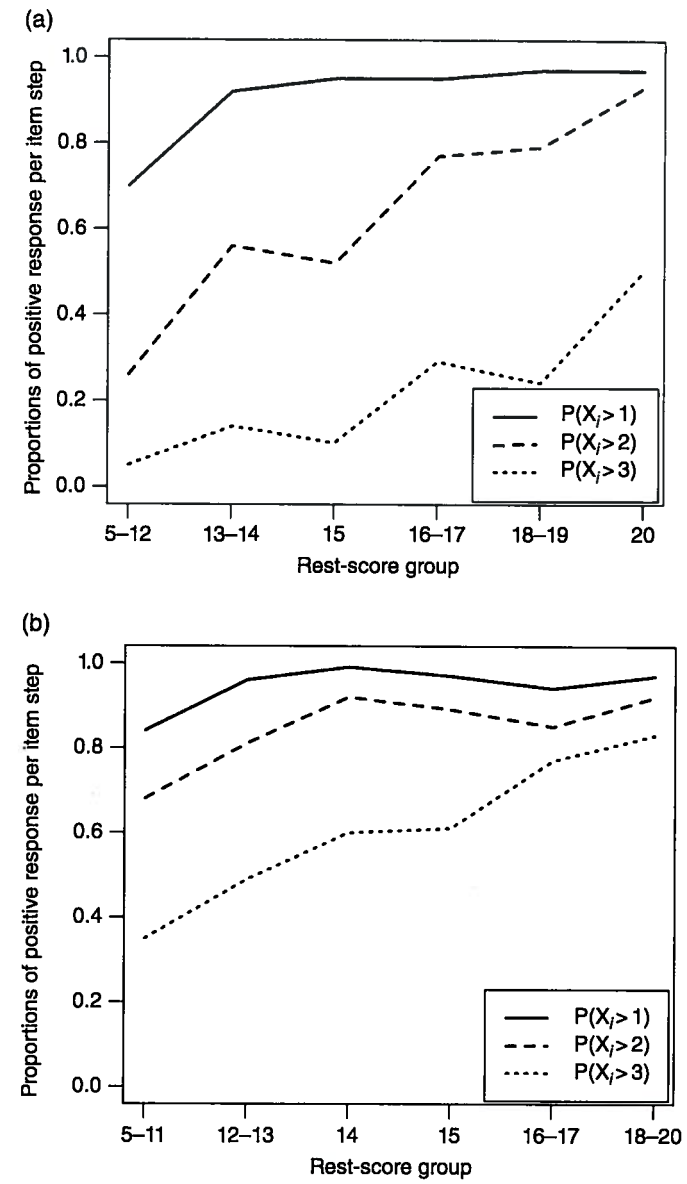

Figure 15.6 (a) Nonparametric ISRFs for item 4 of the Athletic Competence scale of the SPPC. (b) Nonparametric ISRFs for item 5 of the Athletic Competence scale of the SPPC.

model implies that the ordering of the items according to the proportion-correct score is the same for any value of θ. This invariant item ordering (IIO) property may be an interesting property for several applications. For example, it is often assumed but seldom checked that items have the same difficulty ordering for different levels of the latent trait. For example, for many psychological tests for children items are ordered from easy to difficult. When a child does not give the correct answer to, say, three or four subsequent items the test administration is stopped. Here it is assumed that for every child the item difficulty order is the same independently of the trait value. Another example can be

found in all kinds of scales that measure physical functioning. Egberink and Meijer (2011) showed that items from the Physical Functioning of the SF-36 complied with this model. For *polytomous* items the DMM does not imply IIO, but there are several methods proposed to check IIO for polytomous items. The interested reader is referred to Ligtvoet, van der Ark, te Marvelde, and Sijtsma (2010) and Meijer and Egberink (2012); we shall briefly refer to some options later in the chapter. Sijtsma, Meijer, and van der Ark (2011) provide an overview for conducting several steps for a Mokken scale analysis that incorporate both MHM and DMM, and IIO. Next, we briefly discuss some of these methods.

Dimensionality Assessing the dimensionality of the data to be analyzed is an important step in IRT model fitting. Basically, dimensionality of the data concerns the number of different latent variables that determine the scores of each item. Since the IRT models previously discussed are all unidimensional, it is important that relatively homogeneous sets of items are selected prior to attempt fitting an IRT model, whether parametric or not. We next discuss two different approaches to analyze dimensionality.

As mentioned in Sijtsma and Meijer (2007), nonparametric unidimensionality analysis is based on so-called conditional association (CA). For a general definition of CA see Holland and Rosenbaum (1986). One practical implication of CA is that all inter-item covariances within a test should be nonnegative in the sample. Strictly speaking, one negative covariance between a pair of items indicates misfit of the MHM (and of the DMM, since the latter implies the former). However, it is important to observe that all nonnegative inter-item covariances in the data do not imply that the MHM fits. Hence, having nonnegative inter-item covariances is a necessary, but not sufficient, condition for MHM fit.

To investigate Mokken scalability the automated item selection algorithm (AISP) is often used. The AISP is a popular method, although it is sensitive to specific item characteristics (see discussion that follows). The AISP uses the so-called *scalability* coefficient H. H is defined at the item level (H_i) and at the scale level (H). All H coefficients can be expressed as ratios of (sums of) observed covariances and maximum possible covariances. The scalability coefficients play a similar role in the MHM as the slopes of IRFs do in logistic IRT models: The steeper the nonparametric IRF, the larger the scalability indices. The AISP is an iterative algorithm that selects, in each step, the item that maximizes H given the already selected items up to that iteration. Thus, items that have relatively steep IRFs are successively added by the AISP. The procedure continues until the largest item scalability H_i is below a prespecified lowerbound, c. If there are unselected items, the AISP can be run again to create a second item cluster, and so on, until all items have been assigned to some cluster.

For the interpretation of scalability coefficients, Sijtsma and Molenaar (2002, p. 60) give the following guidelines. Item scalability coefficients H_i should be larger than a lowerbound c to be specified ($c = 0.3$ is often used in practice). Also, a scale H coefficient of at least equal to 0.3 is required to ensure that the ordering of persons according to their total score does provide a fair image of the true ordering of the persons on the latent scale (which cannot be directly assessed). More precisely, a scale can be classified as weak ($0.3 \le H < 0.4$), medium ($0.4 \le H < 0.5$), and strong ($H \ge 0.5$) according to the value of H.

Recently, Straat, van der Ark, and Sijtsma (2013) proposed alternatives to this procedure. They tackled the problem that when using the AISP procedure scales may be

Table 15.2 Estimated item parameters and H_i values.

	True parameters		*Estimated parameters and H_i values*		
Item	*a*	*b*	*a (SE)*	*b (SE)*	*H_i (SE)*
1	1.0	−1.0	1.07 (.14)	−0.88 (.11)	0.27 (.03)
2	1.0	−0.5	1.11 (.14)	−0.45 (.08)	0.25 (.02)
3	1.0	0.0	0.98 (.13)	−0.05 (.08)	0.22 (.02)
4	1.0	0.5	1.10 (.14)	0.53 (.09)	0.25 (.02)
5	1.0	1.0	1.03 (.14)	0.98 (.12)	0.27 (.03)

selected that satisfy scaling conditions at the moment the items are selected but may fail to do so when the scale is completed. They proposed a genetic algorithm that tries to find the most optimal division of a set of items into different scales. Although they found that this procedure performed better in some cases, a drawback of this procedure is that a user only gets information about the final result and cannot see which items are being selected during the selection process. So, we recommend using both the AISP and this genetic algorithm when selecting Mokken scales.

Although Mokken scaling has been quite popular to evaluate the quality of empirical datasets, two caveats are important to mention. A first caveat, as explained before, is that Mokken scaling procedures are especially sensitive to forming subscales with items that have high discriminating power and thus are especially sensitive to select items with steep IRFs. This is so, because H_i can be considered a nonparametric equivalent of the discrimination parameter in parametric IRT models. Although one may argue that these types of scales are very useful to discriminate persons with different total scores, compared to parametric models a Mokken scale analysis using the AISP procedure may reject items that may fit a 3PLM or 2PLM but have low discriminating power. In Table 15.2 we show numerical values of H_i under the 1PLM in a five-item test with item location parameters ranging from (−1,1). These values were calculated on the basis of simulated data with θ drawn from N(0,1). As can be seen, estimated item discrimination parameters around 1.0 resulted in H_i values lower than $H_i = 0.30$. Thus, although these six items perfectly fit the Rasch model, all of them will be rejected from the scale in case a researcher uses $c = 0.3$ as the lower bound in the AISP, which is the common choice in practice.

As an alternative, a second procedure to assess dimensionality is the nonparametric DETECT (Dimensionality Evaluation To Enumerate Contributing Traits; Kim, 1994; Stout, Habing, Douglas, & Kim, 1996; Zhang & Stout, 1999) approach. In contrast to the AISP algorithm in Mokken analysis, DETECT is based on covariances between any pair of items, conditional on θ. The LI assumption implies that all these conditional covariances are equal to zero; this condition is known as weak LI. To check weak LI, Stout and coworkers based their method on the observable property that the covariance between any pair of items, say items i and j, must be nonnegative for subgroups of persons that have the same rest score $R_{(-i,-j)}\left(R_{(-i,-j)} = X_+ - \left(X_i + X_j\right)\right)$. Assuming that the items measure Q latent variables to a different degree (i.e., multidimensionality), 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, X_+, or rest scores, $R_{(-i,-j)}$. Both scores summarize test performance but ignore

multidimensionality. Zhang and Stout (1999), however, showed that the sign of $\mathrm{Cov}(X_i, X_j|\theta_q)$ provides useful information about the dimensionality of the data. The covariance is positive when the two items measure the same latent variable and negative when they clearly measure different latent variables. This observation forms the basis of DETECT, allowing a set of items to be divided into clusters that together approach weak LI as well as is possible given all potential item clusters.

Several studies suggested rules-of-thumb that can be used to decide whether a dataset is unidimensional or multidimensional. Stout et al. (1996) considered DETECT values smaller than 0.1 indicating unidimensionality and DETECT values larger than 1 indicating multidimensionality (Stout et al., 1996). Roussos and Ozbek (1996) suggested to use the following rules-of-thumb: $DETECT < 0.2$ displays weak multidimensionality/approximate unidimensionality; $0.2 < DETECT < 0.4$: weak to moderate multidimensionality; $0.4 < DETECT < 1.0$ = moderate to large multidimensionality; $DETECT > 1.0$: strong multidimensionality. Recently, however, Bonifay et al. (2015) discussed that these values are sensitive to the factor structure of the dataset and the relation between general and group factors in the test. They investigated the effect of multidimensionality item parameter bias. The underlying idea was that every dataset is multidimensional to some extent and that it is more important to investigate what the effect is of using a unidimensional IRT model on particular outcome variables (such as item parameter bias) than to investigate whether or not a test is unidimensional. Perhaps, the most important conclusion of their study was that (Bonifay et al., 2015, p. 515);

> when the concern is with parameter bias caused by model misspecification, measuring the degree of multidimensionality does not provide the full picture. For example, in a long test with a reasonably strong general factor and many small group factors, parameter bias is expected to be relatively small regardless of the degree of multidimensionality. Thus, we recommend that *DETECT* values always be considered interactively with indices of factor strength (*ECV*) and factor structure (*PUC*).

Several studies compared the AISP algorithm with *DETECT*. van Abswoude, van der Ark, and Sijtsma (2004), and Mroch and Bolt (2006) showed that *DETECT* was better able to identify unidimensionality than the AISP. van Abswoude et al. (2004) suggested that unidimensionality can best be investigated using *DETECT* and that the best discriminating items can be selected through the AISP. Thus, the reader should be aware of the fact that both methods investigate different characteristics of the data. AISP is in particular sensitive to selecting items with high discriminating power.

Checking monotonicity and invariant item ordering In our discussion on assessing the dimensionality of the data, UD and LI assumptions were already considered. We next discuss how to check the M and IIO assumptions.

The assumption of monotonicity can be fairly easily investigated using graphical methods and eye-ball inspection (as we showed in Figure 15.6), since Junker (1993) proved that UD, LI, and M imply that $P(X_i = 1|R_{(-i)})$ is nondecreasing in $R_{(-i)}$, where $R_{(-i)} = X_+ - X_i$. This property is known as *manifest* monotonicity (MM). A simple statistical test exists to test the statistical significance to violations of MM; violations of MM imply violations of M, but the reverse is not necessarily valid. Both MSP and the *R* package *mokken* allow performing these analyses. *R* package *KernSmoothIRT* (Mazza,

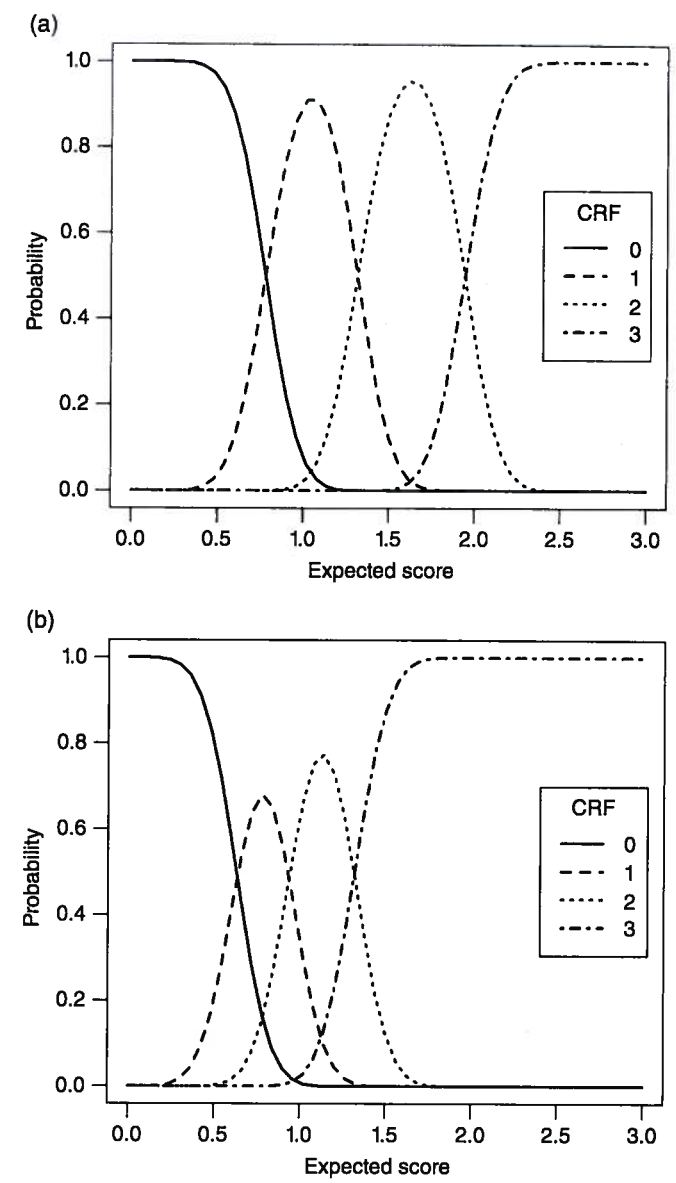

Figure 15.7 (a) Nonparametric CRFs for item 4 of the SPPC estimated using the R KernSmoothIRT package. (b) Nonparametric CRFs for item 5 of the SPPC estimated using the KernSmoothIRT package.

Punzo, & McGuire, 2012) provides an alternative to assess the M assumption which is based on kernel smoothing, a nonparametric regression technique. Figures 15.7(a) and (b) show CRFs estimated using this package for the same items of the SPCC considered previously. It is interesting to compare these CRFs with the parametric CRFs estimated using the GRM (see Figures 15.5a and b). Note that Figure 15.7(a) is very similar to Figure 15.5(a). In Figure 15.7(a) we use expected scores instead of latent trait values.

Figure 15.5(b) seems different than Figure 15.7(b), but note that options 1 and 2 only discriminate in a small area of the expected score. An alternative is to use the program TestGraf (Ramsay, 2000). In TestGraf continuous functions are provided that are based on kernel smoothing and that can be used to investigate the form of the IRFs or CRFs.

Several methods are available to check whether IIO can be safely assumed in the DMM framework. Technical details can be found in Sijtsma and Junker (1996), Sijtsma and Molenaar (2002), and van der Ark (2012). Here we outline the methods that are easily available for practitioners to use. The ***mokken*** package in R offers the ***pmatrix*** and ***restscore*** methods, which are two different variants of the same procedure to inspect IIO for ***dichotomous*** items. Given items i and j with $\bar{X}_i < \bar{X}_j$, and given an unweighed sum score S that does not depend on items i and j, the idea is to verify whether $P(X_i = 1|S = s) \leq P(X_j = 1|S = s)$ for all admissible realizations of S. That is, it is checked whether for every total score s, the probability of answering the easiest item correctly (thus, the proportion-correct score) is larger than answering the more difficult item correctly. Violations of the inequality are tested for their statistical significance. The ***pmatrix*** and ***restscore*** methods are not suitable for polytomous items since DMM does not imply IIO in this case. Ligtvoet et al. (2010) introduced a method (***check.iio*** command in R) that is suitable for both dichotomous and polytomous items. In MSP5.0, the p-matrix and rest-score methods are available for dichotomous items; for polytomous items the ***mokken*** package in R should be used.

Model selection

There are basically two strategies for IRT model selection. In the first strategy, a researcher tries to find the best fitting model with the least number of item parameters. Thus, when the Rasch model can describe the data, a researcher will use the Rasch model and not the 3PLM, and when the 3PLM shows the best fit, this model will be used. A second strategy is to select items for which the responses are in agreement with a prespecified model. For example, the Rasch model may be preferred because the total scores are sufficient statistics for the trait score (i.e., the trait scores can be estimated using the item parameters and the total scores only, no pattern of responses is required). Another argument to use the Rasch model has to do with sample size. As Lord (1980) discussed, if there is only a small group of persons, the a parameter cannot be determined accurately for some items. Lord (1980) conducted a small empirical study and he concluded that "for the 10- and 15-item tests, the Rasch estimator x may be slightly superior to the two-parameter estimator (...) when the number of cases available for estimating the item parameters is less than 100 or 200." Alternatively, the DMM model may be preferred because the ordering of the items according to their difficulty is the same for each person independent of the θ level. In such cases the model can be selected first and then items are selected that can be best described through this model.

For dichotomous data, the 2PLM and the 3PLM are often used because they give an adequate description of many types of data. The 2PLM model may be chosen when there is no guessing involved. Thus, the 2PLM seems to be a suitable model to describe answering behavior on noncognitive questionnaires (personality, mood disorders). The 3PLM can be used when any guessing is involved, as it may happen with cognitive measures (intelligence and educational testing). For polytomous items, as we discussed

previously, the NRM is a valid option for cases in which score categories cannot be necessarily ordered, whereas the (generalized) PCM and the GRM can be considered when the score categories are ordered.

Then there is the question of whether to use a parametric or a nonparametric model. One reason for using nonparametric IRT models is that they are more flexible than parametric models. For example, an IRF may be increasing but not have a logistic structure. A second reason may be sample size. An often-used argument is that when the sample size is relatively modest nonparametric approaches can be used as alternatives to parametric models that, in general, require more persons to estimate parameters. However, recent research showed that researchers should be careful when using small samples. For example, Kappenberg-ten Holt (2014) cautioned that the use of samples of $n = 200$ results in positive bias of the H coefficient, which reduces with increasing sample sizes. In relation to this, a researcher can use standard errors for the H coefficient to obtain an idea about the variability of the coefficient. Also, DETECT input specification file requires a minimum size of 400 persons. Therefore, perhaps the biggest advantage of the nonparametric approach is that it provides some alternative techniques to explore data quality without forcing the data in a structure they may not have.

There are also limitations to the use of nonparametric IRT models. The models are less suited to the construction of computer adaptive tests or when using change scores. Several authors have discussed that change scores are more difficult to interpret using total scores than when using parametric IRT scoring (Brouwer, Meijer, & Zevalkink, 2013; Embretson & Reise, 2000; Reise & Haviland, 2005). A general guide in deciding which model to apply is that nonparametric IRT is an interesting tool to explore data quality, however when trait estimates are needed parametric models must be used.

Alternative approaches: Ideal point models

To analyze polytomous scale data we discussed several models that assume a dominance response process where an individual high on θ is assumed to answer positively with high probability. This approach dates back from Likert's approach to the development and analysis of rating scales. In a recent issue of ***Industrial and Organizational Psychology-Perspectives on Science and Practice***, Drasgow, Chernyshenko, and Stark (2010; see also Weekers & Meijer, 2008) published a discussion paper in which they argued that for personality assessment ideal point test models based on Thurstone scaling procedure are superior over dominance models because the former models provide a better representation of the choice process underlying rating scale judgments. They also discussed that model misspecification can have important consequences in practical test use, such as in personnel selection. In ideal point models the probability of endorsement is assumed to be directly related to the proximity of the statement to the person's standing on the assessed trait. In a series of response papers to this article, several authors criticized or endorsed the claims made by Drasgow et al. (2010) and made suggestions for further research. From these papers, it is clear that still much is unknown about (1) the underlying response process to rating scale data, (2) which test model should be used to describe responses to noncognitive measures, and (3) what the consequences are of model misspecification in practice. We think that future research may shed light on these issues.

Concluding Remarks

In this chapter, we presented an overview of unidimensional IRT modeling. At the start of this chapter we discussed that in scientific journals devoted to test construction and evaluation, IRT is the state-of-the-art technique. In test and questionnaire construction of commercial test batteries our experience is that IRT is not the standard. Evers, Sijtsma, Lucassen, and Meijer (2010) described in the 2009 revision of the Dutch Rating System for Test Quality for the first time IRT criteria to judge whether IRT techniques were in agreement with professional standards. We think there is much to be gained through the application of IRT in test construction. Our experience is that IRT analyses on existing scales show that many scales consist of items and subtests that can be improved through a more rigorous analysis of the quality of individual items. Although IRT is a stronger measurement theory than CTT and estimation of item and person parameters is not easy, for a practitioner there is (sometimes free) software available (see Box 15.1). Hopefully, this chapter contributes to the more wide-spread use of IRT analyses in test construction and evaluation.

Box 15.1: Computer Programs

X Calibre (www.assess.com). One, two, three PL, graded response model, rating scale model, partial credit model.

BILOG-MG (www.ssicentral.com) one, two, three parameter logistic model, differential item functioning.

WINSTEPS and FACETS (www.winsteps.com) Rasch model.

PARSCALE (www.ssicentral.com). Graded response model, partial credit model, generalized partial credit model, generalized partial credit model,

IRTPRO (www.ssicentral.com) One, two, three parameter logistic model, graded, generalized partial credit model, differential functioning

MSP5.0 Mokken models

R package Mokken Mokken models

R package KernSmothIRT Kernel smoothing

TESTGRAF

References

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34*, 42–54.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581–594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques.* New York, NY: Marcel Dekker, Inc.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*(1), 1–16.

Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability.* In F. M. Lord & M. R. Novic (Eds), Statistical theories of mental test scores *(Ch. 17–20)*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*(2), 179–197.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling, 22*, 504–516. DOI: 10.1080/10705511.2014.938596.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). Measuring individual significant change on the BDI-II through IRT-based statistics. *Psychotherapy Research, 23*(5), 489–501.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^P tables. *British Journal of Mathematical and Statistical Psychology, 59*(1), 173–194.

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

Conijn, J. M. (2013). *Detecting and explaining person misfit in non-cognitive measurement* (Unpublished Doctoral Dissertation). University of Tilburg, the Netherlands.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*(4), 465–476.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86.

Dumenci, L., & Achenbach, T. M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20*(1), 55–62.

Egberink, I. J. 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.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure, and results. *International Journal of Testing, 10*(4), 295–317.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273–294.

Glas, C. A. W., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*(2), 87–106.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

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.* 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. *The Annals of Statistics, 14*(4), 1523–1543.

IRTPRO 2.1, Scientific Software International, Lincolnwood, IL.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*(3), 1359–1378.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391–406.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data (Unpublished Doctoral dissertation). University of Illinois at Urbana-Champaign.

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, 70*(4), 578–595.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, 7.*

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517–549.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37*(1), 57–81.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732.

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika, 78*(1), 116–133.

Mazza, A., Punzo, A., & McGuire, B. (2012). *KernSmoothIRT: An R package for kernel smoothing in item response theory.* Accessed 09 April 2012. http://arxiv.org/pdf/1211.1183v1.pdf.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(1), 49–57.

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., & Egberink, I. J. 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. 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*(3), 227–238.

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

Meijer, R. R., & Tendeiro, J. J. (2012). The use of the l_z and the l_z* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics, 37*(6), 758–766.

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. K. (2015). The use of nonparametric item response theory to explore data quality. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment.* (pp. 85–110). London: Routledge.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*(2), 177–195.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton/Berlin: 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.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models.* New York, NY: Routledge/Taylor & Francis Group.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298.

Preston, K., Reise, S., Cai, L., & Hays R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*(3), 523–550.

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

Ranger, J., & Kuhn, J. T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement, 49*(3), 247–268.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford: Nielsen & Lydiche.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of Personality Assessment, 84*(3), 228–238.

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

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

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. (2007). Nonparametric item response theory and special topics. In C. C. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26) (pp. 719–746). Amsterdam: Elsevier.

Sijtsma, K., Meijer, R.R., & van der Ark, L. A. (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 Publications, Inc.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*(3), 331–342.

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

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Methodological artifacts in dimensionality assessment of the hospital anxiety and depression scale (HADS). *Journal of Psychosomatic Research, 71*(2), 116–121.

Suárez-Falcón, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 56*(1), 127–143.

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored internet testing. *Educational and Psychological Measurement, 73*(1), 143–161.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In *Handbook of polytomous item response theory models* (pp. 43–75). New York, NY: Routledge/Taylor & Francis Group.

van Abswoude, A. A. H., van der Ark, L. A., & 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. (2005). Stochastic Ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, *70*(2), 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. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, *48*(5), 1–27.

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

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*(2), 199–218.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*(1), 65–77.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

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

Code Appendix

Code 1 IRTPRO code for nominal response model of athletic competence.

```
Project:
    Name = SPPC Athletic Competence;

Data:
    File = .\SPPC Athletic Competence.ssig;

Analysis:
    Name = Test1;
    Mode = Calibration;
```



```
Title:
Comments:
Estimation:
    Method = BAEM;
    E-Step = 500, 1e-005;
    SE = S-EM;
    M-Step = 50, 1e-006;
    Quadrature = 49, 6;
    SEM = 0.001;
    SS = 1e-005;

Scoring:
    Mean = 0;
    SD = 1;

Miscellaneous:
    Decimal = 2;
    Processors = 4;
    Print CTLD, P-Nums, Diagnostic;
    Min Exp = 1;

Groups:

Group :
    Dimension = 1;
    Items = sp1, sp2, sp3, sp4, sp5, sp6;
    Codes(sp1) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp2) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp3) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp4) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp5) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp6) = 1(0), 2(1), 3(2), 4(3);
    Model(sp1) = Nominal;
    AlphaMatrix(sp1) = Trend;
    GammaMatrix(sp1) = Trend;
    Model(sp2) = Nominal;
    AlphaMatrix(sp2) = Trend;
    GammaMatrix(sp2) = Trend;
    Model(sp3) = Nominal;
    AlphaMatrix(sp3) = Trend;
    GammaMatrix(sp3) = Trend;
    Model(sp4) = Nominal;
    AlphaMatrix(sp4) = Trend;
```

Code 2 IRTPRO code for graded response model of athletic competence.

```
Project:
    Name = SPPC Athletic Competence;

Data:
    File = .\SPPC Athletic Competence.ssig;

Analysis:
    Name = Test1;
    Mode = Calibration;

Title:
Comments:
Estimation:
    Method = BAEM;
    E-Step = 500, 1e-005;
    SE = S-EM;
    M-Step = 50, 1e-006;
    Quadrature = 49, 6;
    SEM = 0.001;
    SS = 1e-005;

Scoring:
    Mean = 0;
    SD = 1;

Miscellaneous:
    Decimal = 2;
    Processors = 4;
    Print CTLD, P-Nums, Diagnostic;
    Min Exp = 1;

Groups:

Group :
    Dimension = 1;
    Items = sp1, sp2, sp3, sp4, sp5, sp6;
    Codes(sp1) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp2) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp3) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp4) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp5) = 1(0), 2(1), 3(2), 4(3);
    Codes(sp6) = 1(0), 2(1), 3(2), 4(3);
    Model(sp1) = Graded;
    Model(sp2) = Graded;
```



```
    Model(sp3) = Graded;
    Model(sp4) = Graded;
    Model(sp5) = Graded;
    Model(sp6) = Graded;
    Mean = 0.0;
    Covariance = 1.0;

Constraints:
```