

## Chapter 8

### Invariant Item Ordering of Transitive Reasoning Tasks

*Klaas Sijtsma and Brian W. Junker*

Utrecht University, Utrecht, The Netherlands, and  
Carnegie Mellon University, Pittsburgh PA, USA

#### 1. Introduction

Transitive inference is a type of reasoning in which an order relation between two objects A and C is inferred from the observed relation between objects within at least two other pairs of objects, one of which contains A and the other C. For example, consider three objects (e.g., sticks, denoted A, B, and C) and the characteristic length ( $X$ ), and assume that A is the shortest and C the longest:  $X_A < X_B < X_C$ . A transitive inference has been drawn if the observations  $X_A < X_B$  and  $X_B < X_C$ , which provide the premise information, lead to the logical conclusion through abstract reasoning that  $X_A < X_C$ . Of course, the number of objects, the amount of available information, and the number of inferences can each be larger. Other examples of characteristics are weight, surface and temperature. Physical objects might be sticks, balls, discs, tubes, cubes, etc.

Research of transitive reasoning has centered around several competing views that have been discussed by Verweij (1994) and Verweij, Sijtsma, and Koops (1996). First, the mode of presentation of the premise information differs across studies. For example, the premises can be presented successively in a random order in such a way that with each premise the others are kept out of sight. Alternatively, the premises can be presented simultaneously in an ordered descending or ascending series, with each of the premises visible in the background while a particular premise is presented. These and other presentation modes have been criticized; see Verweij (1994) for a thorough discussion, and empirical results that favor the successive presentation.

Second, the type of response used to assess transitive reasoning may differ across studies. Responses may be categorized as right or wrong (judgement-only criterion). However, a correct answer could result from a nontransitive (e.g., visual, spatial) process or from guessing. To identify transitive processes, verbal justifications of correct answers are recorded as right or wrong (judgement-plus criterion). This scoring rule has been criticized because it might lead to assessments that reflect a confounding of task performance and verbal ability. Alternatively, chronometric analysis and information processing approaches have been suggested.

Verweij (1994) has constructed measures of the ability of transitive reasoning using item response theory (IRT) modeling. His data resulted from a successive presentation mode of the premises and a scoring procedure of the items that combined the judgement-only and the judgement-plus criterion. The data were analyzed using the Mokken (1971) models, the Rasch (1960) model, and a generalized partial credit model (Verhelst, 1992).

Interpretation of transitive reasoning test results is made easier if an item ordering by difficulty is the same for all individuals taking the test. This is an *invariant item ordering* (Sijtsma & Junker, 1996). We would know, for example, that the same items are difficult in comparison with other items for each child taking the test. Comparison of the performance of children of different ability levels would then be easier because the items answered correctly by the lower ability child would be a subset of the items answered correctly by the other child. Hypotheses about the development of transitive reasoning could be operationalized by means of a set of tasks where the difficulty level corresponds to a developmental stage, and tested by means of an IRT model that assumes an invariant item ordering.

Sijtsma & Verweij (1992) reported that subsets of their transitive reasoning tasks could be invariantly ordered by difficulty both under the judgement-only and the judgement-plus criterion. The goal of this study is to reanalyze these data sets using several methods from nonparametric IRT to investigate the property of an invariant item ordering. The results from these methods are compared and combined into a conclusion about the invariant ordering property of the items.

## 2. Method

The test contained 10 transitivity tasks (Figure 1 shows the tasks in the order of presentation). These tasks were all derived from the literature on transitive reasoning. The tasks pertained to length, weight, and area, and dealt with inequalities, equalities, and mixtures of inequalities and equalities. The subjects were 425 Dutch pupils from 10 primary schools (Grades 2 through 6). Grades were approximately equally represented as were boys and girls in each grade.

Because we investigate an invariant item *ordering*, a nonparametric IRT approach (e.g., Mokken, 1971; Stout, 1987; Junker, 1993) is appropriate (see also chapter 1.6). We therefore assume only unidimensionality, local independence of the item responses given the latent ability  $\theta$ , and a set of item response functions (IRF's) that are nondecreasing and nonintersecting. Nonintersection of IRF's means that  $k$  items can be numbered and ordered such that for the conditional probabilities of obtaining a correct response

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

This is equivalent to an invariant item ordering. First, we consider the investigation of pairs of IRF's and next, of sets of  $k$  IRF's.

*Pairs of IRF's.* Let  $\mathbf{Y}$  be the vector of  $k - 2$  dichotomous item score variables excluding item scores  $X_i$  and  $X_j$ , and let  $h(\mathbf{Y})$  be any function of  $\mathbf{Y}$ . Rosenbaum (1987a) showed that if the IRF's of the items  $i$  and  $j$  ( $i < j$ ) do not intersect, then for any  $h(\cdot)$

$$p[X_i = 1|h(\mathbf{Y}) = c] \leq p[X_j = 1|h(\mathbf{Y}) = c]. \quad (2)$$

The ordering of the items thus is the same in each subgroup defined by a particular value  $c$  of the function  $h(\cdot)$ . An obvious choice for  $h(\cdot)$  is the total score  $S$  on the  $k - 2$  items in  $\mathbf{Y}$ . For unidimensional, locally independent IRT models with nondecreasing IRF's, Grayson (1988) showed that scores like the total "rest score"  $S$  stochastically order  $\theta$ . This implies that with increasing  $S$ ,

$$p[X_i = 1|S = s] \leq p[X_j = 1|S = s] \quad (3)$$

provides us with information about intersections of IRF's at higher regions of the  $\theta$ -scale.

task	concept area	objects	measures	color	format
6	length	3 sticks diameter 11 mm.	12 cm. 11.5 cm. 11 cm.	red (A) green (B) white (C)	$X_A > X_B > X_C$
2	length	4 tubes  diameter 12 mm.	12 cm.	red (A) blue (B) yellow (C) green (D)	$X_A = X_B = X_C = X_D$
8	weight	3 tubes diameter 19 mm.	45 gr. 25 gr. 18 gr.	yellow (A) red (B) blue (C)	$X_A > X_B > X_C$
4	weight	4 cubes  size 5×5×5 cm.	65 gr.	green (A) red (B) yellow (C) white (D)	$X_A = X_B = X_C = X_D$
9	weight	3 balls diameter 3.5 cm.	40 gr. 50 gr. 70 gr.	red (A) blue (B) yellow (C)	$X_A < X_B < X_C$
3	area	3 disks thickness 6 mm.	7.5 cm. 7.0 cm. 6.5 cm.	yellow (A) red (B) green (C)	$X_A > X_B > X_C$
5	length	3 sticks diameter 11 mm.	28.5 cm. 27.5 cm. 27.5 cm.	green (A) red (B) white (C)	$X_A > X_B = X_C$
1	weight	3 balls diameter 3.5 cm.	65 gr. 40 gr. 40 gr.	green (A) yellow (B) red (C)	$X_A > X_B = X_C$
10	length	4 sticks  diameter 11 mm.	12.5 cm. 12.5 cm. 13.0 cm. 13.0 cm.	green (A) yellow (B) red (C) white (D)	$X_A = X_B < X_C = X_D$
11	weight	4 balls  diameter 4 cm.	60 gr. 60 gr. 100 gr. 100 gr.	white (A) green (B) blue (C) red (D)	$X_A = X_B < X_C = X_D$

Note: The original items #7 and #12 did not measure transitive reasoning and have been removed from the data set.

**Figure 1:** Description of the ten transitivity tasks in the order of presentation

Because  $S$  is based on  $k - 2$  items, the division of the latent  $\theta$ -scale into consecutive but partly overlapping regions is relatively reliable. A larger  $k$  yields a more reliable division, but smaller rest score groups and thus smaller power to detect intersections of IRF's. A remedy against small power would be to merge adjacent rest score groups until a satisfying group size is obtained.

Mokken (1971, pp. 132 - 133) proposed a related method based on two square matrices with joint proportions. The  $k \times k$   $P(++)$  matrix contains the joint proportions of subjects that have item  $i$  and item  $j$  correct (all  $i, j; i \neq j$ ); likewise for two incorrect responses in the  $k \times k$   $P(--)$  matrix. Given the ordering of rows and columns according to increasing univariate

proportions correct along the marginals, and given an invariant item ordering, P(++), P(+-), P(-+), and P(--), has nondecreasing rows and columns, and P(--), P(-+), P(+-), and P(++), has nonincreasing rows and columns. It can be shown (e.g., Sijtsma & Junker, 1996) that for pairs of items this method is equivalent to an inspection of inequalities where the conditioning is on  $h(\mathbf{Y}) = \mathbf{X}_f$  ( $f \neq i, j$ ;  $\mathbf{X}_f = 0, 1$ ):

$$p[X_i = 1 | X_f = x_f] \leq p[X_j = 1 | X_f = x_f]. \quad (4)$$

Using one item  $f$ , the latent scale is divided into two partly overlapping regions. The division is relatively unreliable because only one item is used, but the power to detect intersections is relatively high because only two groups are formed. Like the "rest score" approach, this "item splitting" approach scans the whole  $\theta$  scale for intersections provided that the remaining  $k - 2$  items  $f$  ( $f \neq i, j$ ) have a considerable variation in difficulty.

Sijtsma and Junker (1996) proposed a combination of the former two methods by combining their virtues: a reliable division of the  $\theta$ -scale (rest score) and a large power to detect intersections (item score). Subgroups are created using a cut score  $0 \leq s \leq k-2$ , such that a person is assigned to the low group if  $S \leq s$ , and to the high group otherwise. This means that for varying  $s$  "rest score splitting" inequalities should be considered like

$$p[X_i = 1 | S \leq s] \leq p[X_j = 1 | S \leq s]; \text{ and} \\ p[X_i = 1 | S > s] \leq p[X_j = 1 | S > s]. \quad (5)$$

*Sets of k IRF's.* Rosenbaum (1987b) proposed the inspection of orderings of joint conditional proportions of score patterns on  $k$  items. If  $k$  item score variables are divided into two disjoint and exhaustive subsets  $\mathbf{X}_1$  and  $\mathbf{X}_2$  with  $k_1$  and  $k_2$  items, respectively, and the  $k_1$  IRF's in  $\mathbf{X}_1$  are nonintersecting, then  $P[\mathbf{X}_1 = \mathbf{x}_1 | h(\mathbf{X}_2) = c]$  is increasing in transposition for every function  $h(\mathbf{X}_2)$ . Similar results were demonstrated by Sijtsma and Junker (1996) for  $n$ -variate ( $n \leq k$ ) joint proportions that pertain only to scores 1 or 0. Two problems interfere with a straightforward use of ordering properties of such joint proportions to investigate nonintersection of  $k$  IRF's: (1) the number of orderings increases rapidly with  $k$ ; and (2) the sample frequencies needed to estimate joint proportions will often be very small or 0.

A method that is easy to apply is based on a scalability coefficient that was used by Mokken (1971, pp. 148 - 153) to investigate the degree to which persons can be ordered by items. For the ordering of *items by persons* the coefficient is based on the covariances between the scores of two persons,  $a$  and  $b$ , on  $k$  items ( $\text{Cov}_{ab}$ ) and the maximum covariance given their expected proportions of correct answers on the  $k$  items of the test ( $\text{Cov}_{ab}^{\max}$ ). The scalability coefficient is defined as

$$H^T = \frac{\sum_a \sum_{b \neq a} \text{Cov}_{ab}}{\sum_a \sum_{b \neq a} \text{Cov}_{ab}^{\max}}. \quad (6)$$

In addition, coefficients  $H_a^T$ , that relate one individual to the other  $N - 1$  individuals taking the test can be defined. Sijtsma and Meijer (1992) showed that for  $k$  items having nonintersecting IRF's,  $0 \leq H^T \leq 1$ , with equality to 0 if and only if the  $k$  IRF's coincide with the exception of at most one  $\theta$  value. Because nonnegative values of  $H^T$  constitute only a necessary condition for nonintersection, the following rules of thumb were proposed to have more certainty about this

property: If  $H^T \geq 0.3$ , and if the percentage of negative  $H_a^T$  values is less than 10%, we may assume for all practical purposes that the  $k$  IRF's do not intersect. The hypothesis of an invariant item ordering is rejected if one or both requirements are not satisfied.

We used the  $H^T$  methodology in combination with the rest score method, the P(++), and P(--), method, the item splitting method, and the rest score splitting method to investigate whether 10 transitive reasoning items have an invariant item ordering under the two different scoring rules. We used the program MSP (Molenaar, Debets, Sijtsma, & Hemker, 1994) for the  $H^T$ , the rest score, and the P(++), and P(--), approaches; and we used the statistical package S-PLUS (Statistical Sciences Inc., 1992) for the item splitting and the rest score splitting approaches.

### 3. Results

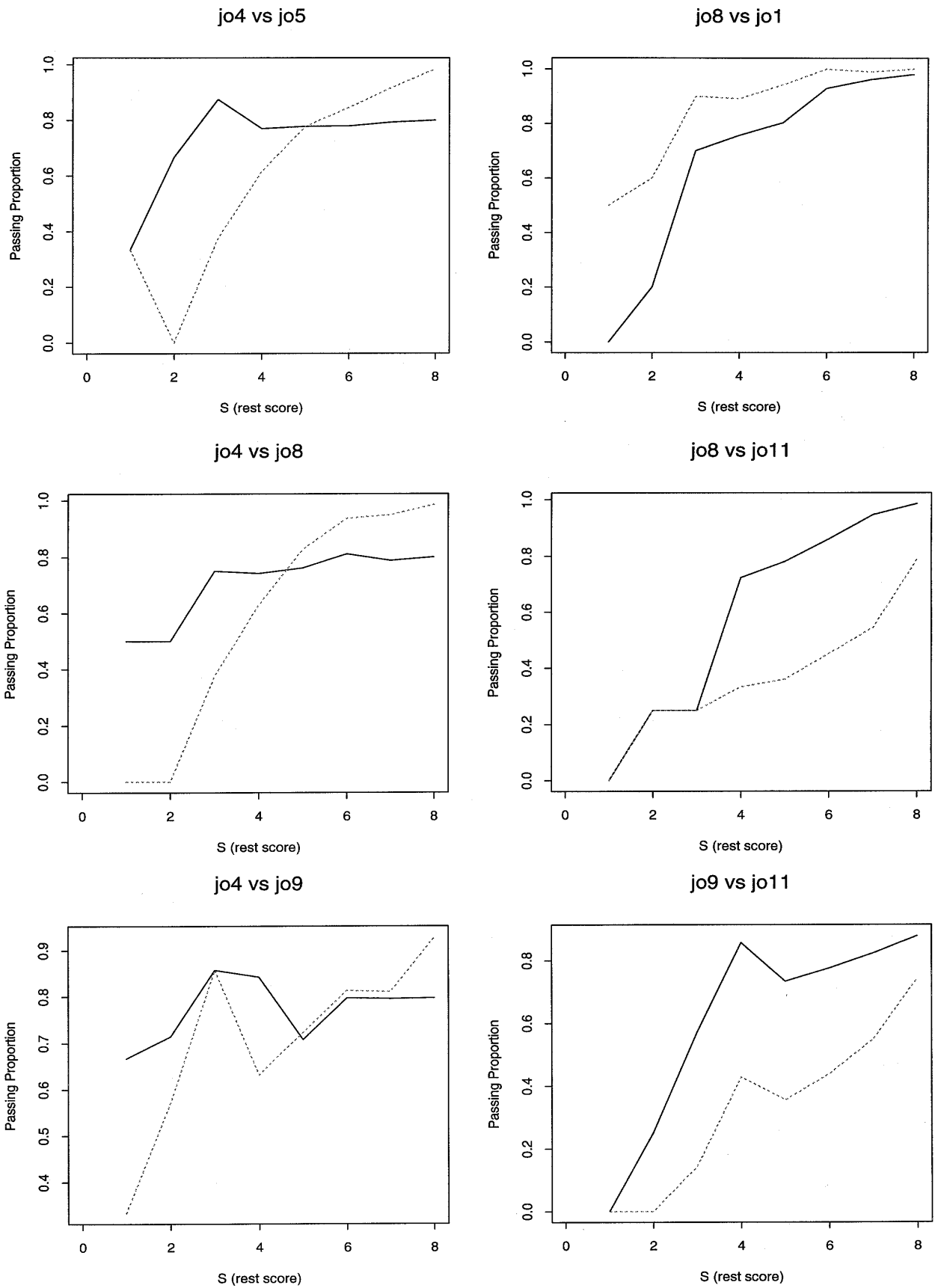
*Judgement-only criterion.* The ordering of the 10 tasks according to their p-values (proportions-correct) is given in Table 1.

rank	1	2	3	4	5	6	7	8	9	10
jo-item	jo10	jo11	jo4	jo9	jo2	jo5	jo8	jo6	jo1	jo3
p-value	.30	.52	.78	.80	.81	.84	.88	.94	.97	.97
jp-item	jp2	jp4	jp10	jp11	jp9	jp6	jp8	jp3	jp5	jp1
p-value	.03	.04	.06	.09	.18	.28	.36	.53	.62	.72

**Table 1:** Ordering of items according to p-values (proportion correct); "jo10" means: judgement-only data; item #10; "jp2" means: judgement-plus data, item #2.

The hypothesis that the 10 IRF's have no intersections should be rejected ( $H^T = 0.47$  and 12% negative  $H_a^T$  values). Information about intersection on the item level can be obtained from 4 sources: (1) rest score grouping; (2) P(++), and P(--), matrices; (3) item splitting; and (4) rest score splitting. Before presenting the results from these methods, three remarks are in order. First, all of the significance tests are based on appropriate z-statistics (approximately normally distributed) with a "significant violation" meaning that a 1.64 cutoff (one-tailed,  $\alpha = 0.05$ ) was exceeded. Second, because of the large number of significance tests, some detailed results may be due to capitalization on chance. Nevertheless, taken together, these significance tests have a real diagnostic value, similar to the value of standardized-residual analysis in linear regression. Third, the results of P(++), and P(--), analysis and item splitting should be the same, except for minor differences in derivations of the test statistics used with each method (compare Molenaar, 1970, with Sijtsma and Junker, 1996).

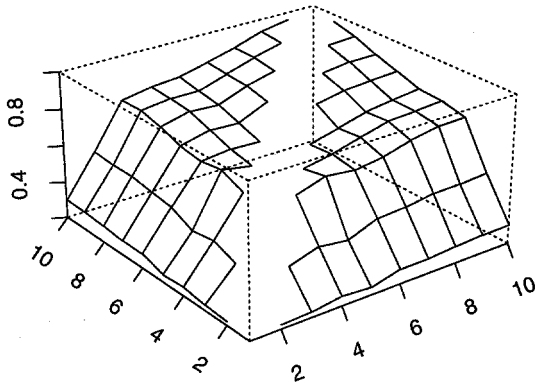
*Rest score grouping* revealed that the item-rest curve of item #4 has a significant reversal of the expected ordering in the lowest rest score group with the item-rest curve of each of the items #5, #8, and #9. These results are based on a merging of adjacent rest score groups such that the minimum group size is at least 20, and a statistical test (Molenaar, 1970; Molenaar et al., 1994) for the null-hypothesis that item-rest score proportions are equal in a particular group against the alternative that they are oppositely ordered compared with the item proportions correct. Figure 2 (left) shows graphs of the item-rest curves based on all rest score groups, thus ignoring minimum group size requirements, for the three item pairs that showed significant reversals.



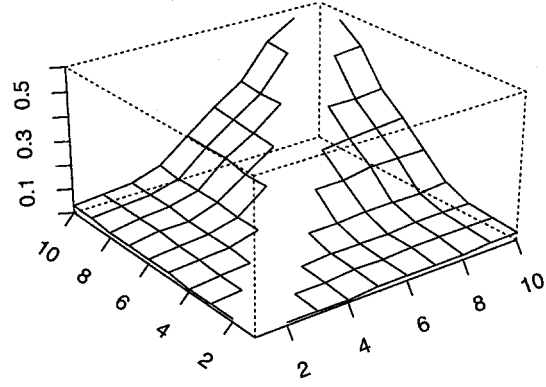
**Figure 2:** Item-rest curves with intersections (left half) and without intersections (right half). „jo4“ means: judgement-only data, item #4; and so forth.

The lower parts of the graphs are based on small numbers of observations. Note, that the curves of the pairs #4 and #5, and #4 and #9 do not monotonely increase. While it is not necessary for them to monotonely increase in order for the IRF's to be nonintersecting, it is typically quite difficult to avoid intersections if one curve is increasing and the other is decreasing, or if one curve covers most of the range from 0 to 1 and the other curve covers only a small part of that range. The former behavior can be seen in the plot of item #4 (jo4; solid line) versus item #9 (jo9; dashed line) in the lower-left plot of Figure 2, and the latter behavior can be seen in the plot of item #4 (jo4; solid line) versus item #8 (jo8; dashed line) in the center-left plot of Figure 2. For the sake of comparison, the right hand side of Figure 2 shows graphs for 3 pairs of items that do not show serious violations of nonintersection of IRF's.

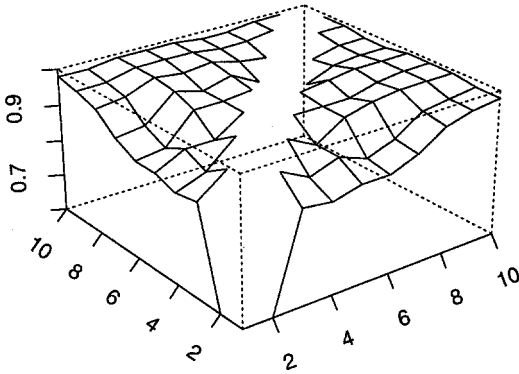
*P(++)* and *P(--)* analysis revealed significant violations (for the statistical test, see Molenaar, 1970; Molenaar et al., 1994) within the item pairs [denoting 1 violation as 1vi, 2 violations as 2vi, and so forth]: (#2, #5; 2vi), (#2, #8; 3vi), (#2, #9; 1vi), (#4, #5; 4vi), (#4, #8; 2vi), and (#4, #9; 3vi). All significant violations occur in the *P(--)* matrix. This can be seen immediately in the lower graph at the left in Figure 3 which displays the **1 - P(--)** matrix; given an invariant item ordering, this graph should slope down from the upper back to the lower front corner. The dips in the slope correspond to observable (not necessarily significant) violations of the expected orderings in the *P(--)* matrix. The *P(++)* graph (upper graph at the left) shows a much more regular slope.



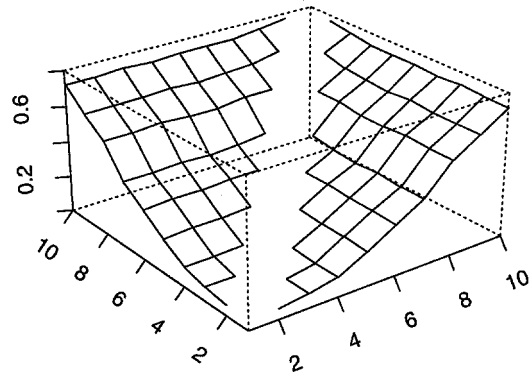
Off-diagonal entries of P(++), judgement only



Off-diagonal entries of P(++), judgement plus



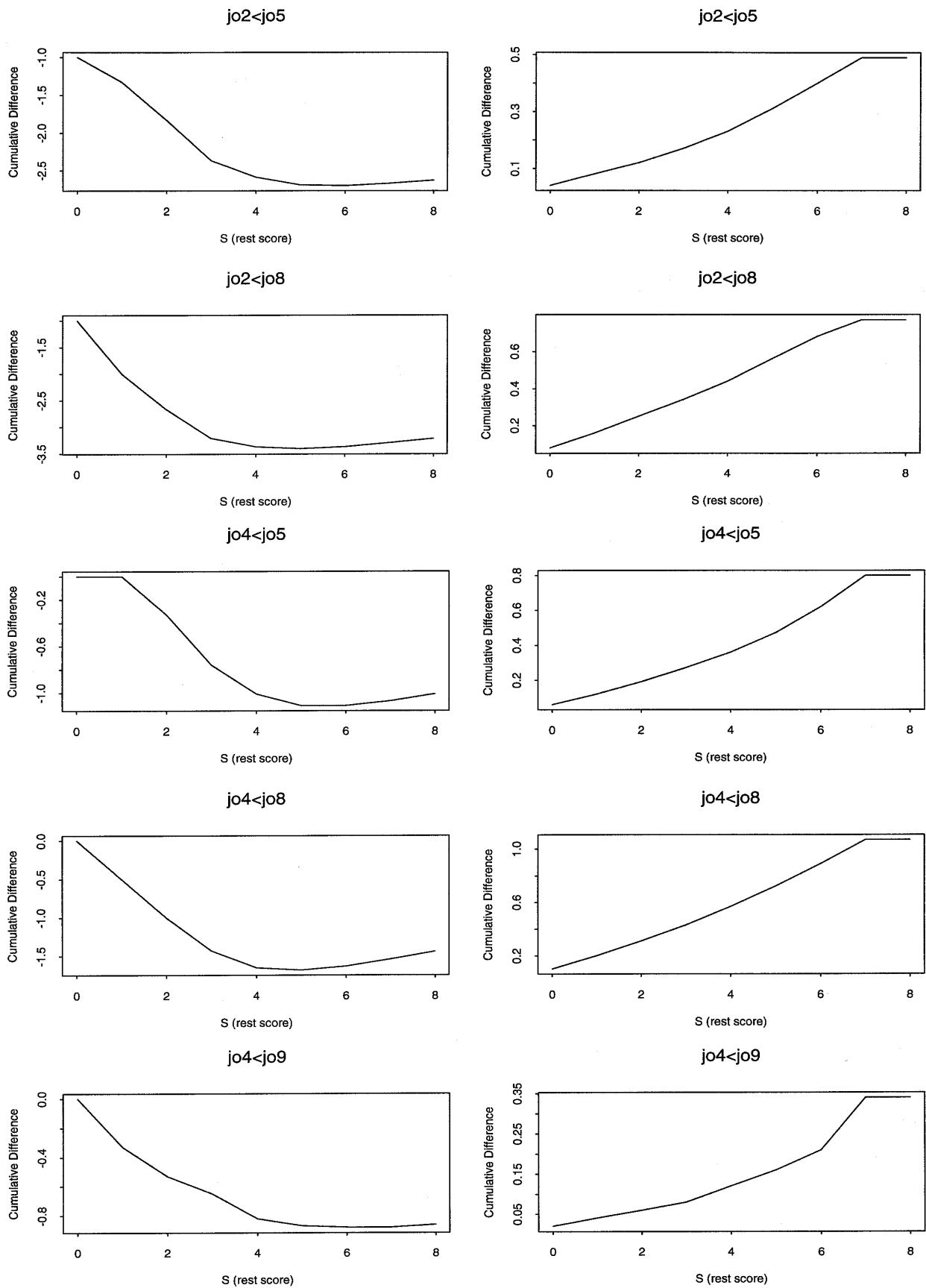
Off-diagonal entries of 1-P(--), judgement only



Off-diagonal entries of 1-P(--), judgement plus

**Figure 3:** Surface plots of the P(++), and P(--), matrices.  
Labels on the X and Y axis are rank orders of item difficulties.

*Item splitting* for  $X_f = 0$ , showed large, significant violations for the item pairs (#2, #5; 2vi), (#2, #8; 3vi), (#4, #5; 5vi), and (#4, #8; 4vi). Smaller, significant violations were found for (#2, #9; 1vi) and (#4, #9; 4vi). Three other pairs showed small, significant violations, but taking the large number of significance tests into account, these need not be taken seriously. Item splitting for  $X_f = 1$  did not reveal any significant violations of the expected orderings.



**Figure 4:** Cumulative difference plots ( $p[X_j = 1] - p[X_i = 1]$ ) across rest score splits for rest score splitting  $S \leq s$  (left half) and  $S > s$  (right half). „jo2“ means: judgement-only data, item #2; and so forth.

*Rest score splitting* for  $S \leq s$  yielded significant violations of expected orderings with the item pairs (#2, #5; 2vi), (#2, #8; 2vi), (#4, #5; 2vi), (#4, #8; 3vi), and (#4, #9; 1vi). These violations are all in the middle rest score splits and appear to be moderate at best. These violations can be conveniently illustrated with the „cumulative difference plots“ in the left half of Figure 4: (1) for each rest score  $s$  the difference  $p[X_j = 1 | S \leq s] - p[X_i = 1 | S \leq s]$  is determined; negative values contradict the expected ordering; (2) the cumulative differences are determined across  $s$ ; and (3) the plot of these cumulative differences against  $s$  should be nondecreasing if IRF's do not intersect. As expected, the cumulative difference plots for the  $S \leq s$  condition for these 5 item pairs (Figure 4, left half) decrease over a substantial range of  $s$ . The decreasing portions of these plots indicate violations of the item ordering at the lower part of the latent scale. The corresponding plots for the  $S > s$  condition (Figure 4, right half) do not show any violations of the item ordering at the higher part of the latent scale. This latter result was found for all item pairs.

These detailed results all suggest that the items #2 and #4, and the items #5, #8, and #9 may be candidate groups for removal. A sensible strategy would include item contents in a decision about this. The items #2 and #4 are the only items involving only equality relations and item #9 has an ordering that is opposite to all other orderings (Figure 1). The items #5 and #8 have similar formats (Figure 1), but they share it with other items that do not come out of the analysis as aberrant. Three analyses were done next: one excluding the items #2 and #4, one excluding the items #5 and #8, and one excluding item #9.

The analysis without the items #2 and #4 yielded  $H^T = 0.65$  and 5.2% negative  $H_a^T$  values. *Rest score grouping* revealed nonsignificant reversals in the lowest rest score groups of the item-rest curves of the items #5 and #9, and #8 and #9. Similar results were obtained in the  $P(-)$  matrix using  $P(++)$  and  $P(-)$  analysis. *Item splitting* only showed a few minor violations. *Rest score splitting*, however, showed 4 significant violations, three involving item #9, and the two largest violations (low rest score groups) involving the item pair #8 and #9.

Deletion of the items #5 and #8 resulted in  $H^T = 0.51$  and 11.8% negative  $H_a^T$  values for the other 8 items. *Rest score grouping* did not provide information about possible intersections.  $P(++)$  and  $P(-)$  analysis revealed small but significant violations in the  $P(-)$  matrix within the pairs (#2, #9; 1vi) and (#4, #9; 2vi). *Item splitting* based on  $X_f = 0$  showed 4 significant violations that all involved item #9. *Rest score splitting* did not reveal significant reversals.

Leaving out item #9 yielded  $H^T = 0.52$  and 11.1% negative  $H_a^T$  values for the remaining 9 items. *Rest score grouping* showed that the item-rest curves of the items #2 and #4 intersect (significant results) with the item-rest curves of the items #5 and #8.  $P(++)$  and  $P(-)$  analysis had the same results as for  $k = 10$ . *Item splitting* based on  $X_f = 0$  revealed 14 significant violations: (#2, #5; 2vi), (#2, #8; 3vi), (#4, #6; 1vi), (#4, #5; 5vi), and (#4, #8; 3vi). *Rest score splitting* showed 10 significant violations, all in low rest score groups, 2 involving each of the item pairs #2, #5 and #4, #8; and 3 involving each of the item pairs #2, #8 and #4, #5.

A possible next step might be the removal of item #9 from either (1) the set without items #2 and #4; or (2) the set without the items #5 and #8. (1) For the remaining 7 items,  $H^T = 0.75$  and 3% of the  $H_a^T$  values were negative. None of the 4 detailed analyses yielded significant

violations. (2) Analysis of the second set of 7 items yielded  $H^T = 0.57$  and 10.4% negative  $H_a^T$  values. Again, no significant violations were found in any of the detailed analyses.

*Judgement-plus criterion.* The ordering of the 10 items under the judgement-plus scoring rule can be found in Table 1. Sijtsma and Verweij (1992) found that the 10 items all had monotonely increasing IRF's. Furthermore, they found a Loevinger's H scalability coefficient equal to 0.76 and item scalability coefficients ranging from 0.52 to 0.82. We found that  $H^T = 0.82$  and that the percentage of negative  $H_a^T$  values was 0.6. None of the detailed analysis on the item level revealed significant evidence of intersections. By way of illustration, Figure 3 (right) shows graphical displays of the P(++ ) and the P(-- ) matrices that are almost perfect, even at the sample level. In particular, the P(-- ) slopes of the judgement-only and the judgement-plus data may be compared. An invariant item ordering has thus been established for all 10 items.

#### 4. Discussion

The combined results for the judgement-only data seem to suggest the removal of the two equality items #2 and #4 from the test. Alternative analysis using Mokken scale analysis and the Rasch model showed that the items #2 and #4 scaled together (Sijtsma & Verweij, 1992). Verweij (1994) has shown that equality tasks are often solved using a reductional strategy, which means that the subject ignores the relations within separate item pairs but rather remembers that all objects are equal. This is not considered transitive reasoning.

Additional deletion of item #9 could be a possibility, but does not seem to be necessary, both on the basis of psychometric and methodological grounds. Psychometrically, the detailed analyses showed that item #9 was involved in a few violations, but the overall analysis using the  $H^T$  methodology was satisfactory. Methodologically, Verweij (1994) has demonstrated that the opposite ordering format of item #9 confused the subjects and thus produced aberrant performance. There was no reason, however, to suspect item #9 of provoking performance that could not, in general, be classified as transitive reasoning. Therefore, the inclusion of this item seems justified and the final scale allowing an invariant item ordering consists of 8 items, excluding the items #2 and #4.

With the judgement-plus data, correct answers due to a reductional strategy (items #2 and #4) and any other strategy that is not a transitive reasoning strategy were scored 0. As a result, the p-value of item #2 dropped from 0.81 to 0.03, and the p-value of item #4 dropped from 0.78 to 0.04 (Table 1). With the judgement-plus scoring, a 1 reflects a correct answer due to transitive reasoning whereas a 0 reflects correct and incorrect answers due to *nontransitive* reasoning and guessing. Scored this way, all items have an invariant ordering.

The use of graphical illustrations of the results of the methods to evaluate the pairs of items helped to summarize the often very large amounts of numerical results from each analysis. For example, with only 45 pairs of items *rest score splitting* produces output both for  $S \leq s$  and  $S > s$ , and for 9 rest score groups  $s$  in each case; this adds up to a total of 810 differences in estimated probabilities and 810 corresponding z-values (of which the 45 pertaining to  $S > 8$  are trivial because the corresponding group size is 0 by necessity). The graphical display of numerous numerical results to facilitate interpretation is an important topic in nonparametric IRT and will be subject to future investigation.

## References

- Grayson, D.A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*, 383 - 392.
- Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, *14*, 1359 - 1378.
- Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Hague: Mouton/Berlin: De Gruyter.
- Molenaar, I.W. (1970). *Approximations to the Poisson, binomial, and hypergeometric distribution functions*. Amsterdam: Mathematical Centre Tracts 31.
- Molenaar, I.W., Debets, P., Sijtsma, K., & Hemker, B.T. (1994). *User's Manual MSP*. Groningen, iecProGAMMA.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Rosenbaum, P.R. (1987a). Comparing item characteristic curves. *Psychometrika*, *52*, 217 - 233.
- Rosenbaum, P.R. (1987b). Probability inequalities for latent scales. *British Journal of Mathematical and Statistical Psychology*, *40*, 157 - 168.
- Sijtsma, K., & Junker, B.W. (1996) A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, *49*, 79-105.
- Sijtsma, K., & Meijer, R.R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*, 149 - 157.
- Sijtsma, K., & Verweij, A.C. (1992). Mokken scale analysis: Theoretical considerations and an application to transitivity tasks. *Applied Measurement in Education*, *5*, 355 - 373.
- Statistical Sciences, Inc. (1992). *S-PLUS, Version 3.1 Release 1 for HP Series 700, HP-UX 8.x*. Seattle, WA: author.
- Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589 - 617.
- Verhelst, N.D. (1992). Het eenparameter logistisch model (OPLM) [The one-parameter logistic model (OPLM)]. Arnhem, the Netherlands: *Cito OPD Memorandum 92-3*.
- Verweij, A.C. (1994). *Scaling transitive inference in 7 - 12 year old children*. Academic dissertation, Vrije Universiteit Amsterdam.
- Verweij, A.C., Sijtsma, K., & Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development*, *19*, 219 - 238.