Comparing of four IRT models when analyzing two tests for inductive reasoning

de Koning, E.; Sijtsma, K.; Hamers, J.H.M.

Published in:
Applied Psychological Measurement

Publication date:
2002

Document Version
Publisher's PDF, also known as Version of record

Comparison of Four IRT Models When Analyzing Two Tests for Inductive Reasoning

Els De Koning, Klaas Sijtsma and Jo H. M. Hamers

Applied Psychological Measurement 2002; 26; 302
DOI: 10.1177/0146621602026003005

Comparison of Four IRT Models When Analyzing Two Tests for Inductive Reasoning

Els de Koning, Leiden University

Klaas Sijtsma, Tilburg University

Jo H. M. Hamers, Utrecht University

This article discusses the use of the nonparametric IRT Mokken models of monotone homogeneity and double monotonicity and the parametric Rasch and Verhelst models for the analysis of binary test data. First, the four IRT models are discussed and compared at the theoretical level, and for each model, methods are discussed for evaluating the fit of the model to test data. Second, each of the four IRT models is used for analyzing the data collected by means of two versions of a test for inductive reasoning. Finally, the results are discussed and recommendations are given about the practical use of each of the IRT models. It is concluded that the simultaneous use of several IRT models for practical data analysis provides more insight into the structure of tests than the rigid use of only one model. *Index terms: double monotonicity model, goodness-of-fit in IRT, IRT model comparison, monotone homogeneity model, nonparametric item response models, one parameter logistic model, parametric item response models, Rasch model.*

Psychological testing aims at measuring individuals on scales for cognitive abilities such as inductive reasoning and divergent thinking, but also personality traits such as introversion and neuroticism. Item response theory (IRT; e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997) provides a set of statistical models for the analysis of the item scores of a sample of persons who responded to the items from a test, aimed at constructing scales for persons and items. For the IRT parameter estimates for persons and items to be useful, the IRT model should fit the person-by-item item score matrix.

The purpose of this study was to compare the usefulness of two nonparametric and two parametric IRT models for the analysis of empirical test data relevant to applied psychological measurement. The IRT models were the nonparametric Mokken (1971; Mokken & Lewis, 1982; related to later work of Stout, 1990) models of monotone homogeneity and double monotonicity and the parametric Rasch (1960) and Verhelst (Verhelst & Glas, 1995) models. Several of these IRT models are nested, each more restrictive model adding one assumption about the response process to the more general model that is closest. Also, each of these IRT models uses its own statistical model-data fit methods implemented in a stand-alone computer program for that particular model. The authors believe it is interesting to users of IRT models, at both the theoretical and the practical level, to learn where the differences between models and their methods are when analyzing test data. In particular, they illustrate how various models and their methods can be combined to obtain more information about one's data than when just one model and its methods were used. Moreover, this study allows researchers to compare the less-well-known nonparametric IRT models with the better known parametric IRT models. The data analyzed here as an example were collected with two versions of

Applied Psychological Measurement, Vol. 26 No. 3, September 2002, 302–320

a test for inductive reasoning (de Koning, Sijtsma, & Hamers, in press). Inductive reasoning is at the core of the intelligence construct.

First, the authors discuss and compare the four IRT models at the theoretical level. Also, for each IRT model, they discuss methods for evaluating the fit of the model to test data. Then, they use each of the four IRT models for analyzing the data collected by means of two inductive reasoning tests. Finally, they discuss and compare the data analysis results and give recommendations on the use of each of the IRT models.

IRT Models

IRT models use the item response function (IRF) for explaining a respondent's probability of answering an item correctly as a function of a latent trait, such as inductive reasoning (de Koning et al., in press). Let θ denote the latent trait and let j be the item index ($j = 1, \dots, J$). Also, let X_j denote the item score variable with realizations x_j, valued 0 (incorrect answer) or 1 (correct answer); and let $\mathbf{X}$ denote a vector with J random variables and $\mathbf{x}$ the vector with J 0,1-realizations of these random variables. The IRF is the conditional probability $P_j(\theta) \equiv P(X_j = 1 \mid \theta)$. A common assumption of the IRT models discussed here is local independence of the J item scores,

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{j=1}^{J} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}. \tag{1}$$

By integrating θ out, one obtains the J-variate distribution of the item scores,

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{j=1}^{J} P_j(\theta)^{x_j}[1 - P_j(\theta)]^{1-x_j}\, dG(\theta), \tag{2}$$

where $G(\theta)$ is the cumulative distribution function of θ. The multivariate distribution $P(\mathbf{X} = \mathbf{x})$ is not restricted in any way without further restrictions on the IRFs, the cumulative distribution of θ, or both (Holland & Rosenbaum, 1986; Junker, 1993; Suppes & Zanotti, 1981). As IRT models differ in the way they restrict the IRFs, the multivariate distribution of $\mathbf{X}$ is differently restricted for different IRT models. To investigate the observable consequences of an IRT model for the purpose of model-data fit, usually only the univariate and the bivariate marginal distributions of this J-variate distribution are studied (e.g., see Sijtsma & Junker, 1996).

Nonparametric IRT models place order restrictions on the IRFs, for example, requiring each IRF to be monotonely nondecreasing in θ. Parametric IRT models define the IRFs to be functions from a particular parametric family, for example, the logistic. As both kinds of models have advantages and disadvantages (Meijer, Sijtsma, & Smid, 1990; Sijtsma, 1998), the authors used both for data analysis and compared the results. First, they discuss the two nonparametric IRT models used here, the monotone homogeneity model (MHM) and the double monotonicity model (DMM) (Mokken, 1971; Mokken & Lewis, 1982; Sijtsma, 1998; similar models were discussed by Stout, 1990; Stout et al., 1996). Second, the authors discuss the parametric IRT models, the Rasch (1960) model (RM), and the Verhelst (Verhelst & Glas, 1995) model, also known as the OPLM (the abbreviation of one parameter logistic model, following Verhelst's terminology). The simple RM was chosen because of its theoretical advantages concerning parameter estimation and population-independent measurement (Fischer & Molenaar, 1995), and the more general OPLM because it allows for different IRF slopes, as does the two-parameter logistic model (e.g., Hambleton & Swaminathan, 1985), while maintaining all the favorable properties of the RM.

Nonparametric IRT Models

Monotone Homogeneity Model and Methods

Theoretical background. The MHM assumes unidimensionality, local independence of the item scores, and monotonicity in θ, that is, nondecreasing IRFs. The shape of an IRF can be anything as long as the curve is nondecreasing, meaning that it could be an irregular and jumpy curve, or a function with many steps, but also a neat convex or concave function, or even a logistic or normal ogive function. Moreover, these possibilities and many others can exist next to each other in the same test. The importance of the MHM is that with its few assumptions it allows for the ordering of respondents on θ using their unweighted number-correct score, defined as $X_+ = \Sigma X_j$ (Hemker, Sijtsma, Molenaar, & Junker, 1997). The observable score X_+ replaces θ, which cannot be estimated due to the nonparametric nature of the MHM.

In this study, the MHM is useful for three reasons. First, although its theoretical foundation is complex (e.g., Hemker et al., 1997; Junker, 1993; Junker & Sijtsma, 2000), the MHM is based on few assumptions, making it robust compared with more complex models as a tool for analyzing test data. This means that it often fits test data when more restrictive IRT models fail (Meijer et al., 1990). Second, the MHM has powerful methods for investigating model-data fit, which makes it interesting as a method for test construction from a data-analysis point of view. Third, the parametric models to be discussed shortly are special cases of the MHM. Thus, an MHM data analysis is a strong and interesting precursor for a data analysis using parametric models (Meijer et al., 1990).

Investigating monotonicity. The program MSP5 for Windows (acronym MSP; which stands for Mokken Scale analysis for Polytomous items; Molenaar & Sijtsma, 2000) was used for investigating monotonicity in θ. For this purpose, Mokken (1971; Mokken & Lewis, 1982) proposed to use the scalability coefficient H_{jk} for pairs of items, the scalability coefficient H_j for an item with respect to the other items in the test, and the scalability coefficient H for the total set of items in the test. Let $Cov(X_j, X_k)$ denote the covariance between X_j and X_k, and let $Cov(X_j, X_k)_{max}$ denote the maximum covariance given the marginal distributions of X_j and X_k. Also, let π_j be the proportion of examinees with a 1-score on item j and let π_{jk} be the proportion with 1-scores on both items j and k. Furthermore, assume that $\pi_j \leq \pi_k$; then,

$$H_{jk} = \frac{Cov(X_j, X_k)}{Cov(X_j, X_k)_{\max}} = \frac{\pi_{jk} - \pi_j \pi_k}{\pi_j(1 - \pi_k)}. \tag{3}$$

In this study, the authors interpret the H coefficients as statistics for slopes of IRFs relative to the spread of the total X_+ score in the group under consideration. Thus, items with high H_j discriminate well in the group in which they are used. This interpretation allows them to compare H_j with IRF slope indices from the RM and the OPLM. The authors now show that H_{jk}, H_j, and H are nondecreasing functions of the variance of X_+.

For this purpose, they write H_{jk} in terms of variances of a total score $X_{j+k} = X_j + X_k$: Let $\sigma^2(X_{j+k}) = \sigma^2(X_j) + \sigma^2(X_k) + 2Cov(X_j, X_k)$; σ_0^2 the variance under marginal independence of X_j and X_k, that is, $\sigma_0^2 = \sigma^2(X_j) + \sigma^2(X_k)$ (note that σ_0^2 depends only on the fixed π_js); and $\sigma_{\max}^2$ the maximum possible variance given the marginal distributions of X_j and X_k, that is, $\sigma_{\max}^2 = \sigma^2(X_j) + \sigma^2(X_k) + 2Cov(X_j, X_k)_{\max}$ (note that $\sigma_{\max}^2$ depends only on the π_js). Then one may write

$$H_{jk} = \frac{\sigma^2(X_{j+k}) - \sigma_0^2}{\sigma_{\max}^2 - \sigma_0^2}. \tag{4}$$

This equation shows that H_{jk} is an increasing function of the variance of the total total score X_{j+k}

when the marginal distributions of the item scores are assumed fixed (which also causes σ_0^2 and $\sigma_{\max}^2$ to be fixed quantities, because they depend only on the π_js).

The item coefficient H_j is defined as the ratio of the sum of all J-1 covariances of fixed item j and the other items $k (k \neq j)$ in the numerator and the sum of J-1 corresponding maximum covariances in the denominator. To write H_j as a ratio of differences of variance terms, one can again use the definitions of variances of sums $X_{j+k} = X_j + X_k$ for all J-1 item pairs j, k with $k \neq j$, such that for fixed item marginals, H_j is an increasing function of the variances $\sigma^2(X_{j+k})$,

$$H_j = \frac{\sum_{k \neq j} Cov(X_j, X_k)}{\sum_{k \neq j} Cov(X_j, X_k)_{\max}} = \frac{\sum_{k \neq j} \left[\sigma^2(X_{j+k}) - \sigma^2(X_{j+k})_0\right]}{\sum_{k \neq j} \left[\sigma^2(X_{j+k})_{\max} - \sigma^2(X_{j+k})_0\right]}. \tag{5}$$

An interesting question is whether H_j is an increasing function of the variance of the total score X_+. Keeping item marginals π_j constant, this variance depends on the $\frac{1}{2}J(J\text{-}1)$ item pair covariances, but only J-1 item pair covariances figure in H_j. If the variance of X_+ increases, this is due to an increase in at least one item pair covariance. If covariances involving item j increase, H_j also increases; otherwise, H_j is not affected. In other words, H_j only picks up some of the increases in $\sigma^2(X_+)$ but not all, and H_j, therefore, is a nondecreasing rather than a strictly increasing function of the total score variance. The relation between H_j and $\sigma^2(X_+)$ is used later on when comparing item indices from different IRT models.

Finally, the H coefficient for J items is the ratio of all $\frac{1}{2}J(J\text{-}1)$ item pair covariances in the numerator and all $\frac{1}{2}J(J\text{-}1)$ maximum item pair covariances in the denominator. Mokken (1971, p. 151) showed that H is a strictly increasing function of $\sigma^2(X_+)$. Mokken, Lewis, and Sijtsma (1986) argued that under the MHM, higher positive H values reflect higher discrimination power of the items and, as a result, more confidence in the ordering of respondents by means of X_+. Because positive H_j values close to 0 imply nearly horizontal IRFs, for practical test construction purposes Mokken (1971, p. 185) recommended to use $H_j = 0.3$ as a lower bound.

MSP estimates an IRF by means of the nonlinear regression of the score of the item j under consideration on the sum score on the other J-1 items. Junker and Sijtsma (2000) called this sum score the restscore, denoted R and defined as $R = \Sigma_{k \neq j} X_k$, because it is the sum of the rest of the item scores not including item j. Under the MHM, this regression must be monotonely nondecreasing in R (Junker, 1993; Junker & Sijtsma, 2000; Rosenbaum, 1984). In empirical data, decreases in the estimate of the item-restscore regression thus may indicate misfit of the MHM and are tested for significance (Molenaar & Sijtsma, 2000).

Investigating dimensionality. For investigating the dimensionality of an item set, MSP contains an automated item selection procedure (e.g., Hemker, Sijtsma, & Molenaar, 1995; Sijtsma, 1998), based primarily on the inter-item covariances and the strengths of the relationships between items and the latent trait(s) as expressed by the item H_j coefficients. Based on such information, clusters of related items measuring a common θ may be identified. For selecting the first item cluster, the item selection procedure starts with the two items having the highest significant positive H_{jk}, and adds items from the remaining items one by one. This is done under the restrictions that (a) items have positive covariances with each of the items already selected in the cluster at a particular point in the selection process; (b) items have an H_j value of at least c ($c > 0$) with the already selected items; and (c) the item selected in a particular selection round maximizes the overall H of this item and the selected items, given all possible choices from the remaining items. The item selection stops when no more items can be selected that satisfy these criteria for inclusion in the cluster. If items remain unselected, using the same selection criteria the selection procedure continues and tries to select a second cluster, a third, and so on, until no items are left that can be clustered.

The end result may be one or more item clusters that each tap another latent trait or latent trait composite, and possibly one or a few items that tap unique latent traits. 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 test structure. The clusters can be the basis for further analysis, such as the fitting of a particular model for each separate cluster. Comprehensive discussions of this item selection procedure are given by Mokken (1971, pp. 170-199), Hemker et al. (1995), and Sijtsma and Molenaar (2002).

Double Monotonicity Model and Methods

Theoretical background. The second nonparametric IRT model is the DMM. This model is based on the same set of assumptions as the MHM and adds the assumption that the IRFs do not intersect. This means that for two arbitrary items j and k, if it is known for one θ_0 that $P_j(\theta_0) < P_k(\theta_0)$, then it follows that for any θ, $P_j(\theta) \leq P_k(\theta)$. This is readily generalized to an ordering of J items. Because the IRFs do not intersect, the item ordering based on the $P_j(\theta)$s is the same, except for possible ties, for each value of θ. Since θ and the conditional probabilities $P_j(\theta)$ are not observable, in practice the proportions of correct answers for each item, the π_js, are used for ordering items. It was shown (Sijtsma & Junker, 1996) that under the DMM this ordering reflects the ordering based on the $P_j(\theta)$s.

Sijtsma and Junker (1996) discussed the importance of an item ordering that is invariant across θ for applications such as differential item functioning (e.g., Holland & Wainer, 1993), person fit analysis (e.g., Meijer & Sijtsma, 2001), and intelligence testing procedures (e.g., Bleichrodt, Drenth, Zaal, & Resing, 1985). In general, each application of a test that assumes that the ordering of the items is the same for different individuals requires the property of an invariant item ordering to hold for the test. See Sijtsma and Junker (1997) for a model-data fit study of the DMM to developmental psychology test data concerning transitive reasoning.

Investigating intersection of IRFs. MSP was used to investigate whether IRFs intersected. The scalability H^T coefficient (Sijtsma & Meijer, 1992) for the J items in a test and the person coefficients H_a^T (a is a person index) were used to evaluate intersection of the J IRFs. The H^T coefficient is similar in mathematical structure to the H coefficient, and the H_a^T coefficient to the H_j coefficient, but H^T and H_a^T use covariances between J item scores of pairs of persons. Sijtsma and Meijer (1992) showed that this role change of items and persons renders the resulting H^T and H_a^T coefficients suitable for investigating the intersection of the IRFs of a set of items. In particular, they recommended that, for all practical purposes, simultaneously $H^T \geq 0.3$ and the percentage of negative H_a^T values < 10 mean that the J IRFs do not intersect.

An additional investigation of the nonintersection of the IRFs compares for each pair of items the item-restscore regressions (here, the restscore was based on J-2 items, excluding the two items, j and k, under consideration: $S = \Sigma_{m \neq j,k} X_m$). If in a particular restscore group the ordering of the items j and k is opposite to the ordering of the items in the total group, the null hypothesis of equality of item difficulties is tested against the alternative that the items have the ordering as found in the restscore group (Molenaar & Sijtsma, 2000).

Parametric IRT Models

Rasch Model and Methods

Theoretical background. Like the DMM, the RM and the OPLM are special cases of the MHM. The RM specializes the MHM by assuming logistic IRFs with a location parameter, denoted δ, and no other item parameters. This implies that the IRFs are parallel curves that do not intersect. The IRF is defined as

$$P_j(\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}. \tag{6}$$

Because its IRFs do not intersect, the RM is a special case of the DMM. The RM has strong statistical properties, for example, sufficiency of total scores for estimation of model parameters and the population independence of these model parameters. Because the model is so well documented, refer to Fischer and Molenaar (1995) for detailed information.

Investigating model-data fit. The authors used the computer program RSP (Rasch Scaling Program; Glas & Ellis, 1993; also see Robin, Xing, & Hambleton, 1999) for investigating fit of the RM to the data. RSP uses the asymptotic chi-square statistic R_1 (Glas, 1988; Glas & Verhelst, 1995) for testing the null hypothesis that J IRFs are logistic with equal slopes against the alternative that they are not, and the asymptotic chi-square statistic R_2 (Glas, 1988; Glas & Verhelst, 1995) for testing the null hypothesis that J items are unidimensional and locally independent against the alternative that they are not. For larger numbers of items, the calculation of R_1 and R_2 may run into trouble (Glas & Ellis, 1993, p. 90), and RSP instead resorts to the approximate chi-square statistics Q_1 and Q_2 (Van den Wollenberg, 1982), which test the same hypothesis as R_1 and R_2, respectively, but are computationally less complex.

In addition to global statistical testing using R_1 / Q_1 and R_2 / Q_2, the authors used local testing by means of the approximate standard normal statistic U_j (Molenaar, 1983), which tests for each separate item the null hypothesis that its IRF is logistic with slope 1 against the alternative that it is not. Because the authors compare U_j with the item scalability coefficient H_j, they give the formal definition of U_j. Let $R = 1, \ldots, J$-1 be the restscore excluding item j, and define cutpoints c_1 and c_2 such that $R \leq c_1$ defines the lowest quartile of the distribution of R and $R \geq c_2$ defines the highest quartile. For frequencies n_{rj} (the number of respondents with a restscore r and a score of 1 on item j) and the expectation under conditional maximum likelihood estimation given the RM, $E(n_{rj}/RM)$, they define differences $\mathit{diff}_{rj} = n_{rj} - E(n_{rj}/RM)$, for all r, which after proper standardization are denoted z_{rj}. Statistic U_j is defined as

$$U_j = \frac{\sum_{r=1}^{c_1} z_{rj} - \sum_{r=c_2}^{J-1} z_{rj}}{(c_1 + J - c_2)^{1/2}}. \tag{7}$$

Positive values of U_j indicate that the IRF is flatter than expected, and negative values indicate that the IRF is steeper than expected.

One Parameter Logistic Model (OPLM) and Methods

Theoretical background. Like the RM, the OPLM has logistic IRFs that vary in location, but unlike the RM, the IRFs of the OPLM also vary in slope. The OPLM does not have a slope parameter, however, but instead requires the researcher to specify an integer slope A_j for each item. As a result, the slope is fixed and the only parameters to be estimated are the location and the ability parameters. The IRF is defined as

$$P_j(\theta) = \frac{\exp[A_j(\theta - \delta_j)]}{1 + \exp[A_j(\theta - \delta_j)]}, A_j \in N^+. \tag{8}$$

Verhelst and Glas (1995) showed that with a user-specified integer slope, the statistical properties of the RM apply for the OPLM. If the model with user-specified slopes is estimated and does not fit the data, new integer values for the slopes may be specified and the model again is estimated and tested for fit to the data. This is repeated until a fitting model is obtained, perhaps after some items have been deleted, and the final slope indices and parameter estimates are interpreted.

Investigating model-data fit. The authors used the computer program OPLM (Verhelst, 1992) for estimating and fitting a logistic IRT model with location parameters and user-specified slope indices. Model fitting according to the OPLM concentrates on the assumptions of monotonicity and sufficiency of a total score based on item scores weighted by their slope indices. OPLM has no test statistics for evaluating unidimensionality and local independence. The null hypothesis of monotonicity and sufficiency is tested by means of a global asymptotic chi-square statistic R_{1c} and by four item-fit statistics, one of which is a chi-square and the other three being comparable to U_j for the RM. Rather than discussing these item-fit statistics, in the Results section the authors report (a) the values of the global R_{1c} before and after slope index specification and (b) the final slope indices A_j used.

Comparison of IRT Models

The OPLM is a special case of the MHM and a liberalization of the RM, but the mutual ordering of the OPLM and the DMM is not clear-cut. The OPLM has logistic IRFs, which is a restriction with respect to the DMM, but the DMM has nonintersecting IRFs, and this can be a strong restriction, especially for longer tests. Thus, the partial ordering of the four IRT models from weak to strong assumptions is MHM–DMM/OPLM–RM.

Method

To illustrate the use of the four IRT models for analyzing relevant psychological test data, the authors used a pretest version and a posttest version of a test for inductive reasoning (de Koning & Hamers, 1995, 1999; de Koning et al., in press), called Test for Inductive Reasoning I (TIR-I) and Test for Inductive Reasoning II (TIR-II), respectively.

Tests, item types. Figures 1a and 1b provide examples of inductive reasoning items. The tests distinguish attribute tasks and relation tasks. Comparing attributes requires the child to simultaneously consider two objects, whereas comparing relations requires simultaneously considering three objects (Klauer, 1989; also see Carpenter, Just, & Shell, 1990). As comparison processes can be aimed at finding similarities, dissimilarities, or both, attribute tasks and relation tasks both can deal with any of these three modes. Finally, tasks with either concrete objects based on daily life experience or geometric objects referring to reasoning at a more abstract level were distinguished (see de Koning et al., in press, for further justification; also Klaver, 1989). To summarize, the TIR-I and the TIR-II each contained 12 types of inductive reasoning items: Attribute or Relation; crossed with Similarities, Dissimilarities, or Both; crossed with Concrete or Abstract. These 12 item types are summarized in Figures 1a (Attribute Items) and 1b (Relation Items). See de Koning et al. (in press) for a detailed description of the 12 item types.

Typical questions posed with different item types are mentioned in the first columns of Figures 1a and 1b. The response mode of each item depends on the item type. Each response was scored as incorrect (score of 0) or correct (score of 1).

The TIR-I and the TIR-II each had 27 unique items and shared 16 anchor items for the purpose of equating (de Koning et al., in press). All 12 item types (Figures 1a and 1b) were represented among the anchor items. Table 1 shows the distribution of the items across the tests and across the item types.

Samples, procedure. The representative samples (stratified using social-economic status) contained 476 third-grade primary school children for the TIR-I and 478 third-grade primary school

Figure 1a
Review of the TIR Item Types: Attribute Items

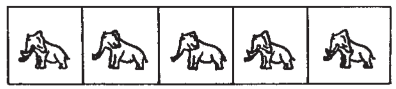
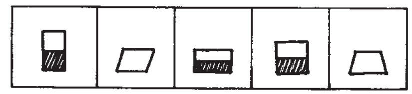
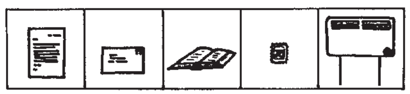
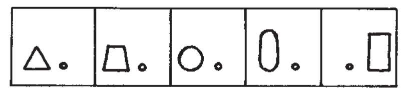
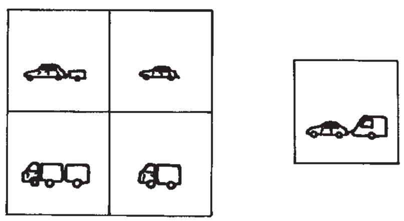
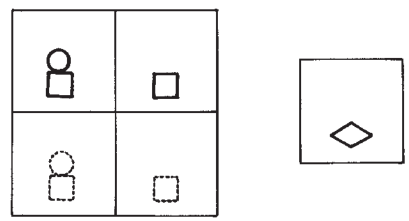
	Concrete Item (real-life objects)	Abstract Item (geometric objects)
Similarities of attributes: (generalization) Make a group (one attribute)		
Dissimilarities of attributes (discrimination) What does not belong to the group? (one attribute)		
(Dis)similarities of attributes (cross-classification) What makes a group? (two attributes)		

Figure 1b
Review of the TIR Item Types: Relation Items

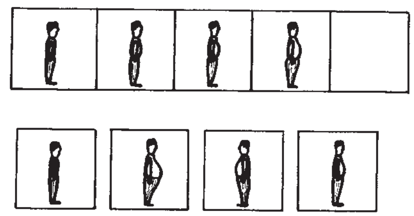
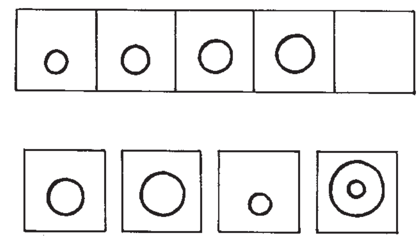
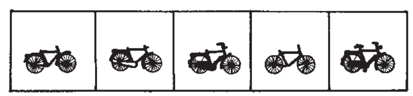
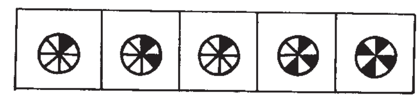
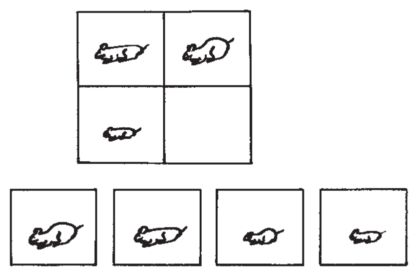
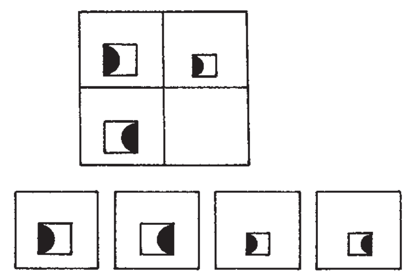
	Concrete Item (real-life objects)	Abstract Item (geometric objects)
Similarities of relations: (seriation) Make a row (one relation)		
Dissimilarities of relations (disturbed seriation) What is wrong in the row? (one relation)		
(Dis)similarities of relations (system construction) Make two rows. (two relations)		

Table 1
Number of Items in TIR-I and TIR-II

	Number of Items					
	Unique TIR-I	Unique TIR-II	Shared TIR-I+II	Total per TIR	Concrete per TIR	Abstract per TIR
Attributes						
Generalization	6	6	3	9	5	4
Discrimination	5	5	2	7	4	3
Cross-classification	3	3	3	6	3	3
Relations						
Seriation	3	3	3	6	3	3
Disturbed seriation	5	5	3	8	4	4
System construction	5	5	2	7	3	4
Total	27	27	16	43	22	21

children for the TIR-II. The tests were administered as group tests in January and June of the same school year, respectively.

Results

Nonparametric IRT Modeling

The TIR as One Test

The monotone homogeneity model. Neither of the TIR item sets had negative item H_j values, but values were low: For the TIR-I, $0.06 \leq H_j \leq 0.33$, and for the TIR-II, $0.08 \leq H_j \leq 0.38$. Furthermore, for the TIR-I, the overall $H = 0.19$ with a percentage of negative H_{jk} values of 6.6%, and for the TIR-II, the overall $H = 0.22$ with a percentage of negative H_{jk} values of 3.4%. Because negative H_{jk}s are in conflict with the MHM (Mokken, 1971, p. 150) and because H_js and Hs lower than 0.3 indicate weak item discrimination, the IRFs of the TIR tests were investigated in greater detail.

For the TIR-I, 14 items had H_j coefficients of 0.15 or lower, and for the TIR-II, this number was 8. Under a fitting MHM, such low values indicate nearly flat but increasing IRFs, and under a nonfitting MHM, such values indicate IRFs that may not be monotonely nondecreasing. Only one item from the TIR-I and none of the items from the TIR-II had item-restscore regressions that violated the monotonicity assumption (MSP combined adjacent restscore groups with scores $R = r, r + 1$, and so on, until each group had at least 20 respondents user-defined; also, testing was done at a nominal Type I error rate of 0.01). Thus, from the combination of nondecreasingness of the IRFs and the low H_js, the authors conclude that in both tests the IRFs are relatively flat curves and, therefore, that most items have weak discrimination power.

The double monotonicity model. For the TIR-I, the authors found $H^T = 0.31$ and a percentage of negative H_a^T values of 0.4, and for the TIR-II, they found $H^T = 0.31$ and a percentage of negative H_a^T values of 0.6. Thus, for practical purposes, the 43 IRFs of each test can be considered to be nonintersecting, and H^Ts close to 0.3 suggested that IRFs are close together when incorrectly ordered (Sijtsma & Meijer, 1992).

An additional investigation of the nonintersection of the IRFs compared for each pair of items the item-restscore regressions. MSP combined adjacent restscore groups, $S = s, s + 1$, and so on, until each group contained at least 20 respondents user-defined. An unexpected ordering (given the ordering based on proportions correct in the total group) of the two items within a restscore group was tested at a nominal Type I error rate of 0.01. MSP counted for each item the total number of reversals with each of the other 42 item-restscore regressions and also the total number of the significant reversals. For the TIR-I, 22 items had no significant reversals with any of the other items. The largest number found was 8 significant reversals (one item), the second largest number was 6 reversals (two items), and the third largest number found was 5 reversals (one item). Given the enormous number of opportunities that 43 curves have for crossing one another, these numbers can be considered low enough to ignore them (Molenaar & Sijtsma, 2000). For the TIR-II, similar results were found: 23 items had no significant reversals with any of the other items, and the largest number of significant reversals found was 6 (one item). Thus, the detailed results supported the results of the global H^T method.

Investigating the Structure of the TIR Tests

As Figures 1a and 1b show, the items can be divided into two subsets measuring either inductive reasoning using pictures of real-life objects or inductive reasoning using abstract geometric objects. The authors call these item subsets Real-Life Objects and Geometric Objects, respectively. Another subdivision comes from the distinction between Attributes of objects and Relations between objects. A finer subdivision is into six item subsets: Generalization items, Discrimination items, and Cross-Classification items (all measuring attributes), and Seriation items, Disturbed Seriation items, and System Construction items (all measuring relations). In this section, the authors take the distinction Real-Life Objects versus Geometric Objects, the distinction Attribute versus Relation, and the distinction between the six item types as the basis for investigating the fit of the MHM and the DMM, respectively.

Fitting the MHM to subscales. Table 2 shows the results of fitting the MHM to the data of the Real-Life Objects and the Geometric Objects subsets, the Attributes and Relations subsets, and each of the six item subsets measuring either attributes of objects or relations between objects.

For both TIR tests, for Real-Life Objects the H values were considerably lower than the minimally acceptable value of 0.3, but for Geometric Objects the H values were close to 0.3. In general, many item H_js were lower than 0.3. For Real-Life Objects, many sample violations of the monotonicity assumption were found, but none was significant at the 1% level (TIR-I and TIR-II). For Geometric Objects, no significant violations were found (TIR-I and TIR-II). For both item subsets and for both the TIR-I and TIR-II data, the authors concluded on the basis of the scalability results and the monotonicity results that the IRFs of the items are increasing with relatively flat slopes.

Table 2 shows for both TIR tests that the subscale Relations had better scalability than the subscale Attributes. However, for Relations, H was only 0.3 and several item H_js were lower than 0.3. For Attributes, several sample violations of the monotonicity assumption were found, but none was significant at the 1% significance level (TIR-I and TIR-II). For Relations, several but not many significant violations were found at the 5% level (TIR-I; not reported in Table 2), but none of the items stuck out in terms of the number of significant results. At the 1% level (reported in Table 2), only one violation was significant, suggesting that there were no serious violations of monotonicity. For the TIR-II at the 5% level, a few significant violations were found, and at the 1% level none. For both item subsets and for both the TIR-I and TIR-II data, the monotonicity results and the scalability results together led to the conclusion that the IRFs are increasing with relatively flat slopes.

For both TIR tests, the three subtests measuring attributes of objects, which together constituted the Attributes subset, had Hs of 0.2 and item H_js of which several were below 0.3. Again no significant decreases in the item-restscore regressions for estimating the IRFs were found. For the

Table 2
TIR-I and TIR-II Fit Results for the MHM and the DMM, Including Scalability Coefficients and Count of Number of Significant Violations (# Sign Viol) of Monotonicity (1% significance level)

		TIR-I					TIR-II				
	J	H	H_j; min, max	# Sign Viol	H^T	% Neg H_a^T	H	H_j; min, max	# Sign Viol	H^T	% Neg H_a^T
Real-Life Objects	22	.14	.07–.25	–	.29	1.5	.18	.08–.41	–	.22	3.6
Geometric Objects	21	.27	.14–.42	–	.37	2.7	.31	.11–.41	–	.46	2.2
Attributes	22	.16	.07–.29	–	.27	4.2	.17	.10–.33	–	.34	2.2
Relations	21	.29	.11–.41	1	.33	2.7	.34	.20–.42	–	.30	5.2
Attributes											
Generalization	9	.20	.09–.30	–	.30	8.5	.25	.18–.47	–	.28	14.6
Discrimination	7	.23	.19–.48	–	.39	4.5	.19	.14–.52	–	.49	2.1
Cross-Classification	6	.20	.12–.29	–	.29	20.8	.21	.15–.30	–	.38	12.5
Relations											
Seriation	6	.31	.25–.37	–	.29	20.6	.40	.30–.47	–	.11[a]	21.8
Disturbed Seriation	8	.34	.20–.46	–	.46	5.6	.33	.25–.45	–	.46	7.0
System Construction	7	.47	.28–.55	–	.18	23.5	.47	.44–.50	–	.04[a]	36.1

[a]Low H^Ts probably due to large numbers (237 and 155, respectively) of respondents whose data could not be used (only 0 or 1 scores; leads to division by 0 when calculating H^T).

three subsets measuring relations between objects, which together constituted the Relations subset, the Hs ranged from 0.3 to 0.5. Only a few items had H_js lower than 0.3, and for two subsets all H_js were higher than 0.3. The three subsets each showed sample violations of the monotonicity assumption, but none was significant, and the IRFs thus seem to be increasing indicators of latent traits as measured by each subtest.

Fitting the DMM to subscales. For each subset of items from the TIR-I and TIR-II, the authors calculated the H^T coefficient and the percentage of negative H_a^Ts. Table 2 shows that for Real-Life Objects from both TIR tests the H^Ts were too low (although close to 0.3 for the TIR-I) and that for Geometric Objects from both TIR tests the requirements with respect to H^T and the percentage of negative H_a^Ts were satisfied. These results suggest that for Geometric Objects, the IRFs do not intersect and that for Real-Life Objects, IRFs have several intersections.

Table 2 shows that for both the TIR-I and the TIR-II tests and for Attributes and Relations, H^T was near 0.30. The percentages of negative H_a^Ts were sufficiently small. Thus, for both subsets one may conclude that these results represent borderline cases when intersection of the IRFs is concerned.

For the six subsets based on attributes of objects and relations between objects, for both TIR data sets, the H^T results were not very consistent, but in general there was much evidence of intersection of IRFs. Only for the Discrimination subset (attributes of objects) and the Disturbed Seriation subset (relations between objects) were the H^T results pointing in the same direction showing evidence of nonintersection of the IRFs.

Searching for subscales under the MHM. The authors used the automated item selection procedure from MSP because this might in an exploratory way lead to new insights into the dimensionality of the datasets. Only the TIR-I data were analyzed because, based on the results found thus far, it was expected that there would not be great differences with the TIR-II data. Following Hemker et al. (1995), the authors tried several values for lowerbound c: $c = 0.0$, 0.3, and 0.4, and monitored the subdivision of the itemset into subsets. Table 3 shows that for $c = 0.0$, 19 of the 21 Relations items were selected into the first subscale along with 9 Attributes items (from each of the three a priori Attributes subscales, 3 items were selected). Also, four other subscales were selected, but none had a clear interpretation. For $c = 0.3$, the first subscale selected had 14 of the 21 Relations items. The other five subscales had small numbers of items and contained items of one or two of the a priori distinguished Attributes subscales. For $c = 0.4$, the set of 43 items was selected into eight small scales, most of which appeared to have no sensible interpretation and, moreover, 15 items remained unscalable.

It was concluded that the Relations items are the best scalable items. The subdivision into six a priori subsets was not found when using the exploratory item selection procedure. Also, the subdivision into Real-Life Objects items and Geometric Objects items did not come out as two clearly different dimensions.

Parametric IRT Modeling

The Rasch model. First, the RM was fitted to the complete set of 43 items of each TIR version. As could be anticipated on the basis of the MHM analyses, the RM did not fit the data for both test versions. For the TIR-I, $R_1 = 476$, df = 168, and p = .00, which rejects the null hypothesis of 43 logistic IRFs with equal slopes, and $Q_2 = 5730$, df = 4300, and p = .00, which rejects the null hypothesis of unidimensionality and local independence. For the TIR-II, the same conclusions were drawn, based on $R_1 = 475$, df = 168, and p = .00, and $Q_2 = 40{,}682$, df = 4300, and p = .00. Because of the heterogeneity of the tests and the clear-cut global test results, no local U_j tests were performed.

Table 3
Items From the TIR-I Using Automatic Item Selection Procedure
Results When Several Lowerbounds c Are Used for Selecting

c	Scale	J	H	Subset: # Items (# Concrete, # Abstract)	
0.0	1	28	.26	Generalization:	3 (3, –)
				Discrimination:	3 (2, 1)
				Cross-Classification:	3 (–, 3)
				Seriation:	6 (3, 3)
				Disturbed Seriation:	6 (3, 3)
				System Construction:	7 (3, 4)
	4 other scales, unclear interpretation; No items left				
0.3	1	14	.41	Seriation:	3 (–, 3)
				Disturbed Seriation:	5 (2, 3)
				System Construction:	6 (2, 4)
	2	4	.43	Discrimination:	4 (3, 1)
	3	5	.36	Generalization:	4 (1, 3)
				Disturbed Seriation:	1 (–, 1)
	4	5	.37	Discrimination:	2 (–, 2)
				Cross-Classification:	3 (–, 3)
	5	3	.34	Generalization:	3 (3, –)
	6	3	.41	Seriation:	2 (2, –)
	9 items unscalable with $c = 0.3$				
0.4	1	9	.51	Seriation:	3 (–, 3)
				System Construction:	6 (2, 4)
	2	3	.62	Discrimination:	3 (2, 1)
	3	4	.51	Disturbed Seriation:	4 (1, 3)
	4	3	.48	Generalization:	2 (–, 2)
				Disturbed Seriation:	1 (–, 1)
	5	2	.56	Generalization:	2 (–, 2)
	6	3	.43	Discrimination:	1 (–, 1)
				Cross-Classification:	2 (–, 2)
	7	2	.44	Generalization:	2 (2, –)
	8	2	.41	Seriation:	2 (2, –)
	15 items unscalable with $c = 0.4$				

Using their knowledge of the a priori subtest structure, in the next step the authors fitted the RM to subtests, exactly as for the MHM and DMM analyses. In addition to global statistical testing using R_1 and R_2 / Q_2, the authors used local testing by means of the approximate standard normal statistic U_j (Molenaar, 1983). Because J standard normal U_j tests were performed, they tested two-sidedly at a 0.2% significance level; thus, $| U_j | \geq 3.08$ led to the rejection of the null hypothesis.

Table 4 shows that for the TIR-I, all R_1 and R_2 / Q_2 test results led to the rejection of the RM assumptions at a 1% significance level (the highest probability of exceedance was .0049 for Cross-Classification). Since the U_j values were almost always between the critical values of –3.08 and 3.08 (Table 4 only gives the two extreme U_j values), these results gave us almost no clues of how to improve the subscales by removing items with either too flat or too steep IRFs. The apparent contradiction between R_1 results and U_j results may suggest that the overall R_1 test may have been too sensitive due to accumulating nonsignificant deviations between observed and expected IRFs across the J items from a test. Also, the U_j results supported the conclusion based on the MHM and DMM analyses that the IRFs are increasing functions with often only few intersections. Moreover, the U_j results provided evidence that these curves can be well approximated by logistic functions.

Table 4
TIR-I and TIR-II Fit Results for the RM. Entries Below R_1 and R_2/Q_2 Must Be Multiplied by 0.0001 to Obtain Probabilities of Exceedance

		TIR-I			TIR-II		
	J	R_1	U_j; min, max	R_2/Q_2[1]	R_1	U_j; min, max	R_2/Q_2[1]
Real-Life Objects	22	3	–1.6; 1.7	–	–	–2.2; 2.5	–
Geometric Objects	21	–	–2.1; 2.9	–	–	–1.5; 2.7	–
Attributes	22	9	–1.6; 1.9	–	1410	–1.4; .8	–
Relations	21	–	–2.3; 3.5	–	–	–1.7; 1.9	–
Attributes							
Generalization	9	26	–1.2; 1.0	–	1285	–1.2; 1.9	84
Discrimination	7	2	–.7; .8	–	9617	–.4; .7	–
Cross-Classification	6	–	–1.9; 2.3	49	1082	–.9; 1.1	92
Relations							
Seriation	6	–	–1.5; 2.2	–	3	–1.4; 2.0	11
Disturbed Seriation	8	–	–2.8; 3.8	–	1	–1.4; 1.5	–
System Construction	7	–	–3.0; 5.2	–	3	–1.7; 2.1	–

[1]For the first two subdivisions, Q_2 was calculated; for the last subdivision into six subtests, R_2 was calculated.

Similar results were found for the TIR-II data, but with the exception of nonsignificant R_1 test results for the three attribute subscales (Generalization, Discrimination, and Cross-Classification) and the total Attribute subscale comprising these three subscales. In combination with the non-significant U_j results, it was concluded that the assumptions of monotonicity and sufficiency were valid here. However, Table 4 shows that the U_j results for the other subscales also were all within the critical region and that, based on this, there was little evidence for rejecting the null hypothesis of monotonicity and sufficiency.

Finally, with a few exceptions the R_2 / Q_2 test results indicated convincing rejections of the null hypothesis of unidimensionality and local independence. This result corroborates the item selection results for the MHM as reported in Table 3, where the authors did not find a clear-cut selection of the items into subsets that ran neatly along the lines of the a priori subdivision followed in Tables 2 and 4, but which indicated multidimensionality that was difficult to interpret.

The OPLM. Table 5 gives the Type I error probability for R_{1c}-RM (slopes of 1, which is the RM) and R_{1c}-A_j (user-specified slope indices A_j). R_1 (Table 4) and R_{1c}-RM (Table 5) are the same statistic, but Tables 4 and 5 give different Type I error probabilities due to somewhat different groupings of restscore R used by RSP and OPLM for calculating the statistics. Table 5 also shows the slope indices, A_j, which were suggested by OPLM on the basis of the misfit of the RM (slopes of 1 for each IRF; for more details, see Verhelst, 1992).

In most cases, R_{1c} could be improved substantially (in a few cases, due to computational problems, parameter estimates could not be obtained). For the Attribute subscale and the three attribute subscales of the TIR-II, for which monotonicity and sufficiency were valid under the RM, slopes of 1 were accepted as final. In almost all other cases, adaptation of the slope indices led to high Type I error probabilities. These probabilities suggested that the choice of the slope indices might have capitalized on chance. The MHM results reported earlier suggested that the IRFs were increasing with rather flat slopes (Table 2), but the OPLM analyses suggest that the slopes of different items show some variation. The interpretation of these A_js and their variation is relative, however, in the sense that replacing each string of A_js in Table 5 with another string of positive integers that is a multiple of the original string would have yielded the same R_{1c}s.

Table 5
TIR-I and TIR-II Fit Results for the OPLM: R_{1c} Probabilities of Exceedance (After Multiplication by 0.0001) for Rasch Model (RM) and OPLM With Imputed Item Slopes (A_j)

		TIR-I			TIR-II		
	J	R_{1c}-RM	A_j	R_{1c}-A_j	R_{1c}-RM	A_j	R_{1c}-A_j
Real-Life Objects	22	–	2 5 2 3 3 3 3 3 3 2 2 2 4 2 4 4 4 3 4 3 3 4	3581	–	3 3 2 2 3 2 4 2 2 3 2 2 3 4 3 3 4 4 4 4 5 3	3867
Geometric Objects	21	–	2 2 2 2 2 2 3 3 3 3 3 4 3 3 3 4 1 5 4 5 4	–	–	1 2 3 2 3 2 3 4 5 3 6 6 5 5 4 5 2 6 6 6 5	955
Attributes	22	21	2 4 2 3 3 3 3 3 2 3 3 3 3 4 3 3 2 2 2 5 5 4	7293	2815	No adaptations	
Relations	21	–	3 2 3 4 4 4 3 3 1 3 3 3 4 1 2 3 3 5 5 5 4	27	–	2 3 2 4 3 3 2 3 3 3 3 2 5 1 4 3 3 5 5 4 3	8909
Attributes							
Generalization	9	31	2 4 3 2 3 5 4 2 2	9176	6316	No adaptations	
Discrimination	7	12	3 2 3 3 3 4 4	3470	8701	No adaptations	
Cross-Classification	6	–	2 2 2 5 4 4	No est.	1935	No adaptations	
Relations							
Seriation	6	4	3 1 3 5 4 4	3820	1	2 3 2 4 4 4	No est.
Disturbed Seriation	8	–	3 3 2 4 3 4 6 1	5078	5	2 3 3 3 3 2 6 2	5934
System Construction	7	–	1 2 3 5 5 4 4	2787	–	3 3 2 4 4 4 2	2057

Note. est. = estimate.

Discussion

The authors' advice for researchers is to use several models for analyzing their data and not just one model. They used four different IRT models that have different measurement properties and different methods for data analysis. These models are like different glasses that one can wear to look at the same phenomenon, one's item response data, and that each offer a somewhat different perspective. The four models used here could be supplemented or even replaced by others, such as multidimensional IRT models or classical methods such as factor analysis or cluster analysis. This depends on the goal of research but also on personal preferences. The basic attitude advocated here is to use multiple methods. In general, the authors advise to start an item analysis with the most liberal models, here the nonparametric MHM and DMM, and then to continue with the more restrictive parametric models, here the RM and the OPLM. A fitting MHM implies an ordinal scale for persons and a fitting DMM in addition implies an ordinal scale for items. The next step is to fit the more restrictive parametric models, which give more profound information about scale and item properties and enable advanced applications such as equating (de Koning et al., in press) and adaptive testing (Hambleton & Swaminathan, 1985).

For this particular study, the simultaneous use of two nonparametric and two parametric IRT models suggests the following conclusions. First, there are differences in the kinds of information given by several statistics and these differences can be used next to each other so that more information can be obtained than would be possible if only one model were used for data analysis. Within the nonparametric IRT context, a higher H coefficient means that more confidence can be held in the ordering of the respondents on θ (Mokken et al., 1986) and a higher H^T means that more confidence can be held in the nonintersection of the J IRFs (Sijtsma & Meijer, 1992). Under the RM, nonsignificant R_1 / Q_1 and R_2 / Q_2 values indicate that the J items have logistic IRFs with the same slopes and that the J items are unidimensional and locally independent, respectively. Of course, the test information function (e.g., Hambleton & Swaminathan, 1985; Van der Linden & Hambleton, 1997), not investigated here, gives local information on the accuracy of person measurement, but H is a convenient and quick global measure of person ordering. Moreover, H^T gives a quick impression on nonintersection of the J IRFs when one is not particularly interested in the logistic shape of these functions or when a nonparametric IRT analysis is done as a precursor for a possible Rasch analysis.

Also at the local item level, statistics from different models give complementary information about fit or misfit. For example, the H_j coefficients give an indication of the discrimination power and relate this information to the total score variance. Under the MHM, a low H_j thus indicates a rather flat IRF and a high H_j a rather steep IRF relative to the group under study. Under the RM, the U_j statistic tells one whether the observed IRF matches the logistic IRF and a high negative value indicates an IRF that is steeper than expected, whereas a high positive value indicates a flatter IRF. Whether the discrimination power of the item is low, intermediate, or high relative to the group under study cannot be derived from such results, however, because under the RM slope parameters are set to 1 just to indicate equality, but any other positive constant would express the same equality property. Thus, under the RM a slope parameter of 1 does not convey information about the discrimination power of an item. Under the OPLM, the A_js indicate the slopes of the logistic IRFs but, as has been seen, these slope indices only have meaning relative to one another. Any set of alternative integer slope indices that is a multiple of the A_js leads to exactly the same fit statistics. To summarize, the H_js give information whether slopes are positive or negative and whether they are flat or steep; the U_js indicate whether IRFs are logistic and flatter or steeper than the "unity" slopes of the RM, and the A_js indicate the relative slopes of different logistic IRFs.

Second, unlike MSP (automated item selection procedure) and RSP (R_2 / Q_2 significance tests), OPLM does not contain methods that allow for a direct evaluation of unidimensionality and local independence. It may be noted that programs for nonparametric IRT analysis such as DIMTEST (Stout, 1990; Stout et al., 1996) and DETECT (Kim, Zhang, & Stout, 1996) are focused on dimensionality analysis by using statistics based on conditional covariances between items and thus could be used supplementary to MSP (or MSP supplementary to DIMTEST and DETECT; this depends largely on one's preference for either method; also see Sijtsma, 1998). Verhelst (personal communication, 1999) explained the absence of a dimensionality statistic in OPLM by the mathematical complexity of such a statistic when the IRF slopes are allowed to vary across the items. For binary data with an unknown dimensionality, varying IRF slopes as indicated by varying H_js, U_js, and A_js may be indicative of multidimensionality as well, but the availability of the R_2 / Q_2 statistics in RSP for the direct investigating of multidimensionality provides a good reason for the use of RSP in addition to the use of MSP and OPLM. Another alternative would be to use a multidimensional IRT model (e.g., Reckase, 1997).

Third, the data analyses made clear that the MHM and DMM models have several easy-to-use statistics at the global (all J items) and local (individual items and pairs of items) analysis levels. Moreover, these models provide indices of measurement quality, in particular the H and H^T coefficients, that concentrate on test and item characteristics in relation to the person distribution, thus allowing statements about the usefulness of the test or an item for measurement in a particular group. This study has illustrated that several statistics for the RM relate the item characteristics to the logistic shape of the IRF (R_1 and U_j; also several item slope statistics used in OPLM but not discussed here), but no information is contained on the strength of the relation between the item and θ. Also, the slope indices A_j from the OPLM give information on the relative slopes of logistic IRFs, but no information on the discrimination power relative to the person distribution. At the level of item analysis, nonparametric IRT thus provides excellent auxiliary information by relating item properties to the person distribution.

Finally, the overall conclusion was that the four IRT models do not fit the complete test data, but also that this misfit informed the authors well about the structure of the TIR tests. This information can be used in at least two ways. First, although some meaningful subdivisions were found, the conceptual distinction into different item types made in the relevant literature (Klauer, 1989) and incorporated in the authors' tests could not be retrieved very convincingly. However, several analyses using different IRT models suggested some form of multidimensionality as being present in the data. This would suggest that a careful conceptual re-analysis of the item types relevant to the measurement of inductive reasoning could be useful to obtain a better understanding of how the concept should be measured. Second, at the psychometric level, ignoring the subset structure altogether and taking all items from the TIR-I and those from the TIR-II as a priori unidimensional tests, it was found that items had low discrimination, meaning that individual items only weakly separated persons with low and high latent traits. This was also found for several a priori identified item subsets. These results would suggest that for measuring inductive reasoning reliably with the types of items used here (which are highly representative of how inductive reasoning is measured traditionally), long tests are needed to obtain sufficient reliability. Another study (de Koning et al., in press) addressed the equating of the two TIR scales after a few of the worst fitting items had been removed, but most items were retained for having sufficient reliability. OPLM was used for this purpose, because it estimates the metric θ parameters that are convenient for equating and because it was more flexible than the RM.

References

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1985). *Revisie Amsterdamse Kinder-Intelligentie Test (RAKIT)* [Revision of the Amsterdam Child Intelligence Test]. Lisse: Swets & Zeitlinger.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404-431.

de Koning, E., & Hamers, J. H. M. (1995). *Programma Inductief Redeneren 1* [Program Inductive Reasoning I]. Utrecht: Utrecht University Press ISOR.

de Koning, E., & Hamers, J. H. M. (1999). Teaching inductive reasoning: Theoretical background and educational implications. In J. H. M. Hamers, J. E. H. van Luit, & B. Csapó (Eds.), *Teaching and learning thinking skills* (pp. 157-188). Lisse: Swets & Zeitlinger.

de Koning, E., Sijtsma, K., & Hamers, J. H. M. (in press). Construction and validation of a test for inductive reasoning. *European Journal of Psychological Assessment.*

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

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

Glas, C. A. W., & Ellis, J. L. (1993). *User's manual RSP: Rasch Scaling Program.* Groningen, The Netherlands: iecProGAMMA.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69-95). New York: Springer-Verlag.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (1995). Selection of unidimensional scales from a multidimensional itembank in the polytomous Mokken IRT model. *Applied Psychological Measurement, 19*, 337-352.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331-347.

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

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

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65-81.

Kim, H. R., Zhang, J., & Stout, W. (1996). *A new index of dimensionality—DETECT.* Internal Report, Department of Statistics, University of Illinois at Urbana-Champaign.

Klauer, K. J. (1989). *Denktraining für Kinder 1. Ein Program zur intellectuellen Förderung* [Inductive reasoning. A programme for the stimulation of inductive reasoning]. Göttingen: Hogrefe Verlag.

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

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*, 283-298.

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

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6*, 417-430.

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to "The Mokken scale: A critical discussion." *Applied Psychological Measurement, 10*, 279-285.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48*, 49-72.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows. User's manual MSP.* Groningen, The Netherlands: iecProGAMMA.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer-Verlag.

Robin, F., Xing, D., & Hambleton, R. K. (1999). Software review: Rasch scaling program (RSP). *Applied Psychological Measurement, 23*, 90-94.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3-31.

Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79-105.

Sijtsma, K., & Junker, B. W. (1997). Invariant item ordering of transitive reasoning tasks. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 100-110). Münster, Germany: Waxmann Verlag.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

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

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

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

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191-199.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

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

Verhelst, N. D. (1992). *Het eenparameter logistisch model (OPLM)* [The one parameter logistic model (OPLM)] (OPD Memorandum 92-3). Arnhem, The Netherlands: Cito.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215-237). New York: Springer.

Author's Address

Send requests for reprints or further information to Klaas Sijtsma, Tilburg University, Department of Methodology and Statistics, FSW, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: k.sijtsma@kub.nl.