Tilburg University

Constructing a transitive reasoning test for 6 - to - 13 year old children

Bouwmeester, S.; Sijtsma, K.

Published in:
European Journal of Psychological Assessment

Publication date:
2006

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

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

Citation for published version (APA):
Bouwmeester, S., & Sijtsma, K. (2006). Constructing a transitive reasoning test for 6 - to - 13 year old children. *European Journal of Psychological Assessment*, 22(4), 225-232.

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.

Constructing a Transitive Reasoning Test for 6- to 13-Year-Old Children

Samantha Bouwmeester¹ and Klaas Sijtsma²

¹Erasmus University, ²Tilburg University, both The Netherlands

Abstract. A new, computerized transitive reasoning test was constructed using 16 well-structured, theory-based tasks. The test was administered to 615 elementary school children. Within-subjects ANOVA showed that task format and presentation form influenced task difficulty level. Mokken scale analysis supported a unidimensional scale that was reliable. Evidence was collected for an invariant task ordering. The misfit of two pseudotransitivity tasks supported discriminant validity.

Keywords: developmental scale, item response theory, Mokken scale analysis, task characteristics, transitive reasoning

Introduction

In everyday life we constantly infer transitive relationships between different agents, such as: If Paris has a sunnier climate than Amsterdam and Madrid is sunnier than Paris, then the combination of these two premises implies that Madrid is sunnier than Amsterdam. Simple as transitive reasoning may seem, it still is unknown when young children are first able to draw transitive inferences or how researchers can reliably measure differences. This study reports on the psychometric properties of a new computer test for transitive reasoning that may be helpful in resolving these problems.

Formally, a transitive reasoning task requires the inference of the unknown relationship R between two agents A and C from their known relationships with a third agent B; that is, (RAB, RBC) → RAC. The relationships RAB and RBC are the premises. When children are capable of drawing a transitive inference from the premises, they are capable of transitive reasoning.

Three Theories on Transitive Reasoning

Piaget's Theory

According to Piaget (1947), children are capable of transitive reasoning once they understand the necessity of using logical rules, know how to use them, and can remember the premises. This allows them to infer any transitive relationship. This understanding is acquired at the concrete operational stage, at approximately 7 years of age. At the preoperational stage, at 2 through 7 years of age, children do not understand the necessity of using logical rules. Instead, objects and their characteristics are considered at a nominal level, that is, unrelated to other objects (Piaget, 1942), and transitive reasoning is not yet feasible.

In the early 1960s, the age boundaries of the developmental stages were criticized, which led to disagreement about the age of emergence of transitive reasoning (see, e.g., Bryant & Trabasso, 1971; Trabasso, 1977). As different kinds of transitive reasoning tasks were used, conflicting results were found both for the age of emergence and the processes involved in transitive reasoning.

Information Processing Theory

Information processing theory posits that the age of emergence of cognitive abilities is not determined by biological maturation but by a child's experience in a specific content area (e.g., Case, 1996). The focus is on the presentation of information and on how it is transformed and stored given limited memory capacity. It is assumed that changes in thinking are induced by a process of continuous self-modification induced by outcomes generated by a child's own activities. Representation of knowledge becomes more abstract with age. These self-modifying processes eliminate the need to account for specific age-defined transition periods (Siegler, 1991).

Bryant and Trabasso (1971) provided evidence that failure to draw transitive inferences at the preoperational stage is the result of a memory deficit rather than logical reasoning limitations. They trained children to memorize the premises and found support that 4- and 5-year-old children were able of transitive reasoning. Trabasso, Riley, and Wilson (1975) showed that children must be able to integrate the premises into a linear ordering, and Trabasso (1977) showed that the transitive relationship is read rather than inferred from this ordering. The efficiency with which encoded information is represented determines whether memory capacity is sufficient to retrieve information.

Fuzzy Trace Theory

Fuzzy trace theory (Brainerd & Reyna, 2004) assumes that information is reduced to the essence, and that a kind of gist is formed to solve a cognitive task. The level of exactness of encoded information varies along a continuum. One end is defined by *fuzzy traces*, which are vague, degenerate representations that conserve only the sense of recently encoded data in a schematic way. The other end is defined by *verbatim traces*, which are literal representations that preserve the content of recently encoded data with exactitude. For example, premise information encoded in verbatim traces is stored literally as "A is longer than B, and B is longer than C." An example of a fuzzy trace is: "things get longer to the left." Because the retention of verbatim traces requires much memory capacity such traces are mostly unavailable; and because fuzzy traces are schematic, longer retention is possible and they are more readily available (Brainerd & Reyna, 2004). Brainerd and Kingma (1984) argued that transitive reasoning is primarily based on the use of fuzzy traces, mainly for reasons of efficiency.

Transitive Reasoning Tasks

Task Characteristics

Based on the formal, logical definition of transitivity a great variety of tasks are possible. Different task characteristics may have differential effects on children's task performance, and are likely to result in different conclusions about transitive reasoning ability. This is known as the criterion problem (see, e.g., Thayer & Collyer, 1978).

Tasks may vary with respect to the property or *content* on which objects are compared. For example, Piaget and Inhelder (1948) used the physical properties of length and weight; Trabasso, Riley, and Wilson (1975), Kallio (1982), and DeBoysson-Bardies and O'Regan (1973) used length; and Piaget (1973) and Verweij, Sijtsma, and Koops (1999) also used size. Riley (1976) used human properties of happiness and niceness and told subjects which object was happier or nicer.

Tasks may also differ with respect to the number of objects: Piaget and Inhelder (1941) and Brainerd (1974) used three objects; Halford and Kelly (1984) used four; Trabasso, Riley, and Wilson (1975), DeBoysson-Bardies and O'Regan (1973) and Perner, Steiner, and Staehelin (1981) used five; and Verweij et al. (1999) used either three, four, or five objects.

Objects may be equal or unequal with respect to content, and within the same task some objects may be equal while others are unequal. Let Y denote the property, and YA the amount object A has; and so on. For example, Piaget (1961) used both inequality ($YA > YB > YC$) and equality tasks ($YA = YB = YC = YD$). Youniss and Murray (1970), and Brainerd (1973) used mixed format ($YA > YB = YC$). The

combination of number of objects and their formal relationships determines task *format*.

Finally, Trabasso, Riley, and Wilson (1975) and Brainerd and Reyna (1990) presented premises in the presence of the other objects in the task (objects were far enough apart so that length differences were not perceptible). This is simultaneous presentation. Chapman and Lindenberger (1988) and Verweij et al. (1999) presented premises successively, only showing the two objects of the pair.

Influence of Task Characteristics on Performance

Piaget's theory assumes that task performance depends only on the execution of logical rules and that differences in performance depend on developmental stage. Information processing theory assumes that content, format, and presentation form influence the encoding of information and the formation of internal representations. Compared to physically perceptible relationships (e.g., length differences) verbally communicated relationships (e.g., differences in happiness) may ask for more articulated levels of formal thinking than is possible before age 12. Thus, encoding of verbal information may be different from encoding of visual information. Further, it may be easier to form an internal representation of the premises when they involve only inequalities instead of both inequalities and equalities (cf. $YA > YB > YC > YD > YE$ and $YA = YB > YC = YD$). Also, the larger the number of premises involving an inequality, the more difficult it may be to represent the task internally (cf. solution of RAE from $YA > YB > YC > YD > YE$ involving four premises with solution of RAC from $YA > YB > YC$ involving two premises). Finally, simultaneously presented information requires less memory capacity than successively presented information (Brainerd & Reyna, 1990).

Fuzzy trace theory assumes that pattern information is more difficult to recognize for mixed format (e.g., $YA = YB > YC = YD$) than for equality format ($YA = YB = YC = YD$), which can be reduced to the gist "all objects are the same." Different task formats ask for differential use of fuzzy and verbatim trace continua (Brainerd & Reyna, 1990). For example, inference of an ordering of objects is more difficult when premises are presented successively instead of simultaneously (also, see Verweij et al., 1999).

Choice of Tasks in the Present Study

We constructed 16 tasks based on the literature on Piaget's theory (Chapman & Lindenberger, 1988; Piaget, 1942), information processing theory (Bryant & Trabasso, 1971; Harris & Bassett, 1975; Murray & Youniss, 1968; Youniss & Furth, 1973), and fuzzy trace theory (Brainerd & Kingma, 1984). These tasks differed with respect to three characteristics (Figure 1).

Format		Simultaneous Presentation (sim)	Successive Presentation (suc)	Test-pair
$Y_a < Y_b < Y_c$ (3)	Physical content (phys)	Task 13	Task 6	Task 6
$Y_a = Y_b = Y_c = Y_d$ (eq)		Task 7	Task 9	Task 9
$Y_a < Y_b < Y_c < Y_d < Y_e$ (5)		Task 4	Task 15	Task 15
$Y_a = Y_b < Y_c = Y_d$ (mix)		Task 11	Task 2	Task 2
$Y_a > Y_b > Y_c$ (3)	Verbal content (verb)	Task 1	Task 12	Task 12
$Y_a = Y_b = Y_c = Y_d$ (eq)		Task 16	Task 3	Task 3
$Y_a > Y_b > Y_c > Y_d > Y_e$ (5)		Task 10	Task 8	Task 8
$Y_a = Y_b > Y_c = Y_d$ (mix)		Task 5	Task 14	Task 14

Figure 1. Tasks of the Transitive Reasoning Test.

Factor Format had four levels: $Y_A > Y_B > Y_C$; $Y_A > Y_B > Y_C > Y_D > Y_E$; $Y_A = Y_B = Y_C = Y_D$; and $Y_A = Y_B > Y_C = Y_D$. In the 3-object task ($Y_A > Y_B > Y_C$), Object A had a higher content level than the other objects; thus, it could be labeled "large". In the 5-object task ($Y_A > Y_B > Y_C > Y_D > Y_E$), Object B had a lower content level than A and a higher content level than C, thus, B could not be labeled uniquely. This was expected to render 5-object tasks more difficult. Factor Content had two levels: Objects were sticks that could differ in length (i.e., physical content) or animals that could differ in age (i.e., verbal content). Age rather than happiness (Riley, 1976) was used because it is more concrete and reduces the risk of error caused by interindividual differences in interpretation. Factor Presentation form had two levels: simultaneous presentation and successive presentation.

Each task was a unique combination of the three factors; thus, there were $4 \times 2 \times 2$ tasks. It was expected that successive presentation was more difficult than simultaneous presentation, verbal content more difficult than physical content, and formats $Y_A > Y_B > Y_C > Y_D > Y_E$ and $Y_A = Y_B > Y_C = Y_D$ more difficult than $Y_A > Y_B > Y_C$ and $Y_A = Y_B = Y_C = Y_D$.

It was investigated whether the 16 tasks constituted a unidimensional scale. If so, tasks represent different difficulty levels, and children can be ordered provided their scale score is reliable. Traditional research has typically distinguished two discrete categories of able and unable

children making it difficult to find a unique age at which transitive reasoning ability emerges. A continuous scale provides information about individual differences between children of the same age and between children of different ages, and may shed a different light on this discussion.

Method

Sample

The sample consisted of 615 children from middle class socioeconomic status families, attending Grade 2 through Grade 6 of six Dutch elementary schools (Table 1).

Instrument

The transitive reasoning computer test was individually administered. A computer test could be better standardized than an *in vivo* test. Moreover, movements and sounds could be implemented to enhance the test's attractiveness and hold the child's attention. An in-depth pilot study considered possible differences between computerized and *in vivo* task presentation. The latter used wooden sticks of different color and length. The administration procedure was the same as that of the computerized administration but took more time. Incorrect/correct responses and verbal

Table 1. Number of children, mean age (*M*), and standard deviation (*SD*) by grade

Grade	Number	Age	
		<i>M^a</i>	<i>SD</i>
2	108	95.48	7.81
3	119	108.48	5.53
4	122	119.13	5.37
5	143	132.81	5.17
6	123	144.95	5.34
Total	615	121.26	18.08

^anumber of months

explanations of these responses did not differ between both administration modes.

Task order was the same for each child. Difficult and easy tasks were alternated to keep children motivated. The same sticks or animals were used in several tasks in order to standardize tasks as much as possible. This could have had the effect of confusing children, but a pilot study showed no evidence of such confusion. Nevertheless, tasks sharing the same objects were alternated as much as possible with tasks having different objects. Tasks were also alternated with respect to task characteristics.

Administration and Scoring Procedures

Children tried three exercises to get used to the program and the tasks. Next, they were administered the 16 transitive reasoning tasks and two pseudotransitivity tasks. With the transitive reasoning tasks, children had to choose (by clicking on a button) the longest stick or the eldest animal from the presented pair, or click on the equality button when they decided that the sticks/animals had the same length/age. The format of the pseudotransitivity tasks was (YA > YB, YC > YD) and (YA = YB, YC = YD), thus they appear like real transitive reasoning tasks but leaving RBC unidentified. Thus, neither task allowed inference of a transitive relationship.

Piaget held the opinion that children were capable of operational reasoning when they could mention aloud all the premises involved (Piaget & Inhelder, 1941; Piaget, Inhelder, & Szeminska, 1948; Piaget, 1961), and verified this by asking children to verbally explain their solutions. Chapman and Lindenberger (1992) assumed children to be able of transitive reasoning when they were able to explain the judgment (e.g., A is longer than C or A and C are equally long). Information-processing theory hypothesized that verbal explanations interfered with cognitive processes (see e.g., Brainerd, 1977). Also, internal representations were not assumed to be necessarily verbal.

The discrepancy of *judgment-only* (i.e., incorrect/correct transitive inferences) and *judgment-plus-explanation* (i.e., incorrect/correct explanations of inferences) approaches can be formulated in terms of type I and type II errors

(Smedslund, 1969). Given the null-hypothesis that transitive reasoning ability is absent, a correct judgment-only response resulting from guessing may evoke a type I error (i.e., a false positive). Likewise, assuming presence of transitive reasoning ability, an incorrect verbal explanation caused by underdeveloped verbal ability may evoke a type II error (i.e., a false negative). Because both types of scoring seem to be both informative and problematic at the same time we decided to collect both (Sijtsma & Verweij, 1999). Judgment-only scores were recorded automatically by the computer, and explanations given by the children after they had clicked on their preferred response were recorded in writing by the experimenter. Judgment-only scores reflected whether the solution was correct (score 1) or incorrect (score 0). Judgment -plus-explanation scores were 1 when a correct explanation of the judgment was given, and 0 when an incorrect explanation or no explanation at all was given. Bouwmeester, Sijtsma, and Vermunt (2004) found that children often gave explanations not providing evidence of transitive reasoning even when they had given a correct judgment.

The pseudotransitivity tasks had no correct answer but an explanation of how a pseudotransitivity task was handled could be incorrect or correct. Hence, these tasks were only included for validation purposes in the judgment -plus-explanation data.

Data Analysis

Within-subjects ANOVA was used to assess the influence of task characteristics on task performance. The Rasch model and two less restrictive nonparametric item-response models were fitted to the judgment -plus-explanation data, the judgment-only-data, and the judgment -plus-explanation data including the two pseudotransitivity tasks.

Results

Influence of Task Characteristics on Task Performance

The *p*-values (sample proportions of correct explanations) ranged from .01 to .86 (Table 2). Proportions of correct explanations given with a correct solution ranged from .75 to 1.00, and proportions of correct solutions given with an incorrect explanation ranged from .14 to .76 (Table 2).

Within-subjects ANOVA showed that all main and interaction effects of the task characteristics were significant ($p < .001$) (Table 3). Effect size was evaluated by means of partial η^2 (Stevens, 1996, p. 1771). Effect sizes were large for Format (partial $\eta^2 = 0.72$) and Presentation form (partial $\eta^2 = .65$), and for Format $\times$ Presentation form (partial $\eta^2 = 0.21$) and Format $\times$ Content $\times$ Presentation form (partial $\eta^2 = 0.32$). Effect sizes were modest for Content (partial

Table 2. Sample proportions of correct explanations, of correct explanation given correct solution, and of correct solution given incorrect explanation of the 16 tasks

Task	Presentation	Format	Content	Proportion		
				Correct explanation	Correct explanation given correct solution	Correct solution given incorrect explanation
1	simultaneous	YA > YB > YC	verbal	.49	.95	.49
13	simultaneous	YA > YB > YC	physical	.57	.97	.34
12	successive	YA > YB > YC	verbal	.39	.96	.39
6	successive	YA > YB > YC	physical	.05	.97	.39
16	simultaneous	YA = YB = YC = YD	verbal	.86	1.00	.55
7	simultaneous	YA = YB = YC = YD	physical	.77	1.00	.43
3	successive	YA = YB = YC = YD	verbal	.44	.98	.76
9	successive	YA = YB = YC = YD	physical	.54	.99	.51
10	simultaneous	YA > YB > YC > YD > YE	verbal	.51	.86	.14
4	simultaneous	YA > YB > YC > YD > YE	physical	.39	.99	.27
8	successive	YA > YB > YC > YD > YE	verbal	.21	.98	.40
15	successive	YA > YB > YC > YD > YE	physical	.07	.88	.45
5	simultaneous	YA = YB > YC = YD	verbal	.14	.75	.16
11	simultaneous	YA = YB > YC = YD	physical	.30	.89	.16
14	successive	YA = YB > YC = YD	verbal	.18	.88	.30
2	successive	YA = YB > YC = YD	physical	.01	1.00	.47

Note: The proportion “correct solution given incorrect explanation” for Task 3 is large (.76). For this task many children explained that they had already seen the test pair during the premise presentation stage. Although this was an incorrect inference, it often led to a correct solution.

Table 3. Tests of within-subjects effects

Effect	df1	df2	F	p	Partial η^2
Format	3	1842	659.76	<.001	0.52
Presentation	1	614	1122.86	<.001	0.65
Content	1	614	68.55	<.001	0.10
Format $\times$ presentation	3	1842	41.80	<.001	0.06
Format $\times$ content	3	1842	31.60	<.001	0.05
Presentation $\times$ content	1	614	90.66	<.001	0.13
Format $\times$ pres $\times$ cont	3	1842	108.08	<.001	0.15

$\eta^2 = 0.10$), and Format $\times$ Content (partial $\eta^2 = 0.12$) and Content $\times$ Presentation form (partial $\eta^2 = 0.13$). Physical content was more difficult than verbal content. Successive presentation was more difficult than simultaneous presentation.

Post hoc analysis was done establishing 95% confidence intervals of the means. Bonferroni adjustment was used to correct the significance level to 0.05/82. Format YA = YB = YC = YD was significantly easier than the other formats. Format YA = YB > YC = YD was the most difficult. Formats YA > YB > YC and YA > YB > YC > YD > YE showed the smallest significant differences. For each format, simultaneous presentation was easier than successive

presentation. The difference between the presentation forms was smaller for format YA = YB > YC = YD than for other formats. Physical content was more difficult than verbal content for formats YA > YB > YC and YA > YB > YC > YD > YE. No significant difference was found for formats YA = YB = YC = YD and YA = YB > YC = YD. Verbal and physical content did not differ significantly for simultaneous presentation. Physical content was more difficult for successive presentation. In particular the combination of physical content and successive presentation made tasks very difficult for formats YA > YB > YC, YA > YB > YC > YD > YE, and YA = YB > YC = YD, but not for format YA = YB = YC = YD.

Item Response Theory Analysis

Rasch Model Analysis

Rasch Model

Let random variable X_j denote the score (0, 1) on task j . The Rasch model assumes that the probability of $X_j = 1$ conditional on a unidimensional latent ability, denoted θ , depends on the task's difficulty level, denoted δ_j :

1 Partial $\eta^2 = 0.01$ was interpreted as small, partial $\eta^2 = 0.06$ as medium, and partial $\eta^2 = 0.14$ as large according to Stevens (1996, p. 177; based on Cohen, 1977, pp. 284–288).

$$P(x_j = 1|\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}.$$

This conditional probability is the item response function. The Rasch Scaling Program (RSP; Glas & Ellis, 1994) uses the asymptotic χ^2 statistic R_1 for testing the null-hypothesis that all item response functions are parallel logistic functions, and the approximate χ^2 statistic Q_2 for testing local independence of the multivariate conditional distribution of the task scores (Glas & Verhelst, 1995). Together, these statistics constitute a full test of the fit of the Rasch model to the data generated by the tasks in the test.

Results

After deletion of cases for which only 0 or 1 task scores were recorded sufficiently large samples remained. The Rasch model was rejected for each of the three data sets: Judgment-plus-explanation data ($R_1 = 94$, $df = 60$, $p = .004$; $Q_2 = 1671$, $df = 520$, $p = .000$), judgment-only data ($R_1 = 217$, $df = 60$, $p = .000$; $Q_2 = 1114$, $df = 520$, $p = .000$), and judgment-plus-explanation data including the two pseudotransitivity tasks ($R_1 = 193$, $df = 68$, $p = .000$; $Q_2 = 1552$, $df = 675$, $p = .000$).

Mokken Scale Analysis

Nonparametric Item Response Models

Unlike the Rasch model, less restrictive nonparametric models define the relationship between $P(X_j = 1|\theta)$ by means of order restrictions (Sijtsma & Molenaar, 2002) instead of a parametric function such as the logistic. The two nonparametric item response models that were used are based on the next three assumptions: Unidimensionality means that one latent ability parameter, θ , explains the data structure; local independence means that given a fixed θ value scores on different tasks are unrelated; and monotonicity means that the item response functions are monotone nondecreasing in θ . These three assumptions constitute the monotone homogeneity model (MHM). The MHM implies the stochastic ordering of persons on θ by means of their sum scores on the tasks (Sijtsma & Molenaar, 2002, p. 22).

Fit of the MHM was investigated using the program Mokken Scale analysis for Polytomous items (MSP; Molenaar & Sijtsma, 2000). Item response functions were estimated and evaluated with respect to monotonicity. Scalability coefficient H for the total test and task scalability coefficient H_j for separate tasks were estimated. Coefficient H is a weighted mean of the H_j s, and provides evidence about the degree to which subjects can be ordered by means of the sum score on the tasks. The MHM implies that $0 \leq H \leq 1$; a scale is considered weak if $0.3 \leq H \leq 0.4$, medium if $0.4 \leq H \leq 0.5$, and strong if $H \geq 0.5$. For individual tasks, a Mokken scale analysis requires that $H_j \geq 0.3$, for all j . Negatively correlating tasks cannot be part of the same scale. See Sijtsma and Molenaar (2002, chap. 5) for more details.

The double monotonicity model (DMM) is more restrictive because of the additional assumption of nonintersection of the item response functions. This assumption is identical to an invariant task ordering. This implies that the tasks have the same ordering for all values of latent ability θ with the exception of possible ties. Such an invariant task ordering greatly enhances the interpretation of test performance. See Sijtsma and Molenaar (2002, chap. 6) for examples.

Nonparametric item response models have been used to construct scales for cognitive abilities (e.g., De Koning, Sijtsma, & Hamers, 2003; Hosenfield, Van den Boom, & Resing, 1997; Verweij, Sijtsma, & Koops, 1996, 1999).

Results for Judgment-Plus-Explanation Data

Because of its extreme p value of .01, Task 2 had negative correlations with Tasks 8 and 15 and was rejected from the analysis. For the other 15 tasks, we found $0.37 \leq H_j \leq 0.66$. The item-restscore regressions, which estimated the item response functions, did not show significant decreases. This supported monotonicity. Overall scalability coefficient H was 0.45, indicating a medium-strength scale.

Cronbach's α was 0.83. Based on H and the H_j s and other analyses (not reported) it was concluded that the 15 tasks formed a unidimensional scale. Thus, all tasks evaluated the same ability and children could be reliably ordered by ability level using the sum score based on the number of correct explanations.

Nonintersection of item response functions was investigated by means of the H -coefficient of the transposed data matrix (which has tasks in the rows and children in the columns), denoted H^T coefficient (Sijtsma & Molenaar, 2002, pp. 107–109). For an invariant task ordering, $H^T > .3$ and the percentage of negative person $H^T_{\alpha s}$ (α is a person index) must not exceed 10. The H^T -coefficient for the scale was 0.52, and the percentage of negative $H^T_{\alpha s}$ -values was 1.6. These results supported nonintersection of item response functions, indicating invariant ordering of the 15 tasks.

Results for Judgment-Only Data

For the 16 tasks, $.01 \leq H_j \leq .25$, and $H = 0.16$. These results indicated that the tasks did not form a practically useful scale. Consequently, the more restrictive DMM was not fitted. Cronbach's α was 0.63, indicating weak reliability.

Results for Judgment-Plus-Explanation Data Including Pseudo-Transitivity Tasks

Both pseudotransitivity tasks had several negative correlations with several transitive reasoning tasks, and were rejected from the analysis. Their H_j s were 0.03 and 0.14, which was another reason for not including them in the scale. A DMM analysis was not useful here.

Discussion

Task format and presentation form influenced task difficulty level. Mixed inequality-equality tasks were more difficult than inequality tasks, and in general equality tasks were easier than inequality tasks. These findings disagree with Piaget's theory but agree with both information processing theory and fuzzy trace theory. Simultaneous presentation was easier than successive presentation. Each of the three theories predict this result, but for different reasons. Piaget's theory assumes that children need functional reasoning acquired in the preoperational stage for inferring relationships when premise presentation is simultaneous. Operational reasoning, acquired in the concrete-operational stage, is needed to infer relationships when presentation form is successive. Instead of two qualitatively different abilities, information processing theory and fuzzy trace theory assume that successive presentation requires more memory capacity than simultaneous presentation, and that this results in more errors. This study gave no evidence of two qualitatively different abilities. Also the combination of task characteristics influenced task difficulty level. In particular, combination of physical content and successive presentation rendered a task difficult.

Because of the misfit of the Rasch model the linear logistic test model (Fischer, 1995), which is a Rasch model with linear restrictions on the task parameters, could not be used to investigate the influence of the task characteristics on task difficulty level. Bouwmeester et al. (2004) showed that different, ordered latent classes could be distinguished in which the task characteristics had differential influence on use of solution strategies.

For the judgment-plus-explanation data, 15 tasks formed a scale on which children can be ordered reliably. The scale also allows an invariant task ordering. This means that the ordering of the tasks by *p*-values is the same for all children and, by implication, all subgroups of children (e.g., grades).

The combination of mixed task-format, physical content, and successive presentation rendered Task 2 extremely difficult. Consequently, the expected covariances of Task 2 with other tasks were approximately 0. Negative covariances were the result of sampling fluctuation. The conclusion was that Task 2 was too difficult for the transitive reasoning test.

The tasks were based on substantive theory about transitive reasoning. The unidimensionality of the data, thus, provided support for convergent validity. The misfit of the pseudotransitivity tasks supported discriminant validity. These validity results are an indication of construct validity. More research supporting construct validity is needed. The tasks were not scaleable under the judgment-only scoring scheme.

Task performance was found to be unidimensional, and three task characteristics influenced task difficulty considerably. Thus, age of emergence of transitive reasoning greatly depends on task difficulty. This conclusion explains

why researchers (e.g., Bryant & Trabasso, 1971; DeBoysson-Bardies and O'Regan, 1973; Youniss and Murray, 1970) who used transitive reasoning tasks with different characteristics reached different conclusions about age of emergence.

Whether the results can be generalized to children younger than 7 years is unknown. Both children of the same age and children of different ages are highly different in transitive reasoning ability. These findings call into question the usefulness of efforts to investigate age of emergence.

References

- Bouwmeester, S., Sijtsma, K., & Vermunt, J.K. (2004). Latent class regression analysis to describe cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology*, 1, 67–86.
- Brainerd, C.J. (1973). Judgments and explanations as criteria for the presence of cognitive structures. *Psychological Bulletin*, 3, 172–179.
- Brainerd, C.J. (1974). Training and transfer of transitivity, conservation, and class inclusion of length. *Child Development*, 45, 324–334.
- Brainerd, C.J. (1977). Response criteria in concept development research. *Child Development*, 48, 360–366.
- Brainerd, C.J., & Kingma, J. (1984). Do children have to remember to reason? A fuzzy trace theory of transitivity development. *Developmental Review*, 4, 311–377.
- Brainerd, C.J., & Reyna, V.F. (2004). Perspectives in behavior and cognition. *Developmental Review*, 24, 396–439.
- Bryant, P.E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, 232, 456–458.
- Case, R. (1996). Changing views of knowledge and their impact on educational research and practice. In D.R. Olson & N. Torrance (Eds.), *The handbook of education and human development* (pp. 75–99). Cambridge, MA: Blackwell.
- Chapman, M., & Lindenberger, U. (1988). Functions, operations, and décalage in the development of transitivity. *Developmental Psychology*, 24, 542–551.
- Chapman, M., & Lindenberger, U. (1992). Transitivity judgments, memory for premises, and models of children's reasoning. *Developmental Review*, 12, 124–163.
- 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, 24–39.
- DeBoysson-Bardies, B., & O'Regan, K. (1973). What children do in spite of adults' hypotheses. *Nature*, 246, 531–534.
- Fischer, G.H. (1995). The linear logistic test model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models, foundations, recent developments, and applications* (pp. 131–155). New York: Springer.
- Glas, C.A.W., & Ellis, J.L. (1994). *Rasch scaling program*. Groningen, The Netherlands: iecProGamma.
- Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models, foundations, recent developments, and applications* (pp. 69–95). New York: Springer.

- Halford, G.S., & Kelly, M.E. (1984). On the basis of early transitivity judgments. *Journal of Experimental Child Psychology*, 38, 42–63.
- Harris, P.L., & Bassett, E. (1975). Transitive inferences by 4-year-old children. *Developmental Review*, 11, 875–876.
- Hosenfield, B., Van den Boom, D.C., & Resing, W.C.M. (1997). Constructing geometric analogies for the longitudinal testing of elementary school children. *Journal of Educational Measurement*, 34, 367–372.
- Kallio, K.D. (1982). Developmental change on a five-term transitivity inference. *Journal of Experimental Child Psychology*, 33, 142–164.
- Molenaar, I.W., & Sijtsma, K. (2000). *User's manual MSP5 for Windows. A program for Mokken Scale analysis for Polytomous items [software manual]*. Groningen, The Netherlands: iecProGamma.
- Murray, J.P., & Youniss, J. (1968). Achievement of inferential transitivity and its relation to serial ordering. *Child Development*, 39, 1259–1268.
- Perner, J., Steiner, G., & Staehelin, C. (1981). Mental representation of length and weight series and transitive inferences in young children. *Journal of Experimental Child Psychology*, 31, 177–192.
- Piaget, J. (1942). *Classes, relations et nombres: Essai sur les groupement logistique et sur la réversibilité de la pensée* [Classes, relations, and names: Essay about logical grouping and the reversibility of thinking]. Paris: Collin.
- Piaget, J. (1947). *La psychologie de l'intelligence* [The psychology of intelligence]. Paris: Collin.
- Piaget, J. (1961). *Les mécanismes perceptives* [The perceptual mechanisms]. Paris: Presses Universitaires de France.
- Piaget, J., & Inhelder, B. (1941). *Le développement des quantités chez l'enfant* [The child's development of quantities]. Neuchâtel: Delachaux et Niestlé.
- Piaget, J., Inhelder, B., & Szeminska, A. (1948). *La géométrie spontanée de l'enfant* [The spontaneous geometric of the child]. Paris: Presses Universitaires de France.
- Riley, C.A. (1976). The representation of comparative relations and transitive inference task. *Journal of Experimental Child Psychology*, 22, 1–22.
- Riley, C.A., & Trabasso, T. (1974). Comparatives, logical structures, and encoding in a transitive inference task. *Journal of Experimental Child Psychology*, 17, 187–203.
- Siegler, R.S. (1991). *Children's thinking, second edition*. New Jersey: Prentice-Hall.
- Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
- Sijtsma, K., & Verweij, A.C. (1999). Knowledge of solution and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement*, 23, 55–68.
- Smedslund, J. (1969). Psychological diagnostics. *Psychological Bulletin*, 71, 237–248.
- Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.
- Thayer, E.S., & Collyer, C.E. (1978). The development of transitive inference: A review of recent approaches. *Psychological Bulletin*, 85, 327–343.
- Trabasso, T. (1977). The role of memory as a system in making transitive inferences. In R.V. Kail, J.W. Hagen, & J.M. Belmont (Eds.), *Perspectives on the development of memory and cognition* (pp. 333–366). Hillsdale, NJ: Erlbaum.
- Trabasso, T., Riley, C.A., & Wilson, E.G. (1975). The representation of linear order and spatial strategies in reasoning: A developmental study. In R.J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 201–229). Hillsdale, NJ: Erlbaum.
- 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.
- Verweij, A.C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development*, 23, 241–264.
- Youniss, J., & Furth, H.G. (1973). Reasoning and Piaget. *Nature*, 244, 314–316.
- Youniss, J., & Murray, J.P. (1970). Transitive inference with non-transitive solutions controlled. *Developmental Psychology*, 2, 169–175.

Samantha Bouwmeester

Institute for Psychology, FSW
Erasmus University
Burgemeester Oudlaan 50
NL-3062 PA Rotterdam
The Netherlands
Tel. +31 10 4-082-795
E-mail bouwmeester@fsw.eur.nl