Latent variable modeling of cognitive processes in transitive reasoning

Bouwmeester, S.

Publication date:
2005

Latent Variable Modeling of Cognitive Processes in Transitive Reasoning

Samantha Bouwmeester

Samantha Bouwmeester

Latent Variable Modeling of Cognitive Processes in Transitive Reasoning

ISBN 90-9019557-2
Printed by PrintPartners Ipskamp, Enschede

Latent Variable Modeling of Cognitive Processes in Transitive Reasoning

(Het Modelleren van Latente Variabelen van Cognitieve Processen in het Transitief Redeneren)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg,
op gezag van de rector magnificus, prof. dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen ten overstaan van een door het college voor
promoties aangewezen commissie in de aula van de Universiteit
op vrijdag 1 juli 2005 om 14.15 uur
door

Samantha Bouwmeester

geboren op 4 november 1976 te Leiden

Promotores: Prof. dr. K. Sijtsma
Prof. dr. W. Koops

Dankwoord

Dit proefschrift is het resultaat van ruim vier jaar werken aan de Universiteit van Tilburg. In deze vier jaren hebben een groot aantal mensen mij, vaak zonder dat ze het zelf wisten, geïnspireerd, gemotiveerd, gestimuleerd en gesteund. Hen wil ik graag bedanken.

Als eerste bedank ik mijn Promotoren. Klaas, vanaf het begin van het project heb je me veel vertrouwen gegeven. Hierdoor kon ik zelfstandig mijn weg zoeken en mijn eigen keuzes maken. Aan de zijlijn was je aanwezig met stimulerende en kritische vragen die mij in staat stelden te groeien. Voor dit alles dank ik je zeer. Willem, ik heb onze afspraken als inspirerend en waardevol ervaren. Na een bezoek aan Utrecht was ik altijd weer enthousiast om verder te gaan.

Jeroen, ik wil je hartelijk danken voor je stimulerende inzet om mijn inhoudelijke vraagstukken te begrijpen en te vertalen naar één van jouw fascinerende latente klassen modellen. Zonder jouw hulp zou dit proefschrift er wezenlijk anders hebben uitgezien. Ton Aalbers van Spits wil ik uitdrukkelijk bedanken voor zijn hulp en de prettige samenwerking bij het programmeren van mijn testprogramma's. Zelfs op onmogelijke tijdstippen wist je een snelle oplossing te realiseren.

Dit proefschrift zou er niet zijn geweest zonder de enthousiaste medewerking van een aantal basisscholen. Ik bedank de leerkrachten en leerlingen van basisschool *Andreas*, *De Oase*, *Houtwijk*, *De Kameleon*, *De Vierboet* en *De Angelaschool*. Mijn speciale dank gaat uit naar basisschool *De Hobbitburcht* en basisschool *De Schapendel*, omdat ik op deze scholen zelfs twee keer mijn data heb verzameld. Daarnaast bedank ik alle kinderen die tijdens de pilot studies hebben meegewerkt aan het onderzoek.

Nina Banens wil ik bedanken voor het deel van de dataverzameling dat zij op zich heeft genomen. Nina, bedankt voor je gedreven en consciëntieuze inzet.

Zonder collega's zou het schrijven van een proefschrift maar saai zijn. Ik bedank daarom al mijn collega's van het departement MTO. Wilco, Sandra, Paqui en Marieke van Onna, bij jullie kon ik als groentje afkijken hoe het moest, promoveren! Joost en Marieke Spreeuwenberg zorgden als nieuwkomers voor nieuwe gesprekken aan de lunch- en koffietafel. Wicher, bedankt voor je antwoorden op al mijn (latex)vragen. Janneke, jij bracht heel veel gezelligheid in onze kamer. Wij hebben over alles gepraat en gelachen, dat maakte mijn UvT-dagen bijzonder de moeite waard!

Het is niet alleen maar leuk geweest de afgelopen vier jaar. Eén tikkie tegen mijn hoofd tijdens het basketballen en m'n leven zag er lange tijd compleet anders uit. Terugknokken, relativeren, vertrouwen en weer doorzetten. Zonder mijn familie en vrienden zou dit ondoenlijk zijn geweest. Ria bedankt, soms was promoveren niet anders dan een wedstrijd. Majida, de Thorung La vormt een mooie metafoor. Bedankt dat je achter me liep en toen het moest voor me. Hanneke, zonder en met woorden begrijpen wij elkaar, dat is genoeg. Evelien bedankt voor alles, en voor veel meer.

Lieve broer, in groep drie was ik al trots dat jij mijn grote broer was, en dat ben ik nog steeds. Lieve ouders, al te mededeelzaam ben ik niet geweest, dat weet ik best. Maar jullie lieten je niet uit het veld slaan. Bedankt voor jullie interesse, steun en onvoorwaardelijk vertrouwen.

Leiden, 20 April 2005,
Samantha Bouwmeester

Contents

Introduction

When I tell you that my brother's cat, Pooky, is older than his dog, Bente, and also that his goldfish, Blub, is younger than his dog, I hope you immediately inferred that Pooky is older than Blub. When you did, you used your ability of drawing a transitive inference, that is, you inferred an unknown relationship (Pooky is older than Blub) from known relationships (Pooky is older than Bente, and Bente is older than Blub). Adults are drawing transitive inferences several times a day, and they do this automatically and unconsciously. However, young children are not capable of drawing such inferences.

Formally, in a transitive reasoning task the unknown relationship, R, between two elements, A and C, can be inferred from their known relationships with a third element, B; that is, $(R_{AB}, R_{BC}) \Rightarrow R_{AC}$. In this example, the relationships R_{AB} and R_{BC} are premises. When children are capable of drawing a transitive inference from the premises, they are capable of transitive reasoning. Cognitive theories disagree about what transitive reasoning is about, which processes are involved, and which kinds of tasks should be used to measure it.

Piaget's Theory

According to Piaget, cognition is constructed by the active, originally sensori-motor, interaction between the child and the external world (Case, 1996; Chapman, 1988; Flavell, 1963). During development the interaction becomes more and more internalized and mental operations can be performed without real interaction with the external environment (Piaget, 1949). Groups of internalized actions form cognitive structures. During development these cognitive structures become less concrete and domain-specific, and more abstract, general and applicable to a broad domain. Piaget constructed cognitive tasks, such as transitive reasoning tasks, to investigate the developmental level of cognition in children (Chapman, 1988; Flavell, 1963). Cognitive development, according to Piaget's theory and research, in principle follows four discrete stages, the sensory motor stage, the preoperational stage, the concrete operational stage, and the formal operational stage. This theoretical framework can be found in any textbook on developmental psychology or cognitive development.

Children are capable of drawing transitive inferences when they understand the necessity of using logical rules. For example, if Y_A stands for the amount object A (e.g., a stick) has of property Y (e.g., length), then $Y_A > Y_B$ and $Y_B > Y_C$ together imply $Y_A > Y_C$. When children know how to use these rules of logic, they are able to solve any transitive relationship as long as they can remember the premises. This understanding is acquired at the concrete operational stage, at about seven years of age (Piaget, 1947), when the cognitive structure of children is for the first time characterized by the reversibility principle (Piaget, 1942, 1947). A transitive inference beautifully demonstrates this reversibility principle: when A is larger than B, B must be smaller than A, and when we know that A is longer than B, and C is shorter than B, then we can use the reversibility principle to conclude that A is longer than C. Children at the preoperational stage, at two through seven years of age (Piaget, 1947), do not understand the reversibility principle. Objects or characteristics of objects are considered in a nominal way, that is, not in relationship to other objects (Piaget, 1942). Due to this nominal thinking, or preoperational thinking in Piagetian jargon, children are not capable of performing internalized operations on objects and they do not understand the necessity of using logical rules. When a cue is provided about the ordering of the objects in a task, an understanding of logical rules may not be necessary to solve the task. For example, the position of the objects can be used for inferring their mutual relationships when all objects are presented simultaneously and ordered on the dimension on which they differ. This kind of reasoning is called functional reasoning. Functional reasoning is typical of the preoperational stage.

Piaget's theory was not meant to be a psychological theory. He was interested in the general, biological development of cognitive structures of the human being in general or the individual child in particular without emphasizing task conditions (Bidell & Fischer, 1992). In accordance with research traditions of their time, Piaget and his colleagues preferred a clinical method to investigate the development of intelligence by using interviews without standardization and statistical data analysis (Flavell, 1963). The Neo-Piagetians maintained the constructivistic assumptions of the theory but attempted to operationalize the constructs empirically by taking variations in tasks and individuals into account (Case, 1992, pp. 166).

Reaction to Piaget

In the early 1960s, the age boundaries of the developmental stages according to Piaget's theory were the first source of criticism of cognitive psychologists. Braine (1959) showed that after the child had learned the premises, (s)he was able to draw transitive inferences at five years of age. His finding evoked a thorough discussion. Braine (1959) argued that remembering the premises was the real problem for young children, not logical reasoning. However, Smedslund (1963, 1965, 1969) argued that Braine's results could be explained alternatively by a *labelling strategy*, according to which children use a nominal label of an object to solve the task. For example, during the premise presentation object A may be encoded as 'short' and object C as 'long'. As a result, the answer that C is longer than A can be inferred from the labels 'long' and 'short', without making use of the relationships within the object pairs A, B and B, C. In their research, Brainerd (1973) and Youniss and Denisson (1971) used Müller-Lyer illusion techniques to prevent children from using this labelling strategy. Youniss and his colleagues (Murray & Youniss, 1968; Youniss & Furth, 1973; Youniss & Murray, 1970) used mixed-format ($Y_A = Y_B > Y_C = Y_D$) relationships. In this kind of tasks, the objects did not have a unique label (object C is both smaller than object B and equally long as object D), so the labelling strategy could not be used. However, Brainerd (1973) argued that illusion and mixed-format tasks confused children and interfered with the reasoning process.

Information Processing Theory

Bryant and Trabasso (1971) used five-objects inequality-format tasks ($Y_A > Y_B > Y_C > Y_D > Y_E$) in which labelling strategies could not be used to solve the transitive relationship R_{BD}. They showed that after an intensive training children were able to draw transitive inferences at the age of five. Bryant and Trabasso (1971) and Riley and Trabasso (1974) explained their results by a linear ordering theory in which children form a symbolic internal representation of the objects and the relationships between the objects. This representation is used to infer the answer. Trabasso (1977) used reaction time to show that the linear ordering theory could explain how an internal representation was formed for drawing inferences without the use of logical rules.

The Neo-Piagetians were not convinced by the results of Trabasso and his colleagues. Perner, Steiner and Staehelin (1981), Perner and Mansbridge (1983), and Perner and Aebi (1985) argued that the visual feedback, the presentation form, and the intensive training lead to specific task conditions in which a labelling strategy could be used to solve the transitive relationship. Chapman (1988) and Chapman and Lindenberger (1992) argued that the simultaneous presentation of the premises provided a positional cue about the ordering of the objects. By means of the intensive training of the premises, children had learned the ordering and drew inferences on the basis of this ordering. This kind of reasoning was functional instead of operational, because children did not need the reversibility principle to solve the transitive relationship.

Although the criticism of information-processing theorists was directed initially at the age boundaries of Piaget's theory, neglect of individual differences, poor experimental setting, and neglect of environmental influences, the most important difference appeared to be the epistemological assumptions of both theoretical approaches. These assumptions led to conflicting requirements of specific task conditions, which explains the gap of two years between the ages at which transitive reasoning first emerged according to the two theories.

Fuzzy Trace Theory

Piaget and the Neo-Piagetians assumed that memory is a necessary but not a sufficient condition for using logical rules. Information-processing theorists assumed that memory of the premises is sufficient for drawing a transitive inference. A strong argument for the hypothesis that memory of the premises is not necessary for drawing a transitive inference is made by fuzzy trace theory (Brainerd & Kingma, 1984, 1985; Brainerd & Reyna, 1993, 2001).

Fuzzy trace theory assumes that human cognition is a parallel encoding mechanism of information at different levels of abstraction (Brainerd & Reyna, 1990, 1995, 2004). 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 degenerated, "fuzzy", way. The other end is defined by verbatim traces, which are literal representations that preserve the content of recently encoded information with exactitude. Because retention of vivid, verbatim traces requires much memory capacity, these traces usually are not available. The information in a fuzzy trace, however, is reduced and schematic, so longer retention is possible and the fuzzy trace is more easily available. People prefer to reason fuzzy rather than verbatim, because the degraded information from the fuzzy-trace is more easily accessible and costs less memory capacity.

The characteristics of a task determine which level of the continuum can be used to solve the transitive relationship. When a cue about the ordering is provided, the fuzzier end of the continuum can be used, which contains a degenerated representation of the objects, for example, "objects get smaller to the left". When cues are absent, it is difficult to reduce information and the verbatim end of the continuum is used. This makes the task more difficult because the literal premise information has to be remembered. When the fuzzy end of the continuum can be used, memory of the premises is not needed. Brainerd and Kingma (1984, 1985) showed that transitive reasoning is primarily based on the schematic information of fuzzy traces.

The Neo-Piagetians Chapman and Lindenberger (1992) argued that fuzzy trace theory only applies to tasks in which a cue is provided about the ordering of objects, that is, tasks which can be solved using functional reasoning. When such cues are not provided, memory of the premises is necessary for applying logical rules, that is, to reason operationally. Brainerd and Reyna (1992) did not distinguish operational and functional reasoning as separate abilities. They argued that reducing information is more difficult when cues about an ordering are absent, and that people attempt to use the fuzziest trace possible.

Issues in Transitive Reasoning

The three theories have different ideas about what cognitive development is and how change in behavior should be measured. Piaget assumed a hierarchical structure in which children are viewed as imperfect adults which have to pass the necessary stages to reach formal thinking. The thinking of children in different stages deviates qualitatively due to the different forms of the cognitive structures.

According to information processing theory, however, the child's thinking deviates from adult's thinking only in a quantitative way. The processing of information is slower and less efficient leading to incomplete, impoverished internal representations of the information. Development, in this respect, is reduced to accumulative learning of internal stimulus-response relations.

Fuzzy trace theory was developed as a reaction to information processing theory's computer-based approach to cognitive development. According to fuzzy trace theory information is processed simultaneously, automatically and unconsciously at a variety of levels which differ in the degree of exactness of the information. Cognitive development is assumed to be the growing capability to retrieve the appropriate level of information given the task requirements. Note that this level is not necessarily a complete or detailed representation of the information involved as is assumed in information processing theory.

The way the theories view development has important consequences for the study of development in transitive reasoning. We not only have to define what development means but also what transitive reasoning is. In this thesis I tried to disentangle the underlying response processes involved in the development of transitive reasoning by taking individual differences and task characteristics into account. I started bottom-up, that is, I did not choose one of the theories as a framework for transitive reasoning but evaluated the different theories by means of the latent structure in empirical data. In the last chapter a top-down approach was followed. Fuzzy trace theory was used as a theoretical model to describe the underlying response process at a detailed level.

Construction of a Scale for Transitive Reasoning

First I constructed a computerized test containing 16 transitive reasoning tasks. Based on earlier research, these tasks were varied on three characteristics which were found to influence the cognitive processes and the accompanying performance. Two pseudo-transitive reasoning tasks were included in the test. They resembled the transitive reasoning tasks, but were different because a transitive relationship could not be inferred from the premise information. The test was administered to a sample of 615 elementary school students ranging from grade two to grade six stemming from six schools in The Netherlands. Both the correct/incorrect answers and the explanations of the answers given by the students were analyzed. Chapter 1 reports the results of a Mokken (1971) scale analysis that was applied to the 16 transitive reasoning tasks in an effort to determine the quality of these tasks and the reliability of the ordering of the students by means of their test score.

Abilities Involved in Transitive Reasoning

Piaget's theory, information processing theory, and fuzzy trace theory posit different ideas about the underlying processes involved in transitive reasoning and the influence of task characteristics on the difficulty of a task. According to Piaget's theory and the Neo-Piagetians, two kinds of reasoning have to be distinguished, functional and operational reasoning, representing qualitatively different abilities. The characteristics of the task determine which type of reasoning is needed. Information processing theory, most extensively elaborated by Trabasso and his colleagues, assumes one underlying ability. Also, the theory assumes that the difficulty of a task is determined by the ease by which the premises are remembered. Fuzzy trace theory also assumes one underlying ability, which is the fuzzy trace ability, but according to this theory task difficulty is determined by the ease by which the ordering of the objects in a task is recognized. Chapter 2 reports an empirical study on the number of abilities involved in transitive reasoning. Three methods are used for this purpose (represented in the computer programs MSP, DETECT, and improved DIMTEST). Multiple regression is used to determine the influence of task characteristics on the difficulty level of the tasks. Moreover, the usefulness of both the correct/incorrect scores and the correct/incorrect explanations is compared.

Continuous or Discontinuous Change?

Another, important topic is whether cognitive development is stage-like, as assumed in Piaget's theory, or continuous without jumpy transitions from one stage to another. When studying a single ability instead of complete cognitive structures, discontinuity can be defined as the existence of a number of modes ordered along the developmental scale which correspond with different rules or strategies that are used to solve particular tasks. In chapter 3, I first discuss a number of research issues typical of the study of developmental change and discontinuity. Then discontinuity is studied in cross-sectional transitive reasoning data. Two statistical mixture models, the binomial mixture model and the latent class factor model, are compared. Unlike the binomial mixture model, the latent class model does not assume binomial distributions, allows task difficulties to be different, and uses the information in the individual's item-score patterns to estimate class probabilities. Next, additional analysis are done to interpret the discontinuity, and this lends meaning to the classes that are distinguished on the basis of the latent class analysis.

Latent Cognitive Variables, Environmental Influences, Cognitive Behavior and Age

In chapter 3 the emphasis is on determining discontinuity in transitive reasoning, and in chapter 4 on the detailed interpretation of latent cognitive classes by means of manifest variables such as age, cognitive behavior, and environmental influences. Again, developmental groups are distinguished but at a more detailed level of sophistication. In this chapter the usefulness of the latent class regression model for studying cognitive developmental phenomena is discussed. Using this model, the relationships between latent and manifest variables can be explained by means of empirical data without the need for strong a priori assumptions made by a cognitive developmental theory. In the latent class regression model a number of classes are distinguished which are characterized by particular cognitive behavior. Task characteristics influence cognitive behavior and this influence varies over different (developmental) classes.

Fuzzy Trace Theory as a Framework for Explaining Individual Differences

Fuzzy trace theory offers a detailed description of the performance on both the memory of the premises and the inference of transitive relationships in transitive reasoning tasks (see Brainerd & Kingma, 1984, 1985; Brainerd & Reyna, 1995). This opens the possibility to test empirically and in great detail the application of the theory in the context of transitive reasoning. In chapter 5 fuzzy trace theory is used as the theoretical framework for modeling both individual differences in performance and task influences on performance on memory test-pairs and transitivity test-pairs. A test is constructed containing four replications of each of three kinds of tasks, each having four memory-of-the-premises items, and three transitive-relationship items. The three task types differ in difficulty with respect to the position of objects and the presentation of the premises. Both the position and the presentation can be ordered or disordered, but the combination of disordered position and disordered presentation is not used because it would render tasks too difficult. The test was administered to a new sample of 409 students ranging in age from 5 to 13 years and stemming from four elementary schools in The Netherlands. Per student 84 responses are used to determine both the verbatim and fuzzy ability levels. Because the retrieval of verbatim and fuzzy traces is dependent on the verbatim and fuzzy ability levels, and the responses to the items of the tasks are dependent on the verbatim and fuzzy traces used, a multilevel latent class model (Vermunt, 2003) is used for data analysis.

Chapter 1

Constructing a Transitive Reasoning Test for Six to Thirteen Year Old Children

1.1 Introduction

The aim of this chapter is to report on the construction of a transitive reasoning test for elementary school students. In a transitive reasoning task, the unknown relationship R between two elements A and C can be inferred from their known relationships with a third element B; that is, $(R_{AB}, R_{BC}) \Rightarrow R_{AC}$. In this example, the relationships R_{AB} and R_{BC} are premises. When children are capable of drawing a transitive inference from the premises, they are capable of transitive reasoning.

1.1.1 *Tasks of the Test*

Researchers used various kinds of tasks for studying the development of transitive reasoning (see, e.g., Bryant & Trabasso, 1971; Chapman & Lindenberger, 1988; Harris & Bassett, 1975; Kallio, 1982;

This chapter has been submitted for publication.

Murray & Youniss, 1968; Perner & Mansbridge, 1983; Perner et al., 1981; Smedslund, 1963; Youniss & Murray, 1970; Verweij, Sijtsma, & Koops, 1999). For our test (see Figure 2.1, chapter 2), we constructed 16 tasks. Each task consisted of objects that had to be compared with respect to a property, such as length. This property was denoted Y, and the value of object A on Y was denoted Y_A, et cetera. Tasks differed with respect to three task characteristics. These characteristics were frequently used by researchers representative of different theoretical approaches (see, e.g., Brainerd & Kingma, 1984; Bryant & Trabasso, 1971; Chapman & Lindenberger, 1988; Harris & Bassett, 1975; Murray & Youniss, 1968; Piaget, 1942; Youniss & Furth, 1973).

The task characteristic *format* determined the kind of transitive relationship. The four levels of *format* were: $Y_A > Y_B > Y_C$; $Y_A = Y_B = Y_C = Y_D$; $Y_A > Y_B > Y_C > Y_D > Y_E$; and $Y_A = Y_B > Y_C = Y_D$. Although the formats $Y_A > Y_B > Y_C$ and $Y_A > Y_B > Y_C > Y_D > Y_E$ differed only in the number of objects involved, they were expected to differ in difficulty. For example, in the 3-object task, object A was always large in comparison with other objects and could therefore be labelled as large. In the 5-object task, object B was small compared with object A and large compared with object C, so that object B did not have a unique label. This difference was expected to produce greater difficulty for 5-object tasks. The task characteristic *presentation* determined whether the premises were presented all together (simultaneously) or one after the other (successively). The task characteristic *content* determined whether the objects that formed the premises were sticks that could differ in length (physical type of content) or animals that could differ in age (verbal type of content). Each task in the test was a unique combination of the three characteristics, such that each of the $4 \times 2 \times 2$ possibilities were represented. The difficulty level of the tasks was determined by the combination of the task characteristics.

The test was administered by computer to 615 students sampled from grade two through grade six in elementary school. First, the students did three exercises to get used to the program, the objects, and the relationships involved. Then, they performed the 16 transitive reasoning tasks and two additional pseudo-transitive reasoning tasks. These latter two tasks resembled the transitive reasoning tasks, but were different because a transitive relationship could not be inferred from the premise information. The format of the two pseudo-transitive reasoning tasks was $(Y_A > Y_B, Y_C > Y_D)$ and $(Y_A = Y_B, Y_C = Y_D)$, in both cases leaving the relationship between B and C unidentified.

Students were asked to click on the longest stick, the eldest animal, or the equality button when they thought that the sticks/animals had the same length/age. In each item, they had to choose one from three options. Children received a 1-score when they correctly explained the transitive relationship, and a 0-score when they gave an incorrect explanation or no explanation at all. Verweij (1994) showed that students often gave non-transitive explanations even when they had chosen the right option. The computer registered the option chosen and the experimenter recorded the verbal explanations.

1.2 Background Analyses

The P-values (sample proportions of correct explanations[1]) of the 16 tasks ranged from 0.01 to 0.86. A within-subject ANOVA showed that all main effects and interaction effects of the task characteristics and combinations of task characteristics were significantly ($p < .001$). Because of the large sample size ($N = 615$) these significant results offered little information about the importance of task characteristics or combinations of them. Partial η^2 (Stevens, 1996, p. 177[2]) was used for expressing effect size. The effect sizes were large for the characteristics *presentation* (partial $\eta^2 = .65$) and *format* (partial $\eta^2 = 0.72$), and for the interactions *presentation*×*format* (partial $\eta^2 = 0.21$) and *presentation*×*format*×*content* (partial $\eta^2 = 0.32$). The effect sizes were modest for the characteristic *presentation* (partial $\eta^2 = 0.1$), and the interactions *presentation*×*content* (partial $\eta^2 = 0.13$)

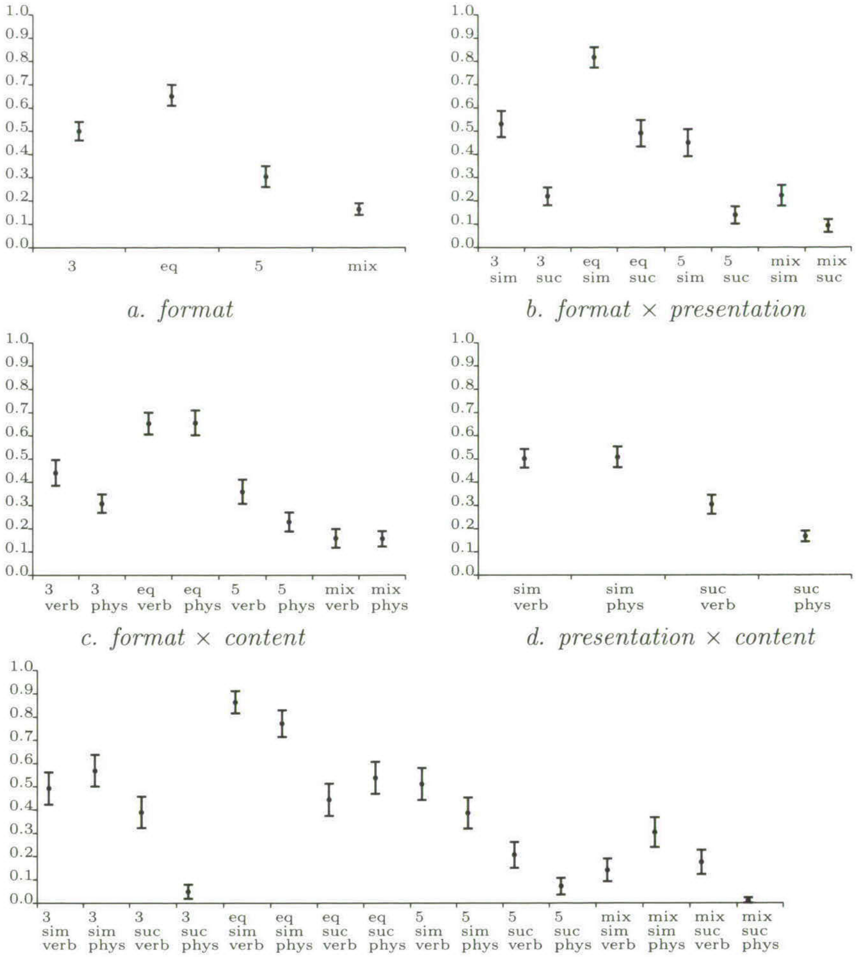
[1] Correct explanations were preceded by correctly chosen options 96% of the time.

[2] Following Stevens (1996, p. 177; based on Cohen, 1977, pp. 284-288), partial $\eta^2 = 0.01$ was interpreted as small, partial $\eta^2 = 0.06$ as medium, and partial $\eta^2 = 0.14$ as large.

and *format*×*content* (partial $\eta^2 = 0.12$). Successive presentation was more difficult than simultaneous presentation. Physical content was more difficult than verbal content. Post hoc analyses were performed to determine to which difference the significant effects could be attributed. The 95% confidence intervals (CIs) of the means are displayed in Figure 1.1 (standard error of the mean based on N=615). Because the number of statistical tests was 82, the significance level was adjusted to 0.05/82 (Bonferroni adjustment).

Figure 1.1a shows that format $Y_A = Y_B = Y_C = Y_D$ is significantly easier than the other formats. Format $Y_A = Y_B > Y_C = Y_D$ is the most difficult, and the formats $Y_A > Y_B > Y_C$ and $Y_A > Y_B > Y_C > Y_D > Y_E$ differ the least but significantly. Figure 1.1b shows that for each format, simultaneous presentation is easier than successive presentation, and that the difference between the two kinds of presentation is smaller for the format $Y_A = Y_B > Y_C = Y_D$ than for the other formats. Figure 1.1c shows that physical content is more difficult for the formats $Y_A > Y_B > Y_C$ and $Y_A > Y_B > Y_C > Y_D > Y_E$, but that there is no significant difference for formats $Y_A = Y_B = Y_C = Y_D$ and $Y_A = Y_B > Y_C = Y_D$. Figure 1.1d shows that verbal and physical content do not differ significantly when presentation is simultaneous, but that physical content is more difficult when presentation is successive. Figure 1.1e shows that in particular the combination of successive presentation and physical content makes the task very difficult for the formats $Y_A > Y_B > Y_C$, $Y_A > Y_B > Y_C > Y_D > Y_E$, and $Y_A = Y_B > Y_C = Y_D$, but not for format $Y_A = Y_B = Y_C = Y_D$.

Table 1.1 gives for each grade the mean test score, the standard deviation, and Cronbach's alpha. The Levene (1960) Test (W) showed that the variances were not equal for the five grades [$W(4, 610) = 3.49$, $p < .01$]. A procedure for comparing means, which takes unequal variances into account (Welch, 1951), revealed that the mean test scores increased with grade level, [$F(4, 610) = 43.66$, $p < .01$]. The 95% CI of the post hoc tests of adjacent grades (using Bonferroni adjustment) showed that only the mean test scores of Grade four and Grade five did not differ significantly (CI: -1.48 - 0.45). A comparison of the alpha coefficients (see Feldt,



a. format

b. format × presentation

c. format × content

d. presentation × content

e. format × presentation × content

Figure 1.1: *95% Confidence Intervals (CIs) for the Item Means Combined for Various Task Characteristics and Combinations of Task Characteristics* ($3 : Y_A > Y_B > Y_C$; *eq* : $Y_A = Y_B = Y_C = Y_D$; $5 : Y_A > Y_B > Y_C > Y_D > Y_E$; *mix* : $Y_A = Y_B > Y_C = Y_D$; *suc: successive; sim: simultaneous; verb: verbal; phys: physical)*

Woodruff, & Salih, 1987) showed that none of the coefficients differed significantly from any of the others.

Table 1.1: *For Each Grade, Mean Test Score, Standard Deviation (SD) and Cronbach's Alpha Based on 15 Tasks*[*]

Grade	*n*	*M*	*SD*	*Alpha*
2	108	3.29	2.74	.79
3	119	4.49	2.96	.77
4	122	6.40	3.67	.84
5	143	6.91	3.04	.76
6	123	7.98	3.06	.77

[*] Task 2 had zero variance in most grades.

1.3 Mokken Scale Analyses

We applied Mokken (1971) scale analysis in an effort to find support for the hypotheses that an increase in test score implies developmental progress, and that the ordering of students by test score is reliable. Mokken scale analysis is based on nonparametric item response theory (IRT; see Sijtsma & Molenaar, 2002). Nonparametric IRT defines the relationship between an observed item score and a latent trait by means of order restrictions, whereas parametric IRT models use a parametric function such as the logistic (Embretson & Reise, 2000).

The nonparametric IRT model that is the basis of a Mokken scale is defined by three assumptions: unidimensionality, local independence and monotonicity. Unidimensionality means that one latent trait parameter θ suffices to explain the data structure. Local independence means that, given a fixed θ value, responses to different tasks are unrelated. Monotonicity means that the item response functions are monotone increasing in θ. This implies an ordering of the students along the scale which, theoretically, is invariant over items. These three assumptions constitute the monotone homogeneity model (MHM). The double monotonicity model (DMM) is a more restrictive model in which a fourth assumption of non-intersection of the item response function is added to the other three. This assumption is identical to an invariant item ordering (Sijtsma & Molenaar, 2002, chap. 6). Several researchers used Mokken scale analysis to construct scales for cognitive abilities (e.g., De Koning, Sijtsma, & Hamers, 2003; Hosenfield, Van den Boom, & Resing, 1997). Verweij, Sijtsma, and Koops (1996, 1999) used Mokken scale analysis to construct a scale for transitive reasoning that used only formal content tasks and item scoring based on a more restricted conceptualization of transitive reasoning.

We used the program MSP (Molenaar & Sijtsma, 2000) to analyze the scalability of our transitive reasoning items. Scalability coefficient H (Mokken, 1971) was used to evaluate the scalability for the total test, and item scalability coefficient, H_j, was used to evaluate separate items. H is a weighted mean of the H_js and provides evidence about the degree to which subjects can be ordered by means of the complete set of tasks. The MHM implies that $0 \leq H \leq 1$; a scale is considered weak if $0.3 \leq H < 0.4$, medium if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$ (Sijtsma & Molenaar, 2002, p. 60). For individual items, a Mokken scale analysis requires that $H_j \geq 0.3$, for all j.

Task 2 was rejected from the analysis, because it had a negative covariance with both tasks 8 and 15 (negative covariances are in conflict with the monotonicity assumption). For the remaining 15 tasks, the task scalability coefficients ranged from 0.37 to 0.66. The overall scalability coefficient H was 0.45, thus indicating a medium scale.

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

The assumption of non-intersection of item response functions was investigated by means of the H-coefficient of the transposed task-person matrix, denoted H^T (Sijtsma & Meijer, 1992). To conclude that the items have an invariant item ordering, Sijtsma and Meijer (1992) recommended that $H^T > .3$ and the percentage of negative person H_a^T must not exceed 10. The H^T-coefficient for the total scale was 0.52, and the percentage of negative H_a^T-values for individuals was 1.6. Together these results support the assumption of non-intersecting item response functions, and this indicated that the tasks could be ordered invariantly.

Next, exploratory Mokken scale analysis was conducted for each grade separately under the restriction that items were only admitted to a scale if their $H_j \geq 0.3$ relative to the other items in that scale. Table 1.2 shows that the scales for Grades two, three, and five, contained nine items, in Grade four, the scale contained 14 items and in Grade six, the scale contained 11 items. The items formed a weak scale in Grade five, a medium scale in the Grades three, four, and six, and a strong scale in Grade two. The H^T values were sufficiently high and the percentages of negative H_a^Ts sufficiently low to conclude that the items had an invariant item ordering.

Table 1.2: *For Each Grade, Number of Tasks in the Scale, Scalability Coefficients H and H^T, and Percentage of Negative H_a^Ts*

Grade	# tasks	H	H^T	% neg.H_a^T
2	9	.54	.57	1.1
3	9	.48	.60	1.0
4	14	.49	.53	.9
5	9	.37	.54	2.2
6	11	.45	.63	.0

Furthermore, we investigated the scalability of the correct-incorrect task scores (these are the task scores that do not take the verbal explanations into account). Based on the 16 tasks, Cronbach's alpha was 0.63, indicating weak reliability. The task H_js varied from 0.01 through 0.25, and the overall scalability coefficient H was 0.16, indicating that the tasks did not form a practically useful scale.

The format of the two pseudo-transitive reasoning tasks was ($Y_A >$

$Y_B, Y_C > Y_D$) and ($Y_A = Y_B, Y_C = Y_D$), in both cases leaving the relationship between B and C unidentified. Thus, these tasks cannot be solved by means of the strategy used to solve real transitive reasoning tasks and, therefore, these tasks were not expected to fit into the transitive reasoning scale. A second Mokken scale analysis was conducted on the data of the 16 tasks and the two pseudo-transitive reasoning tasks to evaluate whether the scale had discriminant validity. The H_js of the two pseudo-transitive reasoning tasks were 0.03 and 0.14. Both tasks had several negative covariances with transitive reasoning tasks and were therefore rejected from the analysis.

1.4 Conclusion

We constructed a test for transitive reasoning containing 16 tasks which were varied systematically with respect to three three task characteristics, and found that in particular the presentation form and the task format influenced the task difficulty level. 15 of the 16 tasks formed a Mokken scale on which the students could be ordered reliable. Also, evidence was collected for an invariant item ordering; that is, an item ordering by means of P-values that is the same for all students and, by implication, all subgroups of students (e.g., grades). The finding that responses to the theory-based tasks were driven by one ability indicated convergent validity. The misfit of the pseudo-transitive reasoning tasks indicated discriminant validity. Together these convergent and discriminant validity results indicate construct validity (Campbell & Fiske, 1959), but more research supporting such a conclusion is needed. An analysis of the correct/incorrect scores without verbal explanations showed showed that the tasks were not scalable. Analyses of the data in separate grades showed a weak scale in one grade, medium scales in three grades, and a strong scale in one grade.

Chapter 2

Measuring the Ability of Transitive Reasoning, Using Product and Strategy Information

Abstract*

Cognitive theories disagree about the processes and the number of abilities involved in transitive reasoning. This led to controversies about the influence of task characteristics on individuals' performance and the development of transitive reasoning. In this study, both product and strategy information were analyzed to measure the performance of 6 to 13 year old children. Three methods (MSP, DETECT, and Improved DIMTEST) were used to determine the number of abilities involved and to test the assumptions imposed on the data by item response models. Nonparametric IRT models were used to construct a scale for transitive reasoning. Multiple regression was used to determine the influence of task characteristics on the difficulty level of the tasks. It was concluded that (1) the qualitatively distinct abilities predicted by Piaget's theory could not be distinguished by means of different dimensions in the data structure; (2) transitive reasoning could be described by one ability, and some task characteristics influenced

the difficulty of a task; and (3) strategy information provided a stronger scale than product information.

* This chapter has been published as: Bouwmeester, S., & Sijtsma, K. (2004) Measuring the Ability of Transitive Reasoning, Using Product and Strategy Information. *Psychometrika, 69, 123–146.*

2.1 Introduction

2.1.1 Definition of Transitive Reasoning

Suppose an experimenter shows a child two sticks, A and B, which differ in length, Y, such that $Y_A > Y_B$. Next, stick B is compared with another stick C which differs in length, such that $Y_B > Y_C$. In this example the length relationships $Y_A > Y_B$ and $Y_B > Y_C$ are the premises. When the child is asked, without being given the opportunity to visually compare this pair of sticks, which is longer, stick A or stick C, (s)he may or may not be able to give the correct answer. When a child is able to infer the unknown relationship ($Y_A > Y_C$) using the information of the premises ($Y_A > Y_B$ and $Y_B > Y_C$), (s)he is capable of *transitive reasoning.*

2.1.2 Theories of Transitive Reasoning

Three general theories on transitive reasoning can be distinguished. They are the developmental theory of Piaget, information processing theory, and fuzzy trace theory. These theories propose different definitions of the transitive reasoning ability and different operationalizations into transitive reasoning tasks. Consequently, the theories led to contradictory conclusions about children's transitive reasoning ability.

Developmental Theory of Piaget

According to Piaget's theory (Piaget, Inhelder, & Szeminska, 1948), children acquire the cognitive operations to understand rules of logic at the *concrete operational stage*, at about six or seven years of age. This understanding implies that an object can have different relationships with other objects. For example, a stick can be longer than a second stick and shorter than a third stick. This understanding is necessary to draw transitive inferences (Piaget & Inhelder, 1941; Piaget & Szeminska, 1941). At the *preoperational stage*, before the concrete operational stage, children think in a nominal way. This means that objects are understood in an absolute form, but not in relationship to other objects. Consequently, at this stage children are incapable of drawing a transitive inference.

Piaget distinguished two kinds of reasoning. To understand a transitive inference, the formal rules of logic had to be acquired and applied to the transitive reasoning problem. This kind of reasoning was called "operational reasoning". A child is able to reason in an operational way at the concrete operational stage. However, Piaget argued that operational reasoning is not necessary in each kind of task. When some kind of spatial cue in the task gives information about the ordering of objects (e.g., when all objects are presented simultaneously), operational reasoning is not required because the information given by the spatial cue can be used to infer the transitive relation; for example, objects become smaller from right to left. In this case, no formal rules have to be understood. Piaget called this kind of reasoning "functional reasoning". Functional reasoning is acquired at the preoperational stage. Piaget was in particular interested in the development of logical comprehension, and therefore used transitive reasoning tasks in which the premises were successively presented to be sure that children had to reason on an operational way. When a successive presentation of the premises is used, spatial cues about the ordering of objects are not available (although other kinds of ordering cues might be available).

Information Processing Theory

Although within information processing theory a broad diversity of ideas about information processing exists, differently oriented researchers on transitive reasoning do not make a distinction between functional and operational reasoning. An understanding of formal logical rules is not a necessary condition for drawing transitive inferences in any version of information processing theory. For example, in their linear ordering theory Trabasso, Riley, and Wilson (1975) and Trabasso (1977) emphasized the linear ordering in which the premise information was encoded and internally represented. Linear ordering was the only ability involved in transitive reasoning rendering it a one-dimensional construct. Task characteristics like presentation form (*simultaneous* or *successive*), task format (e.g., $Y_A > Y_B > Y_C$ and $Y_A = Y_B = Y_C = Y_D$), and content of the task (*physical*, like length; or *verbal*, like happiness) might influence the difficulty to form an internal representation, but the same ability is assumed for all kinds of transitive reasoning tasks.

Sternberg (1980a, 1980b) and Sternberg and Weil (1980) studied the development of linear syllogistic reasoning, a special form of transitive reasoning in which the premise information is presented verbally. Sternberg (1980b) showed that a mixed model, which contains both a linguistic component and a spatial component, could explain linear syllogistic test data (for alternative models, see also Clark, 1969; DeSoto, London, & Handel, 1965; Huttenlocher, 1968; Huttenlocher & Higgens, 1971; Quinton & Fellows, 1975; and Wright, 2001). According to this mixed model, both a verbal and a linear ordering ability are involved in solving linear syllogistic reasoning tasks. Premise information is first encoded linguistically, and then ordered spatially into an ordered internal representation.

Fuzzy Trace Theory

According to fuzzy trace theory (Brainerd & Kingma, 1985, 1984; Brainerd & Reyna, 1995, 2004), 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 information with exactitude. These verbatim traces contain information like: *there is a red object and a yellow object; the objects are vertical bars; and the red bar is longer than the yellow bar*. At the other end of the continuum, the information is stored in a degraded, schematic way;

for example, *objects get longer to the left* (Brainerd & Kingma, 1985; Brainerd & Reyna, 1995). The various levels of the continuum process in parallel; that is, by encoding literal information from a task, at the same time degraded fuzzy information is processed at several levels. Brainerd and Kingma (1984, 1985), and also Brainerd and Reyna (1995) showed that the fuzzy end, containing degraded information about the ordering of objects, was used to draw a transitive inference.

Fuzzy trace theory does not distinguish operational and functional reasoning (Brainerd & Reyna, 1992, see also Chapman & Lindenberger, 1992). It is assumed that task characteristics influence the level of the fuzzy trace continuum that may be used and, consequently, determine the difficulty level of a transitive reasoning task. No logical rules have to be applied and one ability, which is the ability to form and use fuzzy traces, explains an individual's performance on different kinds of tasks, rendering the construct of transitive reasoning a one-dimensional construct.

Comparison of Theories

Number of Abilities Involved The most important point of disagreement is *what* the ability to draw a transitive inference really *is*. Piaget distinguished operational and functional reasoning, two forms of reasoning that were qualitatively different, and acquired at different stages of cognitive development. Trabasso's (1975) linear ordering theory assumes one ability; that is, forming an internal representation of the objects is assumed to be one ability. Sternberg, who studied linear syllogistic reasoning, assumed a mixed model in which both a verbal and a spatial ability are involved. They are assumed to function as two separate abilities. Fuzzy trace theory also assumes one ability; that is, reasoning based on a fuzzy continuum.

From the perspective of Piaget's theory, information processing theory and fuzzy trace theory define transitive reasoning as a functional form of reasoning only applicable to a limited set of transitive reasoning tasks in which a linear ordering of the objects is given by a spatial cue. This functional reasoning does not require an understanding of transitivity, which is only acquired when children are capable of operational reasoning (Chapman & Lindenberger, 1988).

Influence of Task Characteristics on Difficulty Although not all theories make explicit predictions about the influence of task characteristics on the difficulty of a task[1], implications with respect to difficulty can be inferred from the theories' assumptions.

- *Piaget's Theory.* Firstly, because simultaneously presented tasks can be solved by functional reasoning while successively presented tasks must be solved by operational reasoning, from Piaget's theory it can be inferred that simultaneous presentation of the premises of a task is easier than successive presentation. Secondly, because the same logical rules are needed to solve equality, inequality or mixed equality-inequality task formats, the format of the task (e.g., $Y_A > Y_B > Y_C$, or $Y_A = Y_B = Y_C$) does not influence the difficulty of a task. Thirdly, because content of the relationship does not influence the application of logical rules, type of content does not influence the difficulty level of a task. However, Piaget first used length and then other concrete observable relationships to study transitive reasoning. He called the acquisition of understanding of different types of the same ability in different time periods *horizontal décalage* (Piaget, 1942). Therefore, as a fourth prediction it may be hypothesized that inferring a transitive relationship in a physical type-of-content task is easier than in a non-physical type-of-content task.

- *Information Processing Theory.* Firstly, the formation of a linear ordering and the memory of the premises are expected to be easier when the premises are presented simultaneously than when they are presented successively. Secondly, because it is more difficult to form a linear ordering of a mixed format task, it may be expected that mixed inequality-equality tasks are more difficult than equality or

[1] For example, in Piaget's theory the influence of external conditions (like task characteristics) on performance was hardly discussed.

inequality tasks. Although information processing theorists do not use equality format tasks to study transitive reasoning, these tasks may be expected to be easier than inequality-format tasks because the internal representation of an equality task is easier than the internal representation of an inequality task. Thirdly, according to the mixed model of Sternberg (1980b) both a verbal and a spatial ability are needed to solve linear syllogisms. For verbally presented tasks both abilities are required and for physical tasks only the spatial ability is required. Thus, it may be hypothesized that verbal tasks (linear syllogisms) are more difficult than physical tasks.

- *Fuzzy Trace Theory.* Firstly, because the retrieval of a fuzzy trace is easier for simultaneously presented tasks (which contain a spatial-order correlation) than for successively presented tasks (in which the ordering of the premises is less obvious) (Brainerd & Reyna, 1992), successive presentation is expected to be more difficult than simultaneous presentation. Secondly, because it is difficult to reduce the pattern information of the mixed inequality-equality format into a fuzzy trace, it can be hypothesized that the mixed inequality-equality format is more difficult than the equality or the inequality format. Thirdly, when a fuzzy trace is used to infer the transitive relationship only pattern information and no verbatim information (like type of content of tasks) is involved. Thus, different types of contents are not expected to influence the difficulty level.

A summary of the influence of task characteristics on the difficulty level according to the theories is given in Table 2.1.

Responses

Cognitive theories not only disagree about the kinds of tasks that should be used to measure transitive reasoning, but also about the types of responses that are required to verify that a child had really drawn a transitive inference. Piaget asked children to verbally explain their answers to verify whether a child has really used operational reasoning to solve a

Table 2.1: *Comparison of the Theories With Respect to the Number of Abilities and Influence of Task Characteristics on Difficulty Level of Tasks*

Theory	Topic	Predictions
Piaget	NUMBER OF ABILITIES:	two, functional and operational reasoning
	PRESENTATION:	successive more difficult than simultaneous
	FORMAT:	all formats same difficulty
	CONTENT:	verbal content more difficult than physical content
Information Processing	NUMBER OF ABILITIES:	one (linear ordering), two (mixed model)
	PRESENTATION:	successive more difficult than simultaneous
	FORMAT:	equality easier than other formats, mixed more difficult than other formats
	CONTENT:	verbal content more difficult than physical content
Fuzzy Trace	NUMBER OF ABILITIES:	one
	PRESENTATION:	successive more difficult than simultaneous
	FORMAT:	equality easier than other formats, mixed more difficult than other formats
	CONTENT:	physical content and verbal content equally difficult

transitive reasoning task. According to Piaget, children were capable of operational reasoning when they could mention aloud all the premises involved (Piaget & Inhelder, 1941; Piaget et al., 1948; Piaget, 1961). More recently Chapman and Lindenberger (1992) assumed a child to be able to draw a transitive inference when (s)he was able to explain the judgements. However, information processing theory hypothesized that the verbal explanations interfered with the cognitive processes (see e.g., Brainerd, 1977). Also, the internal representation was not assumed to be necessarily verbal. Instead, cognitive processes were measured using reaction times (e.g., Trabasso et al., 1975) or using the performance of children on specific task formats (e.g., Smedslund, 1963; Murray & Youniss, 1968).

When the aim of a study is to construct a transitive reasoning task for determining the age of emergence as exact as possible, using either the judgement or the judgement-plus-explanation may highly influence the result. For example, although a fair comparison between studies using different task formats could not be made, Bryant and Trabasso (1971) found children of only four years of age to be able of transitive reasoning, but

Chapman and Lindenberger (1992) did not find children able of transitive reasoning before the age of seven.

In fact, the discrepancy of judgment and judgment-plus-explanation approaches can be summarized as a choice between type I and type II errors (Smedslund, 1969). Given the null hypothesis that children do not have a transitive reasoning ability, a judgment-only response is prone to evoke a type I error (false positive), assuming that a child is able to draw a transitive inference when in fact it is not. However, when a verbal explanation is required, a type II error (false negative) is likely to occur, by assuming that a child is not able to draw a transitive inference when in fact it is. This inference may be caused by the child's underdeveloped verbal ability. When the aim of the study is to obtain an impression of the processes involved in the development of transitive reasoning, the explanations given by the child are useful, accepting the risk of a type II error and being somewhat conservative about the age of emergence. Using judgment-plus-explanation data, Verweij et al. (1999) showed that several transitive and non-transitive strategies were used to solve different kinds of transitive reasoning tasks. For several task types, different strategies led to correct answers.

2.1.3 Goal of Present Study

The disagreement about the number of abilities involved in transitive reasoning, the type of responses to be recorded, and the influence of task characteristics on task performance led to three hypotheses:

1. H_0: Two qualitatively different abilities, functional and operational reasoning, explain the response patterns on various tasks containing transitive relationships.

   H_A: One ability explains the response patterns on various transitive reasoning tasks. The tasks differ only in difficulty.

2. H_0: The response patterns based on strategy scores provide a better scale than the response patterns based on product scores (see Section 2.2.6, for a description of strategy and product scores).

H_A: Response patterns containing strategy scores and response patterns containing product scores both provide good scales.

3. H_0: The difficulty of transitive reasoning tasks is not influenced by task characteristics or combinations of task characteristics.
   H_A: The difficulty of transitive reasoning tasks is influenced by task characteristics or combinations of task characteristics.

For determining the number of abilities involved in transitive reasoning (first hypothesis), nonparametric item response theory (NIRT) methods (Molenaar & Sijtsma, 2000; Stout, 1993, 1996) were used to investigate the underlying dimensionality of a data set generated by means of a set of tasks having different characteristics. When one ability is involved, the task scores can be explained by one underlying dimension. Then, the transitive reasoning tasks differ only in difficulty as predicted by linear ordering theory (Trabasso et al., 1975) and fuzzy trace theory. When two or more abilities are involved for solving different kinds of tasks, multiple dimensions are needed to describe the responses of children to a set of transitive reasoning tasks.

To investigate which kind of response information gives the most useful insights into transitive reasoning, two kinds of responses were compared (second hypothesis). First, we collected the correct/incorrect judgments children gave on a set of transitive reasoning tasks (quantified as *product scores*). Second, the verbal explanations children gave for the judgments (quantified as *strategy scores*) were recorded. Before comparing the usefulness of both types of responses, the relationship between the two types was investigated. IRT models were used to compare the quality of the product scores and the strategy scores.

The predictions of the theories with respect to the difficulty level of transitive reasoning tasks (Table 2.1) were studied by determining the influence of task characteristics on the difficulty level of the tasks (third hypothesis). For this purpose a multiple regression model were used.

2.2 Method

2.2.1 Operationalization of the Construct

For constructing transitive reasoning tasks, three kinds of task characteristics were used. The first characteristic was presentation form of the premises. According to Piaget's theory, qualitatively different reasoning abilities are involved in successive or simultaneous presentation of the premises, while information-processing theory and fuzzy trace theory assume that one ability is involved in both presentation forms. The second characteristic was task format. Various task formats may have a different influence on the formation of a linear ordering or the use of logical rules. The third characteristic was task content. This characteristic was chosen to measure the influence of different kinds of content of the transitive relationship on performance. According to Sternberg (1980b, 1980a), both a spatial and a verbal representation are involved in solving tasks having a verbal content (linear syllogism) whereas only a spatial representation is involved when the content is physical. The performances on the tasks were both measured by means of the correct/incorrect answers and the verbal explanations of the answers.

2.2.2 Tasks

Three kinds of task characteristics, *presentation form*, *task format*, and *task content* with 2, 4, and 2 levels, respectively, were completely crossed, forming $2 \times 4 \times 2 = 16$ tasks. Figure 2.1 shows the tasks of the transitive reasoning test. Note that the sticks had the colors blue, green, orange, purple, red, and yellow in the computer test. The task characteristics and their levels are:

- Presentation form. The two levels are:

  1. *Simultaneous presentation* (Figure 2.1, tasks 1, 4, 5, 7, 10, 11, 13, and 16). When the premises were presented simultaneously, all the objects were visible simultaneously during the whole task.

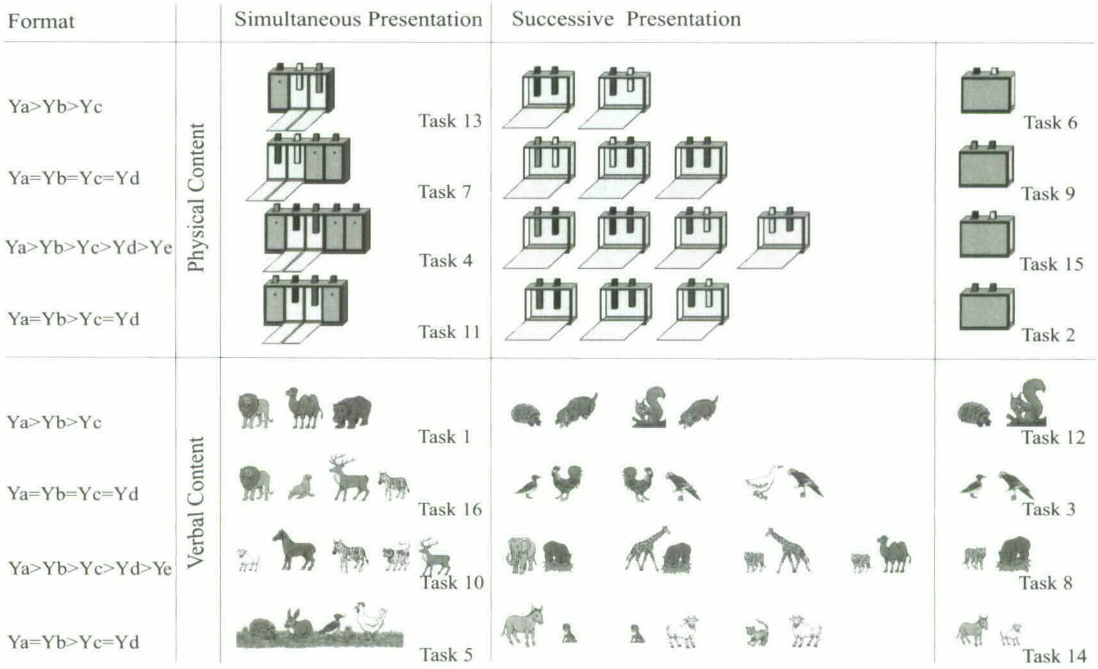


Figure 2.1: *Tasks of the Transitive Reasoning Test*

According to Piaget's theory, this kind of task may be solved using functional reasoning.

2. *Successive presentation* (Figure 2.1, tasks 2, 3, 6, 8, 9, 12, 14, and 15). When the premises were presented successively, in each step of the presentation one pair of objects was visible but the other objects used in the task were not. According to Piaget's theory, this kind of task must be solved using operational reasoning.

- Task format. The four levels are:

  1. $Y_A > Y_B > Y_C$; transitive test pair Y_A, Y_C (Figure 2.1, tasks 1, 6, 12, and 13). In Figure 2.1, Task 1, the lion is assumed to be older than the camel, and the camel is assumed to be older than the hippo.
  2. $Y_A = Y_B = Y_C = Y_D$; transitive test pair Y_A, Y_C (Figure 2.1, tasks 3, 7, 9, and 16). In Figure 2.1, Task 7, all sticks have the same length.
  3. $Y_A > Y_B > Y_C > Y_D > Y_E$; transitive test pair Y_B, Y_D (Figure 2.1, tasks 4, 8, 10, and 15). In Figure 2.1, Task 4, the green stick is longer than the red one, the red one is longer than the purple one, the purple one is longer than the yellow one, and the yellow one is longer than the orange one.
  4. $Y_A = Y_B > Y_C = Y_D$; transitive test pair Y_A, Y_C (Figure 2.1, tasks 2, 5, 11, and 14). In Figure 2.1, Task 5, the hedgehog is assumed to be the same age as the rabbit, the rabbit is assumed to be older than the duck, and the duck is assumed to be the same age as the chicken.

- Type of content. The two levels are:

  1. *Physical content* (Figure 2.1, tasks 2, 4, 6, 7, 9, 11, 13, and 15). When the content of the task was physical, the length relationship between the sticks could be observed visually during the presentation of the premises.

2. *Verbal content* (Figure 2.1, tasks 1, 3, 5, 8, 10, 12, 14, and 16). When the content of the task was verbal, the experimenter told the age relationship between the animals to the child during the presentation of the premises.

2.2.3 Instrument

The transitive reasoning computer program "`Tranred`" (Bouwmeester & Aalbers, 2002) was an individual test, constructed especially for this study. This computer program replaced the normally used *in vivo* presentation of the tasks. The advantage of a computerized test was that the administration of the test was highly standardized. Moreover, movements and sounds could be implemented to enhance the test's attractiveness and hold the child's attention. Finally, the registration of the test scores was done mostly by the program during the test administration. The verbal explanation the child gave after (s)he had clicked on the preferred answer was recorded in writing by the experimenter. The tasks were presented in the same fixed order for every subject (see Figure 2.1 for the task ordering). Relatively difficult tasks were alternated by easier tasks to keep the children motivated. A pilot study showed that the verbal explanations with respect to the same objects appearing in different tasks were hardly ever confused. Nevertheless, to avoid a dependence between the objects of different tasks, tasks sharing the same objects or task characteristics were alternated as much as possible by tasks having different objects or task characteristics.

2.2.4 Procedure

The test was administrated in a quiet room in the school building. The experimenter started a little conversation with the child to put him/her at ease and introduce the task types. Then the child did some exercises to get used to the `Tranred` program. The buttons of the program were explained. It was explained that the colored sticks could have different lengths, which could only be observed when the doors of the box were opened (see Figure 2.1, *physical content*). Also, it was explained that the animals could have different ages, but that this was not observable. After the instructions were given, the test was started.

When the content of the relationship was physical, a box appeared on the screen which either contained all objects (Figure 2.1, *simultaneous presentation of physical content*) or a pair of objects (Figure 2.1, *successive presentation of physical content*). The doors were opened to show the objects of the first premise pair, and the child was asked which stick was longer or whether the sticks had the same length. When the sticks differed in length, the difference could be observed clearly. Then the child clicked on the longest stick, or on the equality button when both sticks had the same length. The doors closed and the doors of the next premise pair opened. The question was repeated for all premise pairs. During the test phase, the doors were closed and the length of the sticks could not be compared visually. The child was asked which of two sticks was longer or whether the sticks had the same length. After the child had clicked on one of the sticks or on the equality button, (s)he was asked to explain the answer. The experimenter wrote down the explanation, the box disappeared from the screen, and the next task started.

When the content of the relationship was verbal, all animals (Figure 2.1, *simultaneous presentation of verbal content*) or a pair of animals (Figure 2.1, *successive presentation of verbal content*) walked onto the screen. For each premise pair, the experimenter told the child which animal was older or that both animals had the same age. The child was asked to click on the oldest animal or on the equality button when both animals had the same age. This was repeated for all premise pairs. In the test phase, the child was asked which of two animals was older or whether both animals had the same age. After the child had clicked on one of the animals or on the equality button, the experimenter asked the child for an explanation of the answer. The experimenter wrote down the explanation, the animals walked off the screen, and the next task started.

The administration of the test took about half an hour, depending on the age of the child. For young children the test took more time and for elder children the test took less time.

2.2.5 Sample

The transitive reasoning test was administered to 615 children ranging in age from 6 to 13 years old. Children came from six elementary schools in the Netherlands. The children came from middle-class social-economic status (SES) families. Table 2.2 gives an overview of the number of children and their mean age within each grade.

Table 2.2: *Number of Children, Mean Age (M) and Standard Deviation (SD) by Grade*

grade	*number*	*age*	
		M[a]	SD
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

[a] number of months

2.2.6 Responses

Product Scores When children clicked on the correct object in the test phase, they received a score of 1. When they clicked on an incorrect object a score of 0 was registered.

Strategy Scores This study builds on previous research on scaling transitive reasoning by Verweij (1994). He found satisfactory inter-rater agreement for two raters who independently coded the verbal explanations given by children who solved transitive reasoning tasks. Figure 2.2 gives an overview of the transitive and non-transitive strategies children used in this study to solve the 16 tasks. The first distinction was made between explanations in which the information of the premises was either used or not. When children did not give an explanation they said that they had either guessed, did not know how they knew the answer, or could not explain their answer. When children gave an explanation but the premise information was not used, children used external information instead to explain their answer (e.g., *the parrot is older because parrots can live more than 40 years*); or they used visual aspects of the task to explain their answer (e.g., *the blue stick is longer because I can see that when I look close*).

When the information of the premises was used correctly, children literally mentioned the premises or reduced the information of the premises. When the premises were mentioned correctly, the child mentioned all the premises involved (e.g., $Y_A > Y_B > Y_C$: *animal A is older than animal C because animal A is older than animal B, and animal B is older than animal C*). This strategy is equivalent to operational reasoning in Piaget's theory. When the information of the premises was reduced correctly, children used a reduction of the premise information, by using the position of the objects (e.g., $Y_A > Y_B > Y_C > Y_D > Y_E$, simultaneous presentation; *all animals are ordered from left to right, the oldest animal first, so animal B is older than animal C*); the time sequence (e.g., $Y_A > Y_B > Y_C > Y_D > Y_E$, successive presentation; *the sticks are ordered in time, stick A was presented first and is the longest, object B was presented before object D, so object B is longer*); a total reduction (e.g., $Y_A = Y_B = Y_C = Y_D$: *all animals have the same age*). When the premises were mentioned incorrectly, children used an incorrect interpretation of the premises (e.g., $Y_A = Y_B > Y_C = Y_D$: *all sticks are equally long, except for stick B, which is longer, so stick A and stick C are equally long*); gave an incomplete explanation (e.g., $Y_A > Y_B > Y_C$: *stick A is longer than stick C because stick B is longer than stick C*); or confused the test-pair with a premise-pair (e.g., $Y_A > Y_B > Y_C$: *stick A is longer than stick C because I have just seen that stick A is longer than stick C*)[2].

[2] In a study by Bouwmeester, Sijtsma, and Vermunt (2004), chapter 4 of this thesis, a nominal variable was used in which all strategies were distinguished to determine the relationships between age, strategy use and task characteristics.

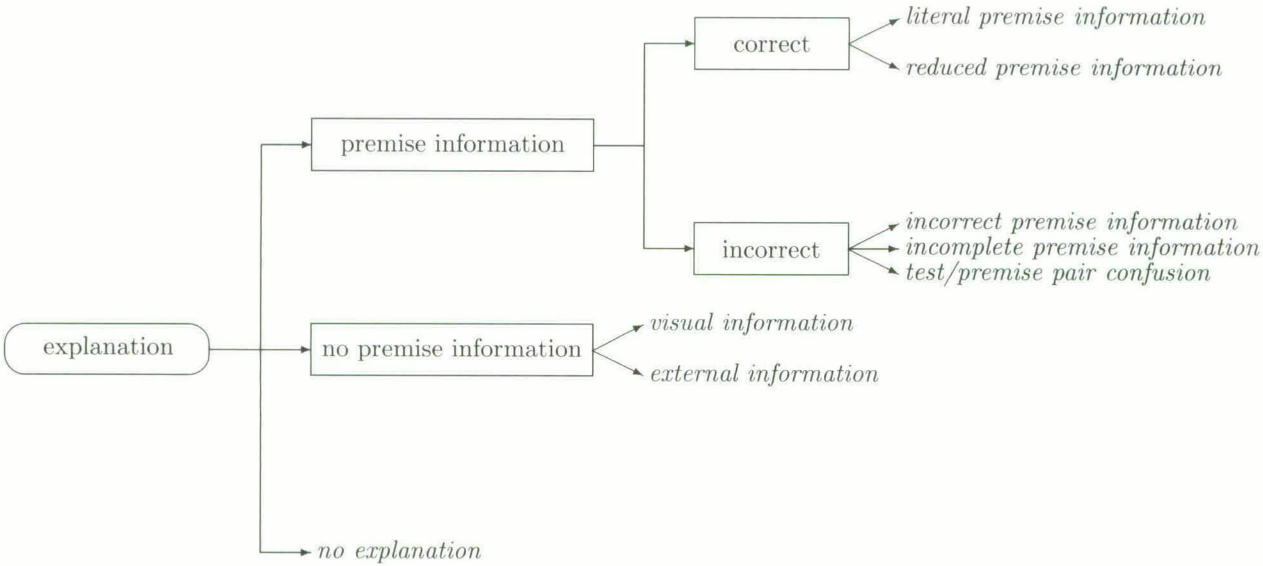

explanation
premise information
correct
literal premise information
reduced premise information
incorrect
incorrect premise information
incomplete premise information
test/premise pair confusion
no premise information
visual information
external information
no explanation

Figure 2.2: *Transitive and Nontransitive Reasoning Strategies*

The strategies in which the premise information was correctly mentioned literally or reduced correctly, were called *transitive reasoning strategies* and received a score of 1. All other strategies received a score of 0. In 0.16% of all cases, the explanation given by the child could not be classified in one of the strategy groups. In those cases a missing value was registered.

2.2.7 Item Response Theory

Our three hypotheses were investigated by means of IRT. Figure 2.3 gives an overview of the successive steps that were followed in this study. We first mention these steps and provide a global description of the rationale behind them. Then we explain the assumptions, methods and models in some detail.

IRT models provide methods to assess the dimensionality of the data, and thus can be used to determine the number of abilities involved in our transitive reasoning test. The program *DETECT* (Stout, 1996), was used to investigate dimensionality using the *local independence* assumption of IRT, and the program *MSP* (Molenaar & Sijtsma, 2000) was used for the same purpose using the *monotonicity* assumption of IRT. DETECT and MSP are exploratory methods. In contrast, the program *Improved DIMTEST* (Stout, 1993) was used to test the hypotheses about the dimensionality resulting from DETECT, MSP, and the theories about transitive reasoning. Our approach is more exploratory than confirmatory, and there is a methodological and a theoretical reason for this. Methodologically, the exploratory methods DETECT and MSP were used instead of a confirmatory method like factor analysis, because factor analysis of dichotomous item scores has problems due to the extreme discreteness of such scores (Nandakumar, Yu, Li, & Stout, 1998; McDonald, 1985; Hattie, Krakowski, Rogers, & Swaminathan, 1996). Van Abswoude, Van der Ark, and Sijtsma (2004) argued that DETECT and MSP do not suffer from these problems. Theoretically, we chose an explorative approach because Piaget's theory is not explicit about the role of task characteristics with respect to the kind of ability (functional or operational) that is involved in transitive reasoning; that is, precise hypotheses about the task loadings on different factors or dimensions can not be posited. However, some less explicit expectations may be derived from the literature. Improved DIMTEST was used to test the expectation that successive tasks are solved by operational reasoning while simultaneous tasks are solved by functional reasoning (Chapman, 1988; Chapman & Lindenberger, 1992).

The results of MSP, DETECT, and Improved DIMTEST were compared and the resulting conclusion answered the first hypothesis about the number of abilities. This conclusion was used as the input for investigating the second hypothesis. This was done by fitting two progressively more restrictive IRT models to the data. First, we fitted the nonparametric monotone homogeneity model (MHM; Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002, chap. 2) to the two data sets. This model implies the ordering of children with respect to ability level. A more restrictive nonparametric model is the double monotonicity model (DMM; Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002, chap. 6). When this model fits, both the children and the transitive reasoning tasks can be ordered, but on separate scales. The linear logistic test model (LLTM; Fischer, 1973, 1995; Scheiblechner, 1972) can be used to model the relationships between task difficulty and task characteristics. However, since the LLTM is a specialization of the Rasch model it is highly restrictive. Because the Rasch model did not fit our data, as an alternative multiple regression on P-values was used (Green & Smith, 1987).

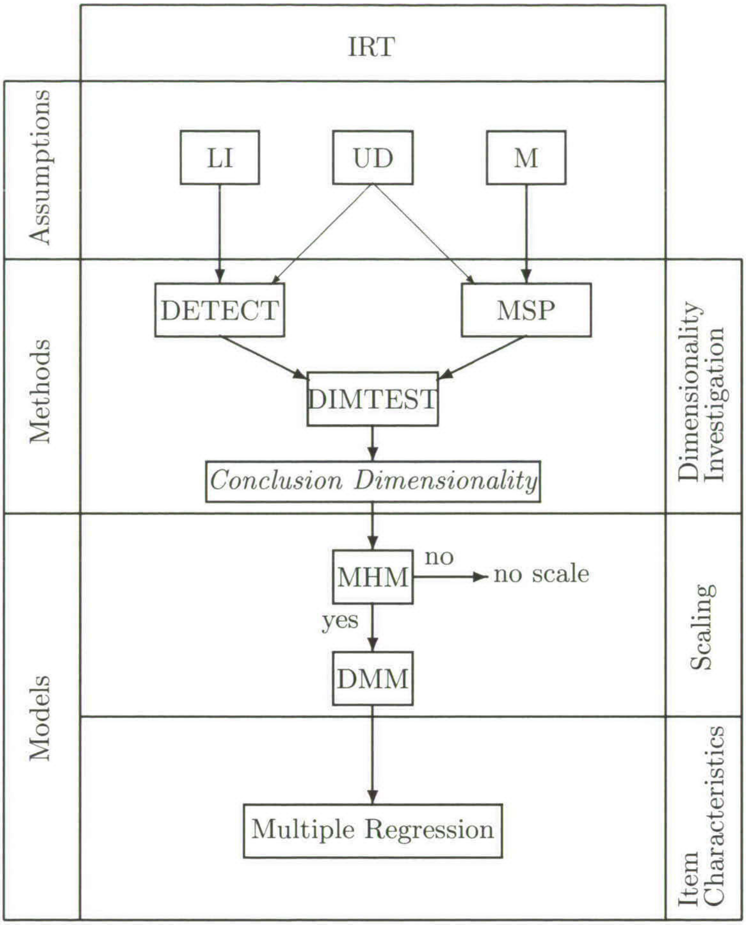


Figure 2.3: *Overview of the Successive Analyses*

Assumptions Common to the IRT Models Used in This Study

Local Independence Let the test consist of J dichotomously scored tasks, and let θ denote the latent ability measured by the J tasks. If the tasks measure more than one ability, we assume W latent ability parameters collected in a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_W)$. Let X_j be the random variable for the score on task j, with $j = 1, \ldots, J$; and let x_j be the realization of this variable, with $x_j = 0, 1$. The task score variables are collected in $\mathbf{X} = (X_1, \ldots, X_J)$, and the realizations in $\mathbf{x} = (x_1, \ldots, x_J)$. Finally, the conditional probability of a 1 score on task j is denoted $P_j(\boldsymbol{\theta})$; this is the item response surface. For scalar θ, $P_j(\theta)$ is the item response function (IRF). The assumption of local independence (LI) is defined as

$$P(\mathbf{X} = \mathbf{x} \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} P_j(\boldsymbol{\theta})^{x_j}[1 - P_j(\boldsymbol{\theta})]^{1-x_j}. \tag{2.1}$$

LI means that a subject's response to a task is not influenced by his/her responses to the other tasks in the test. LI implies that the covariance of two tasks, j and k, given the latent trait composite, $\boldsymbol{\theta}$, is zero; that is, $Cov(X_j, X_k \mid \boldsymbol{\theta}) = 0$. This zero conditional covariance is known as weak local independence, which is important for practical item selection (Stout et al., 1996; Zhang & Stout, 1999a) to be discussed shortly.

Unidimensionality The assumption of unidimensionality (UD) means that the data structure can be explained by a unidimensional latent trait, θ. When UD does not hold, one ability is not enough to explain the variation in the scores on different tasks, and a second ability may be necessary to explain the variability, and perhaps a third, a fourth, and so on. Although UD and LI are mathematically not the same, in practice, the same methods are used to evaluate these assumptions.

Monotonicity For unidimensional θ, we assume that the IRFs are monotone nondecreasing functions. That is, for two arbitrarily chosen fixed values of θ, say, θ_a and θ_b, we have that

$$P_j(\theta_a) \leq P_j(\theta_b), \text{whenever } \theta_a < \theta_b; j = 1, \ldots, J. \tag{2.2}$$

This is the monotonicity (M) assumption. Assumption M also gives information about the dimensionality of the task set, based on the variation in the slopes of the IRFs. Suppose that the task set is multidimensional in the sense that some tasks measure θ_1 and others measure θ_2. Because the slope of an IRF expresses the strength of the relationship of a task with the latent ability or a latent ability composite, it may well be that tasks measuring one ability have steeper IRFs than tasks measuring a different ability. Even if a unidimensional IRT model is incorrectly hypothesized for these multidimensional data, the slopes of the IRFs may provide evidence of this multidimensionality (Hemker et al., 1995; Mokken, 1971; Sijtsma & Molenaar, 2002, chap. 5; Van Abswoude et al., 2004). In this study, we investigated whether all the tasks measure the same θ and, in case of multidimensionality, we tried to identify unidimensional subsets of tasks.

The Monotone Homogeneity Model The MHM (Mokken, 1971, chap. 4; Sijtsma & Molenaar, 2002, chap. 2) is based on the assumptions of LI, UD, and M. The MHM is an NIRT model that orders subjects on the θ scale using their number-correct score, defined as $X_+ = \sum X_j$ (Grayson, 1988; Hemker, Sijtsma, Molenaar, & Junker, 1997). Theoretically, this ordering of persons is the same for each task, and also for a sumscore, $Y_+ = \sum Y_j$, based on the task scores Y_j from any subset of tasks selected from the larger set of tasks that are driven by θ and agree with the MHM. In practice, the number of tasks affects the accuracy of a person ordering estimated by means of the number-correct score X_+.

Methods to Assess the Dimensionality of the Data

We used three methods to assess the dimensionality structure of the two dichotomous data sets. The first method was the item selection procedure in the computer program MSP (Molenaar & Sijtsma, 2000; also, see Sijtsma and Molenaar, 2002, chap. 5). This procedure is used to select the tasks on the basis of assumption M. The second item selection method was DETECT (Zhang & Stout, 1999b). The third method was Improved DIMTEST (Stout, Froelich, & Gao, 2001). This method was used to test the null-hypothesis of UD for the whole task set. Both DETECT and DIMTEST use the assumption of LI to assess UD.

Program MSP. MSP (Molenaar & Sijtsma, 2000) uses scalability coefficient H (Mokken, 1971, pp. 157-169) to assess the discrimination power of individual tasks (i.e., the slopes of the IRFs) and the whole test. The item coefficient H_j is an index of the slope of the IRF relative to the spread of the number-correct score X_+ in the group under consideration. The higher H_j, the better task j discriminates between different X_+ scores. The H coefficient for the whole test of J tasks summarizes the slope information contained in all J item coefficients H_j.

Mokken, Lewis, and Sijtsma (1986) argued that higher positive H values reflect higher discrimination of the whole set of tasks and, thus, a more accurate ordering of subjects. In practical test construction, to have at least reasonable discrimination, a lower bound value for H_j and H of 0.3 is recommended (Mokken, 1971 p. 184). Other guidelines (Sijtsma & Molenaar, 2002, p. 60) for the interpretation of H are: $0.3 \leq H < 0.4$ is a weak scale; $0.4 \leq H < 0.5$ is a medium scale; and $0.5 \leq H < 1.0$ is a strong scale. The MSP item selection procedure has been described in detail by Mokken (1971, pp. 190-194; also see Molenaar & Sijtsma, 2000; and Sijtsma & Molenaar, 2002, chap. 5). It is a bottom-up procedure, that starts by selecting the two items with the highest significantly positive H_{jk} that is at least c ($c > 0$; user-specified). Then the procedure adds tasks one by one, in each step maximizing the total H of the selected items, such that $H_j \geq c$ for all selected items (for possible exceptions, see Sijtsma & Molenaar, 2002, p. 79). After having selected the first scale, the procedure continues by selecting from the unselected items a second scale, a third scale, and so on. Van Abswoude et al. (2004) found that MSP was able to exactly retrieve the true dimensionality from simulated data when latent traits did not correlate highly (say, higher than 0.4). Hemker et al. (1995; see also Sijtsma & Molenaar, 2002, p. 81; Van Abswoude et al., 2004) recommended using a range of c values from $c = 0.00$ to $c = 0.55$ with increments of 0.05, and described sequences of outcomes for increasing c values typical of multidimensionality and unidimensionality.

Program *DETECT*. The computer program DETECT (Zhang & Stout, 1999a, 1999b; Roussos, Stout, & Marden, 1998) contains an item selection algorithm that tries to find the partitioning $\mathcal{P}$ for which the degree to which LI is satisfied is maximal, given all possible partitions of the task set. In contrast to MSP, where assumption M is the basis of the item selection, weak LI is the basis of DETECT. DETECT works best when all individual tasks load on one θ (but not necessarily the same θ for all tasks). This is called approximate simple structure (Zhang & Stout, 1999a). When individual tasks load on different θs, approximate simple structure does not hold and no best partitioning can be determined. Under the assumption of approximate simple structure, the DETECT index is maximal when the underlying structure is correctly represented by the number and the composition of the clusters. When the DETECT value is zero, no best partitioning is possible and the task set is unidimensional. As a rule of the thumb (Zhang & Stout, 1999b), a task set is considered unidimensional when the DETECT value is smaller than 0.1. To evaluate whether approximate simple structure exists, Zhang and Stout (1999b) proposed that index $R \geq 0.8$. When approximate simple structure does not exist, it is difficult to decide how many dimensions are involved. Van Abswoude et al. (2004) recommended to use MSP and DETECT together for analyzing one's data.

Program *Improved DIMTEST*. DIMTEST is a procedure that tests the null hypothesis that a set of items is dimensionally similar to another set of items. Because the DIMTEST procedure does not work for short test, we used the improved DIMTEST procedure (Nandakumar & Stout, 1993). This procedure generates a unidimensional data set using a nonparametric bootstrap method to correct for bias in parameter estimates and to increase the power of the DIMTEST statistic (Stout et al., 2001). The hypothesis is tested that the generated data set has the same dimensionality as the real data set. For example, we tested the hypothesis that the responses to the successively presented tasks are dimensionally distinct from those to the simultaneously presented tasks. We considered the simultaneously presented tasks to be the Assessment Test (AT; see Nandakumar & Stout, 1993) and the successively presented tasks to be the Partition Test (PT; see Nandakumar & Stout, 1993). The items in AT were hypothesized to measure one dominant trait. An asymptotic test statistic denoted T, was used to test whether the items in AT and PT measure the same θ.

IRT Models and Assessment of Fit.

Monotone Homogeneity Model After the dimensionality of the transitive reasoning data was investigated, the computer program MSP (Molenaar & Sijtsma, 2000) was used to investigate the fit of the MHM to the two data sets. To evaluate whether the IRFs of the J tasks were all nondecreasing, subjects were partitioned into J restscore groups on the basis of their restscore, $R_{(-j)} = X_+ - X_j$. The restscore $R_{(-j)}$ is an ordinal estimator of θ (Junker, 1993). To enhance power, small adjacent restscore groups were joined using recommendations given by Molenaar and Sijtsma (2000, p. 100). For each restscore group r the probability of giving a correct answer, $P(X_j = 1 \mid R_{(-j)} = r)$, was estimated, and the hypothesis was tested that these probabilities are nondecreasing in $R_{(-j)}$.

Double Monotonicity Model The DMM adds a fourth assumption to the MHM, which states that the IRFs do not intersect. This fourth assumption equals *invariant item ordering*; that is, the ordering property of the J tasks is the same for different subgroups of subjects (except for possible ties), including individual θs. In particular, for two tasks j and k, if we know for one θ_0 that $P_j(\theta_0) < P_k(\theta_0)$, then it follows that for any θ, we have that $P_j(\theta) \leq P_k(\theta)$. This ordering can be extended to all tasks.

MSP was used to investigate whether the IRFs intersected. The scalability coefficient H^T (Sijtsma & Meijer, 1992) for the J tasks in the test and the person coefficient H_i^T were used to evaluate intersection of the IRFs. As a rule of thumb, if $H^T \geq 0.3$ and the percentage of negative H_i^T values < 10, then the IRFs do not intersect. Three additional methods were used to investigate the nonintersection of IRFs for pairs of tasks. These methods are the *restscore method*, the *restsplit method*, and the inspection of the *P-matrices*, $P(-,-)$ and $P(+,+)$ (Sijtsma & Molenaar, 2002, chap. 6). These methods are based on the same rationale, but use different subgroupings of respondents for estimating the IRFs. The three methods differ in accuracy to estimate the IRFs and in power to detect intersections.

Linear Regression Using P-values In the multiple regression model the proportions correct are regressed on the task characteristics. Because the proportions corrected are bounded between 0 and 1, a logistic transformation of the P-values was used.

2.3 Results

2.3.1 Relation Between Product Scores and Strategy Scores

Table 2.3 shows the proportions of strategy use and the proportions of correct answers given strategy use. The two "correct" strategies (literal and reduced premise information) almost always led to correct answers. The three strategies in which no premise information is used (visual information, external information, and no explanation) have proportions of correct answers close to chance level. Test/premise pair confusion relatively often led to a correct answer, although it is an incorrect strategy. Table 2.3 shows that incorrect strategies often led to correct answers that were produced by chance.

2.3.2 Hypothesis 1: Assessing Dimensionality

Analysis of Product Scores

12 cases were rejected from the analysis because of missing values on one or more tasks. The resulting sample consisted of 603 subjects.

MSP Analysis. Table 2.4 shows the sequence of outcomes of the MSP analysis with increasing c-values. Task 2 was immediately rejected because of negative covariances with other tasks. For lowerbound $c = 0$, two scales were formed containing six and nine tasks, respectively, which suggests that

Table 2.3: *Strategy Use and Proportion of Correct Answers*

STRATEGY	Proportion strategy use	Proportion correct answers
LITERAL COMPLETE PREMISE INFORMATION	.16	.94
REDUCED PREMISE INFORMATION	.21	.97
INCORRECT PREMISE INFORMATION	.19	.23
INCOMPLETE PREMISE INFORMATION	.10	.48
TEST/PREMISE PAIR CONFUSION	.10	.58
VISUAL INFORMATION	.06	.35
EXTERNAL INFORMATION	.03	.36
NO EXPLANATION	.16	.37

the test measures at least two latent abilities. For increasing c-values, Task 3 and Task 6 were also rejected, and a third and a fourth scale were formed, both containing two tasks. For c-values of 0.40 and higher, almost all tasks were rejected and no scale was formed containing more than two tasks. For $c = 0.55$ no scale was formed. On the basis of the guidelines of Hemker et al. (1995), it was concluded that at least two abilities were involved in answering the tasks. One scale contained the tasks 7, 9 and 16 ($H = 0.44$), which all have the format $Y_A = Y_B = Y_C = Y_D$, and another, rather weak ($H = 0.25$) scale contained the tasks 1, 4, 5, 8, 10, 11, 12, 13, 14, and 15, which have the formats $Y_A > Y_B > Y_C$; $Y_A > Y_B > Y_C > Y_D > Y_E$; and $Y_A = Y_B > Y_C = Y_D$.

***DETECT* Analysis.** A random half of the sample was used for the DETECT procedure. The second half of the sample was used for cross-validation. The R index for assessing simple structure was 0.74. This is smaller than the value of at least 0.8 that Zhang et al. (1999b) proposed for approximate simple structure. The maximum DETECT value [denoted $D_\alpha(\mathcal{P}^*)$] was 0.88, which was higher than 0.1, indicating that the task set was not unidimensional. The partitioning with this value had three clusters. For the second half of the sample, using the same partitioning that was found to be optimal for the first data set, we found $D_\alpha(\mathcal{P}^*) =$

Table 2.4: *Item Selection for Increasing c-Values, for Analysis Using Product Scores*

c	scale 1	scale 2	scale 3	scale 4	# tasks rejected
.00	1,3,4,7,9,16	5,6,8,10,11,12,13,14,15			1
.05	1,3,4,7,9,16	5,6,8,10,11,12,13,14,15			1
.10	1,3,4,7,9,16	5,6,8,10,11,12,13,14,15			1
.15	3,4,7,9,16	1,5,8,10,11,12,13,14,15			2
.20	7,9,16	1,4,5,8,10,11,12,13,14,15			3
.25	7,9,16	5,14,10,8,1	4,13	11,15	4
.30	7,9,16	5,10,8,1	4,13		7
.35	7,9,16	5,10,1	4,13		8
.40	9,16	10,1	4,13		10
.45	9,16	10,1	4,13		10
.50	9,16	10,1	4,13		12
.55					16

0.48 and $R = 0.43$. To gain more insight into the dimensionality of the data, 20 random samples of approximately 50% of the subjects were drawn from the original sample and the DETECT value was calculated for each sample. Figure 2.5 shows the number of times that two tasks were in the same cluster. Three (overlapping) clusters can be distinguished. One contained the tasks 3, 7, 9, and 16 (all with format $Y_A = Y_B = Y_C = Y_D$), which were almost always in the same cluster. A second cluster contained the tasks 1, 5, 8, 10, 11, and 14, and a third cluster contained the tasks 2, 4, 12, and 13. Task 6 did not fit well in any of the clusters and Task 15 might belong to either the second or the third cluster.

Improved DIMTEST Analysis Three hypotheses were tested. First, it was tested whether the tasks that were simultaneously presented measured the same ability as the tasks that were successively presented (Piaget's theory). Second, it was tested whether the tasks that had a verbal content measured the same ability as the tasks that had a physical content

Table 2.5: *DETECT Partitioning in Clusters for 20 Random Samples, Product Scores*

	3	**7**	**9**	**16**	**6**	**1**	**5**	**8**	**10**	**11**	**14**	**15**	**2**	**4**	**12**	**13**
3	■	19	19	19	09	04	00	00	00	00	00	01	01	06	00	04
7	19	■	20	20	10	04	00	00	00	00	00	00	00	05	00	03
9	19	20	■	20	10	04	00	00	00	00	00	00	00	05	00	03
16	19	20	20	■	10	04	00	00	00	00	00	00	00	05	00	02
6	09	10	10	10	■	06	05	05	05	03	03	02	00	00	03	00
1	04	04	04	04	06	■	15	16	16	11	09	02	00	01	02	01
5	00	00	00	00	05	15	■	20	20	15	13	05	01	00	02	00
8	00	00	00	00	05	16	20	■	20	15	13	05	01	00	02	00
0	00	00	00	00	05	16	20	20	■	15	12	05	01	00	02	00
11	00	00	00	00	03	11	15	15	15	■	14	09	02	00	00	00
14	00	00	00	00	03	09	13	13	12	14	■	11	05	02	04	02
15	01	00	00	00	02	02	05	05	05	09	11	■	11	07	08	09
2	01	00	00	00	00	00	01	01	01	02	05	11	■	12	13	09
4	06	05	05	05	00	01	00	00	00	00	02	07	12	■	13	17
12	00	00	00	00	03	02	02	02	02	00	04	08	13	13	■	15
13	04	03	03	02	00	01	00	00	00	00	02	09	09	09	15	■

□ 0 — 5 times □ 6 — 9 times ▒ 10 — 15 times ▓ 16 — 20 times

(Sternberg's mixed model). Third, it was tested whether the tasks with an equality format ($Y_A = Y_B = Y_C = Y_D$) measured a different ability than the other tasks, which was the result of MSP and DETECT. The results were:

- *Hypothesis 1:* Statistic T was 1.24 ($p > 0.05$), so we can not conclude that simultaneously and successively presented tasks require different abilities.

- *Hypothesis 2:* Statistic T was 2.51 ($p < 0.05$), so the tasks having a verbal content may measure a different ability than the tasks having a physical content.

- *Hypothesis 3:* Statistic T was 2.85 ($p < 0.05$), so the equality tasks may measure a different ability than the tasks having an inequality or mixed inequality/equality format.

Conclusion About Dimensionality of Product Scores. MSP, DETECT and improved DIMTEST results converged to the conclusion that the structure of the *product scores* is not unidimensional. MSP distinguished at least two dimensions, one defined by tasks with the equality format and the other by the other tasks. DETECT found three partly overlapping clusters, one of which contained the tasks having the equality format. The Improved DIMTEST procedure supported the hypothesis that the tasks having an equality format were dimensionally distinct from the other tasks, and that the tasks having a verbal content were dimensionally distinct from the tasks having a physical content. None of the three methods showed that the successively and simultaneously presented tasks were dimensionally distinct.

Analysis of Strategy Scores

15 subjects were rejected from the analysis because of missing values on one or more tasks. The resulting sample consisted of 600 subjects. Because only six children gave a transitive reasoning explanation for Task 2, this task was rejected from further analysis.

***MSP* Analysis.** Table 2.6 shows the sequence of item selection outcomes with increasing c-values. For $c = 0$, all tasks were selected into the same scale. For higher c-values, all tasks were selected into the same scale until a c-value of 0.40, when Task 12 was rejected from the scale. For $c = 0.45$, a second scale was formed containing the tasks 3, 9, and 14. Considering this sequence of outcomes, it could be concluded that the structure of the strategy scores was unidimensional.

***DETECT* Analysis.** The R ratio on the first half of the sample was 0.68, indicating that there was no approximate simple structure. The max-

Table 2.6: *Item Selection for Increasing c-Values, for Analysis Using Strategy Scores*

c	scale 1	scale 2	# tasks rejected
.00	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.05	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.10	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.15	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.20	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.25	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.30	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.35	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16		
.40	1,3,4,5,6,7,8,9,10,11,13,14,15,16		1
.45	1,4,6,7,8,10,13,15,16	3,9,14	3
.50	2,6,7,9,11,16		6
.55	4,6,8,10,13,16	7,11	5

imum DETECT value $[D_\alpha(\mathcal{P}^*)]$ was 0.57, indicating that the task set was not unidimensional. The partitioning with maximum DETECT value had two clusters. For the cross-validation sample we found that $D_\alpha(\mathcal{P}^*) = 0.24$ and $R = 0.32$. Again, 20 samples of approximately 50% of the original sample size were drawn at random from the original sample and the DETECT values were calculated for each sample. Figure 2.7 shows two overlapping clusters; one cluster containing the tasks 3, 7, 9, and 16, which were almost always in the same cluster, and one cluster containing the other tasks. It could not be decided to which cluster the tasks 4 and 6 belong.

***Improved DIMTEST* Analysis.** The same three hypotheses were tested as was done using the product scores. The results were:

- *Hypothesis 1:* Statistic T was 0.70 ($p > 0.05$), so we could not conclude that simultaneously and successively presented tasks required different abilities.

Table 2.7: *DETECT Partitioning in Clusters for 20 Random Samples, Strategy Scores*

	1	**5**	**8**	**10**	**11**	**12**	**14**	**15**	**13**	**4**	**6**	**3**	**7**	**9**	**16**
1	■	15	16	19	11	13	16	14	07	06	02	01	01	01	01
5	15	■	15	17	14	13	20	16	06	03	04	00	00	00	00
8	16	15	■	17	10	16	16	16	09	07	02	00	01	00	00
10	19	17	17	■	12	14	18	14	08	05	03	00	00	00	00
11	11	14	10	12	■	06	13	13	05	03	08	03	03	03	04
12	13	13	16	14	06	■	13	13	12	07	06	00	00	00	00
14	16	20	16	18	13	13	■	13	06	03	03	00	00	00	00
15	14	16	16	14	13	13	13	■	11	07	04	00	00	00	00
13	07	06	09	08	05	12	06	11	■	16	11	02	03	03	03
4	06	03	07	05	03	07	03	07	16	■	12	07	08	08	08
6	02	04	02	03	08	06	03	04	11	12	■	08	08	09	09
3	01	00	00	00	03	00	00	00	02	07	08	■	19	20	19
7	01	00	01	00	03	00	00	00	03	08	08	19	■	19	18
9	01	00	00	00	03	00	00	00	03	08	09	20	19	■	19
16	01	00	00	00	04	00	00	00	03	08	09	19	18	19	■

□ 0 — 5 times □ 6 — 9 times ▒ 10 — 15 times ▓ 16 — 20 times

- *Hypothesis 2:* Statistic T was 2.26 ($p < 0.05$), so the tasks having a verbal content may measure a different ability than the tasks having a physical content.

- *Hypothesis 3:* Statistic T was 2.30 ($p < 0.05$), so the equality tasks may measure a different ability than the tasks having an inequality or mixed inequality/equality format.

Conclusion About Dimensionality of Strategy Scores. Different methods led to different conclusions about the dimensionality of the data. MSP indicated unidimensionality. Improved DIMTEST suggested distinct abilities for both the equality tasks and tasks having a verbal content. DETECT resulted in two dimensions. One cluster contained the tasks with the equality format and the other cluster contained the other tasks.

The tasks having a verbal content did not form a distinct cluster.

2.3.3 Hypothesis 2: Fitting the NIRT Models

The product scores did not form a unidimensional scale. Therefore, the NIRT models were only fitted to the strategy scores.

Analysis of Strategy Scores

MSP, DETECT and Improved DIMTEST led to different conclusions about the dimensionality structure of the strategy scores. In particular the equality tasks formed a distinct cluster. In the following analyses, 15 transitive reasoning tasks (except Task 2) were used.

MHM Analysis. The H-value of the scale was 0.45, indicating medium strength scale. All H_js were between 0.38 (Task 12) and 0.66 (Task 16). Table 2.8 gives an overview of the P_j-values and the H_j-values. The item-restscore regressions were increasing or non-significantly locally decreasing for each of the 15 tasks. Thus the MHM fitted the 15 tasks.

DMM Analysis. The H^T value was 0.52, and the percentage of negative H_i^T values was 1.4. According to the assessment of intersection via restscore groups, tasks 3 and 10, and tasks 9 and 10 intersected significantly ($z_{3,10} = 1.81; z_{9,10} = 3.05$). Investigating the intersection via restsplit groups, tasks 9 and 10, and tasks 4 and 12 intersected significantly for two dichotomizations (yielding $z_{9,10}$ values of 2.04 and 3.12; and $z_{4,12}$ values of 1.66 and 1.67. The bivariate proportions in the $P(+,+)$ matrix showed an intersection of the tasks 9 and 10.

Summarizing the results of the four methods, the task pair (9,10) had the most serious intersections, but the violations were small. It was concluded that the DMM fitted the strategy data and that an invariant item ordering held for the 15 tasks.

Table 2.8: *P_j-Value and H_j-Value of the Items, Based on Strategy Scores*

Item #	Presentation	Format	Content	P_j	H_j
6	successive	$Y_A > Y_B > Y_C$	physical	.05	.46
15	successive	$Y_A > Y_B > Y_C > Y_D > Y_E$	physical	.07	.47
5	simultaneous	$Y_A = Y_B > Y_C = Y_D$	verbal	.15	.40
14	successive	$Y_A = Y_B > Y_C = Y_D$	verbal	.19	.42
8	successive	$Y_A > Y_B > Y_C > Y_D > Y_E$	verbal	.21	.48
11	simultaneous	$Y_A = Y_B > Y_C = Y_D$	physical	.31	.40
4	simultaneous	$Y_A > Y_B > Y_C > Y_D > Y_E$	physical	.39	.46
12	successive	$Y_A > Y_B > Y_C$	verbal	.40	.38
3	successive	$Y_A = Y_B = Y_C = Y_D$	verbal	.45	.41
10	simultaneous	$Y_A > Y_B > Y_C > Y_D > Y_E$	verbal	.52	.51
9	successive	$Y_A = Y_B = Y_C = Y_D$	physical	.54	.40
1	simultaneous	$Y_A > Y_B > Y_C$	verbal	.56	.46
13	simultaneous	$Y_A > Y_B > Y_C$	physical	.57	.50
7	simultaneous	$Y_A = Y_B = Y_C = Y_D$	physical	.77	.55
16	simultaneous	$Y_A = Y_B = Y_C = Y_D$	verbal	.86	.66

2.3.4 Hypothesis 3: The Influence of Task Characteristics on Difficulty

Multiple Regression

A multiple regression analysis was performed on the 15 tasks to which the DMM fitted. The dependent variable was the logit transformation of the proportion correct of each task. The three task characteristics were the predictor variables. Because the task characteristics were nominal they were transformed to dummy variables. A significant F-value was found: $F_{6,14} = 6.77$ ($p = 0.01$). The adjusted R^2 was 0.71, meaning that the model explained 71% of the variance of the difficulty levels of the 15 tasks. Two regression weights (Table 2.9) significantly deviated from 0. The format $Y_A = Y_B = Y_C = Y_D$ had a positive effect on the easiness of a task. Simultaneous presentation was easier than successive presentation.

Table 2.9: *Estimated Weights of the Multiple Regression Model*

Characteristic	B	SE	β	p-value
(Constant)	-1.980	.740		.028
$Y_A > Y_B > Y_C$	.273	.698	.096	.706
$Y_A = Y_B = Y_C = Y_D$	1.797	.698	.632	.033
$Y_A > Y_B > Y_C > Y_D > Y_E$	.221	.611	.078	.727
$Y_A = Y_B > Y_C = Y_D$	-.957	.631	-.305	.168
Presentation	1.504	.367	.597	.003
Content	.333	.393	.132	.420

Simultaneous presentation form was coded 1, Successive presentation form was coded 0; Verbal type of content was coded 1, Physical type of content was coded 0.

2.4 Discussion

Theories stemming from different epistemological backgrounds used different definitions, operationalizations and methods to study transitive reasoning. This led to disagreement about the number of abilities involved in transitive reasoning, the kind of responses to be collected, and the influence of task characteristics on performance. In this chapter, we first evaluated the hypothesis that different abilities are involved in solving tasks by investigating the dimensionality structure of a task set with various task characteristics. Both the product scores and the strategy scores were analyzed and the results compared. Second, a scale was constructed which measured individual differences in transitive reasoning. Third, the influence of task characteristics on the difficulty level of tasks was determined.

The results of MSP, DETECT and Improved DIMTEST for the product data and the strategy data showed that the dimensionality of successively and simultaneously presented tasks did not differ. Thus, there is no evidence to distinguish between functional and operational reasoning. This result does not support Piaget's theory. With respect to Sternberg's mixed model, it appeared that Improved DIMTEST suggested different abilities for tasks having a verbal content and tasks having a physical content. Although MSP and DETECT did not support this finding, a tentative conclusion might be that there is some evidence that the tasks having a verbal content require an additional verbal ability. A possible explanation for finding the distinct abilities only by means of DIMTEST may be that the verbal content tasks were relatively easy linear syllogisms with respect to the verbal ability component (without negations or marked adjectives; see Sternberg, 1980b). In terms of Sternberg's mixed model this would mean that verbal content tasks require a weak verbal component in addition to the spatial ordering component whereas physical content tasks only require a spatial ordering component.

In contrast to the results of the past four decades of research on cognitive development (see e.g., Brainerd, 1977; Murray & Youniss, 1968; Smedslund, 1963), we found that the strategy scores produced more straightforward and useful findings than the product scores. The data structure of the strategy scores could be explained by one dimension according to MSP, but at least three dimensions were needed to explain the data structure of the product scores. The results of the three methods did not converge to one interpretation. The multidimensionality in the product scores might best be explained by the difference in accuracy and meaning of the two types of responses. A product score of 1 means that the child had clicked on the correct object. A 1 score may therefore not represent true transitive reasoning ability, but instead may be due to additional unimportant skills or tricks. The data structure of the product scores is expected to be fuzzier than the data structure of the strategy scores, for which the meaning of a 0 or 1 score is clearer. This may explain why the product data were multidimensional and the strategy data were unidimensional.

Our population consisted of children of six years and older, which were well capable of explaining their thoughts afterwards. This population was chosen because our aim was to describe the development of transitive reasoning, but not to determine the age of emergence of transitive reasoning. This often was the aim of researchers studying transitive reasoning by young children (Braine, 1959; Smedslund, 1963; Murray & Youniss, 1968;

Bryant & Trabasso, 1971). When younger children are studied, the requirement of verbal explanation may cause many false negatives due to verbal incapacity. Then, product scores are expected to be more useful.

For the strategy scores, DETECT found that the equality-format tasks ($Y_A = Y_B = Y_C = Y_D$) formed a distinct cluster. MSP and Improved DIMTEST did not find a distinct dimension for the equality-format tasks. The equality-format tasks were easy, and they discriminated well between children with low ability levels, and worse between children with higher ability levels. Although the equality-format tasks may not be entirely dimensionally equal to the other tasks, they are useful from a practical point of view because they discriminate well at θ levels not covered by the other tasks but desirable for a transitive reasoning scale.

MSP, DETECT and Improved DIMTEST evaluate dimensionality from different perspectives on the data. The three methods differ in several ways and each has merits and drawbacks. Van Abswoude et al. (2004) concluded that DETECT is the best method to assess true dimensionality. However, the simple structure assumption is a strong assumption which may not be realistic in many psychological settings. MSP is susceptible to locally optimal solutions because it uses a sequential clustering procedure. Further, MSP often does not accurately distinguish highly correlated abilities ($> .4$), but DETECT does. However, by forcing tasks into clusters of highly correlating traits, DETECT is vulnerable to chance capitalization. Also, Van Abswoude et al. (2004) found that DETECT does not reflect well dimensionality when abilities are measured by unequal numbers of tasks. Improved DIMTEST does not reflect true dimensionality well when abilities are measured by unequal numbers of tasks and these task subsets have equal average discrimination. DETECT and Improved DIMTEST both need large sample sizes, and Improved DIMTEST has low power for short tests. Nevertheless, when the methods are used next to each other, they can compensate each other's shortcomings and offer a detailed description of the underlying dimensionality. In future research it would be interesting to sample new data and use the results from the present study for confirmatory analysis. Multidimensional IRT models might be appropriate for this purpose (see e.g., Kelderman & Rijkes, 1994, and Reckase, 1997).

It is important to point out that statistical methods give mathematical definitions of dimensions, and that these dimensions are not equivalent to psychological abilities. The interpretation of the dimensionality of the data is dependent on the operationalization of the construct of transitive reasoning, but not directly on the construct itself. While usually no explicit distinction is made between the operationalization of the construct and the construct itself when interpreting the results, the distinction should not be ignored. In our study, we used a broad operationalization of transitive reasoning by using different kinds of task characteristics. Using this operationalization, we could explain the structure of the strategy data by means of one dimension. When we would have used a narrower operationalization based only on the theory of Piaget (e.g., see Verweij, Sijtsma & Koops, 1999), we probably would have found a different dimensionality structure leading to a different interpretation.

Multiple regression was used to determine the influence of task characteristics on the tasks difficulty level. With respect to presentation form each of the cognitive theories predicted that simultaneous presentation was easier than successive presentation. This was indeed what was found. With respect to the task format, the equality format appeared to be easier than the other formats. This result was correctly predicted by information processing theory and fuzzy trace theory but not by Piaget's theory. Verbal and physical content hardly influenced difficulty level, and this was only predicted correctly by fuzzy trace theory.

This study showed that IRT techniques are not only useful tools to construct tests but also offer a set of methods to investigate psychological theories, in particular the dimensionality of a psychological construct. Now that we know that transitive reasoning can be explained by one dimension, further research should be done to interpret this ability in more detail. In our current research, Bouwmeester et al. (2004) used a latent class regression model (see chapter 4), and found that several latent classes could be distinguished in which children used different patterns of correct and incorrect strategies and the influence of task characteristics on performance was different. From a developmental perspective, it is important to determine whether the development along the ability found in this study is continuous or discontinuous [see e.g., Hosenfield, Van der Maas, & Van den Boom (1997), and Thomas, Lohaus, & Kessler (1999), for studies on discontinuity in other Piagetian tasks].

Chapter 3

Detecting Discontinuity in the Development of Transitive Reasoning: a Comparison of Two Models

Abstract

Since Piaget, the issue of the existence of multiple stages in development is an important topic. In cognitive developmental research, the binomial mixture model is often used to identify discontinuity from empirical data. The binomial distributions that are hypothesized to correspond to the stages are estimated by means of the number of correctly solved tasks in the developmental test. In doing this, the binomial mixture model assumes that all tasks in the test have the same difficulty level. However, the assumption of equal task difficulty may be unduly restrictive for more complicated task types, and the use of the number-correct score ignores valuable information in the pattern of item scores.

Unlike the binomial mixture model, the latent class model does not assume binomial distributions of number-correct scores, allows task difficulties to vary, and uses the information in the individual's item-score pattern to estimate class membership probabilities and item success probabilities

conditional on class membership. In this study, the binomial mixture model and the latent class model were compared at the theoretical level, and applied to data obtained by means of a test for transitive reasoning from a sample of 615 children ranging in age from 6 to 13 years of age. Because the binomial mixture model is nested within the latent class model, the fit of both models could be compared directly. It was concluded that the more general latent class model is more appropriate for identifying multiple stages when tasks differ in difficulty level.

This chapter has been submitted for publication.

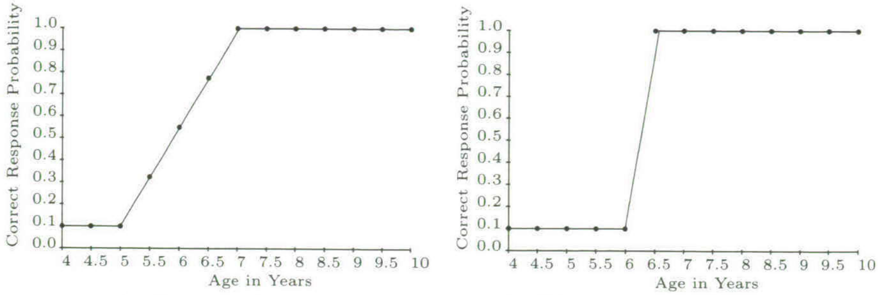
3.1 Introduction

Since Piaget formulated his developmental stage theory, the concept of discontinuity in cognitive development has become an important topic in epistemological, psychological and methodological studies. Discontinuity has been studied for several cognitive developmental abilities (Brainerd, 1978, 1993; Dolan, Jansen, & Van der Maas, 2004; Flavell, 1970; Formann, 2003; Van Geert, 1998; Jansen & Van der Maas, 1997; Thomas, 1989; Thomas & Lohaus, 1993; Thomas et al., 1999; Hosenfield et al., 1997). In this study, the focus was on the detection of discontinuity as reflected by multiple modes in the development of transitive reasoning ability.

In cognitive developmental psychology, discontinuity is used to describe the rather abrupt transition from one mode to another mode. Following Piaget, a mode may be defined as a general cognitive structure or developmental stage (e.g., Chapman, 1988; Flavell, 1985; Piaget, 1947). Alternatively, it may be conceived of as a specific rule or strategy, which is part of a particular ability (e.g., Hosenfield et al., 1997; Thomas et al., 1999; Van der Maas & Molenaar, 1992). Piaget distinguished different stages in cognitive development by means of the assumption that knowledge acquisition develops via cognitive structures which differ qualitatively (see, e.g., Case, 1992; Flavell, 1963). The transition of a cognitive structure into a different cognitive structure explains the discontinuity in the development of knowledge. It was difficult, or even impossible, to translate such general, abstract stages into an empirical setting and investigate them systematically (e.g., Brainerd, 1978; Flavell, 1970, 1985). Commenting on Brainerd (1978), Flavell (1970, p. 187) advised "to give up on grand and sweeping developmental periods that try to find a single, uniform 'deep structure' description of the thinking the child does at a given age". He encouraged to continue studying steps or levels within a single conceptual domain or subdomain. When modes are no longer taken as broad, developmental stages but defined as rules or strategies in a particular domain or subdomain, the transition from one mode to a different mode can be interpreted as discontinuity in the development of the specific ability under study.

Often the term abruptness is used to indicate discontinuity, meaning that the change curve is expected to be jumpy in a particular time interval. For example, Flavell (1970) and Brainerd (1993) agreed that a change curve as in Figure 3.1b is abrupt, showing discontinuity, while a curve as in Figure 3.1a reflects continuous change (they actually disagreed about the validity of the method that was used to determine continuous or discontinuous change). However, deciding on whether or not discontinuity is present is hampered by four problems. First, the slope of a curve depends on the unit of measurement — days, weeks, half-year periods, and so on — and the larger the unit, the steeper the slope. Second, although an observed change curve may show dramatic changes in steepness, the magnitude of the change in steepness needed to decide on discontinuity is arbitrary. Third, aspects of behavior may be related to age, but other variables have presumably a more direct, causal relationship to the behavioral changes found with age (Wohlwill, 1973, p. 26). Finally, chronological age is not a useful variable to study development of behavior, since there are considerable individual differences in rates of development, that is, one child at four years of age may attain a level at some given behavioral dimension which another child may not reach till the age of six (Wohlwill, 1973, p. 26).

Researchers used both cross-sectional and longitudinal designs in study-



a. Continuous change

b. Discontinuous change

Figure 3.1: *(a) Continuous- and (b) Discontinuous Change Curve, Based on Brainerd (1993)*

ing discontinuity in the development of Piagetian concepts. A cross-sectional design (see e.g., Hosenfield et al., 1997) describes discontinuity by means of the existence of multiple modes in the general development of an ability. These modes result from the use of a particular rule or strategy. Because a test measuring the ability of interest usually is administered only once, information about the transition from one mode to a different mode is not available and therefore hypotheses concerning this transition can not be tested. In fact, this approach agrees with Piaget's initial aim to develop a kind of encyclopedia of human cognition in which it is described which cognitive tasks a child is able to solve within a particular age range. Longitudinal designs (see, e.g., Van Geert, 1998) are appropriate when the aim is to describe the transition from one mode or stage to a subsequent mode or stage. The development of a child can be determined by means of repeated measurements during a particular time interval. Markov chain models (e.g., Brainerd, 1979) and catastrophe models (Van der Maas & Molenaar, 1992) were used to study multimodality in longitudinal designs. The choice of a cross-sectional design or a longitudinal design depends on the hypotheses to be tested and the resources available.

In general, it is assumed that an observed discontinuity in the change curve reflects discontinuity in the development of the ability of the child. However, particular properties of the items used to measure the ability may cause an artificial discontinuity in the change curve. For example, when a test contains tasks at two markedly different difficulty levels, the change curve shows discontinuity which actually reflects discontinuity in task difficulty, but not necessarily developmental discontinuity.

The problems of determining the definition, the operationalization, and the measurement of discontinuity may be of concern, in particular, when relating research results to the discussion about discontinuity in general cognitive development, ignoring (1) accounting for the appropriate level of sophistication at which discontinuity is studied; (2) implicit ideas about the size of the change in performance; (3) restrictions of the research design; and (4) properties of the measurement instruments. However, when these issues are considered carefully, conclusions about the development of a particular ability may be well founded. From a pragmatic point of view, well-founded modes offer the possibility to differentiate between groups of children that share relevant cognitive behavior.

3.1.1 Aim of This Study

Several researchers studied discontinuity in Piagetian tasks. For example, Thomas (1989) and Raijmakers, Jansen, and Van der Maas (2004) studied multimodality in classification performance; Thomas and Turner (1991), Thomas and Lohaus (1993), Thomas et al. (1999) and Formann (2003) studied multimodality in performance on the water-level task; Hosenfield et al. (1997) studied multimodality in analogical reasoning; and Van der Maas (1998), and Jansen and Van der Maas (2002) studied multimodality in performance on balance scale tasks. The aim of this cross-sectional study was to determine whether discontinuity reflected by multimodality exists in the development of transitive reasoning ability. In our research on transitive reasoning (Bouwmeester & Sijtsma, 2004; Bouwmeester et al., 2004), expectations about different modes in development arose because of different kinds of explanations children of different ages gave after they had solved transitive reasoning problems. We defined discontinuity as the existence of different strategy groups and not as a particular change in steepness of the developmental growth curve. That is, we expect a relationship between age and strategy use, but we do not expect fixed age periods to define discontinuity.

Fuzzy trace theory (Brainerd & Kingma, 1984; Brainerd & Reyna, 1995, 2001, 2004) predicts that different modes may be expected in performance on transitive reasoning tasks. Fuzzy trace theory assumes a verbatim continuum and a fuzzy continuum. Young children use verbatim information to solve problems (e.g., literal observable information), while older children use progressively more fuzzy information (i.e., degraded, pattern-like information, only holding the gist). Although the underlying verbatim continuum and fuzzy continuum are assumed to be continuous, the performance variable is expected to show at least two groups of children that are characterized by different kinds of strategies reflecting the use of either verbal or fuzzy information.

Several transitive reasoning studies showed that different kinds of strategies were used to solve transitive reasoning tasks. Verweij et al. (1999) showed that children between seven and 12 years of age used different strategies, and Bouwmeester et al. (2004) distinguished seven categories of explanations that children gave to justify their answers to different kinds of transitive reasoning problems.

3.1.2 The Binomial Mixture Model

Thomas and Lohaus (1993), Thomas and Turner (1991), Thomas et al. (1999) and Thomas and Hettmansperger (2001) used the Binomial Mixture Model (BMM) to model discontinuity in performance on the water-level task. In the water-level task children have to draw the water-level in a glass which has a particular angle with the horizontal axis. The tasks differ in the angle of the glass but can be assumed to have equal difficulty (Thomas & Hettmansperger, 2001). Hosenfield et al. (1997) used the BMM to model discontinuity in analogical reasoning.

Assume that a test consists of J tasks, which are scored correct (score 1) or incorrect (score 0). Random variables X_j $(j = 1, ..., J)$ denote these item scores $(X_j = 0, 1)$. The BMM assumes that the frequency distribution of a number-correct score on a test, $X_+ = \sum_{j=1}^{J} X_j$, consists of a mixture of binomial distributions. Further, assume c classes each of which is characterized by a binomial distribution for X_+. Let $Bin(X_+; J, \theta_u)$ be a binomial distribution for X_+, based on J trials each with a success probability θ_u $(0 < \theta_u < 1)$. Let π_u be the marginal probability of belonging to a particular binomial distribution class or component u $(u = 1, ..., c)$ with $0 \leq \pi_u \leq 1$, and $\sum \pi_u = 1$. Then, the c-classes binomial mixture distribution is defined as

$$f(X_+; J) = \sum_{u=1}^{c} \pi_u Bin(X_+; J, \theta_u). \tag{3.1}$$

Although the BMM is the model most commonly used to detect discontinuity in cross-sectional data, it has some serious drawbacks. These are discussed below.

First, Hosenfield et al. (1997) used the BMM to detect multimodality but argued that its fit does not necessarily imply bimodality or multimodality. According to Hosenfield et al. (1997, p. 532) "the presence of bi- (or multi-) modality can only be concluded if the model plot (i.e., the estimated overall frequency distribution of the number-correct score; *the authors*) displays two clearly separable peaks." We think that this conclusion is not tenable in general. A gap between two peaks may be difficult to observe in real data unless the binomial proportions differ markedly in location or the test contains a large number of tasks and, as a result, the peaks of closely located binomial distributions are clearly discernable.

Different kinds of cognitive strategies may lead to a mixture of overlapping distributions which do not clearly show gaps or peaks in the estimated overall frequency distribution. It may be true that multimodality is present when two or more peaks or gaps can be distinguished, but a distribution having one peak may in fact consist of several overlapping distributions (Thomas & Lohaus, 1993). Alternatively, when discontinuity is assumed to be revealed by the existence of multiple rules or strategies, it is not necessary to restrict the concept of discontinuity or multiple modes to observable peaks and gaps in the overall frequency distribution.

Second, the detection of discontinuity does not require a particular shape of the frequency distribution. However, the BMM assumes a mix of binomial distributions and this may unnecessarily restrict the data.

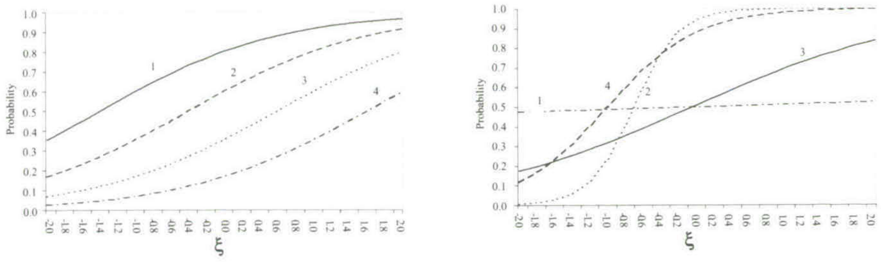
Third, the BMM assumes that the binomial probabilities (i.e., the θ's) are constant among different tasks for all individuals belonging to the same component, u; this is known as task-homogeneity (Formann, 2001, 2003). The assumption of equal task difficulty may be true for the water-level task (although this is also questionable; see Thomas & Hettmansperger, 2001), but unrealistic for many other task types. Moreover, when item difficulty influences the strategy that is used, it is inappropriate to assume that the spread in item difficulty is due to error.

Finally, by using a binomial mixture distribution it is assumed that a particular number-correct score X_+ was produced by just one strategy. For this reason, Thomas et al. (1999, p. 1025) called the BMM conservative: "It will likely find fewer strategies in the population than in fact are represented". They argued, however, that for reasons of parsimony this "is not necessarily an unattractive shortcoming". However, deciding on the existence of some strategies because others can not be revealed by the statistical method seems odd. We argue next that assuming that a particular number-correct score was produced by just one strategy is unrealistic in many practical situations (see also Formann, 2003).

Binomial Mixture Model and Item Response Theory

Many researchers (Hosenfield et al., 1997; Lohaus & Kessler, 1996; Lohaus, Kessler, Thomas, & Gediga, 1994; Thomas, 1989, 1994; Thomas et al., 1999; Thomas & Lohaus, 1993; Thomas & Turner, 1991) who used the BMM for studying discontinuity in cognitive developmental constructs assume that a particular distribution of the number-correct score X_+ or just the mode of this distribution corresponds to a developmental stage or mode. By studying a distinct ability, the mode is interpreted as corresponding to a particular strategy or rule which characterizes the development of the ability. This approach focusses on the number-correct score as the statistic of interest, but not on the individual item scores.

Other researchers (e.g., Bouwmeester et al., 2004; Jansen & Van der Maas, 1997, 2002; Raijmakers et al., 2004; and Van Maanen, Been, & Sijtsma, 1989) focussed at the item scores, and assumed that the use of a



a. Rasch model

b. Two-parameter logistic model

Figure 3.2: *Four Response Functions According to (a) the Rasch Model and (b) the Two-Parameter Logistic Model*

particular strategy implies a correct answer to item j with a probability that is typical of this strategy. For example, the use of Strategy A may imply a probability of 0.9 of having item j correct, and the use of Strategy B a probability of 0.2. Then for Strategy A a score of 1 on item j is the most likely outcome and for Strategy B a score of 0. When applied to each of the J items, this probabilistic approach implies that Strategy A is characterized by a most likely vector of J item scores, and Strategy B by a different vector. However, because the approach is probabilistic, with each strategy each of the 2^J possible item-score vectors has positive probability. This implies that, with each of the strategies, each X_+ score (which is the sum of the item-scores in an item-score vector) occurs with a particular probability. Thus, strategies are characterized by particular distributions of X_+, but the question is whether X_+ provides unequivocal information about strategies.

The idea of the BMM is to identify these X_+ distributions and associate them with strategies. Different strategies may lead to highly distinct most-likely item-score vectors which, however, have the same X_+. For example, assume that $J = 4$, Strategy A's most likely vector is $(1, 1, 0, 0)$ and Strategy B's most likely vector is $(0, 0, 1, 1)$; then for both strategies $X_+ = 2$. Here the BMM approach is likely to fail in identifying different strategies. An approach that focuses on item scores and identifies item-score vectors may be more successful.

Additional evidence supporting the use of item-score vectors comes from modern item response theory (IRT; e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997). Using the conceptual framework of IRT, we demonstrate that (1) IRT models predict that for each individual taking a test, each item-score vector has positive probability; (2) under one rather restrictive IRT model in particular, for each individual, given a fixed X_+ value, one item-score vector clearly has greater probability than the other vectors; (3) for other, more flexible IRT models, for each individual, given a fixed X_+ value, several item-score vectors may have relatively high probabilities; and (4) individuals with different ability levels but the same X_+ value may have different item-score vectors with almost the same probabilities. Together these results support the use of item-score vectors to identify strategies rather than the aggregated X_+ score.

1. *IRT models.* IRT models assume a continuous latent variable, here denoted ξ, instead of discrete components, u. For person s with latent variable location ξ_s, the probability of a correct answer to item j is denoted $P_j(\xi_s)$, and the probability of an incorrect score by $Q_j(\xi_s) = 1 - P_j(\xi_s)$. Using the typical IRT assumption of local independence (Embretson & Reise, 2000, p. 48), the probability of an item-score vector is the product of probabilities $P_j(\xi_s)$ and $Q_j(\xi_s)$. For example, for $J = 4$ the probability of item-score vector (1,1,0,1) is

$$P[(1, 1, 0, 1)|\xi_s] = P_1(\xi_s)P_2(\xi_s)Q_3(\xi_s)P_4(\xi_s). \tag{3.2}$$

Arbitrarily assume that for person s the four probabilities of a correct answer are (0.8, 0.7, 0.8, 0.6). Then, according to Equation 3.2 the observed item-score pattern (1,1,0,1) has probability $0.8 \times 0.7 \times 0.2 \times 0.6 = 0.0672$. Also consider the other three item-score vectors which result in $X_+ = 3$ [i.e., (1,1,1,0), (1,0,1,1), and (0,1,1,1)]. For person s with ξ_s, the probabilities are 0.1792, 0.1152, and 0.0672, respectively. Thus, in general IRT models imply that all item-score vectors have positive probability.

2. *Rasch model.* Let δ_j denote the difficulty parameter of item j. Under the Rasch model items elicit performance according to response function

$$P_j(\xi) = \frac{\exp(\xi - \delta_j)}{1 + \exp(\xi - \delta_j)}. \tag{3.3}$$

For $J = 4$, Figure 3.2a displays the response functions for realistic item parameters $\delta_j = -1.4, -0.4, 0.4, 1.4$ [see Thissen and Wainer (1982) for a reasonable range of location parameters]. For these δs, Table 3.1 shows the probabilities of the $2^J = 16$ item-score vectors for five realistic values of ξ (of which the distribution usually is normed to have a mean equal to 0 and a variance equal to 1; see, e.g., Hoijtink & Boomsma, 1995). Given a fixed ξ value, it can be verified that for each X_+ score one item-score vector has the greatest probability, and other vectors have smaller but non-zero probabilities. This seems to support the "one X_+ value, one strategy" assumption to some extent.

Table 3.1: *Probabilities of Item-Score Vectors Using the Rasch Model, and the Two-Parameter Logistic Model*

Vector No.	Vector	X_+	ξ									
			-2	-1	0	1	2	-2	-1	0	1	2
			Rasch					Two-par. logistic				
1	0000	0	.48	19	.04	.00	.00	.38	.13	.00	.00	.00
2	0001	1	.26	28	.16	.04	.01	.08	.06	.00	.00	.00
3	0010	1	.10	11	.06	.01	.00	.05	**.13**	.01	.00	.00
4	0100	1	.04	05	.03	.01	.00	.00	.04	.03	.00	.00
5	1000	1	.02	02	.01	.00	.00	.34	**.13**	.00	.00	.00
6	0011	2	.05	16	.23	.16	.05	.01	.06	.01	.00	.00
7	0101	2	.02	07	.10	.07	.02	.00	.02	.03	.01	.00
8	1001	2	.01	03	.04	.03	.01	.07	.06	.00	.00	.00
9	0110	2	.01	03	.04	.03	.01	.00	.04	.20	**.14**	.08
10	1010	2	.00	01	.01	.01	.00	.05	**.13**	.01	.00	.00
11	1100	2	.00	00	.01	.00	.00	.00	.04	.03	.00	.00
12	0111	3	.01	04	.16	.28	.26	.00	.02	**.21**	.33	.40
13	1011	3	.00	01	.06	.11	.10	.01	.06	.01	.00	.00
14	1101	3	.00	01	.03	.05	.04	.00	.02	.03	.01	.00
15	1110	3	.00	00	.01	.02	.02	.00	.04	**.20**	.15	.09
16	1111	4	.00	00	.04	.19	.48	.00	.02	.21	.35	.44

boldface: item-score vector probability which is close to another item-score vector probability for item-score vectors with the same X_+.

3. *Two-parameter logistic model.* Next, consider the more flexible, much used two-parameter logistic model (Embretson & Reise, 2000; p. 70). Compared to the Rasch model, this model adds a slope parameter,

denoted α_j, which is comparable to the regression parameter in the logistic regression model (e.g., Agresti, 1990, pp. 85-87). The response function is defined as

$$P_j(\xi) = \frac{\exp[\alpha_j(\xi - \delta_j)]}{1 + \exp[\alpha_j(\xi - \delta_j)]}. \tag{3.4}$$

Figure 3.2b displays four response functions based on $\delta_j = 0.0, 0.7, 1.0, 0.05$ ($j = 1, \ldots, 4$) and $\alpha_j = 0.05, 4.0, 2.0, 0.8$, respectively. This variation in slope parameters reflects variation in strength of relationship of the items with the latent variable, and is assumed to be due to variation in item properties. Different properties may be related to different strategies. Table 3.1 shows the probabilities corresponding to the 16 item-score vectors, using the same ξ values as used for the Rasch model calculations. For $\xi = -1$ and $X_+ = 1$, the item-score vectors 3 and 5 have the highest probabilities, which are close (rounded values of 0.13 and 0.13, respectively). For $\xi = 0$ and $X_+ = 3$, for the item-score vectors 12 and 15 a similar result is obtained (0.21 and 0.20, respectively). Thus, the two-parameter logistic model accommodates the situation that a particular strategy leads to two or more "most-likely" item-score vectors for a particular X_+ score.

4. *Different ξ, same X_+*. People with markedly different ξ values who produced the same X_+ score may have different item-score vectors, giving evidence of different strategies; see Table 3.1, two-parameter logistic model, $X_+ = 2$, $\xi = -1$ with vector (1,0,1,0) and $\xi = 1$ with vector (0,1,1,0), and vector probabilities of 0.14 and 0.13, respectively. Based on the same X_+ value it would be concluded that both respondents used the same strategy. Based on the item-score vectors and given estimates of the latent variable values for both respondents, and assuming that use of different strategies is related to proficiency level, it would be concluded that different strategies had been used.

Several other examples could be constructed showing that using the aggregate number-correct X_+ score for identifying strategies leads to a loss of information, and that the use of the finer-grained item-score vector is better suited for this purpose.

3.1.3 The Latent Class Model

The BMM is a restrictive Latent Class Model (LCM). In an LCM, unidimensionality is not assumed and within a latent class the item parameters are not restricted to be equal. The LCM is a mixture model (Lazerfield & Henry, 1968; see also Hagenaars & McCutcheon, 2002; McCutcheon, 1987), which allows heterogeneity in both individual performance and task difficulty (Formann, 2003). In fact, the binomial model and the BMM are special cases of the LCM with some additional restrictions on the number of classes and the parameters. Classes have prevalence or class probabilities π_u $(0 \leq \pi_u \leq 1, \sum_{u=1}^{c} \pi_u = 1)$. Each class has class-specific parameters, $\theta_{1|u}, ..., \theta_{J|u}$, related to tasks $1, ..., J$, respectively. Let the vector $\mathbf{X}$ contain the item-score variables $[\mathbf{X} = (X_1, ..., X_J)]$, then the LCM is defined as:

$$f(\mathbf{X}) = \sum_{u=1}^{c} \pi(u) \prod_{j=1}^{J} p(\theta_j|u). \tag{3.5}$$

Unlike the BMM, which uses the number-correct score X_+, the LCM uses the vector of scores on the J tasks in the test, $\mathbf{X}$. This difference has important consequences. First, the BMM assumes that the difficulty level of the tasks is equal for individuals in the same component. When the item-score vectors are used, the difficulty level may vary across tasks within a latent class. Second, the same number-correct scores may be based on correct responses to different subsets of tasks. Therefore, different item-score vectors with the same number-correct score may have been produced by different strategies. These strategies are not distinguished by the BMM.

So far the LCM was applied rarely to study discontinuity in cognitive development (for exceptions, see Formann, 2001, 2003; and Jansen & Van der Maas, 2002). The results of an LCM analysis are more difficult to interpret than those of a BMM analysis, mainly because of the varying success probabilities of the tasks. However, Formann (2001, 2003) studied discontinuity in the development of performance on the water-level task, and showed that accepting a well-fitting BMM may be misleading without having additionally evaluated the fit of LCMs.

LCM analysis is not without problems. Van der Maas (1998) and

Raijmakers et al. (2004) argued that an LCM analysis with more than six classes is not feasible using the available techniques. Large sample sizes are needed to prevent sparse frequency tables and p-values associated with the asymptotic χ^2 statistics which can not be trusted. To handle this problem, Formann (2001) used bootstrap procedures to estimate a p-value. A bootstrap procedure is also implemented in the Latent Gold program (Vermunt & Magidson, 2003) which can estimate and fit a large number of LCMs.

3.1.4 Hypotheses

In this study the purposes were to detect discontinuity in transitive reasoning; to determine the relationship of discontinuity and age; and to give a substantive interpretation of the strategy groups. Further, the fit of the BMM and the LCM to transitive reasoning data was evaluated. Before hypotheses about discontinuity could be tested, it had to be determined that the discontinuity in the data could not be attributed to discontinuity by the instrument. The following hypotheses cover these purposes:

1 *The development of transitive reasoning is discontinuous. This is reflected by ordered strategy groups.*

2 *When tasks differ in difficulty, the LCM gives a substantively better explanation of the discontinuity than the BMM, because the LCM accounts for the relationship between task difficulty and strategy used.*

3 *Individual differences in performance on transitive reasoning tasks produce discontinuity within age groups. Clearly defined age periods cannot be distinguished.* Piaget distinguished broad developmental stages which were delimited by age periods. We hypothesized that strategy groups give a clearer description of the discontinuity than age groups because we expect large individual differences in task performance.

4 *Fuzzy trace theory offers an interpretation of the discontinuity.* Fuzzy trace theory offers a framework for interpreting the discontinuity in

transitive reasoning by distinguishing verbatim and fuzzy trace abilities. It was hypothesized that one particular strategy group uses verbatim trace information to solve the tasks and another strategy group uses fuzzy traces.

To determine discontinuity in cross-sectional transitive reasoning data both the BMM and the LCM were fitted to the data. First, to rule out as much as possible the use of different strategies leading to the same number-correct score, the dimensionality of the data collected by means of a transitive reasoning test was determined using a nonparametric version of the Rasch model. This is the double monotonicity model [DMM; Mokken, 1971, pp. 174-176; Sijtsma & Molenaar, 2002, chap. 6; also see this reference for an introduction into nonparametric IRT]. Second, the proportions correct of the items were calculated to determine whether the instrument causes discontinuity, thus inducing method bias. Third, LCMs were fitted to the data to determine whether there was discontinuity and if so, how many classes had to be distinguished (Hypothesis 1). Data from separate age groups were analyzed because important age differences might be masked in a pooled data set (Hypothesis 3). Fourth, the BMM was fitted to the data and compared with the LCM results to assess how much fit was lost when restricting the item parameters to be equal between classes (Hypothesis 2). Fifth, the latent classes were interpreted by means of verbal explanation data in order to determine whether fuzzy trace theory was suited for interpreting discontinuity (Hypothesis 4).

3.2 Method

3.2.1 Sample

The pooled sample consisted of 615 children stemming from Grade two through Grade six of six elementary schools in the Netherlands. Children were from middle class social-economic status families. Table 3.2 gives an overview of the number of children of six age groups, and the mean and the standard deviation of age.

Table 3.2: *Number of Children (n), Mean Age (M) and Standard Deviation (SD) per Age Group*

Age Group[a]	*n*	*M*	*SD*
≤ 96	73	91.78	3.057
97—108	83	103.02	3.138
109—120	126	114.45	3.305
121—132	108	126.70	3.084
133—144	116	138.70	3.005
$\geq$145	59	149.46	3.464

[a] number of months

3.2.2 Material

Transitive reasoning ability was investigated by means of a computerized test containing 16 transitive reasoning tasks (Bouwmeester & Aalbers, 2002). The tasks differed on three task characteristics. The task characteristics had 4, 2, and 2 levels, defining $4 \times 2 \times 2 = 16$ tasks. The task characteristics are summarized in Table 3.3. See Figure 2.1, chapter 2 for an overview of the tasks.

Table 3.3: *Description of the Transitive Reasoning Task Characteristics*

Characteristic	Level	Description
Format	$Y_A > Y_B > Y_C$ $Y_A = Y_B = Y_C = Y_D$ $Y_A > Y_B > Y_C > Y_D > Y_E$ $Y_A = Y_B > Y_C = Y_D$	Defines the logical relationships between the objects involved, e.g., when the relationship is length, $Y_A > Y_B > Y_C$ means that object A is longer than object B, which is longer than object C.
Presentation Form	Simultaneous Successive	Determines whether all objects are presented simultaneously or in pairs during premise presentation.
Content of Relationship	Physical Verbal	Determines whether the relationships can be perceived visually, or are told in words by the experimenter.

3.2.3 Procedure

The transitive reasoning test was an individual test administered in a quiet room in the school building. Before the child was confronted with the actual test tasks, the experimenter explained the different kinds of objects and relationships that were used in the tasks. The administration of the test took approximately half an hour, depending on the age of the child. For more details see chapter 2.

3.2.4 Scoring

For each task the answer was automatically recorded by the computer. A verbal explanation of the answer given by the child was recorded by the experimenter. When the child explained the transitive relationship correctly by mentioning the premises involved or the linear ordering of the objects, the explanation was evaluated to be correct. All other explanations were incorrect. The correct/incorrect explanations were used in the analyses because previous research showed that the explanations were more valid indicators of the underlying ability than the correct/incorrect judgements (Bouwmeester & Sijtsma, 2004; see chapter 2 of this thesis).

3.2.5 Verbal explanation

The correct/incorrect strategy scores were a dichotomization of an original explanation variable having 13 categories. This explanation variable was used to interpret the latent classes. For this purpose we recoded the variable into four categories: (1) children used all the premise information in their explanation (literal premise information), or children give a correct explanation of the ordering (reduced premise information); (2) children used premise information, but incompletely or incorrectly; (3) children used visual information or irrelevant external information in their explanation; and (4) children did not give an explanation (See Figure 2.2, chapter 2).

3.3 Results

3.3.1 Double Monotonicity Model

Because one task was correctly answered by only seven children, it was considered not to be suited for further analysis. The DMM fitted the remaining 15 tasks.[1] Because response functions did not intersect, it was likely that a fixed number-correct score was driven mainly by one strategy (Bouwmeester & Sijtsma, 2004), see chapter 2. Therefore, multiple modes or strategies found by the BMM or the LCM were expected to be ordered along a unidimensional scale.

Table 3.4: *Proportion Correct of the Transitive Reasoning Tasks*

item #	Format	Presentation	Content	P_j
6	$Y_A > Y_B > Y_C$	successive	physical	.05
15	$Y_A > Y_B > Y_C > Y_D > Y_E$	successive	physical	.07
5	$Y_A = Y_B > Y_C = Y_D$	simultaneous	verbal	.15
14	$Y_A = Y_B > Y_C = Y_D$	successive	verbal	.19
8	$Y_A > Y_B > Y_C > Y_D > Y_E$	successive	verbal	.21
11	$Y_A = Y_B > Y_C = Y_D$	simultaneous	physical	.31
4	$Y_A > Y_B > Y_C > Y_D > Y_E$	simultaneous	physical	.39
12	$Y_A > Y_B > Y_C$	successive	verbal	.40
3	$Y_A = Y_B = Y_C = Y_D$	successive	verbal	.45
10	$Y_A > Y_B > Y_C > Y_D > Y_E$	simultaneous	verbal	.52
9	$Y_A = Y_B = Y_C = Y_D$	successive	physical	.54
1	$Y_A > Y_B > Y_C$	simultaneous	verbal	.56
13	$Y_A > Y_B > Y_C$	simultaneous	physical	.57
7	$Y_A = Y_B = Y_C = Y_D$	simultaneous	physical	.77
16	$Y_A = Y_B = Y_C = Y_D$	simultaneous	verbal	.86

The proportions correct (Table 3.4) differed widely. Thus, no distinct

[1] The H-value of the scale was 0.45, and the H^T-value of the scale was 0.52 (see Sijtsma & Molenaar, 2002).

subgroups of items could be distinguished which might cause discontinuity. It was concluded that the instrument may be ruled out as a cause of discontinuity. When discontinuity is found, we will attribute it to a developmental process.

3.3.2 Latent Class Model Analysis

The program Latent Gold 3.0 (Vermunt & Magidson, 2003) was used to estimate the parameters and compute the evaluation statistics of the BMM and the LCM. Two evaluation statistics were computed. First, the likelihood-ratio chi-squared statistic L^2 gives an indication of the fit of the model to the data. The bootstrap p-value, denoted p_{boot}, was used to determine whether the results were significant (using $\alpha = .05$). Second, the Bayesian Information Criterion [defined as $-2LL + \#parameters \times ln(N)$], denoted BIC, served as a selection criterion within the family of models fitted to the same data set. The BIC weights the fit (LL) and the parsimony [$\#parameters \times ln(N)$] of a model: The lower the BIC, the better the model in terms of parsimony.

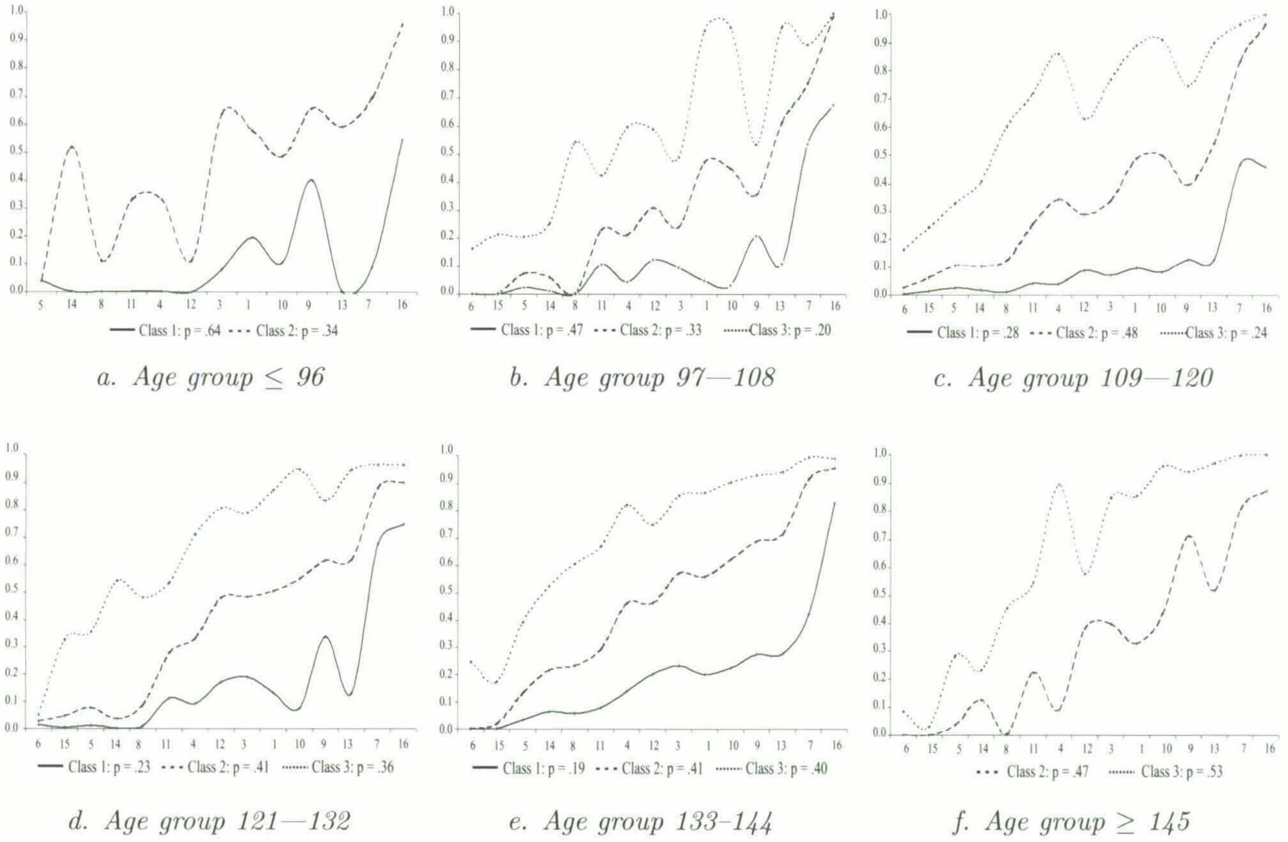
Because the DMM fitted the data, latent class one-factor models were fitted in which the latent classes were ordered on one dimension. In latent class *cluster* models, the latent classes have a nominal measurement level, in latent class *factor* models the latent classes have an interval measurement level. We first fitted latent class factor models having one, two, and three latent classes for each age group. The results of the model fit are shown in Table 3.5. Because two tasks had zero score-variance in the first age group, only 13 tasks were fitted in this group. The most important result in all age groups was the difference in fit of the models having one and two classes. The decrease in L^2 and BIC was large indicating discontinuity in transitive reasoning. Therefore, Hypothesis 1 was accepted. The bootstrap p-values showed that the two-class models could not be rejected in any of the age-groups. A remarkable result was the p-value indicating that the one-class model could not be rejected in age group 133-144, and in age group ≥ 144. The decrease of the L^2 and BIC suggested that the two-class model fitted much better. The difference in fit between the two-class and three-class models was unequal between the different age groups. In particular in the first and last age-groups (≤ 96 months, ≥ 145 months) the gain in fit of the three-class model was small. For the other age-groups the decrease in L^2 and BIC from the two-class model to the three-class model was somewhat larger.

Table 3.5 also shows the fit of the BMM. The BMM was estimated as an LCM with restrictions: restrict the task parameters to be equal within a latent class ($\theta_{j=1,u} = \theta_{j=2,u} = ... = \theta_{j=J,u}$), and the same model can be estimated as when using the number-correct score. The advantage was that the BMM and the LCM could be compared directly.

On the basis of the fit of several BMMs it was difficult to decide whether there was discontinuity in the transitive reasoning data. The one-class model was rejected in all age-groups and the two-class models and the three-class models were also rejected in age-groups $109-120$ and $121-132$. However, the p-values of the models in the other age-groups were rather small ($< .09$). More important, the L^2 and BIC values of the BMMs were much higher than those of the LCMs in all age groups and for all models, indicating that restricting the item parameters deteriorates the fit and masks possible discontinuity in the data structure.

These results showed that ignoring variation in task difficulties within a latent class was inappropriate in the context of transitive reasoning. The estimated success probabilities of the LCM's confirmed this result (see Figure 3.3). Therefore, Hypothesis 2 was accepted.

For the latent class factor model and for each age group, Figure 3.3 shows the success probabilities of the 15 tasks (13 tasks for age group ≤ 96) given class membership. To facilitate the readability of the plots, the order of the tasks is in accordance with their difficulty level (see Table 3.4). The plots show that allowing the tasks to differ in difficulty level resulted in highly varying success probabilities. The marginal success probabilities of being in a particular latent class (printed below panels in Figure 3.3) increased over age groups for the high ability class and decreased over age groups for the low ability class. With respect to the first age group (≤ 96 months), the success probabilities of the two-class model were plotted.



a. Age group ≤ 96

b. Age group 97—108

c. Age group 109—120

d. Age group 121—132

e. Age group 133–144

f. Age group ≥ 145

Figure 3.3: *Estimated Success Probabilities of the Tasks Within Latent Classes in Each of the Age Groups*

Table 3.5: *LCM and BMM Fit Statistics, Per Class and Per Age Group*

			LCM					BMM		
Age group	# classes	L^2	p_{boot}	BIC-value*	#par	# classes	L^2	p_{boot}	BIC-value*	#par
$\leq$ 96	1	388.22	.00	947.70	13	1	644.39	.00	1152.05	1
	2	239.94	.20	859.86	27	2	575.27	.06	1091.57	3
	3	233.31	.15	857.54	28	3	571.65	.08	1096.58	5
97—108	1	559.68	.00	1338.12	15	1	903.07	.00	1618.52	1
	2	375.89	.26	1226.33	31	2	822.65	.05	1547.10	3
	3	362.70	.21	1217.64	32	3	820.72	.06	1554.17	5
109—120	1	978.97	.00	2211.40	15	1	1419.48	.00	2583.65	1
	2	708.24	.07	2018.67	31	2	1260.10	.01	2434.03	3
	3	668.46	.09	1983.77	32	3	1236.89	.03	2420.56	5
121—132	1	894.59	.02	2093.69	15	1	1428.86	.00	2560.47	1
	2	663.10	.16	1939.33	31	2	1315.24	.00	2456.50	3
	3	636.67	.13	1917.71	32	3	1305.87	.01	2456.70	5
133—144	1	909.28	.19	2199.97	15	1	1463.98	.00	2686.63	1
	2	692.07	.35	2060.51	31	2	1349.60	.07	2581.97	3
	3	666.99	.31	2040.29	32	3	1342.31	.08	2584.39	5
$\geq$ 145	1	422.15	.26	1021.39	15	1	835.97	.00	1376.56	1
	2	296.74	.32	963.01	31	2	799.51	.07	1348.48	3
	3	293.02	.25	963.48	32	3	798.37	.08	1355.71	5

*: BIC-value $= -2LL + \#parameters \times ln(N)$

This age group did not have a high ability curve as the other age groups. For the last age group ($\geq$ 145 months), also the two-class model probabilities were plotted. This age group did not have a low ability curve as found in the other age groups. Figure 3.3 shows that different performance groups could be distinguished within age groups, meaning that clearly discernable age periods were inappropriate. This result led to the acceptation of Hypothesis 3.

3.3.3 Interpretation of Discontinuity

Table 3.6 shows the percentages of the explanation categories for the two or three latent factor classes of each age group. Note that in rows percentages sum to 100. The most important result is that the interpretation of the latent classes is the same for all age groups. The two classes of the first age group ($\leq$ 96 months) can be interpreted as the low- and intermediate ability latent classes. The two classes of the last group ($\geq$ 145 months) can be interpreted as the intermediate and high ability latent classes. Figure 3.3 shows that the marginal probabilities of being in a latent class differ between age groups (printed below panels in Figure 3.3).

The percentages showed that in the third latent class most children used correct premise information. That is, they used the literal premise information to infer the transitive relationship or they used the ordering of the premises. For some tasks, children used the premise information, but incorrectly or incompletely. The percentages of the categories external/visual information and no explanation were very small meaning that, in terms of fuzzy trace theory, children mostly used fuzzy trace information to solve the tasks and rarely verbatim trace information.

Children in the second latent class used the premise information, but more often incorrectly than correctly. Moreover, the percentage of no explanation is higher in the second class than in the third class. In terms of fuzzy trace theory, children in the second latent class often used fuzzy trace information but not always the correct trace.

The first class is characterized by a relatively high percentage of external and visual information and no explanation. Children in the first latent class

Table 3.6: *Average Percentage of the Explanation Type that is Used, Per Class and Per Age Group*

Age group	class	correct premise use	incorrect premise use	external or visual	no explanation
	1	9	27	22	41
≤ 96	2	41	35	5	18
	1	12	37	28	24
97—108	2	35	42	5	17
	3	58	35	2	4
	1	9	52	21	18
109—120	2	36	43	8	12
	3	68	22	2	7
	1	16	50	13	21
121—132	2	41	41	6	12
	3	68	23	2	7
	1	18	49	13	21
133—144	2	45	42	3	10
	3	71	23	1	4
≥ 145	2	32	55	5	8
	3	65	29	1	5

sometimes also used the premise information, but incorrectly in most cases. In terms of fuzzy trace theory, children in the first latent class mostly used verbatim trace information but this information does not lead to a correct inference of the transitive relationship.

3.4 Discussion

This study showed that there is discontinuity in the development of transitive reasoning which is reflected by strategy groups. The results of model fitting indicate that development can be described by three ordered classes of low-ability, intermediate-ability and high-ability levels, in which children differ in the kind of information they use to solve the tasks. The classes could be interpreted well by the explanation children gave after they had answered the task. In terms of fuzzy trace theory the classes could be called verbatim-trace class, verbatim/fuzzy-trace class, and fuzzy-trace class. Bouwmeester et al. (2004) used a latent class regression model to investigate the relationships between the explanations children used when answering the tasks, the influence of the task characteristics on performance, and age. They showed that task characteristics (determining item difficulty) had an important influence on strategy use (see chapter 4). A longitudinal study would be better suited to investigate the transition from one mode to another.

This study also showed that the BMM did not result in a useful description of the discontinuity in terms of strategy groups. Although the results of the BMM for age groups led to the conclusion that there was discontinuity, it was difficult to decide how many strategy groups had to be distinguished. Because the transitive reasoning tasks clearly varied in difficulty level, the BMM was too restrictive. Moreover, Bouwmeester et al. (2004, see also chapter 4) showed that, due to task characteristics, the difficulty of tasks influenced the strategy that was used, indicating an interaction between strategy and task characteristics. Thus, ignoring the task difficulty level is not appropriate when studying discontinuity of a cognitive ability that is measured by means of tasks which vary in difficulty.

Fixed age periods that matched useful developmental stages in transitive reasoning could not be identified. Discontinuity is observable in particular in the different strategies that are used. Children of a particular age have a most likely strategy and smaller probabilities of using other strategies. Discontinuity, in this sense, can be interpreted as a probabilistic concept. This result agrees with Wohlwill (1973, pp. 25-27, and chap. 9) who recommends to use other variables than chronological age when describing change in behavior and approaches behavioral change from a differential approach in which fixed age-groups have no meaning.

Our suggestion for future researchers who investigate discontinuity in the development of a particular cognitive ability is to first fit an unrestricted LCM to the data. Next, when an LCM fits and also the DMM model (or the Rasch model) fits, the BMM may be attempted for reasons of parsimony when equal difficulty levels within a class are hypothesized to be realistic.

Chapter 4

Latent Class Regression Analysis for Describing Cognitive Developmental Phenomena: an Application to Transitive Reasoning

Abstract*

The aim of cognitive developmental research is to explain latent cognitive processes or structures by means of manifest variables such as age, cognitive behavior, and environmental influences. In this paper the usefulness of the latent class regression model is discussed for studying cognitive developmental phenomena. Using this model, the relationships between latent and manifest variables can be explained by means of empirical data without the need of strong a priori assumptions made by a cognitive developmental theory. In the latent class regression model a number of classes are distinguished which may be characterized by particular cognitive behavior. Environmental influences on cognitive behavior may vary for different (developmental) classes. An application is given of the latent class

regression model to transitive reasoning data. The results showed that a five-class model best fitted the data and that the latent classes differ with respect to age, strategy use (cognitive behavior) and the influence of task characteristics on the strategy use (environmental influences). The flexibility of the model in terms of mixed measurement levels and treatment of different cognitive variables offers a broad application to several cognitive developmental phenomena.

*This chapter has been published as: 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.*

4.1 Introduction

The general aim of cognitive developmental research is the uncovering of relationships between cognitive processes, environmental influences and age (see e.g., Flavell, 1985; Siegler, 1991). Because cognitive processes can not be observed directly but only inferred from observable variables, observable cognitive behavior is assumed to indicate the latent cognitive processes. In Figure 4.1, a general model is displayed of the relationships between observed and latent variables in the domain of cognitive development. The definition and operationalization of the different aspects and relationships in Figure 4.1 varies for different cognitive developmental theories and the epistemological assumptions about the acquisition of knowledge. Moreover, cognitive developmental theories have different perspectives on the importance of the aspects (Figure 4.1) and how they should be measured.

For example, in the theory of Piaget (see e.g. Flavell, 1963; Chapman, 1988; Bidell & Fischer, 1992), cognitive abilities are assumed to develop in stages which are characterized by a particular kind of knowledge structures. One of the most important purposes of Piaget was to give a broad description of the developing structures. Therefore his theory was domain-general without paying much attention to the influence of external

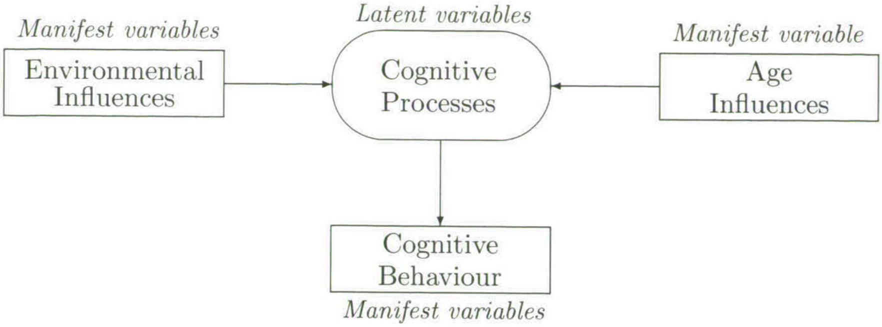


Figure 4.1: *A General Model for the Relationship Between Manifest and Latent Variables in the Domain of Cognitive Development*

conditions (Case, 1992). In information processing theory (see e.g. Kail & Bisanz, 1992), however, development is defined as cumulative learning without qualitative change. External experiences make it possible to acquire knowledge, that is, to learn and develop cognitively.

Dependent on the theoretical perspective, assumptions are made about the unobservable (latent) processes and how these processes should be measured using observable variables. Given the assumptions, relationships between observable variables such as age, task conditions and cognitive behavior, and unobservable variables, such as cognitive processes, are modeled. By studying the observable variables empirically or by means of computer simulation, one wants to reveal the latent cognitive processes and the relationships between these cognitive processes, environmental influences and age.

However, it is difficult to test a model empirically in which both the observed and the unobserved variables are represented, that is, to estimate and test relationships between observable and unobservable variables without the need of strong cognitive theoretical assumptions. Nevertheless, statistical models in which latent variables can be defined using manifest variables do exist and can be used to study relationships between age, environmental influences and cognitive processes (Embretson, 1985, 1991; Fischer, 1995; Kelderman & Rijkes, 1994; Mislevy & Verhelst, 1990; Sijtsma

& Verweij, 1999).

In modern test theory, for example, the observed responses to a number of tasks (e.g., arithmetic problems), which measure a particular ability (e.g. arithmetic ability) are used to determine the number of latent abilities needed for explaining the observable data structure, and the strength of the relationships between the item scores and these latent abilities. Thus, modern test models, also known as item item response theory (IRT) models (see, e.g., Hambleton & Swaminathan, 1985; Sijtsma & Molenaar, 2002), make it possible to reveal and statistically test a latent structure for explaining the data without the need to posit an a priori theoretical structure stipulated by cognitive theory.

In IRT models the latent variable is continuous, whereas latent class models (e.g., Hagenaars & McCutcheon, 2002) assume latent abilities to be discrete consisting of two or more nominal or ordered classes. In particular when studying cognitive development these latent class models are useful to distinguish groups of children on a developmental scale which are characterized by a pattern of specific cognitive behavior. The cognitive behavior in a specific latent class may differ, in a particular aspect, from the cognitive behavior in other latent classes. Latent class models allow the estimation of the classes of the latent variable from the data instead of assuming them on the basis of a cognitive theory. However, latent class models can also be used in a confirmatory way by testing the latent class structure assumed by a cognitive theory (chapter 5).

In the domain of cognitive developmental theory, age is hypothesized to have influence on the formation of the latent classes. One may expect that a particular latent class, which is characterized by specific cognitive behavior, may fit better for children of a particular age range than for children outside this range. Latent class analysis makes it possible to empirically determine the influence of age (as a covariate) on the formation of the latent classes.

A division into classes does not necessarily imply a cognitive stage theory. In contrast, the cognitive behavior typical of a latent class may be an expression of the same underlying ability continuum. The classes may be ordered and it depends on the level of description of the observed variables whether the interpretation of the latent classes differs quantitatively or qualitatively. For example, Bouwmeester and Sijtsma (2004) found that the response patterns of children on a set of transitive reasoning tasks could be explained by one ability, but that a broad variety of explanations were used to motivate the responses. Possibly, on a more detailed level, the transitive reasoning ability can be divided into a number of classes which are characterized by a specific pattern of cognitive behavior.

The power of the latent class model is that specific behavior patterns can be distinguished and the influence of age determined without a priori cognitive theoretical assumptions. However, it is possible to test a cognitive stage theory using latent class models. Jansen and Van der Maas (1997) used a latent class model to empirically study the different stages of reasoning on the balance scale task (Inhelder & Piaget, 1958; Siegler, 1976) and found that the theoretical stages were, together with some others classes, represented by the latent classes.

An additional possibility of latent class models is to describe the classes in more detail by assessing the influence of certain external conditions on cognitive behavior in a particular class and compare classes with respect to the influence of external conditions on cognitive behavior in a set of classes. For this purpose, we used a latent class regression model (Wedel & DeSarbo, 1994; Vermunt & Magidson, 2000) in which a multiple regression function is estimated for a number of classes. The formation of the latent classes is influenced by the covariate age. For every latent class, the influence of external conditions on the cognitive behavior can be determined. This latent class regression model is a very general and flexible model which can be applied to a broad range of cognitive developmental phenomena. Examples are the development of reasoning on the balance scale task (see e.g., Jansen & Van der Maas, 1997), transitive reasoning (see e.g., Verweij, 1994), inductive reasoning (see e.g., De Koning, 2000), and analogical reasoning (see e.g., Hosenfield, 2003).

Both the covariate, the dependent and the predictor variables can have different measurement levels. For example, instead of age in months, grade level can be used as a covariate or other child characteristics like gender, cultural background or social economic status. Cognitive behavior may be operationalized as correct/incorrect responses, strategy information, verbal explanations, or reaction times. Predictors may be all kinds of external conditions. For example, tasks may vary in specific task characteristics, or the experiment may take place on different locations or at different times. In the next section an application of the latent class regression model is given in the context of transitive reasoning development.

4.2 An Application

In a transitive relationship, the unknown relationship R between two elements A and C can be inferred from their known relationships with a third element B; that is $(R_{AB}, R_{BC}) \Rightarrow R_{AC}$. In this example, the relationships R_{AB} and R_{BC} are premises. In the research on transitive reasoning a number of different task characteristics are used to study the ability of transitive reasoning. Different kinds of transitive and non-transitive strategies appeared to be used to draw transitive inferences in tasks having different task characteristics (Perner & Mansbridge, 1983; Verweij, 1994; Bouwmeester & Sijtsma, 2004). In the last decades a discussion has been taken place about which kinds of cognitive behavior are really expressions of transitive reasoning; which kinds of tasks should be used to measure transitive reasoning; and what really develops when studying transitive reasoning (see, e.g., Smedslund, 1969; Trabasso, 1977; Brainerd & Reyna, 1992; Chapman & Lindenberger, 1992). Therefore, it is important to reveal the relationships between age, cognitive behavior, and external conditions, when studying the development of this cognitive developmental phenomenon.

4.3 Method

4.3.1 Instruments

Bouwmeester and Sijtsma (2004) investigated transitive reasoning by constructing a computer test containing 16 transitive reasoning tasks. The tasks differed on three important external conditions, called task characteristics. The task characteristics had 4, 2, and 2 levels defining $4 \times 2 \times 2 = 16$ tasks. A description of the task characteristics is given in Table 4.1. (See also Figure 2.1, chapter 2.)

Table 4.1: *Description of the Transitive Reasoning Task Characteristics*

CHARACTERISTIC	Level	Description
FORMAT	$Y_A > Y_B > Y_C$ $Y_A = Y_B = Y_C = Y_D$ $Y_A > Y_B > Y_C > Y_D > Y_E$ $Y_A = Y_B > Y_C = Y_D$	Defines the logical relationships between the objects involved, e.g., when the relationship is length, $Y_A > Y_B > Y_C$ means that object A is longer than object B, which is longer than object C.
PRESENTATION FORM	Simultaneous Successive	Determines whether all objects are presented simultaneously or in pairs during premise presentation.
CONTENT OF RELATIONSHIP	Physical Verbal	Determines whether the relationships can be perceived visually, or are told in words by the experimenter.

4.3.2 Strategies

For each task both the correct/incorrect responses and the verbal explanations were recorded. The verbal explanations associated with the correct/incorrect responses showed that children used a broad variety of explanations but that this differentiation could not be discovered by considering only the correct/incorrect responses. Moreover, Bouwmeester and Sijtsma (2004) showed that correct/incorrect responses to the tasks of the transitive reasoning test did not form one reliable ability scale. Thus, we used the verbal explanations data in this study. These verbal explanations were categorized into seven strategies, which are displayed in Table 4.2.

4.3.3 Sample

The sample consisted of 615 children stemming from Grade two through Grade six. Children came from six elementary schools in the Netherlands. They were from middle class social-economic status (SES) families. Table

Table 4.2:
Description of the Seven Strategies Used to Solve the Transitive Reasoning Tasks.

Name	Description	Example
LITERAL	All necessary premise information is used to explain the transitive relationship.	Object A is longer than object C because object A is longer than object B and object B is longer than object C.
REDUCED	The premise information is used in a reduced form.	Animals get older to the right, so the horse is older than the cow because it is positioned before the cow.
INCORRECT	Premise information is incorrectly used, or incorrect premise information is used.	The lion is older than the camel because the hippo and the lion have the same age.
INCOMPLETE	Premise information is used correctly but incompletely.	the blue stick is longer than the red stick because the blue stick is longer than the green stick.
FALSE MEMORY	The test pair is confused with a premise pair.	I've just seen that the blue stick is longer than the red stick, so that will still be the case.
EXTERNAL & VISUAL	Visual or external information is used to explain the transitive relationship, no premise information is used.	The parrot is older than the duck because parrots can become very old; When I look very well, I can see that the blue stick is longer than the red stick.
NO EXPLANATION	No explanation is given.	I guessed, I just don't know.

4.3 gives an overview of the number of children per grade, and the mean age and the standard deviation of age within each grade.

4.3.4 Data

A representation of the input data file for the latent class regression analysis is shown in Table 4.4. Each of the 615 children performed 16 tasks (in the table indicated as replications). Each task was defined by a combination of three task characteristics. For example, Task 1 had format $Y_A > Y_B > Y_C$, *simultaneous* presentation form, and *verbal* type of content. Each child used one of the seven strategies, and the same child could use different strategies for different tasks.

Table 4.3: *Number of Children, Mean Age (M) and Standard Deviation (SD) per Grade*

Grade	*n*	*Age*	
		M^a	*SD*
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

[a] number of months

Table 4.4: *Input Data File for the Latent Class Regression Analysis; 16 Lines per Case, Each Line Representing a Transitive Reasoning Task*

Replication.	Case Id	Grade	Format*	Presentation*	Content*	Strategy
1	1	2	1	1	1	3
2	1	2	2	1	1	2
3	1	2	3	1	1	6
.	1	2	.	.	.	.
.	1	2	.	.	.	.
15	1	2	3	2	2	5
16	1	2	4	2	2	3
1	2	3	1	1	1	5
2	2	3	2	1	1	1
.	.	.	.	.	.	.
16	2	3	4	2	2	6
.	.	.	.	.	.	.
.	.	.	.	.	.	.
1	615	6	1	1	1	4
.	615	.	.	.	.	.
.	615	.	.	.	.	.
16	615	6	4	2	2	3

* Format, Presentation, and Content were the three task characteristics; for a detailed description see Bouwmeester and Sijtsma (2004), chapter 2

4.4 Analysis

4.4.1 Parts of the Model

It was expected that the strategy responses of the children on the 16 transitive reasoning tasks could be divided into a number of classes that were ordered along a developmental scale and differed with respect to specific strategy use for different kinds of transitive reasoning tasks. The formation of the latent classes was expected to be influenced by age.

The first part of the latent class regression model is defined by the probability (π) of being in a particular latent class (realization x of latent variable X), given grade level (realization z^c, of covariate Z^c (where ^{c} stands for *covariate*), that is,

$$\pi(x|z^c). \tag{4.1}$$

These marginal probabilities of being in a specific class given a value on the covariate, add to 1 over the latent classes x:

$$\sum_x \pi(x|z^c) = 1. \tag{4.2}$$

In the second part of the model, the probabilities are estimated of using a particular cognitive behavior given the latent class and the value(s) on one or more external conditions. In this application the dependent variable "cognitive behavior" is the discrete variable "strategy" (Y, with realizations y) that has seven categories. The predictor variables "external conditions" are three "task characteristics" ,Z_1^p, Z_2^p, Z_3^p, with realizations z_1^p, z_2^p, z_3^p (where ^{p} stands for *predictor*) having also a discrete measurement level:

$$f(y|x, z_1^p, z_2^p, z_3^p). \tag{4.3}$$

For each task (which consists of a combination of the three task characteristics) a multinomial probability function is estimated for the use of a strategy in a latent class, and this is done for each combination of a strategy and a latent class. In a fixed latent class, these probabilities add to 1 over strategies (y), that is,

$$\sum_y f(y|x, z_l^p, z_2^p, z_3^p) = 1. \tag{4.4}$$

Because there are 16 tasks, there are 16 of these probability functions for each latent class.

Then, Equations 4.1 and 4.3 combine into the latent class regression model. The model is defined by the product of a summation over latent classes of the marginal probability of being in a latent class, given the grade level and the product of multinomial probabilities for each task (denoted t, 16 combinations of task characteristics):

$$f(\mathbf{y}|z^c, \mathbf{z}_1^p, \mathbf{z}_2^p, \mathbf{z}_3^p) = \sum_x \pi(x|z^c) \prod_t f(y_t|x, z_{1t}^p, z_{2t}^p, z_{3t}^p). \tag{4.5}$$

Because there are 16 observations per case, the dependent variable $\mathbf{Y}$ is a vector containing the 16 strategy responses and the predictor variables $\mathbf{Z}_l^p$ are also vectors containing the levels of the task characteristics.

4.4.2 Parameters

To calculate the multinomial probabilities of being in a latent class given grade level ($\pi(x|z^c)$ in Equation 4.5), two kinds of parameters have to be estimated, denoted by γ_x^0 and $\gamma_{z^c x}^1$. Parameters γ_x^0 are the intercepts for the latent class variable and parameters $\gamma_{z^c x}^1$ are the covariate effects on the latent class variable. The first part of Equation 4.5 is modeled by a multinomial probability, which is defined as a logistic regression function:

$$\pi(x|z^c) = \frac{exp(\eta_{x|z^c})}{\sum_x exp(\eta_{x|z^c})}. \tag{4.6}$$

The linear term $\eta_{x|z^c}$ equals

$$\eta_{x|z^c} = \gamma_x^0 + \gamma_{z^c x}^1. \tag{4.7}$$

To estimate the multinomial probability function of using a particular strategy given the latent class and a combination of task characteristics (i.e., $f(y_t|x, z_{1t}^p, z_{2t}^p, z_{3t}^p)$, in Equation 4.5), again two kinds of parameters have to be estimated, denoted by β_{xy}^1 and $\beta_{xz_{lt}^p}^2$. Parameters β_{xy}^1 are the class-specific intercepts. For all strategies in every latent class there is a β_{xy}^1 parameter. Parameters $\beta_{xz_{lt}^p}^2$ are the class-specific regression coefficients. For all levels of the task characteristics there is a parameter for

Table 4.5: *Number of Parameters to be Estimated*

Classes	γ^0_x	$\gamma^1_{z^c x}$	β^1_{xy}	$\beta^2_{xz^p_{lt}}$	Total
1	1-1=0	1-1=0	$(7-1) \times 1 = 6$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 1 = 30$	36
2	2-1=1	2-1=1	$(7-1) \times 2 = 12$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 2 = 60$	74
3	3-1=2	3-1=2	$(7-1) \times 3 = 18$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 3 = 90$	112
4	4-1=3	4-1=3	$(7-1) \times 4 = 24$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 4 = 120$	150
5	5-1=4	5-1=4	$(7-1) \times 5 = 30$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 5 = 150$	188
6	6-1=5	6-1=5	$(7-1) \times 6 = 36$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 6 = 180$	226
7	7-1=6	7-1=6	$(7-1) \times 7 = 42$	$(7-1) \times [(4-1)+(2-1)+(2-1)] \times 7 = 210$	264

all strategies in every latent class. The multinomial probability function is again a logistic regression function:

$$f(y_t|x, z^p_{1t}, z^p_{2t}, z^p_{3t}) = \frac{exp(\eta_{y|x,z^p_{1t},z^p_{2t},z^p_{3t}})}{\sum_y exp(\eta_{y|x,z^p_{1t},z^p_{2t},z^p_{3t}})}. \tag{4.8}$$

The linear term $\eta_{y|x,z^p_{1t},z^p_{2t},z^p_{3t}}$ equals

$$\eta_{y|x,z^p_{1t},z^p_{2t},z^p_{3t}} = \beta^1_{xy} + \beta^2_{z^p_{1t}xy} + \beta^2_{z^p_{2t}xy} + \beta^2_{z^p_{3t}xy}. \tag{4.9}$$

The number of parameters to be estimated increases rapidly with an increasing number of latent classes. Table 4.5 shows the number of parameters to be estimated for models with one through seven latent classes.

4.4.3 Fit of the model

The program Latent Gold (Vermunt & Magidson, 2003) was used to estimate the parameters and calculate the fit of the model. The program gives evaluation statistics, estimates of the parameters and the accompanying standard errors and z-values.

In the program Latent Gold a number of evaluation statistics are provided to choose a plausible model. First, the log-likelihood statistics are calculated which express the fit for models with a user-specified number of latent classes. The amount of reduction of the log-likelihood statistic for models with an increasing number of classes can be considered to choose the best fitting model. Because of sparse frequency tables, the asymptotic p-values associated with the χ^2 statistics often can not be trusted. Therefore, a p-value can be estimated by means of bootstrapping (Efron

& Tibshirani, 1993) which is implemented in the program. The bootstrap L^2 procedure involves generating a certain number of replication samples from the maximum likelihood solution and re-estimating the model with each replication sample. L^2 is a test statistic or fit measure. The bootstrap p-value is the proportion of replication samples with higher L^2 than in the original sample. For example, when 40% of the replication samples has a L^2 value higher than the L^2 value of the original sample, the bootstrap p-value is 0.40. However, a conditional bootstrap procedure, in which the fit of models with different classes can be compared has not yet been implemented in the Latent Gold program.

Secondly, the BIC values are calculated. The lower the BIC value the more parsimonious the model (McLachlan & Peel, 2000). Thirdly, the proportions of classification errors are provided. This proportion indicates how well the model can predict latent class membership given the value on the covariate and the dependent variable (Andrews & Currim, 2003). This proportion is not a fit measure, but it is an important measure to evaluate the distinctiveness of different classes.

Fourth, the class sizes and interpretation of the classes were used to choose a model. Although the evaluation statistics calculated by the program provided useful guidelines to choose the best-fitting model, the final decision was based on the interpretableness of the classes and the class-size.

4.5 Results

Analysis of variance with number-correct score on all 16 tasks as dependent variable and school and grade as independent variables showed no significant effects for the same grades of different schools. Therefore, it was concluded tentatively that school had no influence on a child's transitive reasoning ability.

4.5.1 Model Fit and Number of Classes

Seven models were fitted with an increasing number of classes ranging from one to seven. Table 4.6 shows the evaluation statistics which were used to choose a final model.

Although a number of fit-statistics which evaluate different aspects of the model can be used to choose a plausible model, the choice of a final model also depends on substantive considerations, previous research results, considerations of parsimony, and so on. This can be compared with factor analysis, where the choice of the final factor solution also depends on considerations other than statistical ones. It remains difficult, maybe practically impossible to determine the exact number of latent classes.

The log-likelihood statistics showed a reduction of at least 37% in magnitude from the log-likelihood statistics from the one-class model to the five-class model. The reduction of the log-likelihood statistic from the five-class model to the six-class model was only 12%. On the basis of the log-likelihood statistics the five-class model would be chosen.

The six-class model was the most parsimonious model in terms of BIC values. The proportion of classification errors first increases from the one-class model through the four-class model. This can be explained by the fact that correct classifying is more difficult with a higher number of classes. The result that the proportion of classification errors decreases with the five-class model and then increases again indicates that the five-class model may be preferred above the six-class model.

On the basis of the evaluation statistics provided by the program, the five- and six-class models are most plausible. For this application, the bootstrap procedure was not informative to choose the best fitting model. On the basis of the class sizes and the interpretation of the classes, the five-class model was chosen as the final model. The six-class model had three relative small classes (marginal probability < 0.10). Moreover, the smallest class did hardly differ from another class with respect to the interpretation.

4.5.2 The γ Parameters: Class Size and Influence of Grade

Table 4.7 shows the γ parameters. The γ^0 parameters are intercept parameters which were used to calculate class size. The $\gamma^1_{xz^c}$ parameters were all significant ($z > 1.96$). This means that the covariate grade had a significant influence in all classes. When these γ parameters are inserted in

Table 4.6: *Model Fit Statistics for Six Latent Class Models*

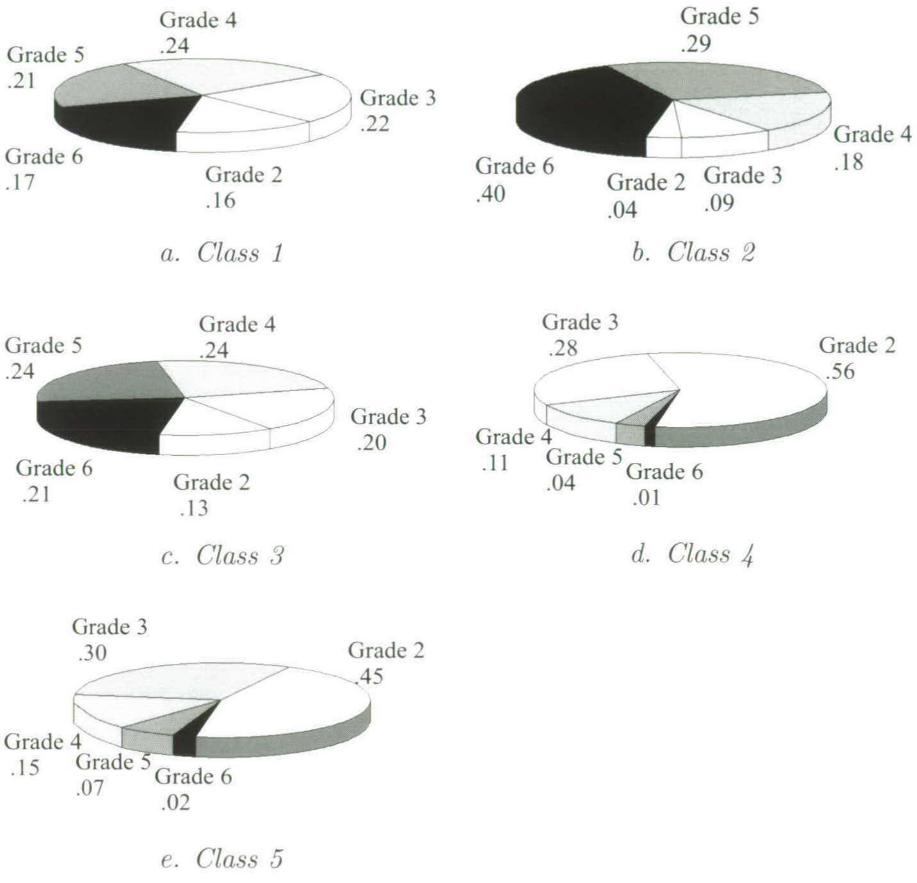
Number of Classes	L^2 Value	BIC Value*	Number of Parameters	Proportion of Classification Errors
1	25338.59	31320.69	36	.00
2	22867.01	29093.12	74	.03
3	21824.83	28294.97	112	.05
4	21176.73	27890.43	150	.05
5	20777.86	27736.04	188	.05
6	20430.21	27632.41	226	.06
7	20204.84	27651.06	264	.06

*: BIC-value = -2log-likelihood + # parameters * ln(N)

Equations 4.7 and 4.6, respectively, the marginal probabilities (class size, see Table 4.7) and the probability distribution of grade given the latent class can be calculated. Figure 4.2 shows the probabilities of grade for each class. In particular in class two Grade six had high probability. Also in classes one and three, higher grades had higher probability than lower grades. For the classes four and five, lower Grades two and three had higher probability than higher Grades four, five and six.

Table 4.7: γ*-Parameter Estimates and Class Size for the Five-Class Model*

Class	γ_x^0	$\gamma_{xz^c}^1$	Class Size
1	-.251	.220	.381
2	-4.386	.796	.266
3	-1.866	.325	.146
4	3.957	-.812	.106
5	2.545	-.530	.101



a. Class 1

b. Class 2

c. Class 3

d. Class 4

e. Class 5

Figure 4.2: *Probability Distribution of Grade Given Latent Class*

4.5.3 The β Parameters: Strategy Use and Influence of Task Characteristics

Table 4.8 shows the class-specific intercepts, the β^1_{xy}-parameters. A non-significant β^1_{xy} parameter estimate does not significantly deviate from zero, which means that there is no effect for this strategy in the particular class. The β^1_{xy}-parameters can be used to calculate the probability distribution of using a strategy given the latent class $\pi(y|x)$ ($\pi(y|x) = \frac{exp(\beta^1_{xy})}{\sum_y exp(\beta^1_{xy})}$). Figure 4.3 shows the probabilities of using a particular strategy for each class.

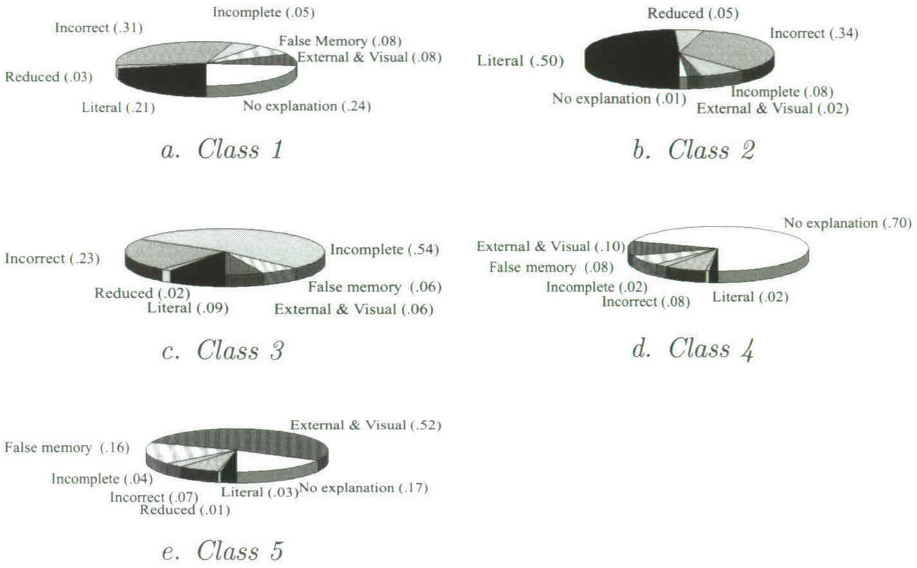
Table 4.8: *The β^1_{xy}-Parameter Estimates for the Five-Class Model*

Strategy	Class 1	Class 2	Class 3	Class 4	Class 5
LITERAL	.705	2.260	*.788*	-1.137	-.910
REDUCED	*-1.421*	*-.026*	*-.871*	*-2.320*	*-2.351*
INCORRECT	1.098	1.900	1.707	*.433*	*.038*
INCOMPLETE	-.704	*.392*	2.559	*-.770*	*-.651*
FALSE MEMORY	*-.273*	-2.367	*.409*	*.440*	.887
EXTERNAL & VISUAL	*-.236*	-.926	*.410*	*.726*	2.040
NO EXPLANATION	.831	-1.232	*-5.001*	2.629	.946

Italics: effect is not significant($p > .05$)

Children in class one in particular use INCORRECT premise information, LITERAL premise information and NO EXPLANATION. Children in class two are characterized by the use of LITERAL premise information and INCORRECT premise information. Children in class three are characterized by the use of INCOMPLETE premise information and INCORRECT premise information. Children in class four in particular do NOT give an EXPLANATION or use EXTERNAL & VISUAL information. Children in class five are characterized by the use of EXTERNAL & VISUAL information, FALSE MEMORY and NO EXPLANATION.

There are 280 class-specific regression coefficient parameters ($\beta^2_{xz^p_{lt}}$), that is, one for each strategy (7) in each class (5), for every level of the task



a. Class 1

b. Class 2

c. Class 3

d. Class 4

e. Class 5

Figure 4.3: *Probability Distribution of Strategy by Latent Class*

characteristics (4+2+2). It is beyond the scope of this chapter to interpret these parameters in detail, but we will give a global interpretation of the influence of the task characteristics on the strategy use in the latent classes by describing the size of the effect of the parameters. Table 4.9 gives the interpretation of the strength of the influences. TASK FORMAT has some influence on strategy use in the Classes one and two but hardly in the classes three, four and five. PRESENTATION FORM has a strong effect on strategy use in all classes except class three. CONTENT of the relationship has a strong effect on strategy use in the classes one and two, some effect in the classes three and four and hardly any effect in class five.

4.6 Discussion

When studying cognitive development of transitive reasoning using a latent class regression model we found that a number of classes can be distinguished which differ with respect to cognitive behavior. Using the grade level distribution over the classes, an ordering of the classes became visible.

Table 4.9: *Size of the Effect of Influence of the Task Characteristics on Strategy Use*

Characteristic	Class 1	Class 2	Class 3	Class 4	Class 5
TASK FORMAT	some	some	hardly	hardly	hardly
PRESENTATION FORM	strong	strong	hardly	strong	strong
CONTENT	strong	strong	some	some	hardly

Classes Four and Five which contained in particular lower-grade children were characterized by superficial cognitive behavior using almost no task information but rather directly observable task characteristics or unimportant information from the external world. Several times, children gave no explanation at all. In the classes one and three, in which children from all grades were represented but in particular from Grade three, four, and five, children often knew that they had to use the task information but they did not have a complete or correct representation of the task space. Class two contained in particular higher-grade children which were able to use the task information, understood the underlying pattern and were able to form a complete internal representation of the task space in most cases.

By treating age as a covariate which influenced the formation of the classes, the developmental ordering of the classes was not assumed to be known a priori. The results showed that there is a developmental ordering, but that children from the highest grade are (marginally) represented in lower-ability classes, while children from lower grades are represented in the higher-ability class. The model thus gives opportunities to diagnose children which deviate from the age-related criterion and to interpret the deviation in detail.

An interesting finding of this application, which is difficult to reveal when no latent classes are distinguished, is the differential influence of task characteristics on strategy use in a latent class. In addition to a general overview of the strategy use in a particular latent class, the latent class regression model makes it possible to explain or predict the influence of external conditions at a detailed level, or even the influence of interactions of external conditions (which was not done in this application).

In interesting product of this method is that specific cognitive behavior can be better interpreted in relation to other cognitive behavior. For example, it is difficult to interpret the NO EXPLANATION strategy when there is no further information. When children do not give an explanation, they may simply not know how to solve the problem; they may know that the premises information has to be used, but they do not know how; or they may simply not know how to explain their answer to other people. The distribution of the strategies over the classes gives information on how to interpret this NO EXPLANATION strategy. In class one and class four children often used NO EXPLANATION, but children in class one used some more-proficient strategies besides the NO EXPLANATION strategy, while children in class four in particular used low-proficiency strategies. It appears that children in class four had absolutely no idea how to solve the tasks, while children in class one understood that they had to use the premises but did not know how to use them.

The analysis of this application was explorative. We did not assume a particular cognitive theory which was tested. However, it is also possible to test a cognitive theory in terms of the latent class regression model, that is, to perform a confirmative analysis. Assuming Piaget's theory, we could have tested whether empirical data fitted in the cognitive developmental stages Piaget assumed. Then, it should be tested whether the data could be explained by a number of latent classes which represented the cognitive developmental stages. Using the latent class regression model, it is also possible to model a priori hypotheses about developmental stages. It may be expected from a developmental theory that a particular condition has no effect in one particular latent class, or an equal effect in two or more different classes. By imposing restrictions on the model, effects can be set to zero, or can be set equal for different classes.

It has to be emphasized that we used data from a cross-sectional design, that is, children of different ages were tested once. This design makes it possible to interpret development in terms of differential classes, but we can only speculate about an individual child's transition from one class to another. A longitudinal study is necessary to study this transition. The latent class regression model can also be used to study such a longitudinal design.

In this chapter we introduced the latent class regression model for studying cognitive developmental phenomena. The most important value of the model is the possibility to empirically test the presence or absence of latent classes without the need of strong cognitive theoretical assumptions about the latent variable(s). In the application of the model to transitive reasoning data, a very large number of parameters had to be estimated, making the model relative complex. The large number of parameters was caused by a nominal dependent variable, having seven categories, and nominal independent variables. Models with other types of dependent and independent variables will contain substantially fewer parameters.

The flexibility of the model in terms of mixed measurement levels and treatment of different cognitive variables further offers a broad application to a number of cognitive developmental phenomena, such as conservation, symbolic analogies, verbal analogies, inductive reasoning, reading comprehension, or problem solving.

Chapter 5

Development and Individual Differences in Transitive Reasoning: a Fuzzy Trace Theory Approach

Abstract

Individual differences in transitive reasoning were investigated in 4 to 13 year-old children. The performance on three kinds of tasks which mainly differed with respect to their presentation ordering and position ordering was studied in an effort to determine the use of fuzzy trace theory (Brainerd & Kingma, 1984) as a framework for explaining the development of transitive reasoning. The results from a sample of 409 children ranging in age from 64 to 159 months showed that the two-dimensional classification of performance patterns agreed with the expected distinction of performance groups according to fuzzy trace theory. Task format had a stronger effect on performance on transitivity test-pairs than on memory test-pairs. Furthermore, the developmental effects showed more improvement in fuzzy ability than in verbatim ability.

This chapter has been submitted for publication.

5.1 Introduction

A transitive reasoning task requires the inference of an unknown relationship between two objects from the known relationships between each of these objects and a third object. For example, let three sticks, A, B, and C, differ in length, denoted Y, such that $Y_A > Y_B > Y_C$, then given $Y_A > Y_B$ and $Y_B > Y_C$ the relationship between A and C can be inferred from these two relationships. Together, the given relationships are the premises.

A transitive reasoning task consists of a presentation stage and a test stage. At the presentation stage, the premise pairs, A and B and B and C, are presented, and the child is given the opportunity to memorize the premises, $Y_A > Y_B$ and $Y_B > Y_C$.

During the test stage, the child has to reproduce the premises. At this stage, because the premises are reproduced, the premise pairs are called memory test-pairs. The object pair of which the relationship has to be inferred from the memory test-pairs, here (A,C), is called the transitivity test-pair, because it tests the ability to infer a transitive relationship from the available premise information.

In Piaget's theory, transitive-reasoning tasks are used to study the understanding of operational reasoning (Piaget, 1942; Piaget et al., 1948). Classical Piagetian theory assumes that children are capable of transitive reasoning at the concrete operational stage (from approximately seven through 13 years) in which they have acquired the ability to infer *logically* an unknown relationship from two or more premises. When children do not use the premise information for their explanation, they are assumed to reason functionally. Functional reasoning is characteristic of the preoperational stage (from approximately two through seven years).

Bryant and Trabasso (1971; see also Breslow, 1981; Riley & Trabasso, 1974; Thayer & Collyer, 1978; Trabasso, 1977; Trabasso et al., 1975) hypothesized that not the understanding of logical rules, but memory of the premises is crucial in transitive reasoning. They trained four and five year-old children and showed that they were able of transitive reasoning and that their transitivity test-pairs performance could be explained completely by their performance on the memory test-pairs. Therefore, these researchers concluded that understanding of logical rules was not required and memory of the premises was sufficient to infer the unknown relationship.

A few years later Brainerd and Kingma (1984) showed that nor an understanding of logical rules nor memory of the premises was necessary to infer transitive relationships. They used fuzzy trace theory to explain their results (Brainerd & Reyna, 1992, 1990, 1995, 2001, 2004; Reyna, 1992, 1996; see also Chapman & Lindenberger, 1992).

Fuzzy trace theory assumes that incoming information is processed simultaneously in different traces. These traces contain different features of the incoming information. The traces stem from an underlying continuum. On the one hand there is a fuzzy continuum containing vague, pattern-like information in a degenerated form at different fuzzy trace-levels, only holding the gist of information. The fuzzier the trace, the more reduced and vague the information. On the other hand there is a verbatim continuum containing literal and detailed information in different traces about, for example, color, shape or size. The more verbatim the trace, the more details it contains about the information.

According to fuzzy trace theory, information is encoded and processed in a number of traces, simultaneously and automatically. The kind of task determines which trace is the most appropriate to retrieve. For example, when a cognitive task requires memory of color, shape, or size of previously presented information, a verbatim trace is required which contains this detailed information. However, when the task requires inferences between objects, a fuzzy trace is required which contains pattern-like information. Because the structure of verbatim traces is more complex than the pattern-like structure of fuzzy traces, the verbatim traces are only available for a limited amount of time, while the fuzzy traces can be retrieved much longer (Reyna & Brainerd, 1990; Brainerd & Kingma, 1984; Brainerd & Reyna, 2004).

Brainerd and Kingma (1984) showed that the application of fuzzy trace theory to transitive reasoning needs the unitary trace model. In the unitary trace model both the memory test-pairs and transitivity test-pairs are inferred from the same fuzzy trace. They showed that this unitary trace model could well explain children's performance. However, the performance on memory and transitivity test-pairs in tasks with different kinds of manipulations was far from perfect. Apparently, several children were not able to retrieve the appropriate fuzzy trace but retrieved other, less efficient traces to solve the transitive relationship. The average scores used in Brainerd and Kingma's (1984) study did not take into account individual differences and did not allow to distinguish different strategy groups. However, Brainerd and Kingma explained that performance may be influenced by temporal and spatial position effects. With respect to verbatim traces, a *temporal position effect* may occur when memory is overloaded and a child is not able to retrieve the verbatim trace containing the complete premise information. Then performance on the memory test-pairs presented first or last (or both) is better than on the midterms. With respect to fuzzy traces, a *spatial position effect* may occur when the most appropriate fuzzy trace for the cognitive task is not used. For example, in a 5-object transitive-reasoning task, such as $Y_A < Y_B < Y_C < Y_D < Y_E$, large objects are on the right and small objects are on the left leaving the midterms undefined. Performance on one or both of the end-anchors is better than on the midterms. In their study, Brainerd and Kingma (1984) concentrated in particular on the influence of task manipulations on performance. In the present study, both individual differences of children and the influence of task manipulations on performance were taken into account.

In a study on the strategies children used for solving transitivity test-pairs, Bouwmeester et al. (2004) found that when children are not able to use a fuzzy trace to infer transitive relationships, they use a verbatim trace which mostly does not lead to a correct answer. Bouwmeester et al. (2004) also found that for tasks in which few cues were given about the ordering, either due to format ($Y_A = Y_B > Y_C = Y_D$) or presentation form, literal premise information was used to solve the transitive relationship. However, the information was often incorrectly remembered, which rendered those tasks difficult. This literal premise information can be assumed to stem from a verbatim trace. Children who scored high on the ability scale mostly used ordering information to solve the tasks, which stemmed from a fuzzy trace.

5.1.1 Aim of This Study

Based on the results from previous research and the theoretical framework of fuzzy trace theory, we expect that (1) both verbatim and fuzzy traces are involved in de development of transitive reasoning, and (2) the development of transitive reasoning can be described by a changing interaction of these two kinds of traces in time when a shift takes place from verbatim to fuzzy thinking (Brainerd & Reyna, 1995). Based on these expectations, the aim of this study is to reveal the development of both verbatim and fuzzy traces in the context of transitive reasoning, by distinguishing groups of children that differ in their use of verbatim and fuzzy traces when responding to memory test-pairs and transitivity test-pairs in different kinds of transitive-reasoning tasks. The relationship of age and the strategy groups is expected to reveal whether the development of transitive reasoning is characterized by a shift from verbatim to fuzzy thinking.

Verbatim Ability

The trace-levels that are retrieved by the child depend on his/her verbatim and fuzzy ability levels. Figure 5.1 shows the relationships between verbatim ability, verbatim traces and performance on memory test-pairs. When applied to transitive reasoning, a verbatim ability level induces verbatim traces according to a particular probability structure (in the context of transitive reasoning three verbatim traces were hypothesized). The probability distribution is defined as $P(\text{trace}|\text{ability})$. This is the probability of using a particular trace given ability level. Note that both the ability and the trace variables are unobservable, that is, they are latent variables. It is hypothesized that $P(\text{guessing}|\text{ability})$ decreases as a function of ability, and is maximal when ability level is low; $P(\text{temporal position}|\text{ability})$ first increases and then decreases as a function of the ability and is maximal when the ability level is intermediate; and $P(\text{complete memory}|\text{ability})$ increases as a function of ability and is maximal when the ability level is high.

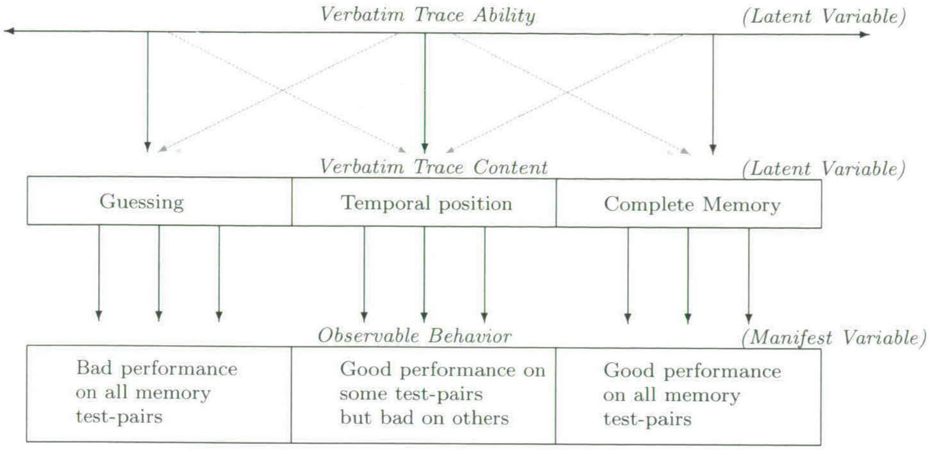


Figure 5.1: *Relationship Between Continuous Latent Ability and Discrete Manifest Behavior Based on a Verbatim Trace*

In Figure 5.1 the black arrows between the latent variable levels indicate high probability, the grey arrows indicate lower probability and the light grey arrows indicate low probability.

A verbatim trace corresponds to a particular probability to answer a memory test-pair correctly. The success probability of a correct answer to a test-pair given the verbatim trace level is denoted as $P(\text{test-pair}|\text{trace level})$. The response variable is an observed variable (i.e., a manifest variable). When trace level is low, a child has a probability approximately at chance level to answer a test-pair correctly. When trace level is high, a child has a probability close to 1 to answer a test-pair correctly.

The retrieval of verbatim traces enables characteristic performance on the memory test-pairs. Guessing is likely to lead to bad performance on memory test-pairs. A temporal position effect is likely to produce good performance on memory test-pairs presented first or last in a sequence, and worse performance on the memory test-pairs presented in between. For example, in a transitive-reasoning task (e.g., $Y_A > Y_B > Y_C > Y_D > Y_E$) in which the premises are presented in an ordered form (first $Y_A > Y_B$, second

$Y_B > Y_C$, third $Y_C > Y_D$, and finally, $Y_D > Y_E$), children remember the premise relationships $Y_A > Y_B$ and $Y_D > Y_E$ better than $Y_B > Y_C$ and $Y_C > Y_D$. When children retrieve a verbatim trace with the complete premise information, performance is likely to be good on all memory test-pairs. Note that verbatim ability level (and accompanying trace levels) is assumed to have no effect on the performance on transitivity test-pairs.

Fuzzy Ability

Figure 5.2 shows the relationships between fuzzy ability, fuzzy traces and the performance on memory and transitivity test-pairs. When applied to transitive reasoning, a fuzzy ability level induces fuzzy traces according to a particular probability structure (in the context of transitive reasoning three fuzzy traces were hypothesized). It is hypothesized that $P(\text{guessing}|\text{ability})$ decreases as a function of ability, and is maximal when ability level is low; $P(\text{spatial position}|\text{ability})$ first increases and then decreases as a function of ability and is maximal when the ability level is intermediate; and $P(\text{complete ordering}|\text{ability})$ increases as a function of ability and is maximal when ability level is high. In Figure 5.2 the black arrows indicate high probability, the grey arrows indicate lower probability and the light grey arrows indicate low probability.

A fuzzy trace corresponds with a particular probability to answer a test-pair correctly. For example, when pattern information is not available, children guess on all test-pairs which is likely to result in bad performance[1]. When children retrieve a spatial position fuzzy trace, they have higher success probabilities on the end anchors of the spatial representation than on the midterms. For example, in a transitive-reasoning task (e.g., $Y_A > Y_B > Y_C > Y_D > Y_E$), in which the objects are positioned from large to small but the premises are not presented in an ordered way (e.g., first $Y_B > Y_C$, second $Y_D > Y_E$, third $Y_A > Y_B$, and finally, $Y_C > Y_D$) children use the fuzzy trace "small objects are on the right and large objects are on

[1] To keep the explanation comprehensive here, we assumed that children will not resort to verbatim information. In the complete model, this problem is resolved by taking both verbatim ability and fuzzy ability into account.

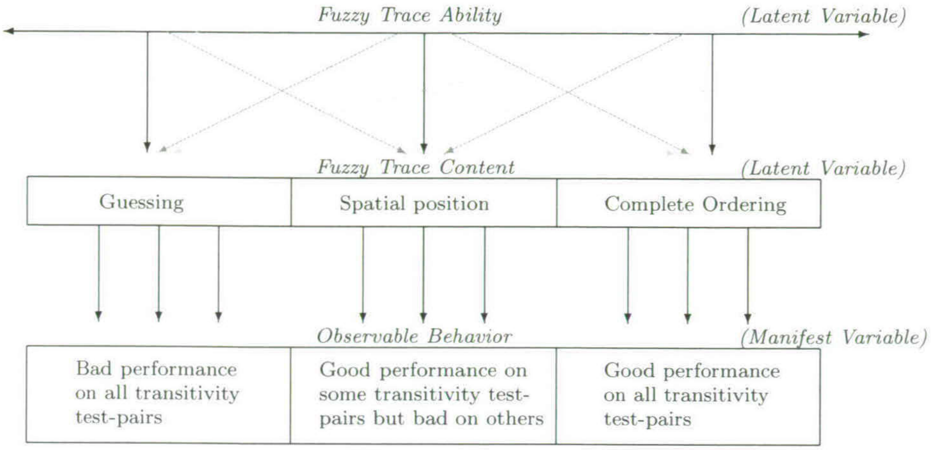

Fuzzy Trace Ability
(Latent Variable)
Fuzzy Trace Content
(Latent Variable)
Guessing
Spatial position
Complete Ordering
Observable Behavior
(Manifest Variable)
Bad performance on all transitivity test-pairs
Good performance on some transitivity test-pairs but bad on others
Good performance on all transitivity test-pairs

Figure 5.2: *Relationship Between Continuous Latent Ability and Discrete Manifest Behavior Based on a Fuzzy Trace*

the left". They are likely to perform better on the end anchor test-pairs of the spatial representation ($Y_A > Y_B$, $Y_D > Y_E$, $Y_A > Y_C$, $Y_C > Y_E$) than on the midterm test-pairs ($Y_B > Y_C$, $Y_C > Y_D$, $Y_B > Y_D$). The effect may also occur at only one end of the ordering; for example, the fuzzy trace is "left-side objects are large". According to the unitary trace model (Brainerd & Kingma, 1984), this spatial position effect occurs both on transitivity test-pairs and memory test-pairs. When a child uses the appropriate fuzzy trace, performance is expected to be good on all test-pairs.

Development of Verbatim and Fuzzy Abilities

According to Brainerd and Kingma (1984, 1985), Reyna and Brainerd (1990), and Reyna (1992) the development of verbatim ability is rather fast and reaches full development at approximately five years of age. Fuzzy ability develops slower and is not expected to reach full development during childhood (Reyna & Brainerd, 1990; see also Liben & Posnansky, 1977; Marx, 1985b, 1985a; Perner & Mansbridge, 1983; Stevenson, 1972). Figure 5.3 gives a schematic impression of the development of fuzzy and verbatim ability.

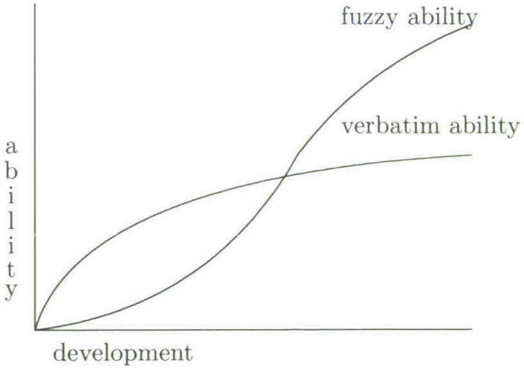


Figure 5.3: *Schematic Display of The Development of Fuzzy and Verbatim Trace Abilities*

Both abilities are expected to play an important role in the performance on the memory and transitivity test-pairs in a transitive-reasoning task. When crossed completely the three verbatim and fuzzy trace levels lead to nine theoretical combinations, each of which is characterized by its own expected performance on the test-pairs. The characteristics of the task are expected to influence the retrieval of verbatim and fuzzy traces.

5.1.2 Transitive-Reasoning Tasks and Task Manipulations

Transitive-reasoning tasks differ with respect to the cues they provide about the ordering of the objects. For example, when the objects are *positioned* in a linear order and also *presented* in a linear order, the cues about the ordering of the objects are obvious. As a consequence, the required fuzzy ability level is not as high as when, for example, the objects are not positioned in a linear order or presented in a linear order. In this study, we used three kinds of tasks, that may be characterized as (1) ordered position, ordered presentation ($O_{pos}O_{pres}$); (2) ordered position, disordered presentation ($O_{pos}D_{pres}$); and (3) disordered position, ordered presentation ($D_{pos}O_{pres}$). The combination of "disordered position, disordered presentation" was not used because it was expected to be too difficult even for adults (see Brainerd & Reyna, 1992). In this study, in every task-type four premise pairs were presented. Next, the child was confronted with the test-pairs. Each task had four memory test-pairs and three transitivity test-pairs.

Ordered Position, Ordered Presentation Tasks ($O_{pos}O_{pres}$)

In $O_{pos}O_{pres}$ tasks, the objects are ordered from small to large or large to small. The presentation of the premises is also ordered. Thus, first premise pair (A, B) is presented, followed consecutively by premise pairs (B, C), (C, D) and (D, E). Ordered presentation of the ordered objects makes the use of pattern information from fuzzy traces rather easy. Figure 5.4 shows an example of the four premises of an $O_{pos}O_{pres}$ task. Box 1 presents the first premise pair, box 2 presents the second premise pair, and so on. The "test-pair" box shows an example of the first memory test-pair.

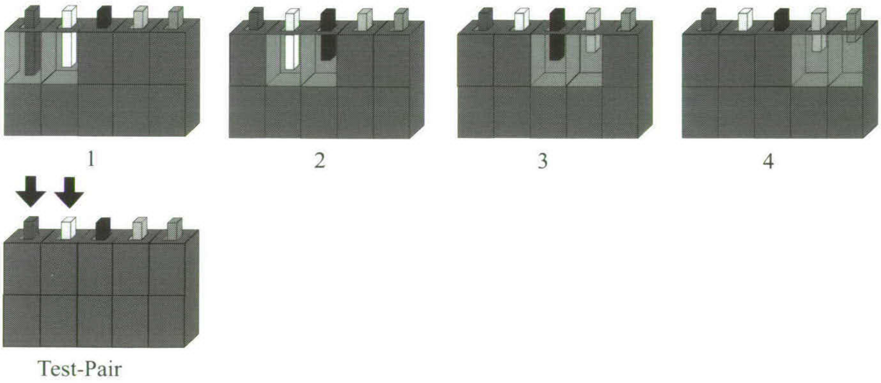


Figure 5.4: *Example of the Premise Presentation of an "Ordered Position, Ordered Presentation" Task*

The expected performance patterns for combinations of verbatim and fuzzy trace levels on the memory and transitivity test-pairs of $O_{pos}O_{pres}$ tasks are shown in Table 5.1. When fuzzy trace level is *intermediate* or *high*, the expected performance on all test-pairs is good, because pattern information can easily be used to infer the relationships in both the memory

Table 5.1: *Expected Performance on the Test-Pairs of $O_{pos}O_{pres}$ Tasks for Nine Combinations of Trace Levels*

Verbatim	Fuzzy	Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3
low	*low*	○	○	○	○	○	○	○
	intermediate	●	●	●	●	●	●	●
	high	●	●	●	●	●	●	●
intermediate	*low*	★	○	○	★	○	○	○
	intermediate	●	●	●	●	●	●	●
	high	●	●	●	●	●	●	●
high	*low*	●	●	●	●	○	○	○
	intermediate	●	●	●	●	●	●	●
	high	●	●	●	●	●	●	●

○: bad performance; ★: moderate performance; ●: good performance

test-pairs and the transitivity test-pairs. When fuzzy trace level is *low*, the combination with (1) *low* verbatim trace level is expected to lead to guessing, yielding success probabilities at approximately chance level on all test-pairs; (2) *intermediate* verbatim trace level is expected to lead to temporal position effects, yielding moderate performance on the first and last memory test-pairs (M_1 and M_4) and bad performance on all other test-pairs; and (3) *high* verbatim trace level is expected to lead to complete memory of the memory test-pairs, yielding high success probabilities on the memory test-pairs and low success probabilities on the transitivity test-pairs.

Ordered Position, Disordered Presentation Tasks ($O_{pos}D_{pres}$)

In $O_{pos}D_{pres}$ tasks, the objects are ordered from small to large or large to small. The presentation of the premises is disordered; for example, in Figure 5.5 first (C, D) is presented, followed consecutively by (A, B), (D, E), and (B, C). The midterm relationships are always presented first and last, and the end anchors are always presented in between so as to be able to distinguish a temporal position effect from a spatial position effect (see also Brainerd & Kingma, 1984). Disordered presentation makes the use of fuzzy traces more difficult than ordered presentation because it is more difficult to recognize the ordering of the objects. The "test-pair" box in Figure 5.5 shows the first transitivity test-pair. The expected performance patterns

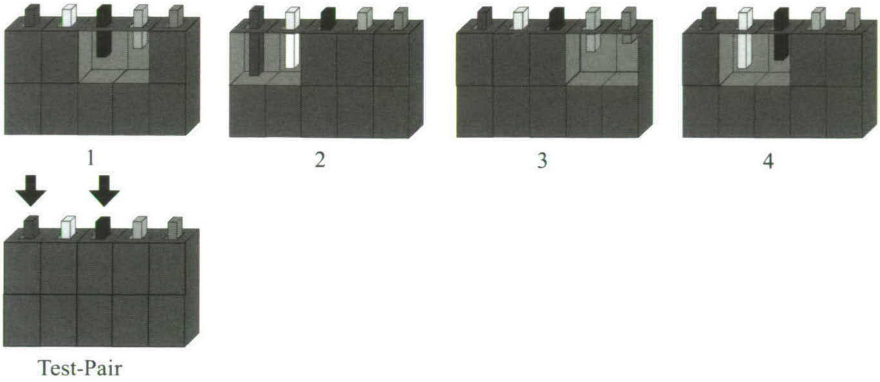


Figure 5.5: *Example of the Premise Presentation of an "Ordered Position, Disordered Presentation" Task*

for combinations of verbatim and fuzzy trace levels on the memory and transitivity test-pairs of $O_{pos}D_{pres}$ tasks are shown in Table 5.2. When verbatim trace level is *low*, the combination with (1) *low* fuzzy trace level leads to bad performance on all test-pairs; (2) *intermediate* fuzzy trace level leads to a spatial position effect resulting in moderate performance on the end anchors, (Y_D, Y_E), (Y_A, Y_B), (Y_A, Y_C), (Y_C, Y_E), and bad performance on the other test-pairs; and (3) *high* fuzzy trace level leads to good performance on all test-pairs because the ordering information can be used to solve both memory and transitivity test-pairs. When verbatim trace level is *intermediate*, the combination with (1) *low* fuzzy trace level leads to temporal position effects, yielding moderate performance on the first and last memory test-pairs (M_1 and M_4) and poor performance on all other test-pairs; (2) *intermediate* fuzzy trace level leads to both spatial and temporal position effects resulting in moderate performance on the test-pairs except T_1; and (3) *high* fuzzy trace level leads to good performance on all test-pairs. When verbatim trace level is *high*, the combination with (1) *low* fuzzy trace level leads to complete memory of the memory test-pairs, resulting in high success probabilities on the memory test-pairs and low success probabilities on the transitivity test-pairs; (2) *intermediate* fuzzy

Table 5.2: *Expected Performance on the Test-Pairs of $O_{pos}D_{pres}$ Tasks for Nine Combinations of Trace Levels*

		hypothesized probabilities						
Verbatim	Fuzzy	Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3
low	*low*	○	○	○	○	○	○	○
	intermediate	○	⋆	⋆	○	○	⋆	⋆
	high	●	●	●	●	●	●	●
intermediate	*low*	⋆	○	○	⋆	○	○	○
	intermediate	⋆	⋆	⋆	⋆	○	⋆	⋆
	high	●	●	●	●	●	●	●
high	*low*	●	●	●	●	○	○	○
	intermediate	●	●	●	●	○	⋆	⋆
	high	●	●	●	●	●	●	●

○: bad performance; ⋆: moderate performance; ●: good performance

trace level leads to complete memory and a spatial position effect resulting in good performance on all memory test-pairs and moderate performance on the end-anchored transitivity test-pairs (T_2 and T_3); and (3) *high* fuzzy trace level leads to good performance on all test-pairs.

Disordered Position, Ordered Presentation Tasks ($D_{pos}O_{pres}$)

In $D_{pos}O_{pres}$ tasks, the objects are positioned disorderly. That is, in Figure 5.6 stick A is in the third position in the box, while stick B is in the first position. The presentation of the premises is ordered. That is, in Figure 5.6, first premise pair (A,B) is presented, followed consecutively by premise pairs (B,C), (C,D) and (D,E). A disordered position requires both high verbatim and fuzzy ability levels, because positional cues about the ordering of the objects are not provided. Consequently, not only the ordering has to be recognized but also the premise information has to be remembered. The "test-pair" box, in Figure 5.6 shows the first memory test-pair.

The expected performance patterns for combinations of verbatim and fuzzy trace levels on the memory and transitivity test-pairs of $D_{pos}O_{pres}$ tasks are shown in Table 5.3. When verbatim trace level is *low*, the perfor-

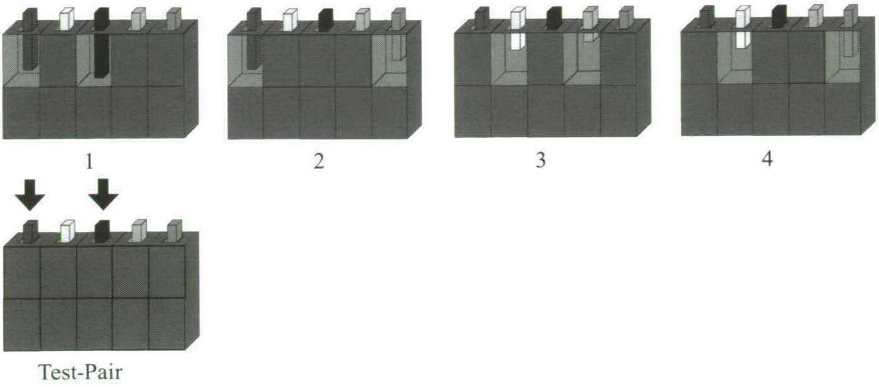


Figure 5.6: *Example of the Premise Presentation of a "Disordered Position, Ordered Presentation" Task*

mance is expected to be bad for all test-pairs, independent of fuzzy trace level. That is, for $D_{pos}O_{pres}$ tasks at least intermediate verbatim ability is needed to remember the premises or recognize the ordering of the objects. When verbatim trace level is *intermediate*, the combination with (1) *low* and *intermediate* fuzzy trace level leads to temporal position effects resulting in moderate performance on the first and last presented memory test-pairs (M_1, M_4); and (2) *high* fuzzy trace level leads to spatial position effects yielding moderate performance on the end-anchors (M_1, M_4, T_1, and T_3). When verbatim trace level is *high*, the combination with (1) *low* and *intermediate* trace levels leads to complete memory resulting in good performance on the memory test-pairs but bad performance on transitivity test-pairs; and (2) *high* fuzzy trace level leads to good performance on all test-pairs.

5.1.3 Theoretical Model

The theoretical model of fuzzy trace theory with respect to transitive reasoning is displayed in Figure 5.7. The correct and incorrect responses to the memory (M) and transitivity (T) test-pairs of the three kinds of tasks are at the lowest level of analysis. These responses are determined by the retrieval of verbatim and fuzzy traces. These traces are at the second level.

Table 5.3: *Expected Performance on the Test-Pairs of $D_{pos}O_{pres}$ Tasks for Nine Combinations of Trace Levels*

		hypothesized probabilities						
Verbatim	Fuzzy	Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3
	low	○	○	○	○	○	○	○
low	*intermediate*	○	○	○	○	○	○	○
	high	○	○	○	○	○	○	○
	low	★	○	○	★	○	○	○
intermediate	*intermediate*	★	○	○	★	○	○	○
	high	★	○	○	★	★	○	★
	low	●	●	●	●	○	○	○
high	*intermediate*	●	●	●	●	○	○	○
	high	●	●	●	●	●	●	●

○: bad performance; ★: moderate performance; ●: good performance

The use of the traces is governed by probability processes conditional on the verbatim and fuzzy ability levels, which are at the third level.

In Figure 5.7 responses to the test-pairs are manifest (i.e., observable) variables, and the verbatim and fuzzy abilities and verbatim and fuzzy trace variables are latent (i.e., unobservable) variables. The latent ability variables are continuous and the latent trace variables are ordered categorical variables. In this study, the structure of this theoretical model was tested empirically. When the theoretical model fits, the estimated probability structure must agree with the hypothesized probabilities in Tables 5.1, 5.2 and 5.3. When this is the case we are able to distinguish groups of children that differ in their use of verbatim and fuzzy traces when responding to memory test-pairs and transitivity test-pairs.

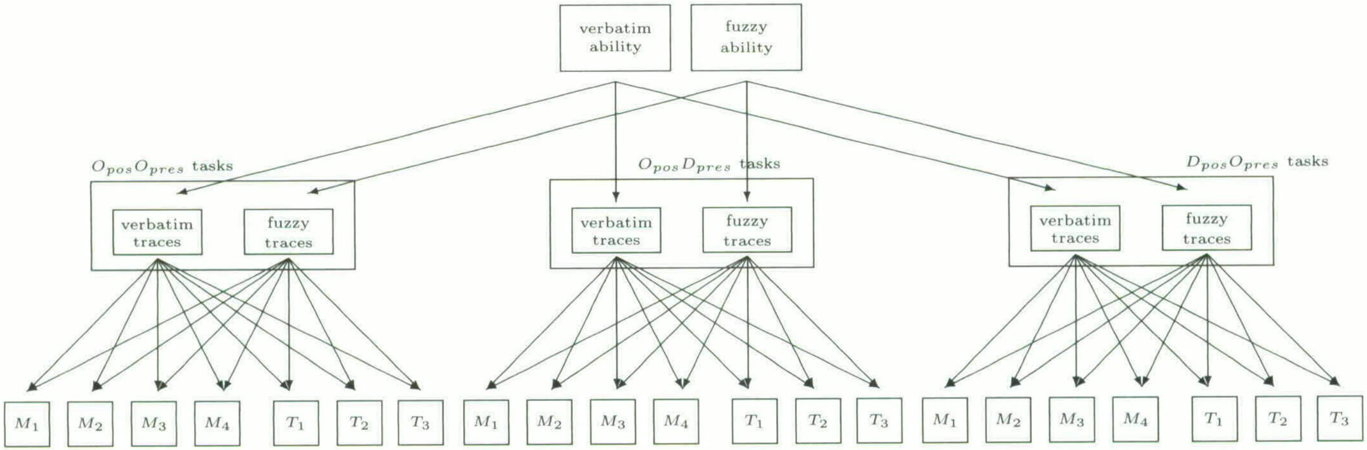


Figure 5.7: *Theoretical Model of Transitive Reasoning*

5.1.4 Hypotheses

The hypotheses to be tested were divided into three categories. The first concerns the structure of the theoretical model, the second the interpretation of the abilities, and the third the relationship between age and ability level. Together the hypotheses were a test of the fit of the theoretical model to the empirical data.

I. Structure of Theoretical Model

Hypothesis I_1***:*** TWO ABILITIES EACH INFLUENCING THREE ORDINAL TRACE LEVELS EXPLAIN PERFORMANCE BETTER THAN ONE ABILITY INFLUENCING A LIMITED NUMBER OF ORDINAL TRACE LEVELS.

Fuzzy trace theory explains the performance on test-pairs by means of *two* abilities. Ability level explains the differential use of verbatim and fuzzy trace levels. Combination of trace levels governs nine classes of typical performance. This model is hypothesized to reflect the data structure better than alternative models which posit one ability governing either one, two, three, four, or five[2] ordinal trace levels yielding one through five typical performance classes.

Hypothesis I_2***:*** THREE VERBATIM TRACE LEVELS AND THREE FUZZY TRACE LEVELS ARE THE OPTIMAL NUMBERS TO DISTINGUISH DIFFERENT PERFORMANCE GROUPS IN TRANSITIVE REASONING.

Both verbatim and fuzzy abilities are continuous. Children differ considerably with respect to ability level. However, only a limited number of verbatim and fuzzy trace levels are needed to distinguish typical performance groups. Thus, children close on verbatim ability are expected to have the same or nearly the same probability distribution for use of verbatim traces resulting in typical performance on the test-pairs. We hypothesized that three verbatim trace levels and three fuzzy trace levels are optimal. This hypothesis is tested against models having either two or four trace levels.

[2] Experience with latent class analysis has shown that fitting more than five ordered classes does not improve the fit of the model anymore (Van Onna, 2002)

Hypothesis I_3***:*** THE MODEL IN WHICH ABILITY HAS — VIA THE TRACE LEVELS — AN INDIRECT EFFECT ON PERFORMANCE EXPLAINS PERFORMANCE BETTER THAN A MODEL IN WHICH ABILITY HAS A DIRECT EFFECT ON PERFORMANCE.

According to the theory ability level influences the kind of trace level that is retrieved, and trace level influences the performance on the test-pairs. The hypothesis of three levels — ability - trace - performance — is tested against a model in which trace level (second level in Figure 5.7) is left out, indicating a direct effect of ability level on performance.

II. Interpretation of the Abilities

Hypothesis II_1***:*** THE TWO ABILITIES ARE VERBATIM ABILITY AND FUZZY ABILITY.

It is hypothesized that the verbatim ability influences, via the retrieval of verbatim traces, the performance on memory test-pairs but not the performance on transitivity test-pairs. The higher the verbatim trace level, the better the performance on memory test-pairs. The fuzzy ability level influences, via the retrieval of fuzzy traces, both the performance on memory test-pairs and transitivity test-pairs (this is the unitary trace model; Brainerd & Kingma, 1984).

Hypothesis II_2***:*** THE PERFORMANCE ON THE TEST-PAIRS OF DIFFERENT TASK TYPES AGREES WITH THE PERFORMANCE PREDICTED IN TABLES 5.1, 5.2 AND 5.3.

The performance on the test-pairs can be predicted by combinations of verbatim and fuzzy trace levels. Characteristics of the task influence how easily a trace can be used, and this is reflected by the expected performance patterns (Tables 5.1, 5.2 and 5.3).

III. Relationship Age and Ability

Hypothesis III***:*** AGE IS POSITIVELY RELATED TO VERBATIM AND FUZZY ABILITY LEVELS.

The higher the age, the higher the probability of a high ability level for both verbatim and fuzzy abilities. We hypothesized that the development of verbatim ability is fast, in particular, during the first years of life. After the first five years development progresses slowly and not remarkably. Fuzzy ability development is hypothesized to progress later and continue even into adulthood (see Figure 5.3).

5.2 Method

5.2.1 Instruments

A computer test for transitive reasoning, called `Tranred2`, was constructed (Bouwmeester & Aalbers, 2004). Tranred2 is an individual test. The registration of the test scores during test administration was done by the program. There were four versions of the test in which the tasks were presented in different order. These four versions were used to control for order effects of the task presentation. Based on their order of entry in the investigation, children were assigned to one of the four versions.

5.2.2 Sample

The transitive reasoning test was administered to 409 children ranging from 5 to 13 years of age. Children came from four elementary schools in the Netherlands. They were from middle class social-economic status (SES) families. Table 5.4 gives an overview of the number of children per grade, and the mean age and the standard deviation of age within each grade.

5.2.3 Design

The three kinds of tasks described earlier were used. Four versions of each task type were administered. Tasks of the same type differed in the colors of the sticks and the direction of the ordering or the presentation; that is, sticks could be ordered from left to right or from right to left, and sticks could be presented from small to large, or from large to small. One type of task was always followed by a different kind of task. After the premises

Table 5.4: *Number of Children, Mean Age in Months (M) and Standard Deviation (SD) in Each Grade*

Grade	*Number*	*Age*	
		M	*SD*
Kindergarten	39	73.67	4.70
1	65	86.15	4.81
2	70	100.16	5.85
3	60	111.80	5.80
4	63	123.44	5.52
5	58	140.31	7.69
6	54	146.18	6.61

were presented, first the four memory test-pairs were presented and next the three transitivity test-pairs. The ordering of the memory test-pairs was always the same as the ordering in which the premises were presented. A 1-score was assigned when the child clicked on the correct stick; and a 0-score otherwise. So for each child, 7 (test-pair) × 3 (task-type) × 4 (task-type-version) = 84 scores were assigned.

5.2.4 Procedure

The test was administered in a quiet room in the school building. The experimenter started a short conversation with the child to put her/him at ease. The child started doing two introductory tasks in which it was explained that (s)he had to click on the longest stick every time. Next, the experimenter explained that there were 13 tasks and that the child had to do them on his/her own. The child did not know that the first of the 13 tasks was another introductory task of which the purpose was to let the child get used to the idea that (s)he had to work on her/his own now.

5.2.5 Analyses

From Theoretical Model to Statistical Model

The theoretical model including the latent and manifest variables was fitted to the test-pair data by means of a multilevel latent class model (Vermunt, 2003). This model was preferred over an analyses of variance (ANOVA) model for three reasons. Firstly, the manifest, dependent variables are binary [correct (score 1)/incorrect (score 0)], whereas ANOVA assumes interval measurement level for the dependent variable. Secondly, the theoretical model contains dependent observations at two levels. At the first level, the seven test-pair scores within a task — four memory test-pairs and three transitivity test-pairs — are dependent due to the combinations of trace levels that are retrieved. For example, when a child retrieves the fuzzy trace "objects become smaller from right to left", (s)he is able to infer all memory test-pairs and transitivity test-pairs correctly. At the second level, the combination of verbatim and fuzzy traces that is used for solving a particular task is dependent on the child's verbatim and fuzzy ability levels. A multilevel model incorporates these dependencies, whereas a within-subject ANOVA is unable to do this. Thirdly, the theoretical model encompasses both manifest and latent variables. An ANOVA model cannot deal with latent variables, but a multilevel latent class model can. To summarize, a multilevel latent class models formed an appropriate model to evaluate the fit and the interpretation of the theoretical model.

An upgraded version of the program Latent Gold (Vermunt & Magidson, 2003) was used to estimate the parameters of the model and calculate fit statistics. For evaluating the fit, the sample was randomly split into two halves. The first half was used to evaluate the improvement of the fit of different models. Next the fit of the hypothesized model estimated in the first half of the sample was compared with the fit of that same model in the second half. When the fit statistics in both halves are close, the degree of chance capitalization is small and considerable negligible.

5.3 Results

5.3.1 Background Analysis

An ANOVA was performed to determine whether the order of the tasks had an effect on the number-correct score, that is, the number of correct answers to 84 items [7 (test-pair) × 3 (task-type) × 4 (task-type-version)]. An ANOVA with number-correct score as dependent variable and test-version as independent variable showed that the four test versions did not differ significantly [$F(3, 401) = 1.32, p > .05$]. Thus the presentation order of the tasks had no effect on number-correct.

A within-subject ANOVA was used to test whether the four replications of the three different task-types differed with respect to the number-correct score. The means (aggregated over test-pairs) and the 95% confidence intervals are given in Table 5.5. For $O_{pos}O_{pres}$ tasks the replications differed significantly [$F(2.75, 1112.58) = 2.93, p = .037$]. The partial η^2 (for effect size; Cohen, 1977) was low (.007). The confidence intervals of the replications all overlapped. For $O_{pos}D_{pres}$ tasks the replications differed significantly [$F(2.641, 1069.671) = 15.60, p = .000$]. The partial η^2 was low (.037). The confidence intervals of the replications all overlapped. For $D_{pos}O_{pres}$ tasks the replications differed significantly [$F(2.978, 1202.226) = 3.46, p = .016$]. The partial η^2 was low (.007). The confidence intervals of the replications overlapped.

Although there were some replications which differed significantly with respect to their average performance, the effect sizes were small and the confidence intervals showed that the differences between replications were small in all cases[3]. Therefore, we used all replications to estimate the model structure.

[3] Note that a significant overall F-value does not guarantee that individual groups differ significantly (Stevens, 1996, pp. 163–164)

Table 5.5: *Means (M, Aggregated over Test-Pairs), Standard Errors (SE) and 95% Confidence Intervals (CI) for the Replications (A, B, C and D) of Each of the Three Task Types*

	$O_{pos}O_{pres}$			$O_{pos}D_{pres}$			$D_{pos}O_{pres}$		
Rep.	*M*	*SE*	*95% CI**	*M*	*SE*	*95% CI**	*M*	*SE*	*95% CI**
A	.75	.02	.71-.78	.72	.01	.68-.75	.57	.01	.54-.60
B	.78	.01	.74-.81	.69	.01	.66-.72	.59	.01	.56-.62
C	.79	.01	.76-.82	.65	.01	.62-.69	.55	.01	.52-.58
D	.78	.01	.75-.82	.75	.01	.72-.79	.57	.01	.54-.60

* Bonferroni adjustment

5.3.2 Hypotheses Testing

I. Structure of Theoretical Model

All hypotheses with respect to structure — two abilities or one, three trace levels optimal, two-level dependencies — were confirmed[4]. See Table 5.6 for the results. Model B, consisting of one ability and one latent class, fitted worse than model C, which had one ability influencing five ordered trace levels [see BIC, AIC3, and decrease in LL (taking into account the difference in number of parameters) in Table 5.6]. Moreover, the decrease in LL of Model A relative to Model C, given the increase in number of parameters, was substantial. This indicates that Model A fits better than Model C (a formal significance test is hazardous, however; therefore the BIC values were compared). In Model D the latent trace levels were omitted leading to a direct effect of ability level on performance. The fit of model D was worse than the fit of model A in terms of BIC, AIC3, and decrease in LL. Therefore it was concluded that the trace levels could not be omitted. The three-trace-level model (A) fitted better than the two-trace-level model (E) in terms of BIC, AIC3, and decrease in LL.

[4] One particular pattern of outliers, constituting 0.4% of all patterns, was found which negatively influenced the fit of the models. For children who produced this pattern the answers were scored as if they were missing.

Table 5.6: *Fit Measures for the Estimated Models*

Model	Description	LL	#Par	BIC	AIC3
A	2 abilities, 3 ordered trace levels	-8880.33	69	18298.45	17967.67
B	1 ability, 1 trace level	-10043.71	43	20422.56	19958.42
C	1 ability, 5 ordered trace levels	-9014.70	45	18395.71	18164.49
D	2 abilities, no trace levels	-9259.08	63	19009.18	18707.18
E	2 abilities, 2 trace levels	-8914.07	67	18350.34	18029.14
F	2 abilities, 4 ordered trace levels	-8848.05	71	18249.47	17909.10
Cross Validation	2 abilities, 3 ordered trace levels	-9090.06	69		

BIC: $-2LL + \#parameters \times ln(N)$ (for N=204)

AIC3: $-2LL + 3 \times \#parameters$

The four-trace-level model (F) fitted better than the three-trace-level model in terms BIC, AIC3, and decrease in LL. The interpretation of the four-class model showed that two of the four classes did not differ conceptually. Therefore, it was concluded that three trace levels were optimal to distinguish relevant groups.

Chance capitalization was evaluated by fitting Model A to the second random half of the sample (see Table 5.6). Because the numbers of records (subjects × items per subject) was not exactly the same in both subsamples (2426 records and 2434 records) due to missing values, we compared the log-likelihood per record: For the first sample the LL per record equalled -3.66, and for the second sample it equalled -3.73. This means that Model A fitted almost equally well in both samples.

It can be concluded that the hypothesized model fitted the data well in comparison with alternative models. However, a fitting model can only be accepted when the interpretation of the parameters agrees with the underlying theory. This interpretation follows below.

II. Interpretation of Estimated Model

Hypothesis II_1: Table 5.7 shows the structure of the estimated success probabilities for the seven test-pairs per combination of verbatim and fuzzy trace levels. The estimated success probabilities are summarized in three categories to keep the presentation of the results orderly. Notation ○ means a success probability lower than 0.65; ⋆ means a success probability between 0.65 and 0.80; and • means a success probability higher than 0.80. With respect to the first trace (given low second trace level), in general, the success probabilities of the memory test-pairs (M_1, M_2, M_3, and M_4) are low when the trace level is low (rows 1, 2 and 3), higher when the trace level is intermediate (rows 10, 11, 12) and high when the trace level is high (rows 19, 20, 21). This pattern was found with the memory test-pairs but not with the transitivity test-pairs (T_1, T_2, T_3). Therefore, the first latent trace can be interpreted as the verbatim trace.

Table 5.7: *Estimated Success Probability for the Test-Pairs of Three Task-Types, for Nine Combinations of Latent Trace Levels*

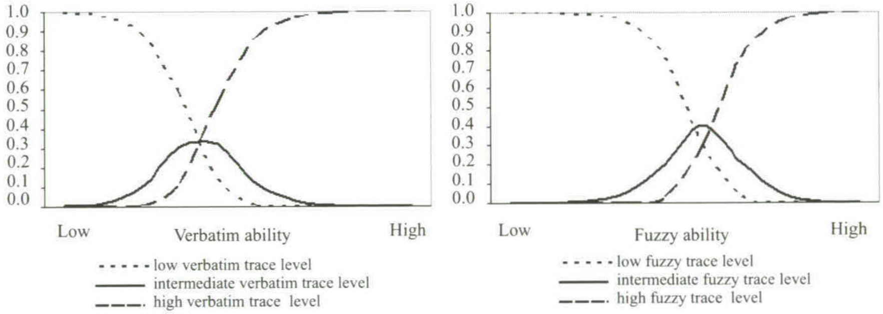
			Estimated probabilities						
First Trace	Second Trace	Task Type	Memory				Transitivity		
			M_1	M_2	M_3	M_4	T_1	T_2	T_3
low	low	$O_{pos}O_{pres}$	○	○	○	○	○	○	○
		$O_{pos}D_{pres}$	○	○	○	○	○	○	○
		$D_{pos}O_{pres}$	○	○	○	○	○	○	○
	Interm.	$O_{pos}O_{pres}$	●	●	●	●	●	●	●
		$O_{pos}D_{pres}$	★	★	●	●	★	★	●
		$D_{pos}O_{pres}$	○	○	○	○	★	○	○
	high	$O_{pos}O_{pres}$	●	●	●	●	●	●	●
		$O_{pos}D_{pres}$	●	●	●	●	●	●	●
		$D_{pos}O_{pres}$	★	○	○	○	★	○	★
Interm.	low	$O_{pos}O_{pres}$	●	●	●	●	○	○	○
		$O_{pos}D_{pres}$	●	●	●	★	○	○	○
		$D_{pos}O_{pres}$	●	★	○	○	○	○	○
	Interm.	$O_{pos}O_{pres}$	●	●	●	●	●	●	●
		$O_{pos}D_{pres}$	●	●	●	●	★	●	●
		$D_{pos}O_{pres}$	●	●	★	★	★	○	★
	high	$O_{pos}O_{pres}$	●	●	●	●	●	●	●
		$O_{pos}D_{pres}$	●	●	●	●	●	●	●
		$D_{pos}O_{pres}$	●	●	★	●	●	○	★
high	low	$O_{pos}O_{pres}$	●	●	●	●	○	○	○
		$O_{pos}D_{pres}$	●	●	●	●	○	○	○
		$D_{pos}O_{pres}$	●	●	●	●	★	○	○
	Interm.	$O_{pos}O_{pres}$	●	●	●	●	●	●	●
		$O_{pos}D_{pres}$	●	●	●	●	★	●	●
		$D_{pos}O_{pres}$	●	●	●	●	●	★	★
	high	$O_{pos}O_{pres}$	●	●	●	●	●	●	●
		$O_{pos}D_{pres}$	●	●	●	●	●	●	●
		$D_{pos}O_{pres}$	●	●	●	●	●	★	●

○: $< .65$; ★: $.65 - .79$; ●: $> .79$

With respect to the second latent trace (Table 5.7, second column), the success probabilities of the transitivity test-pairs (T_1, T_2, and T_3) are low when the trace level is low (rows 1, 2, 3 — 10, 11, 12 — 19, 20, 21), (in general) higher when the trace level is intermediate (rows 4, 5, 6 — 13, 14, 15 — 22, 23, 24) and (in general) high when the trace level is high (rows 7, 8, 9 — 16, 17, 18 — 25, 26, 27). Therefore, the second trace can be interpreted as the fuzzy trace.

Standard errors of the estimated success probabilities (not tabulated here) were between 0.000 and 0.077 (mean = 0.027, standard deviation = 0.02). This means that the confidence intervals were relatively small.

Figure 5.8a shows the distribution of latent verbatim trace levels, given a child's latent verbatim ability. Figure 5.8a shows that the probability of using a low verbatim trace level decreases as a function of verbatim ability and is maximal when ability level is low. The probability of using an intermediate verbatim trace level first increases and then decreases as a function of ability and is maximal when the verbatim ability level is intermediate.



a. Relationship verbatim ability and verbatim trace levels.

b. Relationship fuzzy ability and fuzzy trace levels.

Figure 5.8: *Distribution of Latent Verbatim Trace Levels Given Latent Verbatim Ability (Panel a) and Distribution of Latent Fuzzy Trace Levels Given Latent Fuzzy Ability (Panel b)*

Note that along the verbatim ability, the probability of using an intermediate verbatim trace level is never higher than the probabilities of using a low or a high verbatim trace level. The probability of using a high verbatim trace level increases as a function of ability and is maximal when ability level is high.

Figure 5.8b shows the distribution of latent fuzzy trace levels, given a child's latent fuzzy ability. The interpretation of the distribution of the fuzzy trace levels is the same as the interpretation of the verbatim trace levels. Note that in Figure 5.8b there exists a small region on the fuzzy ability where the intermediate fuzzy trace level has higher probability than the low and high fuzzy trace levels.

Hypothesis II_2: In Table 5.8, both the hypothesized and the estimated success probabilities of the test-pairs of $O_{pos}O_{pres}$ tasks are shown. The majority of the estimated success probability patterns agreed with the hypothesized success probability patterns. The pattern in the fourth row differed in estimated and hypothesized success probabilities with respect to the memory test-pairs.

Table 5.8: *Estimated Success Probability for the Test-Pairs of Tasks $O_{pos}O_{pres}$ for Nine Combinations of Latent Trace Levels*

		Hypothesized probabilities							Estimated probabilities						
Verbatim	Fuzzy	Memory				Transitivity			Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3	M_1	M_2	M_3	M_4	T_1	T_2	T_3
low	*low*	○	○	○	○	○	○	○	.59	.48	.38	.49	.55	.56	.50
	interm.	●	●	●	●	●	●	●	.94	.97	.95	.96	.97	.94	.97
	high	●	●	●	●	●	●	●	.99	1.0	1.0	1.0	1.0	.99	1.0
interm.	*low*	★	○	○	★	○	○	○	*.93*	*.95*	*1.0*	*.99*	.49	.46	.48
	interm.	●	●	●	●	●	●	●	.99	1.0	1.0	1.0	.97	.91	.96
	high	●	●	●	●	●	●	●	1.0	1.0	1.0	1.0	1.0	.99	1.0
high	*low*	●	●	●	●	○	○	○	.99	1.0	1.0	1.0	.43	.36	.46
	interm.	●	●	●	●	●	●	●	1.0	1.0	1.0	1.0	.96	.87	.96
	high	●	●	●	●	●	●	●	1.0	1.0	1.0	1.0	1.0	.99	1.0

○: $< .65$; ★: $.65 - .79$; ●: $> .79$

It was hypothesized that the *intermediate* verbatim trace level and the *low* fuzzy trace level lead to a temporal position effect, predicting moderate probabilities for the memory test-pairs presented first and last (M_1 and M_4) and low probabilities for the test-pairs in between (M_2 and M_3). However, the results showed complete memory for premises when verbatim trace level is intermediate and fuzzy trace level is low.

Table 5.9 shows that for $O_{pos}D_{pres}$ tasks the majority of the estimated success probability patterns agreed with the hypothesized success probability patterns. Two patterns (in rows 2 and row 4) differed in hypothesized and estimated success probabilities. It was hypothesized that *intermediate* fuzzy trace level in combination with *low* verbatim trace level leads to a spatial position effect resulting in higher success probabilities for the end-anchored test-pairs (Table 5.9, row 2; note that the end-anchored test-pairs were M_2, M_3, T_2 and T_3) than for the mid-term test-pairs. However, the estimated success probabilities show that this spatial position effect is only active at one end-anchor leading to high success probabilities for the test-pairs M_3 and T_3. Further, it was hypothesized that *intermediate* verbatim trace level in combination with *low* fuzzy trace level leads to a temporal position effect (Table 5.9, row 4). However, the estimated success probabilities showed a temporal position effect for the first memory test-pairs (M_1 and M_2), but not for the last.

Table 5.9: *Estimated Success Probability for the Test-Pairs of Tasks $O_{pos}D_{pres}$ for Nine Combinations of Latent Trace Levels*

		Hypothesized probabilities							Estimated probabilities						
Verbatim	Fuzzy	Memory				Transitivity			Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3	M_1	M_2	M_3	M_4	T_1	T_2	T_3
low	*low*	○	○	○	○	○	○	○	.45	.41	.50	.46	.47	.44	.46
	interm.	○	★	★	○	○	★	★	.71	.74	.87	.82	.79	.79	.89
	high	●	●	●	●	●	●	●	.88	.92	.98	.96	.94	.95	.99
interm.	*low*	★	○	○	★	○	○	○	.95	.85	.80	.73	.45	.54	.55
	interm.	★	★	★	★	○	★	★	.98	.96	.97	.94	.78	.85	.92
	high	●	●	●	●	●	●	●	.99	.99	1.0	.99	.94	.96	.99
high	*low*	●	●	●	●	○	○	○	1.0	.98	.94	.89	.43	.63	.64
	interm.	●	●	●	●	○	★	★	1.0	.99	.99	.98	.77	.89	.94
	high	●	●	●	●	●	●	●	1.0	1.0	1.0	1.0	.93	.97	.99

○: $< .65$; ★: $.65 - .79$; ●: $> .79$

Table 5.10 shows that for $D_{pos}O_{pres}$ tasks, four estimated success probability patterns agreed with the hypothesized success probability patterns.

Five patterns (in rows 3, 4, 5, 6, 8) differed in hypothesized and estimated success probabilities. Firstly, low success probabilities on all test-pairs were hypothesized when verbatim trace level was *low* and fuzzy trace level was *high* (Table 5.10, row 3). However, the estimated probabilities showed a spatial position effect resulting in moderate and high success probabilities for the test-pairs M_1, M_4, T_1, and T_3. Secondly, a temporal position effect was hypothesized when verbatim trace level was *intermediate* and fuzzy trace level was *low* or *intermediate* (Table 5.10, rows 4 and 5). However, the estimated success probabilities only showed this effect on the first memory test-pair but not on the last memory test-pair when fuzzy trace level was *low*. A spatial position effect was active (in particular at one side) when fuzzy trace level was *intermediate*. Thirdly, for *intermediate* verbatim trace level and *high* fuzzy trace level, a spatial position effect was hypothesized (Table 5.10, row 6). The estimated probabilities showed a spatial position effect at only one end-anchor but not at both. Finally, it was hypothesized that *high* verbatim trace level and *intermediate* fuzzy trace level would lead to high memory test-pair probabilities and low transitivity test-pair probabilities (Table 5.10, row 8). The estimated probabilities for the transitivity test-pairs were high for the end-anchors.

Table 5.10: *Estimated Success Probability for the Test-Pairs of Tasks $D_{pos}O_{pres}$ for Nine Combinations of Latent Trace Levels*

		Hypothesized probabilities							Estimated probabilities						
Verbatim	Fuzzy	Memory				Transitivity			Memory				Transitivity		
		M_1	M_2	M_3	M_4	T_1	T_2	T_3	M_1	M_2	M_3	M_4	T_1	T_2	T_3
low	*low*	○	○	○	○	○	○	○	.50	.24	.25	.19	.50	.57	.47
	interm.	○	○	○	○	○	○	○	.61	.38	.32	.39	.66	.56	.62
	high	○	○	○	○	○	○	○	*.71*	.54	.41	.64	*.79*	.56	*.75*
interm.	*low*	★	○	○	★	○	○	○	.88	.71	.63	*.52*	.61	.63	.53
	interm.	★	○	○	★	○	○	○	*.92*	*.83*	.71	.75	*.75*	.63	.67
	high	★	○	○	★	★	○	★	*.94*	*.90*	*.78*	*.89*	.85	.63	.78
high	*low*	●	●	●	●	○	○	○	.98	.95	.90	.84	.71	.69	.58
	interm.	●	●	●	●	○	○	○	.99	.97	.93	.93	*.82*	*.69*	*.71*
	high	●	●	●	●	●	●	●	.99	.99	.95	.98	.90	.69	.82

○: < .65; ★: .65 – .79; ●: > .79

III. Relationship Age and Ability

Hypothesis III: Figure 5.9 displays the scatterplots of the verbatim and fuzzy ability by age. The fit of linear, quadratic and cubic regression curves did not differ significantly. Thus, the curvature of the hypothesized developmental curves in Figure 5.3 was not supported by the data. The percentages of explained variance of the linear models were .08 for verbatim ability and .20 for fuzzy ability.

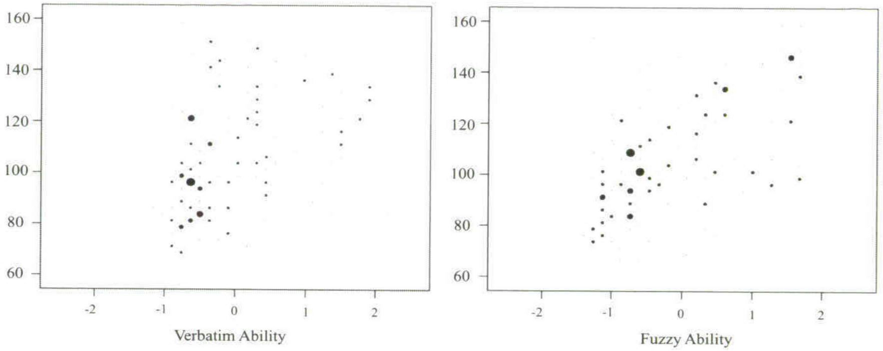


Figure 5.9: *Scatterplots of Verbatim and Fuzzy Ability Scores and Age in Months (the Larger the Bullets, the More Data Points on the Same Position)*

5.4 Discussion

In this study, fuzzy trace theory was applied to transitive reasoning. A theoretical model was set up in which the performance on memory test-pairs and transitivity test-pairs was explained by the use of verbatim and fuzzy traces, which were dependent on the verbatim and fuzzy ability levels, respectively. Age was hypothesized to be related to both abilities. A multilevel latent class model was used to handle the dependencies between ability level and trace retrieval on the one hand, and trace retrieval and performance on the test-pairs on the other hand. Fitting the model had two aspects. Firstly, we investigated the structure of the empirical data and concluded that two abilities had to be distinguished. Secondly, we investigated whether these abilities could be interpreted as verbatim and fuzzy abilities, and concluded that this was justified by the results.

This study showed that a high ability to remember premises is not enough to correctly infer transitive relationships. An important result was that children who have a high verbatim ability level but a low fuzzy ability level performed well on the memory test-pairs but at chance-level on the transitivity test-pairs. This finding disagrees with Trabasso's linear ordering theory which assumes that memory of the premises is enough to infer the transitive relationship. Moreover, the results did not agree with Piaget's theory. Piaget's theory assumes that memory for the premises is a prerequisite for the capacity of using logical rules and inferring transitive relationships. The format of the task was not expected to influence the use of logical rules when the premises could be remembered. We found that memory for the premises was not a prerequisite for inferring the transitive relationship and that the format of the task had strong influence on the success probability of inferring the transitive relationships, even when the memory test-pairs were correctly remembered. However, the initial aim of Piaget was not to give such a detailed description of transitive reasoning, making a comparison between his theory and the present study disputable.

Some relevant deviations from the expected probability patterns of combinations of verbatim and fuzzy trace levels were found. These deviations in particular concerned the finding of temporal position effects only at the information presented first instead of the information presented first and last. Spatial position effects in $D_{pos}O_{pres}$ tasks were found in particular at one side of the ordering (containing the longest sticks) but not on both. This result may be explained by a marking effect. That is, linguistic factors played a role in the end-anchoring. During the premise presentation children had to click on the longest stick, which may explain that their representation of the long-end-anchor is better than the short-end-anchor (see Riley & Trabasso, 1974; Trabasso et al., 1975; Sternberg, 1980b).

Brainerd and Kingma (1984) showed that the unitary trace model could well explain performance on memory and transitivity test-pairs. This model assumes that both memory and transitivity test-pairs are solved by means of fuzzy traces. We approached the data from a different angle by distinguishing strategy groups instead of fixed age groups, and concluded that different groups can be distinguished which are characterized by differential use of verbatim and fuzzy traces. For children having high fuzzy ability levels, indeed the unitary trace model can explain both performance on memory and transitivity test-pairs when verbatim ability level is intermediate or high. However, for children having intermediate or low fuzzy ability level the verbatim trace has a strong influence on the performance on memory test-pairs, indicating that there is a changing orientation from the use of verbatim traces to both kinds of traces and, finally, to fuzzy traces. For tasks in which the position of the objects is not ordered, as in $D_{pos}O_{pres}$ tasks, both high verbatim and fuzzy trace levels were required to infer the transitive relationship.

We determined the influence of age by means of the relationship between age and ability level. We used a different perspective than Brainerd and Kingma (1984), who assumed fixed age groups and investigated the differences between various age groups in performance. In Brainerd and Kingma's (1984) study individual differences within age groups were ignored. We showed that the correlation between age and verbatim ability was low and between age and fuzzy ability moderate. This result indicated that age influences performance but that the effect is not strong. Therefore, it seems more appropriate to study development by distinguishing strategy groups instead of fixed age groups. In his book, Wohlwill (1973, pp 26-28), when summarizing Kessen's (1960) objections to the use of age as a variable in behavioral research, already argued that chronological age is not a useful variable in statements of functional relationships to behavior, since there are considerable differences in rates of developmental change.

The results of this study also have implications for the discussion about developmental stages. With respect to transitive reasoning even five-year old children may have a substantial probability to retrieve high-level fuzzy traces and thus infer the complete ordering of a task. Also, 12-year old children may have a substantial probability to retrieve the lowest trace level and thus do not recognize any ordering in the task. In other words, it is not possible to distinguish clear-cut developmental stages in the development of transitive reasoning (see also Bouwmeester & Sijtsma, submitted, chapter 3 of this thesis). Because we used a cross-sectional design, no conclusions could be drawn about the transition from one ability level to another. A longitudinal design is needed to study such transitions. This requires an extra level in the multi-level structure to model the dependencies within individual children's data over time.

Appendix

Model formulation

Let test pairs be indexed $k = 1, .., 7$; tasks $i = 1, .., 12$; and children $j = 1, .., N$. Response variable $Y_{ijk} = 1$ when child j gives a correct response to test-pair k in task i, and $Y_{ijk} = 0$ otherwise. The scores of child j on task i are collected in the vector $\mathbf{Y}_{ij}$, and $\mathbf{Y}_j$ denotes the scores of child j on all 12 tasks.

The variant of the multilevel latent class (LC) model we used contains two ordinal latent variables denoted by X_{ij} and Q_{ij} representing the verbatim and fuzzy traces, respectively, for a particular task i. These two mutually independent latent variables are assumed to have discrete realization between 0 and 1, with equal distances between categories. With three classes per dimension, $x = 0.0$, 0.5, or 1.0, and $q = 0.0$, 0.5, or 1.0. This yields an LC model with multiple latent variables that Magidson and Vermunt (2001) called an LC factor model. If we assume that the various tasks performed by a child are independent of one another, the relevant LC factor model for $\mathbf{Y}_{ij}$ is of the form

$$P(\mathbf{Y}_{ij}) = \sum_{x} \sum_{q} P(X_{ij} = x) P(Q_{ij} = q) \prod_{k=1}^{7} P(Y_{ijk} | X_{ij} = x, Q_{ij} = q). \qquad (5.1)$$

This equation reveals the basic assumption of a LC model: the scores on the 7 test-pairs are mutually independent given the latent verbatim and fuzzy trace levels of child j at task i.

Because of the nesting of tasks within children, the standard assumption of independent observations is not correct for our data. The multiple tasks performed by a child can, however, be assumed to be mutually independent given the child's latent verbatim and fuzzy abilities. These two continuous latent variables, which are denoted by W_j and V_j respectively, with realization w and v, have the role of random effects in the models for X_{ij} and Q_{ij} (Vermunt, 2003). The abilities or random effects W_j and V_j modify the model for $\mathbf{Y}_{ij}$ described in Equation 5.1 as follows:

$$\begin{aligned} P(\mathbf{Y}_{ij}|W_j = w, V_j = v) &= \sum_x \sum_q P(X_{ij} = x|W_j = w)\, P(Q_{ij} = q|V_j = v_j) \\ & \prod_{k=1}^{7} P(Y_{ijk}|X_{ij} = x, Q_{ij} = q). \end{aligned} \tag{5.2}$$

As can be seen, X_{ij} is assumed to depend on W_j, and Q_{ij} on V_j. Moreover, the effects of the continuous latent abilities on the responses are assumed to be fully mediated by the discrete latent trace levels.

The probability associated with all responses of an individual, denoted by $P(\mathbf{Y}_j)$, is obtained by taking the product of $P(\mathbf{Y}_{ij}|W_j = w, V_j = v)$ over the 12 tasks and integrating the two latent ability variables out of the equation. This yields:

$$P(\mathbf{Y}_j) = \int_w \int_v f(W_j = w)\, f(V_j = v) \left[\prod_{i=1}^{12} P(\mathbf{Y}_{ij}|W_j = w, V_j = v)\right] d\,w\, d\,v. \tag{5.3}$$

Note that $P(\mathbf{Y}_{ij}|W_j = w, V_j = v)$ has the form described in Equation 5.2, and $f(W_j = w)$ and $f(V_j = v)$ are standard normal univariate distributions.

The three types of model probabilities appearing in Equation 5.2 – $P(X_{ij} = x|W_j = w)$, $P(Q_{ij} = q|V_j = v_j)$, and $P(Y_{ijk}|X_{ij} = x, Q_{ij} = q)$ – are parameterized as logit models. The probability of a correct response of child j on test-pair k of task i is restricted by a standard binary logit model of the form

$$P(Y_{ijk} = 1|X_{ij} = x, Q_{ij} = q) = \frac{\exp(\beta_{0ki} + \beta_{1ki} \cdot x + \beta_{2ki} \cdot q + \beta_{3ki} \cdot x \cdot q)}{1 + \exp(\beta_{0ki} + \beta_{1ki} \cdot x + \beta_{2ki} \cdot q + \beta_{3ki} \cdot x \cdot q)}, \tag{5.4}$$

where β_{0ki} is an intercept, β_{1ki} and β_{2ki} are the main effects of verbatim trace level and fuzzy trace level, respectively, and β_{3ki} is the interaction effect of verbatim and fuzzy trace level. The indices k and i indicate that these parameters differ across test-pairs and tasks. This is, however, not fully correct since the parameters were restricted to be equal for all four replications of the same task-type (e.g., $\beta_{0k,i+3} = \beta_{0k,i}$). This implies that we have to estimate only three sets of free β parameters.

The other two parts of the model, capturing the relative sizes of the verbatim and fuzzy trace levels given the verbatim and fuzzy ability levels, are modeled as

$$P(X_{ij} = x|W_j = w) = \frac{\exp(\gamma_{0x} + \gamma_1 \cdot x \cdot w)}{\sum_x \exp(\gamma_{0x} + \gamma_1 \cdot x \cdot w)},$$

and

$$P(Q_{ij} = q|V_j = v) = \frac{\exp(\gamma_{3q} + \gamma_4 \cdot q \cdot v)}{\sum_q \exp(\gamma_{3q} + \gamma_4 \cdot q \cdot v)}.$$

These are adjacent-category ordinal logit models similar to the ones used in partial-credit models, which are IRT models for ordinal items. The γ parameters are assumed to be equal across the 12 tasks.

The multilevel latent class models were estimated by means of maximum likelihood using an adapted version of the EM algorithm (Vermunt, 2003, 2004). This procedure is implemented in version 4.0 of Latent GOLD (Vermunt & Magidson, 2003), a Windows-based program for LC analysis, that is available at www.statisticalinnovations.com.

Epilogue

I started this thesis project believing that the development of transitive reasoning could be studied by simply letting children perform a transitive reasoning task and ask them to explain their answer. When children mentioned the premises necessary for the transitive inference, this was taken as evidence that they were capable of transitive reasoning; and when they did not mention the premises, they were incapable of transitive reasoning.

This was a simple and equally naive idea which I rejected after having seen two children perform a transitive reasoning task. These children explained their answers in several ways, which included information about either the ordering of the sticks, the colors of the sticks, aspects of the environment, or the premise information. Some of these explanations were incorrect, having nothing to do with the task. Among the correct explanations some used the premisses, but others used a strategy that did not include the premises.

For me, this was the first serious confrontation with a difficult problem: what exactly *is* transitive reasoning? Piaget used transitive reasoning tasks only as tools to study whether children were capable of operational reasoning. According to his theory, children had to understand and apply logical rules in concrete tasks like transitive reasoning tasks. However, in practice it appeared that children used different strategies to infer the transitive relationship and often these strategies led to correct inferences. As a result, it seemed implausible to conclude that these children were incapable of transitive reasoning.

What is transitive reasoning? How does it develop, and how is development characterized? What is the role of environmental influences? These issues formed the fundamental questions of cognitive development according to Wohlwill (1973).

Wohlwill (1973, pp. 40-42) claimed that the discovery and synthesis of developmental dimensions was the first step in studying cognitive developmental concepts. In his book "the study of behavioral development", Wohlwill (1973) extensively discussed the questions to be asked when studying cognitive development and the methods to be used for answering these questions. He explained that methods available at the time when he wrote his book were suited primarily for analyzing data collected in an experimental context and, therefore, often inappropriate for studying developmental change. According to Wohlwill, developmental psychology requires a differential approach in which changes in behavior are described within the natural environment in which the emphasis is on response patterns and individual differences.

Modern test theory or item response theory has grown substantially over the past decades, now offering appropriate and sophisticated analysis methods to handle differential questions of the type discussed by Wohlwill. In this thesis, a few of Wohlwill's developmental issues were discussed in the context of transitive reasoning and item response theory was used to clarify these issues.

Often, developmental theories are rather vague or unspecified with respect to the underlying dimensions of constructs and the influence of task characteristics on children's performance. Moreover, dimensionality does not have an absolute meaning and is valuable only to the degree in which the research is based on a clear and unambiguous operationalization of the construct of interest. Many theories lack this clarity. Furthermore, the definition of a psychological dimension does not have a one-to-one relationship with a mathematically defined dimension as represented in statistical methods such as item response models. The methods used in this thesis for investigating the dimensionality, assumed slightly different mathematical definitions of dimensionality which led to somewhat different results and interpretations. This taught us that psychological dimensionality can be approached from different statistical perspectives which, when used together, may give a rather complete picture of the psychological dimensionality.

According to Wohlwill (1973, e.g., p. 40), the next step in describing developmental change was to determine whether behavior changes are quantitative or qualitative and, corresponding with this, how to interpret continuity or discontinuity in development. Wohlwill (1973, p. 59) emphasized that the answer to this question is mainly determined by the level of analysis. Change can be analyzed at many levels of sophistication, each of which leads to different conclusions about continuity or discontinuity. Moreover, Wohlwill (1973, p. 25) argued that chronological age is inappropriate for detecting discontinuity due to the probabilistic character of change in behavior.

Nowadays, continuity and discontinuity can be studied effectively by means of latent class analysis. This method can be used to distinguish groups of children, which differ with respect to their response patterns to transitive reasoning tasks. In this thesis, discontinuity was studied on the basis of the data structure without a priori assuming fixed age groups. Moreover, latent class analysis made it possible to study relationships between environmental influences, cognitive behavior, and age in different latent classes. We emphasize that latent classes identified from the data only have relative meaning, primarily dependent on the operationalization of the construct, the level of analysis, and the particular statistical method used. Without a highly accurate level of specification of the developmental theory the statistical model cannot offer useful results.

Wohlwill (1973) adviced to study individual differences in development by means of the changes in individuals' score patterns produced in response to the tasks. In chapter 5, fuzzy trace theory was used to explain individual differences in the development of transitive reasoning in detail. Brainerd and Kingma (1984, 1985) elaborated fuzzy trace theory but used an experimental design to test different aspects of the theory. In their research, these authors were unable to study individual differences, and development could not be investigated because average age scores were used as the level of analysis. The availability of new and advanced statistical methods enabled us to analyze response patterns and predict the responses processes on different kinds of transitive reasoning tasks assuming distinct verbatim and fuzzy ability levels. The recently developed multi-level latent class model is a sophisticated and powerful tool for testing the hypothesized structure of the theoretical model and for describing the development of transitive reasoning at a detailed level of analysis. In future research, this method may be used in the context of a longitudinal design for studying developmental transition processes.

The developmental issues discussed in this thesis are not specific for transitive reasoning. The issues of dimensionality, qualitative or quantitative change, influence of environmental factors, and development in individual response patterns can be generalized to other developmental - often Piagetian - concepts. Verweij (1994), De Koning (2000), Jansen (2001), and Hosenfield (2003) already made fruitful contributions. With respect to transitive reasoning a longitudinal study would be the next step to study the transition in developmental change.

What did we learn about developmental psychology from the hundreds of children who performed transitive reasoning tests in this study? In the introduction, I mentioned the most important differences between Piaget's theory about cognitive development, information processing theory, and fuzzy trace theory. In this thesis, the hierarchical nature of Piaget's theory, which views children as imperfect adults progressing through the necessary stages, starting from the sensory-motor stage and ending at the stage of formal adult thinking, was not supported by the empirical observations. Thus, the development of children's reasoning was not characterized by a shift from functional to operational thinking. Moreover, it was found that development was neither characterized by an increase in the completeness of a quantitative, symbolic representation of incoming information nor the efficiency to form such a representation, as is assumed by information processing theorists. Instead, we found that development seems to be characterized by a growing ability to retrieve information which adequately matches the task requirements. According to fuzzy trace theory, for solving a cognitive task children learn to use the fuzziest trace that leads to success. During development people learn that pattern information is often better suited than verbatim information because pattern information can be retrieved longer than verbatim information and new information can be inferred from the pattern information. However, for some cognitive tasks, requiring detailed verbatim information, children perform a better job than many adults!

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention, criteria for finite mixture logit models. *Journal of Marketing Research, 40*, 235–243.

Bidell, T. R., & Fischer, K. W. (1992). Beyond the stage debate. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 100–140). Cambridge: Cambridge University Press.

Bouwmeester, S., & Aalbers, T. (2002). TRANRED. Tilburg: Tilburg University.

Bouwmeester, S., & Aalbers, T. (2004). TRANRED2. Tilburg: Tilburg University.

Bouwmeester, S., & Sijtsma, K. (2004). Measuring the ability of transitive reasoning, using product and strategy information. *Psychometrika, 69*, 123–146.

Bouwmeester, S., & Sijtsma, K. (submitted). Detecting discontinuity in the development of transitive reasoning: A comparison of two models.

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.

Braine, M. D. S. (1959). The onthogeny of certain logical operations: Piaget's formulation examined by nonverbal methods. *Monographs for the Society for Research in Child Development, 27*, 41–63.

Brainerd, C. J. (1973). Judgments and explanations as criteria for the presence of cognitive structures. *Psychological Bulletin, 3*, 172–179.

Brainerd, C. J. (1977). Response criteria in concept development research. *Child Development, 48*, 360–366.

Brainerd, C. J. (1978). The stage question in cognitive-developmental theory. *Behavioral and Brain Sciences, 2*, 173–213.

Brainerd, C. J. (1979). Markovian interpretations of conservation learning. *Psychological Review, 86*, 181–213.

Brainerd, C. J. (1993). Cognitive development is abrupt (but not stage-like). *Monographs for the Society for Research in Child Development, 58*, 170–190.

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., & Kingma, J. (1985). On the independence of short-term memory and working memory in cognitive development. *Cognitive Psychology, 17*, 210–247.

Brainerd, C. J., & Reyna, V. F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*, 3–47.

Brainerd, C. J., & Reyna, V. F. (1992). The memory independence effect: What do the data show? What do the theories claim? *Developmental Review, 12*, 164–186.

Brainerd, C. J., & Reyna, V. F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100*, 42–67.

Brainerd, C. J., & Reyna, V. F. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*, 1–75.

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. *Advances in Child Development and Behaviour, 28*, 41–99.

Brainerd, C. J., & Reyna, V. F. (2004). Perspectives in behavior and cognition. *Developmental Review, 24*, 396–439.

Breslow, L. (1981). Reevaluation of the literature on the development of transitive inferences. *Psychological Bulletin, 89*, 325–351.

Bryant, P. E., & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature, 232*, 456–458.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Case, R. (1992). Neo-Piagetian theories of child development. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 161–196). Cambridge: Cambridge University Press.

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. (1988). *Constructive evolution: Origins and development of* P*iaget's thought.* Cambridge: Cambridge University press.

Chapman, M., & Lindenberger, U. (1988). Functions, operations, and decalage 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.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Journal of Educational Psychology, 76*, 387–404.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

De Koning, E. (2000). *Inductive reasoning in primary education.* Unpublished doctoral dissertation, Utrecht University, The Netherlands.

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.

DeSoto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Social Psychology, 2*, 513–521.

Dolan, C. V., Jansen, B. R. J., & Van der Maas, H. L. J. (2004). Constrained and unconstrained multivariate normal finite mixture modeling of piagetian data. *Multivariate Behavioral Research, 39*, 69–98.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman Hall.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), *Test design: developments in psychology and psychometrics.* Orlando FL: Academic Press.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495–515.

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

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93–103.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

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-Verlag.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget.* New York: Litton Educational Publishing.

Flavell, J. H. (1970). Stage-related properties of cognitive development. *Cognitive Psychology, 2*, 421–453.

Flavell, J. H. (1985). *Cognitive development.* Englewood Cliffs, N J: Prentice-Hall, Inc.

Formann, A. K. (2001). Misspecifying latent class models by mixture binomials. *British Journal of Mathematical and Statistical Psychology, 54*, 279–291.

Formann, A. K. (2003). Modeling data from water-level tasks: A test theoretical analysis. *Perceptual motor skills, 96*, 1153–1172.

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics, 12*, 369–381.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.* Cambridge: Cambridge University press.

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

Harris, P. L., & Bassett, E. (1975). Transitive inferences by 4-year-old children. *Developmental Review, 11*, 875–876.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20*, 1–14.

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

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331–348.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models, foundations, recent developments, and applications* (pp. 53–68). New York: Springer-Verlag.

Hosenfield, B. (2003). *The development of analogical reasoning in middle childhood.* Unpublished doctoral dissertation, Universiteit Leiden, The Netherlands.

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.

Hosenfield, B., Van der Maas, H. L. J., & Van den Boom, D. C. (1997). Detecting bimodality in the analogical reasoning performance of elementary schoolchildren. *International Journal of Behavioral Development, 20*, 529–547.

Huttenlocher, J. (1968). Constructing spatial images. *Psychological Review, 75*, 550-560.

Huttenlocher, J., & Higgens, E. T. (1971). Adjectives, comparatives and syllogisms. *Psychological Review, 78*, 487–504.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Jansen, B. R. J. (2001). *Development of reasoning on the balance scale task.* Unpublished doctoral dissertation, Universiteit van Amsterdam, The Netherlands.

Jansen, B. R. J., & Van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321–357.

Jansen, B. R. J., & Van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383–416.

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

Kail, R., & Bisanz, J. (1992). The information-processing perspective on cognitive development in childhood and adolescene. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 229–260). Cambridge: Cambridge University Press.

Kallio, K. D. (1982). Developmental change on a five-term transitivity inference. *Journal of Experimental Child Psychology, 33*, 142–164.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scores items. *Psychometrika, 59*, 149–176.

Kessen, W. (1960). Research design in the study of developmental problems. In P. H. Mussen (Ed.), *Handbook of research methods in child development* (pp. 257–262). New York: Wiley.

Lazerfield, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics. essays in honor of* H*arold* H*otelling* (pp. 278–292). Stanford: Stanford University Press.

Liben, E. H., & Posnansky, C. J. (1977). Inference on inference. The effects of age, transitive ability, memory load, and lexical factors. *Child Development, 48*, 490–497.

Lohaus, A., & Kessler, T. (1996). Zwischen Lösungsverhalen und verbalisiertem Verständnis des Lösungsprinzips bei der Wasserspiegelaufgabe. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 28*, 316–335.

Lohaus, A., Kessler, T., Thomas, H., & Gediga, G. (1994). Individuelle Unterschiede bei räumlichen Fähigkeiten im Kindesalter. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 26*, 373–390.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology, 31*, 223–264.

Marx, M. H. (1985a). More retrospective reports on event-frequency judgments: shift from multiple traces to strength factor with age. *Bulletin of the Psychonomic Society, 24*, 183–185.

Marx, M. H. (1985b). Retrospectives reports on frequency judgments. *Bulletin of the Psychonomic Society, 23*, 309–310.

McCutcheon, A. L. (1987). *Latent class analysis.* California: Sage.

McDonald, R. P. (1985). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.

McLachlan, G. J., & Peel, P. (2000). *Finite mixture models.* New York: John Wiley and Sons, Inc.

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

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

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

Molenaar, I. W., & Sijtsma, K. (2000). *User's manual* MSP*5 for windows. A program for* M*okken* S*cale analysis for* P*olytomous 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.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41–68.

Nandakumar, R., Yu, F., Li, H. H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement*, *22*, 99–115.

Perner, J., & Aebi, J. (1985). Feedback-dependent encoding of length series. *British Journal of Developmental Psychology*, *3*, 133–141.

Perner, J., & Mansbridge, D. G. (1983). Developmental differences in encoding. *Child Development*, *54*, 710–719.

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.* Paris: Collin.

Piaget, J. (1947). *La psychologie de l'intelligence.* Paris: Collin.

Piaget, J. (1949). Le problème neurologique de l'intériorization des actions en opérations réversibles. *Archives de Psychologie*, *32*, 241–258.

Piaget, J. (1961). *Les méchanicismes perceptives.* Paris: Presses Universitaires de France.

Piaget, J., & Inhelder, B. (1941). *Le développement des quantités chez l'enfant.* Neuchatel: Delachaux et Niestlé.

Piaget, J., Inhelder, B., & Szeminska, A. (1948). *La géométric spontanée de l'enfant.* Paris: Presses Universitaires de France.

Piaget, J., & Szeminska, A. (1941). *La genèse du nombre chez l'enfant.* Neuchatel: Delachaux et Niestlé.

Quinton, G., & Fellows, B. (1975). "perceptual" strategies in the solving of three-term series problems. *British Journal of Psychology*, *66*, 69–78.

Raijmakers, M. E. J., Jansen, B. R. J., & Van der Maas, H. L. J. (2004). Rules and development in triad classification task performance. *Developmental Review, 24*, 289–321.

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

Reyna, V. F. (1992). Reasoning, remembering, and their relationship: Social, cognitive, and developmental issues. In M. L. Howe, C. J. Brainerd, & V. F. Reyna (Eds.), *Development of long-term retention* (pp. 103–132). New York: Springer-Verlag.

Reyna, V. F. (1996). Conceptions of memory development with implications for reasoning and decision making. *Annals of Child Development, 12*, 87–118.

Reyna, V. F., & Brainerd, C. J. (1990). Fuzzy processing in transitivity development. *Annals of Operations Research, 23*, 37–63.

Riley, C. A., & Trabasso, T. (1974). Comparatives, logical structures, and encoding in a transitive inference task. *Journal of Experimental Child Psychology, 17*, 187–203.

Roussos, L. A., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35*, 1–30.

Scheiblechner, H. (1972). Das Lernen und Lösen complexer Denkaufgaben (learning and solving complex thought problems). *Zeitschrift für experimenteler und angewandte Psychologie, 19*, 481–520.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*, 481-520.

Siegler, R. S. (1991). *Children's thinking, second edition.* New Jersey: Prentice-Hall,Inc.

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

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

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. (1963). Development of concrete transitivity of length in children. *Child Development, 34*, 389–405.

Smedslund, J. (1965). The development of transitivity of length: a comment on Braine's reply. *Child Development, 36*, 577–580.

Smedslund, J. (1969). Psychological diagnostics. *Psychological Bulletin, 71*, 237–248.

Sternberg, R. J. (1980a). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology, 109*, 119–159.

Sternberg, R. J. (1980b). The development of linear syllogistic reasoning. *Journal of Experimental Child Psychology, 29*, 340–356.

Sternberg, R. J., & Weil, E. M. (1980). An aptitude × strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology, 72*, 226–239.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Stevenson, H. W. (1972). *Children's learning.* New York: Appleton-Century-Crofts.

Stout, W. (1993). DIMTEST. Urbana-Champaign: The William Stout institute for Measurement.

Stout, W. (1996). DETECT. Urbana-Champaign: The William Stout institute for Measurement.

Stout, W., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357–375). New York: Springer.

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

Thayer, E. S., & Collyer, C. E. (1978). The development of transitive inference: a review of recent approaches. *Psychological Bulletin, 85*, 327–1343.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397–412.

Thomas, H. (1989). A binomial mixture model for classification performance: A commentary on Waxman, Chambers, Yntema and Gelman (1989). *Journal of Experimental Psychology, 48*, 423–430.

Thomas, H. (1994). Mixture decomposition when the components are of unknown form. In A. Von Eye & C. C. Clogg (Eds.), *Latent variable analysis* (pp. 313–328). Thousand Oaks, CA: Sage.

Thomas, H., & Hettmansperger, T. P. (2001). Modelling change in cognitive understanding with finite mixtures. *Applied Statistics, 50*, 435–448.

Thomas, H., & Lohaus, A. (1993). Modeling growth and individual differences in spatial tasks. *Monographs for the Society for Research in Child Development, 58*, (9, serial No. 237).

Thomas, H., Lohaus, A., & Kessler, T. (1999). Stability and change in longitudinal water-level task performance. *Developmental Psychology, 35*, 1024–1037.

Thomas, H., & Turner, G. F. W. (1991). Individual differences and development in water-level task performance. *Journal of Experimental Child Psychology, 51*, 171–194.

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.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study on test dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3–24.

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

Van der Maas, H. L. J. (1998). The dynamical and statistical properties of cognitive strategies: Relations between strategies, attractors, and latent classes. In K. M. Newell & P. C. M. Molenaar (Eds.), *Applications of nonlinear dynamics to developmental process modeling* (pp. 161–176). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van der Maas, H. L. J., & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99,* 395–417.

Van Geert, P. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotski, and beyond. *Psychological Review, 4,* 634–677.

Van Maanen, L., Been, P., & Sijtsma, K. (1989). The lineair logistic test model and heterogeneity of cognitive strategies. In E. E. C. I. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–287). Berlin: Springer-Verlag.

Van Onna, M. J. H. (2002). Bayesian estimation and model selection on ordered latent class models for polytomous items. *Psychometrika, 67,* 519-538.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33,* 213–239.

Vermunt, J. K. (2004). An EM algorithm for the estimation of parametric hierarchical nonlinear models. *Statistica Neerlandica, 58,* 220–233.

Vermunt, J. K., & Magidson, J. (2000). *Latent Gold: user's guide.* Belmont: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2003). *Latent Gold 3.0.* Belmont, M A: Statistical Innovations Inc.

Verweij, A. C. (1994). *Scaling transitive inference in 7-12 year old children.* Unpublished doctoral dissertation, Vrije Universiteit Amsterdam, The Netherlands.

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.

Wedel, M., & DeSarbo, W. A. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (p. 352-388). Cambridge, MA: Blackwell.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika, 38*, 330–336.

Wohlwill, J. F. (1973). *The study of behavioral development.* New York: Academic Press.

Wright, B. C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. *Developmental Review, 21*, 375–422.

Youniss, J., & Dennison, A. (1971). Figurative and operative aspects of children's inference. *Child Development, 42*, 1837–1847.

Youniss, J., & Furth, H. G. (1973). Reasoning and Piaget. *Nature, 244*, 314–316.

Youniss, J., & Murray, J. P. (1970). Transitive inference with nontransitive solutions controlled. *Developmental Psychology, 2*, 169–175.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129–152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–349.

Summary

Transitive reasoning is an important construct in developmental psychology. According to Piaget operational reasoning is required to infer a transitive relationship. This operational reasoning is characteristic of the concrete operational stage, one of the four stages in Piaget's theory.

In a transitive reasoning task the unknown relationship between two elements (transitive relationship) can be inferred from their known relationships (premises) with a third element. According to Piaget children have to be capable to understand and apply rules of logic to infer transitive relationships.

Piaget's theory about transitive reasoning evoked much of discussion and research was initially focussed at the age of emergence of transitive reasoning. Later on, attention shifted towards the underlying processes involved in transitive reasoning. Researchers from different theoretical backgrounds used different definitions and operationalisations of the construct leading to different conclusions about the processes involved. The most important purpose of this dissertation was to disentangle the cognitive processes involved in transitive reasoning and to compare three leading theories.

Chapter 1 describes the construction of a transitive reasoning test containing 16 transitive reasoning tasks that differed with respect to the presentation form of the premises, the content of the task and the kind of relationship between the objects used in the tasks. Previous research showed that these task characteristics influence the difficulty of a task. The test was administered to 615 children ranging in age from 6 to 13 years old. 15 of the 16 tasks formed a reliable Mokken scale on which the children could be ordered reliably according to their number-correct score.

In chapter 2 an empirical study is described in which the three leading theories were compared with respect to dimensionality of the construct of transitive reasoning and the influence of task characteristics on the difficulty level of the task. Moreover, it was investigated whether the correct / incorrect explanations the children gave after responding to the task led to

more valid information about transitive reasoning ability than the correct / incorrect responses. Different nonparametric item response techniques were used to determine the dimensionality of the data. The ability could be described by one dimension when the correct / incorrect explanations were used while at least three dimensions were required when the correct / incorrect responses were used. It was concluded that the correct / incorrect explanations gave more unambiguous and accurate information about the transitive reasoning ability than the correct / incorrect responses.

Moreover, the results showed that the dimension could better be interpreted by information processing theory and fuzzy trace theory than by Piaget's theory. The distinction between functional and operational thinking (typical of Piaget's theory) was not reflected by the results. The difficulty level of the tasks was especially determined by the degree to which the premisse information could be reduced into a more patternlike form (typical of information processing theory and fuzzy trace theory).

In chapter 3 it was investigated whether the development of transitive reasoning is continuous or discontinuous. First, a number of aspects involved in studying discontinuity were discussed. Second, two latent class models were compared. The results showed that the binomial mixture model, which is a common model to study discontinuity in cross-sectional research, fitted worse than the latent class factor model. Both models showed that the development of transitive reasoning was discontinuous. At least two classes could be distinguished in the ability of transitive reasoning which could be interpreted by fuzzy trace theory.

In chapter 4 the relationships between age, strategy use and task characteristics were investigated. A latent class regression model was used to describe the influence of task characteristics on strategy use. Five latent classes were distinguished in which the influence of task characteristics on strategy use differed. Young children in particular used irrelevant details of the task to infer (mostly incorrectly) the transitive relationship. Task characteristics had little influence on strategy use. For elder children task characteristics influenced the strategy use considerably.

In chapter 5 fuzzy trace theory was used to described the performance of children on three kinds of transitive reasoning tasks in detail. The tasks differed with respect to the ordering of the objects and the presentation of the premises. According to fuzzy trace theory information is processed simultaneously at multiple levels. Fuzzy trace theory distinguishes a verbatim ability and a fuzzy ability. According to the theory, only fuzzy ability is needed to infer transitive relationships. Using this theory the performance of children on different kinds of task given their verbatim and fuzzy ability level could be predicted well.

A multi-level latent class model was used to determine whether the theoretical model fitted the empirical data. The results showed that the theoretical model fitted well. Both the verbatim and fuzzy ability were reflected in the data structure and the predicted performance agreed with the estimated performance to a large extent.

In the epilogue it was concluded that the development of measurement methods and techniques over the past decades enabled us to study developmental issues in a differential way. According to Wohlwill (1973) a differential approach forms the essence of developmental psychology. However, at the time he wrote his book no adequate statistical methods were available.

The stages formulated in Piaget's theory were not supported by our empirical observations on transitive reasoning. Moreover, it was found that cognitive development was neither characterized by an increase in the amount of information processed as assumed by information processing theory. The results of this thesis in particular showed that cognitive development is characterized by an increasing ability to process information at different levels and to retrieve information that adequately matches task requirements.

Samenvatting (Summary in Dutch)

Transitief redeneren is een belangrijk begrip in de ontwikkelingspsychologie. Volgens Piaget is operationeel redeneren nodig om een transitieve relatie af te kunnen leiden. Operationeel redeneren is kenmerkend voor het concreet-operationele stadium, één van de vier ontwikkelingsstadia uit Piaget's theorie.

In een transitieve redeneertaak kan een onbekende relatie (transitieve relatie) tussen twee elementen worden afgeleid uit twee bekende relaties (premissen) tussen deze twee elementen en een derde element. Volgens Piaget moeten kinderen in staat zijn om logische regels te begrijpen en toe te passen om de onbekende relatie af te kunnen leiden.

Naar aanleiding van Piaget's theorie over transitief redeneren is er veel onderzoek gedaan. Dit onderzoek was in eerste instantie vooral gericht op de vraag op welke leeftijd kinderen voor het eerst in staat zijn tot transitief redeneren. Later verschoof de aandacht vooral naar de onderliggende processen betrokken bij transitief redeneren. Onderzoekers uit verschillende onderzoekstradities gebruikten verschillende definities van transitief redeneren en verschillende operationaliseringen van het begrip in transitieve redeneertaken. Het belangrijkste doel van deze dissertatie was om de cognitieve processen die een rol spelen bij het transitief redeneren in kaart te brengen en op deze manier theorieën over transitief redeneren op een aantal aspecten met elkaar te vergelijken.

Hoofdstuk 1 beschrijft de constructie van een transitieve redeneertest met zestien transitieve redeneertaken die verschillen wat betreft de aanbieding van de premissen, de context, en het soort van relatie tussen de premissen. Eerder onderzoek heeft aangetoond dat deze taakkenmerken de prestatie sterk beïnvloeden. De test werd voorgelegd aan 615 basisschool leerlingen van groep vier tot en met groep acht. Het dubbele monotonie model van Mokken paste op vijftien van de zestien taken. Hieruit kon geconcludeerd worden dat de taken betrouwbaar geordend konden worden volgens de totaalscore van de test en dat de taken een invariante ordening hadden.

In Hoofdstuk 2 wordt een empirische studie beschreven waarin drie

theorieën werden vergeleken op de dimensionaliteit van het construct transitief redeneren en de invloed van taakkenmerken op de moeilijkheid van de taak. Ook werd onderzocht in hoeverre de juist / onjuist verklaringen die kinderen gaven na het beantwoorden van een taak validere informatie opleverden dan alleen de juist / onjuist antwoorden. Verschillende non-parametrische item response technieken werden gebruikt om de dimensionaliteit van de test te beoordelen. Het bleek dat het theoretisch construct redelijk goed met één dimensie beschreven kon worden wanneer de juist / onjuist verklaringen werden gebruikt. Om het construct met behulp van de juist / onjuist antwoorden te beschrijven waren drie dimensies nodig. Hieruit kon geconcludeerd worden dat de juist / onjuist verklaringen eenduidigere en accuratere informatie opleverden over transitief redeneren dan de juist / onjuist antwoorden.

Ook bleek dat de gevonden dimensie beter geïnterpreteerd kon worden volgens de informatie-verwerkings theorie en de fuzzy-trace theorie dan volgens Piaget's theorie. Een onderscheid tussen functioneel en operationeel redeneren (zoals in Piaget's theorie) werd niet gevonden. De moeilijkheid van de taken bleek vooral af te hangen van de mate waarin de gedetailleerde informatie in taken gereduceerd kon worden tot patrooninformatie (zoals in informatie-verwerkings theorie en fuzzy-trace theorie).

In Hoofdstuk 3 werd de vraag onderzocht of de ontwikkeling van het transitief redeneren continu of discontinu verloopt. Eerst werden verschillende onderzoekskwesties die een rol spelen bij het meten van discontinuïteit besproken en vervolgens werden twee latente klassen modellen met elkaar vergeleken. Het bleek dat het binomiale mixture model, dat doorgaans wordt gebruikt om discontinuïteit bij cross-sectioneel onderzoek vast te stellen, slechter paste dan het latent klassen factor model. Beide modellen lieten zien dat de ontwikkeling van transitief redeneren discontinu was. De vaardigheid van het transitief redeneren bleek op z'n minst uit twee latent klassen te bestaan die geïnterpreteerd konden worden met behulp van fuzzy trace theorie.

In Hoofdstuk 4 werd de relatie onderzocht tussen leeftijd, strategiegebruik en taakkenmerken. Een latente klassen regressie model werd gebruikt om de invloed van taakkenmerken op strategiegebruik te beschrijven. Er werden vijf latente klassen onderscheiden waarbij de relatie tussen strategiegebruik en taakkenmerk verschilden. Het bleek dat jonge kinderen vaak met behulp van irrelevante details de transitieve relatie probeerden af te leiden. De taakkenmerken hadden nauwelijks invloed op het strategiegebruik. Bij oudere kinderen hadden taakkenmerken daarentegen een belangrijke invloed op het strategiegebruik.

In Hoofdstuk 5 tenslotte werd fuzzy-trace theorie gebruikt om een gedetailleerde beschrijving te geven van de prestaties van kinderen op drie transitieve redeneertaken. Deze taken verschilden wat betreft de ordening van de objecten in de taak en de presentatie van de objecten. Volgens de fuzzy-trace theorie wordt informatie op tal van niveaus tegelijkertijd verwerkt. Fuzzy-trace theory onderscheidt een vaardigheid in het verwerken van gedetailleerde informatie en een vaardigheid in het verwerken van patrooninformatie. Volgens de theorie speelt bij het afleiden van transitieve relaties vooral het gebruik van patrooninformatie een rol. Vanuit het theoretische model konden voorspellingen worden gedaan over de prestaties van kinderen met een bepaald detailvaardigheidsniveau en patroonvaardigheidsniveau op verschillende taken. Een multi-level latente klassen model werd gebruikt om de te bepalen of de voorspellingen op basis van het theoretisch model werden teruggevonden in de geobserveerde data. De resultaten lieten zien dat het theoretische model goed paste bij de empirische data; de twee soorten vaardigheden werden teruggevonden en de voorspelde prestatie van kinderen kwam goed overeen met de geobserveerde prestatie.

In de Epiloog werd geconcludeerd dat met de nieuw ontwikkelde meetmethoden en analysetechnieken ontwikkelingsvraagstukken op differentiële wijze kunnen worden beantwoord. Volgens Wohlwill (1973) vormt de differentiële benadering de essentie van de ontwikkelingspsychologie maar ontbraken in de tijd dat hij zijn boek schreef adequate statistische methoden.

Daarnaast werd geconcludeerd dat de stadia geformuleerd in Piaget's theorie niet worden teruggevonden bij het transitief redeneren. Ook werd niet gevonden dat de cognitieve ontwikkeling wordt gekenmerkt door een steeds completere verwerking van informatie zoals wordt aangenomen door de informatie-verwerkings theorie. Uit de resultaten beschreven in dit proefschrift blijkt vooral dat cognitieve ontwikkeling wordt gekenmerkt door een groeiende vaardigheid om informatie te verwerken op verschillende niveaus en weer te gebruiken op een niveau dat optimaal aansluit bij hetgeen de taak vereist.