

Tilburg University

Construction and Validation of a Test for Inductive Reasoning

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

Published in:
European Journal of Psychological Assessment

Publication date:
2003

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

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
de Koning, E., Sijtsma, K., & Hamers, J. H. M. (2003). Construction and Validation of a Test for Inductive Reasoning. *European Journal of Psychological Assessment*, 19(1), 24-39.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Construction and Validation of a Test for Inductive Reasoning*

Els de Koning¹, Klaas Sijtsma², and Jo H.M. Hamers³

¹Leiden University, ²Tilburg University, ³Utrecht University, The Netherlands

Keywords: inductive reasoning, item response models, item response model comparison, test for inductive reasoning

Summary: We present in this paper a test for inductive reasoning (TIR), which consists of two versions that can be used to assess the inductive reasoning development of third-grade pupils in primary education. The test versions can also be used in combination with a training program for inductive reasoning. Two experiments using samples of 954 and 145 pupils were carried out to investigate the psychometric properties of the tests, including validity. Item response theory (IRT) analyses revealed that the scores on the two TIR tests gave meaningful inductive reasoning summaries. This was supported by analyses of the convergent and divergent validity of the TIR tests. IRT analyses were used to equate the two TIR test versions such that the scores can be compared on a common scale. Possible explanations for the misfit of items that were deleted from the TIR tests are discussed.

Introduction

In this paper a new test for inductive reasoning (TIR) is presented, which consists of two versions that can be used to assess development of third-grade pupils in primary education. The test versions can also be used in combination with a teaching program for inductive reasoning (De Koning & Hamers, 1995). This program is applied to the third grade of Dutch primary schools (6-, 7-, and 8-year-olds) with mainly low socio-economic status (SES) pupils (De Koning, 2000; De Koning, Hamers, Sijtsma & Vermeer, 2002). One version of the TIR can be used as a pretest to determine a baseline before the program is applied, and the other version can be used as a posttest for evaluating the learning effects of the program. The items of the TIR and the tasks in the training program refer to the same underlying inductive reasoning construct.

Inductive reasoning is considered to be the general (g) part of human intelligence (Carpenter, Just, & Shell, 1990; Carroll, 1993; Snow, Kyllonen, & Marshalek, 1984; Vernon, 1971). It is supposed to underlie perfor-

mance on complex tasks from diverse content domains (Csapó, 1999; De Koning, 2000; De Koning & Hamers, 1999; Sternberg, 1998; Sternberg & Gardner, 1983). Spearman (1927) considered inductive reasoning processes to comprise the educative ability, that is, the ability to generate the “new” – the productive characteristic of human beings. Educative ability contrasts reproduction, which relies on the ability to process the “known/familiar.”

At the core of the operationalization of inductive reasoning lie comparison processes (Carpenter, et al., 1990; Mulholland, Pellegrino, & Glaser, 1980; Sternberg, 1998). Carpenter et al. (1990) investigated the solution processes underlying inductive reasoning items of Raven’s Standard Progressive Matrices (Raven, 1958). They found that the basic solution process comprised a pairwise comparison of the elements (e.g., the geometric patterns) and their attributes (e.g., the components of the geometric pattern). Comparison is described as an incremental, reiterative process, resulting in a stepwise induction of all the transformations of elements and their attributes. Klauer (1989) specified the comparison processes such that specific types of inductive reasoning could be

* The original data upon which this paper is based are available at <http://www.hhpub.com/journals/ejpa>

defined. These types can be used to design tasks for measuring and training the inductive reasoning ability.

Operationalization of Inductive Reasoning

Klauer (1989) defined inductive reasoning as the systematic and analytic comparison of objects aimed at discovering regularity in apparent chaos and irregularity in apparent order. Regularities and irregularities at the nominal level are recognized by comparing attributes of elements, for example, shape or color. Comparisons at the ordinal and the ratio level involve relationships among elements, for example, with respect to size and number. Comparing attributes or relationships can be directed at finding similarities, dissimilarities, or both. This resulted in six (two types of level crossed with three types of comparisons) formal, interrelated content-independent types of inductive reasoning tasks.

Tasks requiring finding similarities or dissimilarities of attributes of objects are called *generalization* and *discrimination* tasks, respectively. Tasks that demand the simultaneous induction of similarities and dissimilarities are called *cross-classification* tasks. Tasks meant to find similarities, dissimilarities, or both in the relationships between objects are called *Seriation*, *Disturbed Seriation*, and *System Formation* tasks, respectively. Klauer (1989) operationalized the comparison processes in tasks with concrete objects used in daily life (i.e., knowledge-based), and in tasks with geometric patterns referring to reasoning at a more abstract level. Crossing these two content types with the six task types resulted in 12 item types that were included in the TIR tests.

Test for Inductive Reasoning (TIR) Items

Figures 1a and 1b show the 12 types of items used in the TIR tests. In Figure 1a, the three rows contain examples of attribute items that demand pupils to inspect objects with respect to their similarities (generalization; abbreviated gen), dissimilarities (discrimination; dis), or both (cross-classification; cc). In Figure 1b, the rows contain examples of relation items that require pupils to search for similarities (Seriation; ser), dissimilarities (Disturbed Seriation; dser), or both (System Formation; sys). In both Figures, for each of the six item types an example of a knowledge item is given in the second column (i.e., picture item) and an example of a geometric item is given in the third column.

In the test each item is administered on a separate page. Typical questions accompanying the tasks are printed in the first column.

Geometric items were constructed using simple, easy-

to-perceive elements like circles, ellipses, squares, parallelograms, triangles, and simple transformations of their attributes and relations. The transformations were not hidden or misleading, yet they did not result in patterns that were easy to perceive. Carpenter et al. (1990) showed that easy to perceive patterns elicit perceptual processes rather than inductive reasoning processes. Elements were transformed only once in order to prevent subjects from storing and retrieving results of subsequent transformations in working memory. The maximum number of elements in each item entry and the maximum number of transformations of the attributes or relations was three, which matches the number of schemes our participants of 6 to 8 years of age were assumed to be able to activate simultaneously (Case, 1974; Pascual-Leone, 1970).

Knowledge items comprised of only familiar objects like animals, clothes, or articles for everyday use. They were pictured with little detail to prevent distraction by irrelevant features. Two methods were used to increase the transformation difficulty in knowledge items. First, the most comparable with the geometric transformations was the change of the number of attributes or relations of objects or parts of objects. Second, a more common method used in creating knowledge-based reasoning items in intelligence tests (e.g., the WISC-R) is to gradually introduce more abstract transformations. This reflects the intellectual development that is thought to rely initially on perceptual features. Because of a growing ability to abstract from time- and space-bound perception, children are supposed to induce more generalized attributes and relations among objects (Carey, 1985; Piaget, 1970). It was assumed that the abstract knowledge was present in the age range chosen in our sample. In the second column of Figures 1a and 1b, the first two items are examples of perceptual and more abstract knowledge items.

Goals

Two experiments were carried out to investigate the psychometric properties of the tests, including their validity. The first experiment served to calibrate the TIR-I (pretest) and TIR-II (posttest) items by means of item response models (Sijtsma & Molenaar, 2002; Van der Linden & Hambleton, 1997). The calibrated TIR scales were investigated with respect to differential item functioning among three SES samples and two gender samples. The convergent validity was inspected by comparing the TIR to the SPM Raven. Finally, the scales of both TIR versions were equated such that they could be used to assess the pupils' inductive reasoning ability by comparing their score changes on a common scale.

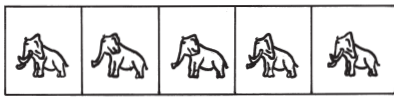
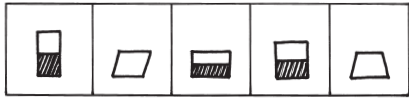

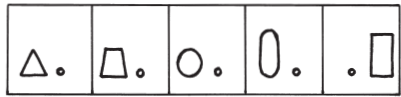
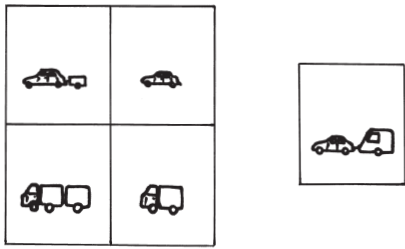
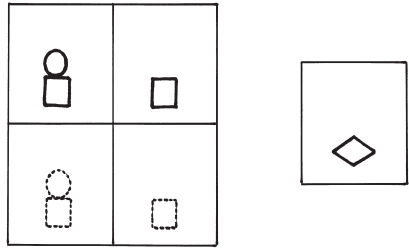
	Picture Item (real-life objects)	Abstract Item (geometric objects)
<p>Similarities of attributes: (generalization)</p> <p>Make a group</p> <p>(one attribute)</p>		
<p>Dissimilarities of attributes (discrimination)</p> <p>What does not belong to the group?</p> <p>(one attribute)</p>		
<p>(Dis)similarities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

Figure 1a. Review of the TIR item types: attribute items.

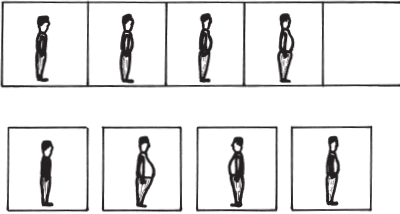
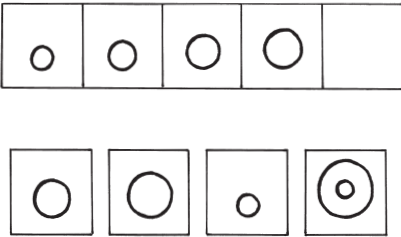
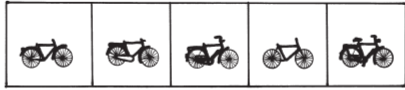

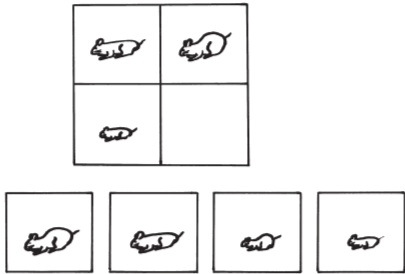
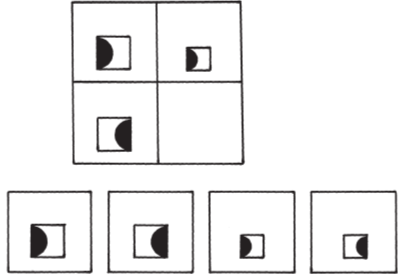
	Picture Item (real-life objects)	Abstract Item (geometric objects)
<p>Similarities of relations: (seriation)</p> <p>Make a row</p> <p>(one relation)</p>		
<p>Dissimilarities of relations (disturbed seriation)</p> <p>What is wrong in the row?</p> <p>(one relation)</p>		
<p>(Dis)similarities of relations (system formation)</p> <p>Make two rows</p> <p>(two relations)</p>		

Figure 1b. Review of the TIR item types: relation items.

The second experiment served to further investigate the convergent and divergent validity of the two TIR tests. The validation procedure aimed at checking whether knowledge items (i.e., pictures) and geometric items both measured reasoning, that is, the production of knowledge rather than memory (reproduction) of knowledge. The TIR was compared to the SPM Raven to investigate its convergent validity, and to a vocabulary test to investigate its divergent validity. It was expected that the TIR would have a high correlation with the SPM Raven, which measures the production of knowledge; and a low correlation with the vocabulary test, which measures the reproduction of knowledge. Finally, the TIR was compared to a listening comprehension test. Since listening comprehension requires both vocabulary and reasoning ability, it was expected that the correlation of the TIR and listening comprehension would be positioned between the correlations of the TIR and the SPM Raven on the one hand, and the TIR and vocabulary on the other hand.

Experiment 1

Method

Population and Sample

Because the TIR tests and the Program Inductive Reasoning (De Koning & Hamers, 1995) are mainly used in primary schools with many low-SES pupils, the concept of backwardness was important. Backwardness was quantified as the school score, which is a formally implemented measure in the Dutch school system. This score reflects the number of pupils visiting the school, weighted by SES, language (Dutch versus a foreign language), and profession of the parents of individual pupils. The school score determines the extra facilities schools are entitled to. The weights are 1.25 for Dutch working-class children, 1.40 for bargee's children (i.e.,

children of parents who operate a freight ship) not living with their parents, 1.90 for children having at least one non-Dutch parent (being limited in terms of educational and professional levels as well), and 1.00 for all other children (Sijstra, 1992). The stratification boundaries for schools were set at 1.05 and 1.15, the cut-off scores guaranteeing a reasonable distribution of pupils over the weight categories of 1.00, 1.25–1.40, and 1.90, respectively (Wijnstra, 1987). A systematic selection of schools, following a randomly chosen starting point in a list of schools not ordered according to the stratification criterion (school score), completed the sampling. Table 1 shows the number of schools and the number of pupils involved in the investigation.

The total sample contained 954 pupils from the third grade. Of this sample, 478 pupils were tested in January and 476 pupils in June. The January sample comprised 230 boys and 248 girls. The mean age was 85 months and the standard deviation was 5.77. The June sample contained 238 boys and 238 girls. The mean age was 88 months and the standard deviation was 4.86.

Test Design

The TIR-I and the TIR-II each had 43 items, of which 16 items were common to both tests. The overlap consisted of items from each of the item types (see Table 2). The TIR-I was administered to the January sample, the TIR-II to the June sample.

Instruments

Apart from the TIR tests, the Standard Progressive Matrices (Raven, 1958) was administered for the purpose of investigating the convergent validity of the TIR Tests. Much research confirmed that the SPM Raven is a valid and reliable test of inductive reasoning (Carpenter et al., 1990; Snow et al., 1984). However, the Raven items are not based on an explicit operationalization of comparison processes such that subtypes of inductive reasoning can be distinguished. The Raven consists of 60 items

Table 1. Number of schools and pupils (boys (b) and girls (g)) per stratum.

	TIR-I								Total	TIR-II								Total
	No. schools		Pupils				No. schools			Pupils								
	1.00		1.25–1.40		1.90		1.00			1.25–1.40		1.90						
	b	g	b	g	b	g	b	g	b	g	b	g						
Stratum 1: 1.00–1.05	6	70	61	7	5	3	2	148	5	74	84	1	2	161				
Stratum 2: 1.06–1.15	7	53	61	14	14	4	8	154	2	55	54	13	11	12	8	153		
Stratum 3: 1.16–1.90	5	12	11	7	16	58	69	176*	6	40	36	14	15	29	28	162		
Total	18	135	133	28	35	65	79	478*	13	169	174	28	28	41	36	476		

* Three pupils were not labeled by their "pupil-weight"

Table 2. Number of items in TIR-I and TIR-II.

	Unique TIR-I	Unique TIR-II	Number of items		Pictures per TIR	Geometric per TIR
			Shared TIR-I + II	Total per TIR		
Attributes:						
Generalization	6	6	3	9	5	4
Discrimination	5	5	2	7	4	3
Crossclassification	3	3	3	6	3	3
Relation:						
Seriation	3	3	3	6	3	3
Disturbed seriation	5	5	3	8	4	4
System formation	5	5	2	7	3	4
Total	27	27	16	43	22	21

divided into five subsets (set A to set E) of increasing difficulty. Each item takes one page, and each of the 60 pages is divided into two half-pages. On the upper half, a matrix of figures is depicted containing a missing element. This element has to be detected among the six (sets A and B) or eight (sets C, D, and E) alternatives printed at the bottom of the page. Many researchers (e.g., Bereiter & Scardamalia, 1979; Hunt, 1974; Willmes, Heller, & Lengfelder, 1997) have tried to explain the varying difficulty of the Raven items. The main distinction between items refers to the kinds of cognitive processes that supposedly underlie the correct solution of the items. Willmes et al. (1997) hypothesized a dichotomy between the first items (A1–B7), only requiring visual comparison processes, and the other items, which demand the application of inductive reasoning processes. Bereiter and Scardamalia (1979) quantified 48 of the 60 SPM Raven items in terms of mental demand, which they defined as increasing from one to five (MD1–MD5).

Procedure

The class administration of the SPM Raven took 45 minutes. Each of the six main item types of the TIR tests required separate instruction. The administration of the 43 TIR items took 60 minutes. These time limits allowed for power conditions.

Statistical Analysis

The quality of the TIR test items was evaluated in three analysis phases. In the *first phase*, the item response functions (IRFs), showing the participant's probability of answering a particular item correctly as a function of inductive reasoning, and the dimensionality of the tests were investigated using four item-response models (De Koning, Sijtsma, & Hamers, 2002). We made use of the advantages of two nonparametric item response models, which are the models of monotone homogeneity (MHM) and double monotonicity (DMM) (Mokken, 1971,

1997), and two parametric item-response models, the Rasch (1960) model and the one parameter logistic model (Verhelst & Glas, 1995; hereafter called the Verhelst model). All four models provide global methods (for all 43 items simultaneously) and local methods (for each item separately) to investigate whether the IRFs are monotone increasing functions and whether all items measure the same latent trait of inductive reasoning.

The nonparametric MHM and DMM in particular provide information about reliability of person ordering [H coefficient (global), and H_j and H_{jk} coefficients (local)] and the nonintersection of the IRFs [H^T coefficient (global) and H^T_a coefficient (local)]. The R_I statistic (global test) and the U_j statistic (local item test) for the Rasch model, and the R_{Ic} statistic (global test) and the M_j statistics (local item tests) for the Verhelst model relate the item characteristics to the logistic shape of the IRF.

Like the Rasch model, the Verhelst model has logistic IRFs that vary in location; unlike the Rasch model the IRFs of the Verhelst model also vary in slope. The Verhelst model does not have a slope *parameter*, however, but rather requires the researcher to *impute* an integer slope A_j for each item. Verhelst and Glas (1995) showed that, with an imputed integer slope, the statistical properties of the Rasch model apply for the Verhelst model.

Explicit procedures for evaluating unidimensionality are absent in the software for investigating the MHM and the DMM (program MSP; Molenaar & Sijtsma, 2000) and the Verhelst model (program OPLM; Verhelst, 1992). However, the Rasch methods can be used to check whether the item sets satisfy unidimensionality. Following Sijtsma's (1983) methodology, we used Andersen's (1973) Likelihood Ratio (LR) test for this purpose.

In the *second phase*, we investigated the invariance of the item parameters among equal-ability pupils from the three SES groups. Also, invariance of item parameters was investigated for boys and girls. Glas and Ouborg (1993) described a procedure to detect biased items; that is, items with different parameters in different SES

groups or gender groups. They used the Verhelst model for this purpose. Analyses of correlation patterns of the TIR tests and the SPM Raven provided information on the convergent validity.

In the *third phase*, we used the Verhelst model to equate the scales of both TIR versions, that is, to calibrate all items of the TIR-I and the TIR-II on the same scale, using the common items as an anchor for relating the unique items to the same metric.

Results

Phase 1: Data Analyses with Four IRT Models

The global test results for the MHM (TIR-I: $H = 0.19$; TIR-II: $H = 0.22$), the Rasch model (TIR-I: $R_I = 476.42$, $df = 168$, $p < .001$; TIR-II: $R_I = 475.53$, $df = 168$, $p < .001$) and the Verhelst model (TIR-I: $R_{Ic} = 380.65$, $df = 126$, $p < .001$; TIR-II: $R_{Ic} = 456.58$, $df = 126$, $p < .001$) revealed that the models did not fit the data. The H values (MHM) suggested that the IRFs had relatively flat slopes, meaning there was a weak relation between the

item scores and the latent trait. The global and local test results of the DMM showed that the data allowed for invariant ordering of items (TIR-I: $H^T = 0.31$, percentage of negative H_a^T values = 0.4%; TIR-II: $H^T = 0.31$, percentage of negative H_a^T values = 0.6%).

The local test results for the four models suggested which items could be left out in order to create item sets that models would better fit. First, the items that could not be fitted by any of the four models were removed. Because removal of one item may change the statistics of others, items were left out one by one on the basis of low H_j values or significant U_j or M_j values. Furthermore, apart from psychometric considerations, the representation of item types was considered before leaving out items. Tables 3 and 4 show the results of the analyses.

For the MHM, Tables 3 and 4 show that the scalability coefficient H was close to the lowerbound value of 0.3 (Mokken, 1971, p. 153) (TIR-I: $H = 0.29$; TIR-II: $H = 0.30$). The TIR-I Generalization items (gen4 and gen11), with relatively low H_j coefficients of 0.15 each, were not left out because otherwise too few items of this type

Table 3. TIR-I global and local test results of four models: The model of monotone homogeneity (MHM), the model of double monotony (DMM), the Rasch model and the Verhelst model.

item	MHM*			DMM*		RSP*	Verhelst*			
	H_j ≤ 0.15	$ Z_{max} $ ≥ 1.96	Zsig ≥ 1	$ Z_{max} $ ≥ 1.96	# Zsig ≥ 1	$ U_j $ ≥ 1.96	A_j	$ M_{1j} $ ≥ 1.96	$ M_{2j} $ ≥ 1.96	$ M_{3j} $ ≥ 1.96
gen2							3			
gen4	0.15					2.11	1			
gen9							2			
gen11	0.15			2.49	2		1			
dis18							2			
dis24							2			
dis29				2.47	2		3			1.98
cc39							3			
cc41							3			
cc43							3			
ser46							3			
ser50							3			
ser53							4			
ser55							4			
dser69							3			
dser72				2.11	2		3			
dser75				2.49	1		3			
dser77							3			
dser78							4			
dser80							2			
sys85							3			
sys88						-2.03	6	-2.08		
sys90				2.47	2		6	-2.03		
sys91							6			
sys93							4			

* MHM: $H = 0.29$ (4% negative H_{jk} values); DMM: $H^T = 0.46$ (1.9% negative H_a^T values); Rasch: $R_I = 210.99$, $df = 96$, $p = 0.000$; Verhelst: $R_{Ic} = 76.71$, $df = 72$, $p = 0.33$; gen = Generalization, dis = Discrimination, cc = Cross-Classification, ser = Seriation, dser = Disturbed Seriation, sys = System Formation

Table 4. TIR-II global and local test results of four models: the model of monotone homogeneity (MHM), the model of double monotony (DMM), the Rasch model and the Verhelst model.

item	MHM*			DMM*		RSP*	Verhelst*			
	H_j ≤ 0.15	$ Z_{max} $ ≥ 1.96	Z_{sig} ≥ 1	$ Z_{max} $ ≥ 1.96	# Z_{sig} ≥ 1	$ U_j $ ≥ 1.96	A_j	$ M_{1j} $ ≥ 1.96	$ M_{2j} $ ≥ 1.96	$ M_{3j} $ ≥ 1.96
gen1										2
gen3										2
gen8										3
gen10										1
gen13										2
dis18										3
dis25										3
dis28										1
dis30				2.10	1			2.73	2.51	3
cc40										3
cc42				2.07	1			2.30		4
cc43				2.42	1					3
ser46										3
ser49		2.38	1	2.07	3			-2.23		4
ser52										5
ser53								-2.23		4
ser56										5
dser68										3
dser69								2.32		3
dser71										4
dser72										3
dser76				1.97	2					3
dser79				2.42	3					5
dser81										2
sys83										4
sys86										4
sys87										4
sys90										5
sys94				2.10	1					4

*MHM: $H = 0.30$ (2.2% negative H_{jk} values); DMM: $H^T = 0.41$ (1.5% negative H^T_o values); Rasch: $R_I = 102.07$, $df = 85$, $p = 0.10$; Verhelst: $R_{Ic} = 68.54$, $df = 84$, $p = 0.89$; gen = Generalization, dis = Discrimination, cc = Cross-Classification, ser = Seriation, dser = Disturbed Seriation, sys = System Formation

would remain, and the inductive reasoning construct would not be represented well enough. For the TIR-II only one item violated the model (ser49). The combination of low H values with only one significant violation of the model could be explained by the relatively flat IRF slopes. Despite a few significant Z values (values in the fifth column, number of significant values in the sixth column of Tables 3 and 4), indicating intersections of the IRFs, the H^T values of 0.46 and 0.41 justified the conclusion that at a global level the item sets complied with the DMM.

For the Rasch model, the R_I test result of the TIR-I ($R_I = 210.99$, $df = 96$, $p < .001$) suggested that this model did not fit the data. U_j test results (seventh column of Table 3) showed that only two items violated the Rasch model's assumptions significantly (gen4 and sys88). Analyses of the TIR-II test data did not show significant R_I or U_j test results ($R_I = 102.07$, $df = 85$, $p = .10$), indicating that the Rasch model fitted the TIR-II data.

The Verhelst model: Discrimination indices. The H values of both the TIR-I and the TIR-II indicated that a few items had flat IRFs, and the Verhelst model was used to study the numerical values of the slopes of the IRFs. The discrimination indices (denoted A_j) are displayed in the eighth column of Tables 3 and 4. The Verhelst model complied with both the TIR versions (TIR-I: $R_{Ic} = 76.71$, $df = 72$, $p = .33$; TIR-II: $R_{Ic} = 68.54$, $df = 84$, $p = .89$). Only a few minor M_j test violations were found at the item level. We concluded that the IRFs approached the logistic function. The discrimination indices of the TIR-I varied from one to six. Not surprisingly, the least discriminating items (gen4 and gen11) had the lowest H_j values. The items with the highest discrimination index of 6 were system formation items. The two items that did not comply with the Rasch model (gen4 and sys88) were found in the lowest and highest part of the A_j index range, respectively. Although leaving out these items might have resulted in a Rasch item set, for reasons of repre-

Table 5. Number of initial and deleted items in the TIR-I and the TIR-II.

		Number of items					
		Initial per test	Picture Items TIR-I deleted	TIR-II deleted	Initial per test	Geometric Items TIR-I deleted	TIR-II deleted
Attributes	(Picture + Geometric: 22)	12	9	8	10	3	2
Relations	(Picture + Geometric: 21)	10	5	1	11	1	3
Total		22	14	9	21	4	5

sentation of the inductive reasoning concept and the discrimination power, it was decided to maintain these items in the test. As expected, the range of discrimination index values of items from the TIR-II was narrower (1 through 5) than from the TIR-I. The item ser49 marginally violated three models, the MHM, the DMM and the Verhelst model, but it was kept in the test because it had high discrimination power.

The discrimination indices of both TIR tests showed lower values for the Generalization and Discrimination items, and higher values for the Cross-Classification items, the Seriation items, the Disturbed Seriation items and the System Formation items. The relative size of the discrimination indices reflected the relative size of H and H_j values of the various item subsets (De Koning et al., 2002).

Careful inspection of the items left out of the TIR-I and the TIR-II revealed that the majority were attribute items and picture items (see Table 5). The percentages of deleted items from the TIR-I that were attribute items or relation items were 55 (12 out of 22) and 29 (6 out of 21), respectively. For the TIR-II, these percentages were 45 (10 out of 22) and 19 (4 out of 21), respectively. The percentages of deleted items from the TIR-I that were picture items or geometric items were 64 (14 out of 22) and 19 (4 out of 21), respectively. For the TIR-II, these percentages were 41 (9 out of 22) and 24 (5 out of 21), respectively. After deletion of 18 items in the TIR-I and 14 items in the TIR-II, the two TIR versions each still consisted of 12 types of items. The TIR-I contained 25 items (43–18) and the TIR-II contained 29 items (43–14).

The Rasch Model: Unidimensionality. To investigate the assumption of unidimensionality, we used Andersen's (1973) LR test. The sample was divided into two halves based on the correct and incorrect answering of a splitter item (Sijtsma, 1983; Van den Wollenberg, 1982; other methods are discussed by Glas & Verhelst, 1995, and by Ponocny, 2001). Systematic differences between the estimates of the item parameters in the two groups indicate a violation of unidimensionality. The test was done at a significance level $\alpha = 0.001$, as recommended by Glas

and Ellis (1993; the test is sensitive to small deviations, and a low α avoids falsely rejecting the null hypothesis to some degree). Several splitter items were used to obtain valid conclusions. The items dis29, dis30, and dser72 were used here for illustration purposes. Items dis29 and dis30 were designed as parallel items for the TIR-I and the TIR-II. Item dser72 was shared by both tests.

The choice of splitter items was based on a proportion-correct of approximately 0.50, which produces almost equal estimation accuracy in the two subsamples. Van den Wollenberg (1982) recommends using splitter items that are suspected to measure latent traits different from those measured by several of the other items in the test. Because our tests have six item types by definition, this may induce multidimensionality. Item contents thus seems to be a sensible *a priori* criterion for choosing any of the items as a splitter item, and this agrees with Van den Wollenberg's (1982) recommendation.

Figure 2 shows that Andersen's test was not significant, meaning that item parameters in both subgroups based on the discrimination splitter items were equal (TIR-I: LR = 46.77; $df = 23$, $p = .002$; TIR-II: LR = 24.68, $df = 27$, $p = .592$). This result suggests that the item sets were unidimensional. The Andersen test results for the disturbed Seriation items were significant (TIR-I: LR = 53.68; $df = 23$, $p < .001$; TIR-II: LR = 59.98, $df = 27$, $p < .001$). However, because the displayed item parameter estimates did not reveal clear subdivisions of the items into subsets, an obvious criterion for a practically useful subdivision was not available. Moreover, in such cases test users prefer to consider the total item set to be dominated by one latent trait and ignore so-called nuisance traits (Sijtsma & Molenaar, 2002); at the mathematical level the reader may want to consult Stout's (1990) concept of essential unidimensionality. Other splitter items did not produce significant results or results that could be interpreted clearly. For example, item dser78 (TIR-I) had LR = 65.92, $df = 23$, and $p < .001$; and item dser68 (TIR-II) had LR = 49.81, $df = 27$, and $p = .005$. Graphical displays did not result in clearly interpretable results. Thus, for practical purposes all items together were considered to cover the same inductive reasoning construct reasonably well.

Table 6. R_{1c} tests of the TIR-I and the TIR-II for the whole sample and for the subsamples based on SES and gender.

	TIR-I			TIR-II		
	R_{1c}	df	p	R_{1c}	df	p
Whole sample	76.71	72	.330	68.54	84	.889
SES	329.26	164	.003	326.42	308	.226
Gender	183.92	168	.190	197.56	196	.459

Both TIR test scores were reliable: Cronbach's α coefficients were 0.82 and 0.84 for the TIR-I and the TIR-II, respectively.

Phase 2: Validity of TIR-I and TIR-II

Differential item functioning. The Verhelst model was used to inspect the invariance of the item parameters among equal ability participants from the three SES groups and from the two gender groups. The first step

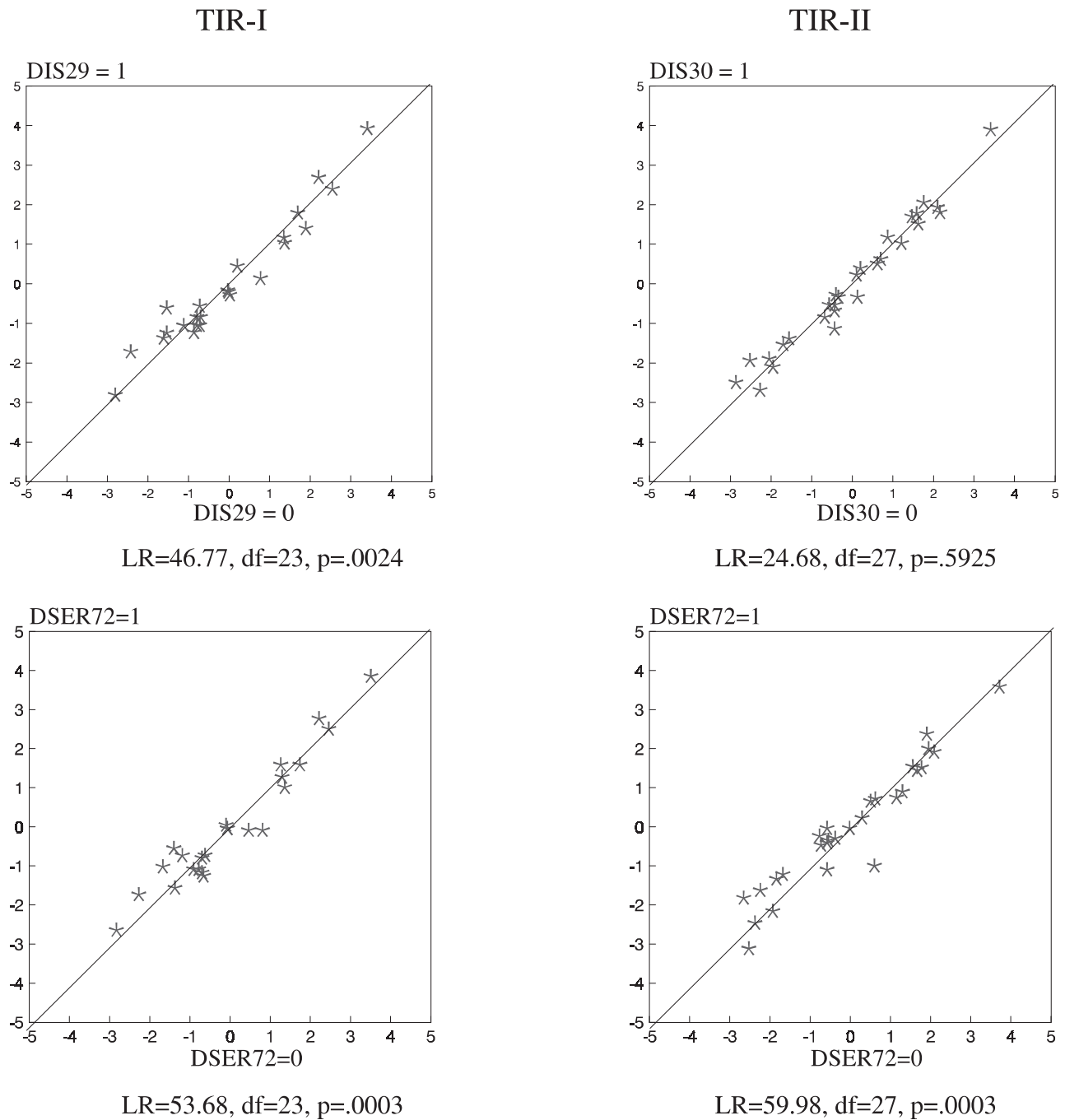


Figure 2. Presentation of item parameter estimates and Andersen's likelihood ratio test results of the TIR-1 (left figures) and the TIR-II (right figures) after splitting the sample into two parts based on answers of the items discrimination 29, 30 (upper figures) and disturbed seriation 72 (lower figures).

Table 7. Correlations of the TIR-I, the TIR-II and the SPM-Raven (total sets and subsets).

	Total set	B8-E12	MD1	SPM Raven			
				MD2	MD3	MD4	MD5
TIR-I	.67	.63	.41	.61	.63	.54	-.05
TIR-II	.67	.66	.38	.65	.63	.54	.11
Raven (total) (January sample)	1.00	.96	.48 ^x	.78 ^x	.79 ^x	.69 ^x	.02 ^x (n.s.)
Raven (total) (June sample)	1.00	.98	.42 ^x	.79 ^x	.82 ^x	.75 ^x	.20 ^x

Correlations are significant at the .001 level unless otherwise stated; ^x Raven total score corrected for MD score

checked whether the item parameters were the same in the various subgroups. In the second step, detailed information was obtained about the standardized differences between observed and expected frequencies of participants in subgroups of equal ability. Table 6 displays the results of the first step.

None of the R_{Jc} test results in Table 6 exceeded the significance level of .001, but for the TIR-I the SES groups came close. An explanation is that the lowest SES group contained participants who had not yet mastered sufficient ability in the Dutch language necessary to understand the instructions of the TIR. Also, the picture items could have involved objects, attributes, or relations not yet known to these children. Because biased items may lead to conclusions about a low level of inductive reasoning – when in fact language deficiency is responsible for such a low score – these items had to be detected and removed from the tests.

In the second part of the analysis, the sample was ordered with respect to the sum of item scores weighted with the corresponding discrimination indices. Subsequently, the sample was divided into four homogeneous weighted-sumscore groups of approximately equal size, with every sumscore group containing pupils from the three SES groups. For every item the standardized differences between observed and expected frequencies of participants in the twelve subgroups were computed. The sign of the standardized difference reveals whether there were more observations or fewer observations in a group than expected on the basis of the Verhelst model. Since only 16 out of 300 (i.e., 25 items by 12 subgroups) statistical tests showed significant results, there was no convincing evidence that the items were biased. Moreover, the significant deviations did not consistently appear in the lowest SES group.

Convergent validity. Table 7 shows that both TIR tests correlated highly with the total score on the SPM Raven items. The TIR-I and the TIR-II had higher correlations with the MD2, MD3, and MD4 item subsets than with the MD1 and MD5 subsets. We found the same correlation pattern for the Raven total scores with these item subsets, both for the January sample and the June sample. The TIR-I and the TIR-II had slightly higher correlations

with the total set of SPM Raven items than with the subset of SPM Raven items (B8-E12), which demands inductive reasoning rather than visual perception. The same correlation pattern was found for the SPM Raven total set with this subset (B8-E12), both for the January and the June sample. This indicated that the TIR tests and the SPM Raven showed similar correlation patterns.

Phase 3: Linked Design Calibration and Equation

Since the Verhelst model complied with both TIR tests resulting from the item analyses, this model was used to calibrate the items of the TIR-I and the TIR-II together. This is known as *equating* (Engelen & Eggen, 1993). In the combined item set, the discrimination indices ranged from 1 through 6, and only three items showed significant M_j test results (cc42, ser49, sys88). The global test result was not significant ($R_{Jc} = 172.52$, $df = 162$, $p = .27$). It could be concluded that the combined item set satisfied the assumptions of the Verhelst model. The item parameters on the equated TIR-scale are shown in the first three columns of Table 8.

The item parameter estimates (β) and discrimination indices (A) were used to evaluate the items with respect to differences between (a) the TIR-I and the TIR-II; (b) picture items and geometric items; and (c) attribute items and relation items. Student's t -tests for equality of mean β s and mean A s, and F tests for equality of variances of β s and A s revealed no significant differences between the two TIR versions, showing that both versions had the same mean and variance in difficulty ($t = .813$, $df = 52$, $p = .42$; $F_{1,52} = 1.07$, $p = .31$) and in power to discriminate pupils on θ ($t = .228$, $df = 52$, $p = .82$; $F_{1,52} = .01$, $p = .92$). For picture items and geometric items we found no significant differences with respect to the mean difficulty ($t = -.05$, $df = 45$, $p = .96$) and the discrimination power ($t = -1.05$, $df = 45$, $p = .30$). The attribute items had significantly lower mean difficulty ($t = -5.00$, $df = 45$, $p < .01$) and discrimination power ($t = -4.91$, $df = 45$, $p < .01$) than the relation items.

Analysis of variance was used to test which item sets comprising the attribute items (Generalization, Discrimination, Cross-Classification) and relation items (Seriation, Disturbed Seriation, System Formation) were re-

Table 8. Estimated item location parameters (β) and person parameters (latent traits: θ) on the Equated TIR Scale, and their standard errors. Discrimination indices (A) are in (brackets). Only a limited number of person parameter estimates are given.

Item	Item parameter		Score	Person parameter	
	Estimate	St. error		Latent trait	St. error
gen10	-3.444 (1)	0.353	TIR-I 0	-2.510	1.468
dis28	-3.030 (1)	0.291	3	-1.162	0.447
gen4	-2.174 (1)	0.181	8	-0.625	0.266
dis18	-1.220 (2)	0.094	12	-0.409	0.217
gen13	-1.068 (2)	0.126	16	-0.253	0.191
gen3	-0.899 (2)	0.109	20	-0.126	0.174
gen1	-0.830 (2)	0.103	25	0.009	0.161
dis24	-0.584 (2)	0.074	29	0.106	0.153
dis25	-0.556 (3)	0.086	33	0.195	0.147
gen9	-0.426 (2)	0.067	37	0.277	0.141
gen11	-0.383 (2)	0.066	42	0.371	0.135
gen2	-0.251 (3)	0.051	46	0.440	0.132
cc40	-0.087 (3)	0.052	50	0.505	0.130
dser69	-0.085 (3)	0.034	54	0.569	0.131
cc39	-0.067 (3)	0.044	59	0.650	0.135
dser75	-0.052 (3)	0.044	63	0.720	0.142
cc43	-0.025 (3)	0.033	67	0.800	0.155
dser76	-0.016 (3)	0.049	71	0.899	0.176
cc41	0.004 (4)	0.037	76	1.084	0.230
cc42	0.033 (4)	0.042	80	1.367	0.350
ser49	0.064 (4)	0.041	83	2.131	0.927
ser55	0.083 (4)	0.035			
ser56	0.147 (5)	0.035			
gen8	0.181 (3)	0.042			
ser50	0.188 (3)	0.039			
ser46	0.243 (3)	0.029			
ser52	0.247 (5)	0.032	TIR-II 0	-4.835	2.210
dser79	0.284 (5)	0.031	4	-1.428	0.463
ser53	0.287 (4)	0.024	9	-0.790	0.291
dser77	0.307 (3)	0.038	13	-0.526	0.231
dser78	0.417 (5)	0.028	18	-0.315	0.188
dser71	0.455 (4)	0.032	23	-0.168	0.163
sys87	0.493 (4)	0.032	28	-0.053	0.147
sys91	0.570 (6)	0.026	33	0.044	0.137
sys88	0.580 (6)	0.026	37	0.113	0.132
sys86	0.607 (4)	0.031	42	0.194	0.128
sys83	0.633 (4)	0.031	47	0.272	0.126
dis29	0.644 (3)	0.037	52	0.348	0.125
sys90	0.644 (6)	0.020	57	0.425	0.126
dser68	0.725 (3)	0.037	61	0.488	0.128
dser72	0.726 (3)	0.027	66	0.569	0.132
sys94	0.728 (4)	0.031	71	0.655	0.138
dis30	0.822 (2)	0.051	76	0.751	0.149
sys93	0.897 (4)	0.036	81	0.866	0.169
sys85	0.983 (3)	0.042	85	0.987	0.198
dser81	1.458 (2)	0.063	90	1.237	0.291
dser80	1.745 (2)	0.084	94	1.944	0.821

responsible for the significant differences in difficulty and discrimination power that we found using Student's *t*-tests. The Bonferroni correction, adjusting the significance level for multiple comparisons, showed that the difference between attribute items and relation items on β and A was caused by significant differences between the Generalization items and Discrimination items, on the one hand, and the three relation sets, on the other hand.

The results showed that the data were suited for a horizontal equating procedure (Engelen & Eggen, 1993; Veldhuijzen, Godebeld, & Sanders, 1993), because the TIR versions consisted of the same types of items, they were unidimensional, and they did not show differential item functioning among groups (SES, gender) of participants. Furthermore, the TIR versions had the same mean and variance in difficulty and discrimination power. Based on the item parameter estimates, the Verhelst model was used to estimate for every weighted sumscore a person parameter (θ). These estimates were equated, and the result is shown in the last three columns of Table 8. The scores on the TIR-I range from 0 through 83, and on the TIR-II from 0 through 94. For reasons of brevity, the table shows only the scores (and the latent traits) for every fifth percentile. The Verhelst model provides a caution index ζ for every participant, indicating the extent to which the item score pattern is expected given the item parameters. With only 2% (18 out of 954) of the participants having unexpected patterns, we used the estimated person parameters to standardize the scores on TIR-I and TIR-II. For both TIR tests the estimated θ s were ordered and, subsequently, centiles and quartiles (10, 25, 50, 75, 90) were computed. These cut-off points can be used for normalizing individual scores: For every participant it is possible to compare the TIR-scores with the population distribution. Thus, it is possible to measure the progress in inductive reasoning of every third-grade pupil.

Experiment 2

Method

Population and Sample

From 103 schools, each having more than 80% pupils with a (SES) weight factor of 1.90, six school classes in the third grade (6-, 7-, and 8-

Table 9. Correlations of the TIR-I, the TIR-II, the SPM-Raven, the Listening Comprehension Test, and the Vocabulary Test.

		TIR-I			TIR-II			SPM Raven	List. Comp.	Vocab.
		total	pict	geo	total	pict	geo			
TIR-I	total	1.00	.76 ^{xx}	.96 ^{xx}	.68 ^{xx}	.58 ^{xx}	.65 ^{xx}	.61 ^{xx}	.48 ^{xx}	.29 ^{xx}
	pict.		1.00	.55 ^{xx}	.49 ^{xx}	.41 ^{xx}	.47 ^{xx}	.41 ^{xx}	.40 ^{xx}	.30 ^{xx}
	geom.			1.00	.66 ^{xx}	.57 ^{xx}	.63 ^{xx}	.61 ^{xx}	.45 ^{xx}	.24 ^{xx}
TIR-II	total				1.00	.90 ^{xx}	.93 ^{xx}	.72 ^{xx}	.44 ^{xx}	.31 ^{xx}
	pict.					1.00	.66 ^{xx}	.65 ^{xx}	.39 ^{xx}	.27 ^{xx}
	geom.						1.00	.67 ^{xx}	.40 ^{xx}	.29 ^{xx}
SPM Raven								1.00	.40 ^{xx}	.21 ^x
List. Comp.									1.00	.64 ^{xx}
Vocabulary										1.00

^x Correlation is significant at the .05 level (2-tailed), ^{xx} Correlation is significant at the .01 level (2-tailed)

year-olds) were randomly selected. The sample consisted of 145 pupils (82 boys and 63 girls). The mean age was 86 months, and the standard deviation was 6.12.

Instruments

The TIR-I and the TIR-II, the SPM Raven, a vocabulary test (Verhoeven, 1996), and a listening comprehension test (CITO, 1995) were administered. The Vocabulary Test and the Listening Comprehension Test are widely used in Dutch primary education to compare the achievements of individual pupils and groups of pupils.

The Vocabulary Test consists of four pictures per item. The pupils have to indicate one picture that fits the description the teacher reads out. The test consists of 50 items.

The Listening Comprehension Test consists of 44 statements and short stories the teacher reads out. The pupils have to indicate the picture that matches the statement or the short story. That is, they have to induce the meaning by linking parts in the statements and stories that are connected. This requires an adequate vocabulary, awareness of grammar, and inductive reasoning. Thus, the test measures memory of knowledge and production of knowledge.

Procedure

The classroom administration of the TIR-I and the TIR-II took 60 minutes, that of the SPM Raven 45 minutes. The Vocabulary Test and the Listening Comprehension Test each took 90 minutes.

Statistical Analysis

Correlations were computed to inspect the relations of the TIR-I, the TIR-II, the SPM Raven, the Listening Comprehension Test, and the Vocabulary Test. Linear regression analyses were done to examine whether the hypothesized decreasing relation strength of the TIRs with the SPM Raven, the Listening Comprehension Test, and the Vocabulary Test, respectively, could be confirmed.

Results

Table 9 shows the correlations of the TIR-I, the TIR-II, the SPM-Raven, the Listening Comprehension Test, and the Vocabulary Test. The TIR-I and the TIR-II correlated highly with the SPM Raven (0.61 and 0.72, respectively). These correlations were comparable with values found in the first experiment (see Table 7, first column; 0.67 in both cases). As hypothesized, the correlations of the TIR-I and the TIR-II with Listening Comprehension were lower (0.48 and 0.44, respectively), and correlations were lowest with Vocabulary (0.29 and 0.31, respectively). The correlations of the subsets of geometric and picture items from the TIR tests with the SPM Raven were moderate to high (TIR-I: 0.41 and 0.61, respectively; TIR-II: 0.65 and 0.67, respectively). Their correlations with Listening Comprehension were slightly lower (TIR-I: 0.40 and 0.45, respectively; TIR-II: 0.39 and 0.40, respectively), and they were lowest with Vocabulary (TIR-I: 0.30 and 0.24, respectively; TIR-II: 0.27 and 0.29, respectively). The correlations of the SPM Raven with Listening Comprehension and Vocabulary showed a similar correlation pattern (0.40 and 0.21, respectively). This indicated that the TIR tests and the SPM Raven showed similar relation patterns with other tests.

Regression analyses with each of the TIR tests as dependent variable and the SPM Raven, Listening Comprehension and Vocabulary as independent variables, showed that most variance of the TIR-I could be explained by the SPM Raven (37%), and that Listening Comprehension explained an additional 7% ($F_{2,144} = 55.90, p < .01$). Vocabulary did not contribute uniquely to the explanation of the TIR-I variance. Neither Listening Comprehension nor Vocabulary contributed to the explanation of the TIR-II variance, after SPM Raven had been selected (52% explained variance; $F_{1,144} = 157.58, p < .01$). As the TIR-II was administered 6 months later than the TIR-I, this indicated that for older participants the scores relied more on reasoning and less on the knowl-

edge of vocabulary and grammar than for younger participants.

General Discussion

We used four IRT models to scale 12 types of inductive reasoning items. The total scores on the two TIR tests give meaningful inductive reasoning summaries collected under power conditions. The convergent and divergent validity results supported the IRT analyses in that the TIR scores reflect inductive reasoning ability. The testing procedures provided by the four IRT models resulted in the deletion of misfitting items. The majority of the deleted items were attribute items and picture items.

The results from the splitter-item method showed that the tests were not entirely unidimensional. However, we decided not to follow a purely statistical line of reasoning and also keep items in the test that deviated mildly from others to maintain good coverage of the different aspects of the inductive reasoning ability. More support for this decision came from the practical observation that pure unidimensionality is a theoretical ideal and that real tests are multidimensional to at least some degree, even if the test constructor explicitly pursued unidimensionality (also see Nunnally, 1978). The distinction between a dominant latent trait and nuisance traits was made at the theoretical level by, for example, Stout (1990). Here, we ignored the subtleties of Stout's (1990) argument, but noted that the inductive reasoning items left in our tests, even when representing different types, probably have enough in common in terms of underlying cognitive processes to be in the same test. Other arguments came from test practice, where small deviations from unidimensionality are tolerated because trait coverage often is considered more important. Finally, splitting our tests into substantively purer subtests would yield short tests with inaccurately estimated latent traits. The usefulness of working with one TIR score was further corroborated in a study that evaluated the effectiveness of training programs (De Koning, Hamers, Sijtsma, & Vermeer, 2002).

With respect to the deletion of the attribute items, we suggest the following explanation: The maximum number of attributes and relations to be induced in each TIR item is three, which matches the number of schemes our participants theoretically were supposed to be able to activate simultaneously (Case, 1974; Pascual-Leone, 1970). According to Klauer (1989) the comparison of attributes requires persons to attend simultaneously to two objects. In contrast, comparing relations is possible only if three objects are simultaneously investigated. Because the basic comparison process is limited to two elements (Carpenter et al., 1990), relation items might

require more extensive mental coordination for decomposing the items. This involves high-level strategic processes that probably resemble the executive assembly and control processes described by Marshalek, Lohman, and Snow (1983). The stronger demand of mental coordination resulted for the relation items in a higher power to discriminate participants than for the attribute items. To construct attribute items that better discriminate participants, it seems necessary to increase the maximum number of attributes to be more than three. This higher maximum will impose an additional load on working memory as it will require the participants to keep track of the variation associated with already induced transformations while inducing new transformations (Mulholland et al., 1980).

The content of the deleted items mostly concerned the picture items and not the geometric items. Geometric items have the advantage that they are easily decomposable into characteristics that influence processing, for example, the number of attributes and the number of transformations. Therefore, geometric content is used by many cognitive researchers (e.g., Evans, 1968; Mulholland et al., 1980) to model the inductive reasoning solution processes. Test developers (e.g., Hosenfeld, Van Den Boom, & Resing, 1997) used geometric items to predict the inductive reasoning test scores. For picture items there is a risk that they tap *memory* of knowledge (Spearman's reproductive ability) rather than *reasoning* about knowledge (Spearman's productive ability). Richardson (1996), for example, changed item elements and transformations of the SPM Raven tasks into social situations. This was criticized by Roberts and Stevenson (1996), who argued that the problem-solvers were given too many clues, which undermined the requirement for reasoning. Goswami's (1991) review of many studies of inductive reasoning revealed that children are able to properly apply inductive reasoning processes if they have the knowledge about the relations involved. Sternberg and Gardner (1983) compared geometric, verbal, and schematic pictorial inductive reasoning items (classifications, series, and analogies) and concluded that highly similar process steps are used in solving the tasks, but that these process steps operate on different knowledge stores and possibly different forms of representation of the different contents. Thus, content more than task type served as a greater source of individual differences in induction problems. We hypothesize that this variation interacted with inductive reasoning, and that our deleted items measured this interaction.

Despite the difficulties we experienced in designing picture items, for two reasons we would like to include these types of items in an inductive reasoning test for pupils. First, from an ecological perspective (Sternberg, 1998), it is not valid to limit the productive characteristic

of human beings to the mental manipulation of meaningless geometric material. Second, from a developmental perspective, many researchers now take an integrated approach examining the development of reasoning strategies by studying the interaction of knowledge and reasoning skills (Zimmerman, 2000). This integrated approach is an attempt to solve the debate about what actually drives development. The primacy of *knowledge* is reflected in research that stresses the knowledge base as the conceptual system upon which the reasoning mechanisms operate (Vosniadou, 1989). The primacy of *reasoning* is reflected in the view that the knowledge base plays a subordinate part in development. The mixture of content in the TIR test items reflects the integrated approach.

The method of combining a training program with tests to precisely measure an ability has been used by several researchers (e.g., Brown, Campione, Reeve, Ferrara, & Palinscar, 1991; Feuerstein, Rand, Jensen, Kaniel, & Tzuriel, 1987; Palinscar & Brown, 1988). The group under study in our research belonged to the low SES category in which the inductive reasoning ability is less well developed than expected (De Koning, 2000; Hamers, De Koning, & Sijtsma, 1998). Since inductive reasoning underlies the learning in various domains, including school domains (Csapó, 1999; De Koning 2000; De Koning & Hamers, 1999; Klauer, 1997, 1999), it is important to know what the potential development is of pupils' use of domain-independent inductive reasoning procedures. By using the combination of program and tests we may be able to detect pupils and specific inductive reasoning tasks that might need more of our attention.

The TIRs can also be used without the training program. The tests provide standardized scores for assessing the individual development. Klauer's (1989) operationalization of inductive reasoning into separate task types clarifies the similarities with our test tasks that are taught in the regular curriculum. This means that the relationship between inductive reasoning as measured by the test and school-domain tasks becomes understandable. For teachers this is very important since they are supposed to include the underlying inductive reasoning processes in their instruction of domains like mathematics and reading comprehension (De Koning et al., 2002). By teaching these processes, they assume that pupils will become aware of widely applicable strategies and, subsequently, will become able to flexibly apply these strategies in other domains (De Koning, 2000).

Acknowledgments

This research was supported by SVO grant, project number 95600. The Hague Center of Education assisted greatly in the collecting of the data.

References

- Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123–140.
- Bereiter, C., & Scardamalia, M. (1979). Pascual-Leone's M construct as a link between cognitive-developmental and psychometric concepts of intelligence. *Intelligence*, 3, 41–63.
- Bidell, T.R., & Fischer, K.W. (1992). Beyond the stage debate: Action, structure, and variability in Piagetian theory and research. In R.J. Sternberg & C.A. Berg (Eds.), *Intellectual development* (pp. 100–140). Cambridge: Cambridge University Press.
- Brown, A.L., Campione, J.C., Reeve, R.A., Ferrara, R.A., & Palinscar, A.S. (1991). Interactive learning and individual understanding: The case of reading and mathematics. In L.T. Landsmann (Ed.), *Culture, schooling, and psychological development* (pp. 136–170). Norwood, NJ: Ablex Publishing Corporation.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- 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.
- Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, 6, 544–573.
- CITO. (1995). *Luistertoets. Handleiding*. [Listening Comprehension Test. Manual]. Arnhem: Author.
- Csapó, B. (1999). Improving thinking through the content of teaching. In J.H.M. Hamers, J.E.H. van Luit, & B. Csapó (Eds.), *Teaching and learning thinking skills* (pp. 37–63). Lisse: Swets & Zeitlinger.
- De Koning, E. (2000). *Inductive reasoning in primary education. Measurement, teaching, transfer*. Zeist: Kerckbosch.
- De Koning, E., & Hamers, J.H.M. (1995). *Programma Inductief Redeneren 1* [Program Inductive Reasoning 1]. 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., Hamers, J.H.M., Sijtsma, K., & Vermeer, A. (2002). Teaching and transfer of inductive reasoning in primary education. *Developmental Review*, 22, 211–241.
- De Koning, E., Sijtsma, K., & Hamers, J.H.M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement*, 26, 302–320.
- Dodwel, P.C. (1960). Children's understanding of number and related concepts. *Canadian Journal of Psychology*, 14, 191–205.
- Engelen, R.J.H., & Eggen, T.J.H.M. (1993). Equivaleren [Equating]. In T.J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* [Psychometrics into practice] (pp. 309–348). Arnhem: CITO Instituut voor Toetsontwikkeling.
- Evans, T.G. (1968). A program for the solution of a class of geometric analogy intelligence test questions. In M. Minsky (Ed.),

- Semantic information processing* (pp. 271–353). Cambridge, MA: MIT Press.
- Feurstein, R., Rand, Y., Jensen, M.R., Kaniel, S., & Tzuriel, D. (1987). Prerequisites for assessment of learning potential: The LPAD model. In C.S. Lidz (Ed.), *Dynamic Assessment: An interactional approach to evaluating learning potential* (pp. 35–51). New York: Guilford.
- Glas, C.A.W., & Ouborg, M.J. (1993). Vraagzekerheid [Differential item functioning]. In J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* [Psychometrics into practice] (pp. 349–370). Arnhem: CITO Instituut voor Toetsontwikkeling.
- 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.
- Glas, C.A.W., & Ellis, J.L. (1993). *User's manual RSP. Rasch Scaling Program*. Groningen, The Netherlands: iecProGAMMA.
- Goswami, U. (1991). Analogical Reasoning: What develops? A review of research and theory. *Child Development*, 62, 1–22.
- Grigorenko, E.L. & Sternberg, R.J. (1998). Dynamic testing. *Psychological Bulletin*, 124, 1, 75–111.
- Hamers, J.H.M., De Koning E., & Ruijsenaars, A.J.J.M. (1997). A diagnostic program as learning potential assessment procedure. *Educational and Child Psychology*, 14, 73–82
- Hamers, J.H.M., De Koning, E., & Sijtsma, K. (1998). Inductive reasoning in the third grade: Intervention promises and constraints. *Contemporary Educational Psychology*, 23, 132–148.
- Holland, J.H., Holyoak, H.J., Nisbett, R.E., & Thagard, P.R. (1986). *Induction. Processes of inference, learning and discovery*. Cambridge, MA: MIT Press.
- Hosenfeld, B., Van den Boom, D.C., & Resing, W. (1997). New Instrument. Constructing geometric analogies for the longitudinal testing of elementary school children. *Journal of Educational Measurement*, 34, 4, 367–372.
- Hunt, E.B. (1974). Quote the Raven? Nevermore! In L.W. Gregg (Ed.), *Knowledge and cognition* (pp. 129–158). Hillsdale, NJ: Erlbaum.
- Klauer, K.J. (1989). *Denktraining für Kinder 1. Ein Program zur intellektuellen Förderung* [Inductive reasoning. A program for the stimulation of inductive reasoning]. Göttingen: Hogrefe.
- Klauer, K. J. (1990). A process theory of inductive reasoning tested by the teaching of domain-specific thinking strategies. *European Journal of Psychology of Education*, 5, 191–206.
- Klauer, K.J. (1997). Lässt sich die Strategie des induktiven Denkens auf schulisches Lernen transferierbar lehren? [Can the strategy to reason inductively be taught such that it transfers to learning of school-type material?] *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 29, 225–241.
- Klauer, K.J. (1999). Über den Einfluss des induktiven Denkens auf den Erwerb unanschaulich-generischen Wissens bei Grund- und Sonderschülern. [On the impact of inductive reasoning on the acquisition of abstract generic knowledge with elementary school and with learning disabled children.] *Psychologie in Erziehung und Unterricht*, 46, 7–28.
- Marshalek, B., Lohman, D.F., & Snow, R.E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, 7, 107–127.
- 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*. The Hague: Mouton/Berlin: De Gruyter.
- Mokken, R.J. (1997). Nonparametric models for dichotomous responses. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York: Springer-Verlag.
- Molenaar, I.W., & Sijtsma, K., (2000). *MSP5 for Windows. User's manual*. Groningen, The Netherlands: iecProGAMMA.
- Mulholland, T.M., Pellegrino, J.W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12, 252–284.
- Nisbett, R.E. (1993). *Rules for reasoning*. Hillsdale, NJ: Erlbaum.
- Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Palinscar, A.S., & Brown, A.L. (1988). Teaching and practical thinking skills to promote comprehension in the context of group problem solving. *RASE: Remedial and Special Education*, 9, 1, 53–59.
- Pascual-Leone, L. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica*, 32, 4, 301–345.
- Pennings, A.H., & Hessels, M.G.P. (1996). The measurement of mental attentional capacity: A Neo-Piagetian developmental study. *Intelligence*, 23, 1, 59–78.
- Piaget, J. (1970). Piaget's theory. In P.H. Mussen (Ed.), *Carmichael's handbook of child development* (pp. 703–732). New York: Wiley.
- Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, 66, 437–460.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche.
- Raven, J.C. (1958). *Standard Progressive Matrices*. London: Lewis.
- Richardson, K. (1996). Putting Raven into context: A response to Roberts & Stevenson. *British Journal of Educational Psychology*, 66, 533–538
- Roberts, M.J., & Stevenson, N.J. (1996). Reasoning with Raven – with and without help. *British Journal of Educational Psychology*, 66, 519–532.
- Sijtsma, K. (1983). Rasch-homogeniteit empirisch onderzocht [Rasch homogeneity empirically examined]. *Tijdschrift voor Onderwijsresearch*, 8, 104–121.
- Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- Sijtsma, J. (1992). *Balans van het taalonderwijs halverwege de basisschool* [Evaluation of language education half-way primary school]. Arnhem: CITO Instituut voor Toetsontwikkeling.
- Snow, R.E., Kyllonen, P.C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R.J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 47–103). Hillsdale, NJ: Erlbaum.
- Spearman, C. (1927). *The abilities of man*. New York: Macmillan.
- Sternberg, R.J. (1998). When will the milk spoil? Everyday induction in human intelligence. *Intelligence*, 25, 3, 185–203.
- Sternberg, R.J., & Gardner, M.K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112, 1, 80–116.

-
- Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293–325.
- Van den Wollenberg, A.L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. *Applied Psychological Measurement*, 6, 83–91.
- Van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Veldhuijzen, N.H., Godebeld, P., & Sanders, P.F. (1993). Klassieke testtheorie en generaliseerbaarheidstheorie. [Classic test theory and generalizability theory]. In T.J.H.M. Eggen & P.F. Sanders (Eds.), *Psychometrie in de praktijk* [Psychometrics into practice] (pp. 33–82). Arnhem: CITO Instituut voor Toetsontwikkeling.
- 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-Verlag.
- Verhoeven, L. (1996). *Woordenschattoets I. Handleiding*. [Vocabulary Test I. Manual]. Arnhem: CITO.
- Vernon, P.E. (1971). *The structure of human abilities*. London: Methuen.
- Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge acquisition: A developmental perspective. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 413–437). Cambridge: Cambridge University Press.
- Wijnstra, J. (1987). *De samenstelling van de schoolbevolking in het basisonderwijs*. [The composition of the school population in primary education.] Arnhem: CITO Instituut voor Toetsontwikkeling.
- Willmes, K., Heller, K.A., & Lengfelder, A. (1997). Testrezenion zu Standard Progressive Matrices. [A review of the Standard Progressive Matrices (SPM).] *Zeitschrift für Differentielle und Diagnostische Psychologie*, 18, 117–120.
- Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, 20, 99–149.
-
- Els de Koning
Department of Education and Youth Studies, FSW
Leiden University
P.O. Box 9555
2300 RB Leiden
The Netherlands
Tel. +31 71 527-3400
Fax +31 71 527-3619
Email koninge@fsw.leidenuniv.nl
-
- Klaas Sijtsma
Department of Methodology and Statistics, FSW
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands
Tel. +31 13 466-3222
Fax +31 13 466-3002
Email k.sijtsma@kub.nl
-
- Jo H.M. Hamers
Department of Special Education, FSW
Utrecht University
P.O. Box 80140
3508 TC Utrecht
The Netherlands
Tel. +31 30 253-4611
Fax +31 30 253-7731
Email j.hamers@fss.uu.nl
-