Latent Class Modeling of Phases in the Development of Transitive Reasoning

Samantha Bouwmeester
Erasmus University, Rotterdam, The Netherlands

Klaas Sijtsma
Tilburg University, Tilburg, The Netherlands

Fuzzy trace theory posits that during development the use of verbatim information for solving transitive relationships shifts to the use of gist information. In cognitive developmental research that uses a cross-sectional design, the binomial mixture model is often used to identify such shifts. Because the binomial mixture model assumes equal task difficulty and uses the number of correctly solved tasks for data analysis, it may be too restrictive and the more flexible latent class model is adopted as an alternative. This model allows varying task difficulty and uses the pattern of task scores as input for data analysis. The binomial mixture model and the latent class model are compared theoretically, and applied to transitive reasoning test data obtained from a cross-sectional sample of 615 children. The latent class model is found to be more appropriate for identifying multiple phases. Three phases are distinguished which can be interpreted well by means of fuzzy trace theory. These phases do not encompass fixed age periods.

INTRODUCTION

Since Piaget formulated his developmental stage theory, the study of discontinuity in cognitive development has become important in psychological and methodological studies. More recently, influenced by catastrophe theory (Thom & Fowler, 1975), discontinuity has been called phase transition, a terminology adopted here and explained in more detail in the next section. Phase transition

Correspondence concerning this article should be addressed to: Samantha Bouwmeester, Institute for Psychology, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: bouwmeester@fsw.eur.nl

has been studied by many researchers under this or different names for several cognitive developmental abilities (Brainerd, 1978, 1993; Dolan, Jansen, & Van der Maas, 2004; Flavell, 1970; Formann, 2003; Hosenfeld, Van der Maas, & Van den Boom, 1997; Jansen & Van der Maas, 1997, 2001; Thomas, 1989; Thomas & Lohaus, 1993; Thomas, Lohaus, & Kessler, 1999; Van Geert, 1998). This study focused on the detection of phases in the development of transitive reasoning ability. An individual is capable of transitive reasoning if (s)he is able to infer an ordinal relationship between two objects from other ordinal relationships in which these objects are involved. For example, if one knows that stick A is longer than stick B, and that B is longer than C, the correct inference that A is longer than C gives evidence of transitive reasoning ability.

In cognitive developmental psychology, phase transition is manifest in two or more states. There are two definitions of such states, in cross-sectional designs often called modes. First, Piaget defined a state as a general cognitive structure or a developmental stage (e.g., Chapman, 1988; Flavell, 1985; Piaget, 1947). He distinguished different phases in cognitive development by assuming that knowledge acquisition develops via cognitive structures which differ qualitatively (e.g., Case, 1992; Flavell, 1963). The transition of a cognitive structure into a different cognitive structure is an intuitively appealing example of what we would call phase transition in the development of knowledge. However, it has been found difficult, or even impossible, to translate such general, abstract cognitive structures into an empirical setting and investigate them systematically (e.g., Brainerd, 1978; Flavell, 1970, 1985).

Second, other researchers defined a phase as a specific rule or strategy, which is part of a particular ability. In this context, Flavell (in Brainerd, 1978) 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", as is typical of the Piagetian approach. Instead he encouraged the definition of phases as rules or strategies in a particular domain or subdomain, and to interpret the transition from one phase to a different phase as discontinuity in the development of a specific ability. For example, Jansen and Van der Maas (1997, see also Jansen & Van der Maas, 2001, 2002) distinguished four rules in solving the balance scale task. In this context, each phase represents the use of a specific rule, and subsequent modes represent increasingly better or more adequate rules (for other examples, see Hosenfeld et al., 1997; Siegler & Chen, 2002; Thomas et al., 1999; Van der Maas & Molenaar, 1992).

Cross-Sectional Versus Longitudinal Design

Before we discuss phase transition in development of transitive reasoning, it may be useful to say something about the abruptness (e.g., manifest as sudden

jumps) with which transitions from one phase to the next phase occur and, in particular, the design necessary to investigate this abruptness.

Catastrophe theory (Thom & Fowler, 1975) offers an opportunity to model developmental phase transition in cognitive development using longitudinal designs. Development can be monitored by means of repeated measurements during a particular time interval. Longitudinal designs (see, e.g., Van Geert, 1998) are appropriate when the aim is to describe the transition from one phase to a subsequent phase in much detail; then sudden jumps, if any, may become manifest. Markov chain models (e.g., Brainerd, 1979) and catastrophe models (Van der Maas & Molenaar, 1992) have been used to study such discontinuity.

Several researchers studied phases in performance on Piagetian tasks using a cross-sectional design. For example, Thomas (1989) and Raijmakers, Jansen, and Van der Maas (2004) studied phases in classification performance; Thomas and Turner (1991), Thomas and Lohaus (1993), Thomas et al. (1999) and Formann (2003) studied phases in performance on the water-level task; Hosenfeld et al. (1997) studied phases in analogical reasoning; and Van der Maas (1998) and Jansen and Van der Maas (2002, 2001) studied phases in performance on balance scale tasks. Cross-sectional designs do not provide information about the transition from one phase to a different phase and, as a result, hypotheses concerning this transition—for example, is it sudden or gradual?—cannot be tested. The presence of multiple phases in data collected by means of cross-sectional designs is seen as an indicator of transitions in development (Jansen & Van der Maas, 2001; Van der Maas & Molenaar, 1992).

The choice of a cross-sectional design or a longitudinal design depends on the hypotheses to be tested and the resources available. In this study a cross-sectional design was used because our aim was to detect phases in the development of transitive reasoning and to interpret phases by means of fuzzy trace theory, to be explained next.

Phases in Development of Transitive Reasoning

Our approach is based on fuzzy trace theory (Brainerd & Kingma, 1984; Brainerd & Reyna, 2001, 2004; Reyna & Brainerd, 1995). Fuzzy trace theory predicts that different phases may be expected in performance on transitive reasoning tasks. According to this theory, young children tend to use verbatim information to solve problems (i.e., they use literal observable information), while older children tend to use gist information (i.e., they use degraded, pattern-like information, only holding the gist). Thus, fuzzy trace theory assumes an ability for verbatim reasoning and another ability for gist reasoning. According to fuzzy trace theory these levels of information process in parallel.

Given particular levels of verbatim ability and gist ability, the theory assumes that children use both verbatim and gist traces but with different probabilities. These traces may be seen as kinds of strategies. For example, when children have low gist-ability level they tend to have high probability to use verbatim information to solve a transitive relationship. However, this is expected not to lead to a correct inference because the verbatim information used is not relevant for the relationships between the objects (e.g., the child has correctly remembered the colors of the objects or that the objects were sticks and not triangles). When children pass a threshold on the gist ability "scale" they have higher probability to use pattern information that enables the correct inference of a transitive relationship (e.g., they may use the strategy: "sticks become shorter from left to right; therefore, stick A is longer than stick C"). When they reach a high gist-ability level they have a high probability to use pattern information also for transitive reasoning tasks in which it is difficult to recognize the ordering in the information presented (e.g., let Y represent length; then a difficult task format is, for example, $Y_A = Y_B > Y_C = Y_D$).

Bouwmeester and Sijtsma (2004) and Bouwmeester, Sijtsma, and Vermunt (2004) showed that young children give verbal explanations of their performance that indicate predominant use of verbatim strategies. Older children give explanations indicating predominant use of gist strategies. Bouwmeester et al. (2004) showed that task characteristics influenced the use of particular strategies more for older than for younger children.

Thus, fuzzy trace theory predicts that during development children shift from verbatim processing to gist processing. This shift is assumed to be reflected in the strategies that are used predominantly to solve transitive reasoning tasks. Thus, the shift actually is a transition from one distribution of strategies stemming from verbatim traces—provided multiple strategies are used—to another distribution of strategies stemming from gist traces.

At least three phases are expected in the development of transitive reasoning. In the first, low-performance phase, children tend to solve transitive reasoning tasks using verbatim information, which leads to incorrect answers to all the types of tasks used. In the second, intermediate-performance phase, children have passed a threshold on the gist ability "scale" and tend to use pattern information. Children in this phase are expected to correctly infer the transitive relationship for tasks in which the ordering of the objects is obvious (e.g., $Y_A > Y_B > Y_C$; all three objects are presented simultaneously) but not to be able to correctly infer relationships in tasks in which the ordering is less obvious (Brainerd & Reyna, 1990; Verweij, 1994). Children in the high-performance phase are expected to have high probability to correctly infer the transitive relationship for all tasks.

Following fuzzy trace theory, our approach defines a phase by means of a particular probability distribution of using a range of available strategies. Thus,

children who are in a particular phase have a probability typical of that phase of using strategy S_1, another probability of using strategy S_2, and so on for other available strategies. Children who are in another phase have a different probability distribution of using these strategies, and so on for a possible third phase, a fourth, and so on. This study is aimed at identifying these phases from cross-sectional data and establishing phase transition in the development of transitive reasoning. Because this is a cross-sectional study, no information is available on whether transitions are sudden or gradual.

Finally, unlike fuzzy trace theory, information processing theory (Bryant & Trabasso, 1971; Riley & Trabasso, 1974; Trabasso, Riley, & Wilson, 1975) does not assume different phases but, based on Bryant's and Trabasso's (1971) linear ordering theory, hypothesizes that transitive reasoning can be explained by the ability to remember the premise information. This ability develops as a process of cumulative learning of stimulus-response relationships and forming internal representations of these relationships. This process is unidimensional and develops quantitatively without qualitatively different phases and transitions from one phase to the other. We will also investigate this possibility as an alternative to fuzzy trace theory.

Phases and Age

Transitive reasoning ability increases with age (Brainerd & Kingma, 1984, 1985; Reyna, 1992; Reyna & Brainerd, 1990) but rate of development need not be the same for each child. Also, different children may learn to use different strategies for transitive reasoning. These observations agree with Wohlwill (1973) who argued that chronological age is not a useful variable in statements of functional relationships with behavior. Thus, it may be appropriate to ignore fixed-age groups and study development by distinguishing groups on the basis of their probabilities to use particular strategies.

DATA ANALYSIS MODELS

The Binomial Mixture Model

The presence of multiple phases—sometimes called multimodality—in cross-sectional data is studied usually by means of the binomial mixture model (BMM). Thomas and Turner (1991), Thomas and Lohaus (1993), Thomas et al. (1999) and Thomas and Hettmansperger (2001) used the BMM to study phases in performance on the water-level task. This is a task in which children have to draw the water level in a glass which has a particular angle with a horizontal axis. The angle varies across tasks but it is assumed that this does not affect

task difficulty (Thomas & Hettmansperger, 2001). Hosenfeld et al. (1997) used the BMM to analyze phases in analogical reasoning.

Assume that a test consists of J tasks, which are either correctly (score 1) or incorrectly solved (score 0). Let random variables X_j ($j = 1, \ldots, J$) represent these task scores ($X_j = 0, 1$), and let $X_+ = \sum_{j=1}^{J} X_j$ be the number of correctly solved tasks, also called the number-correct score. Realizations of X_+ are denoted x_+. Assume c classes, which are indexed u ($u = 1, \ldots, c$). In the BMM context such classes are called components, but because we will compare the BMM to the latent class model (LCM) we will use the term "class" throughout. In each class it is assumed that tasks are solved with a constant success probability, θ_u, such that the tasks can be conceived of as J independent trials. Hence, X_+ follows a binomial distribution, denoted $Bin(X_+|J, \theta_u)$, which may be different across classes depending on θ_u. The BMM assumes that the frequency distribution of X_+ in the whole group consists of a mixture of c binomial distributions. Let π_u be the marginal probability of belonging to a particular class, with $\sum_{u=1}^{c} \pi_u = 1$, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_c)$; then the c-classes binomial mixture distribution is defined as

$$
\begin{aligned}
f(X_+ = x_+|J, \boldsymbol{\theta}) &= \sum_{u=1}^{c} \pi_u Bin(X_+ = x_+|J, \theta_u) \\
&= \sum_{u=1}^{c} \pi_u \binom{J}{x_+} \theta_u^{x_+}(1 - \theta_u)^{J-x_+},
\end{aligned}
\tag{1}
$$

for $x_+ = 0, \ldots, J$. Despite its frequent use for studying phases in cross-sectional data, the BMM has some serious drawbacks.

First, Hosenfeld et al. (1997) argued that a fitting BMM does not necessarily imply presence of multiple modes. According to these authors (ibid., p. 532), and adopting their terminology, "the presence of bi- (or multi-) modality can only be concluded if the model plot (i.e., the estimated overall frequency distribution of X_+; *the authors*) displays two clearly separable peaks." The number-correct scores between these two peaks have relatively low frequencies, and are therefore assumed to give evidence of what is called an inaccessible region. According to catastrophe theory the presence of an inaccessible region is another indicator of phase transition (Jansen & Van der Maas, 2001).

In real-data analysis, at least three types of problems may obscure the finding of peaks and gaps that result from the use of different cognitive strategies. First, if the binomial probabilities θ_u differ only little between classes the resulting binomial distributions overlap to a high degree and peaks nearly coincide. Second, if a small number of tasks is used the binomial distributions show so little detail—meaning that they are very coarse—that their peaks cannot be distinguished unless they are far apart. Third, there is no compelling reason

why different cognitive strategies would always lead to different binomial distributions that are located far apart. In fact, a mixture of largely overlapping distributions may result and the estimated overall frequency distribution does not show clearly visible gaps or peaks (Thomas & Lohaus, 1993). Thus, taking observable peaks and gaps as evidence for phase transition may often lead to the wrong conclusion. Alternatively, phase transition due to the use of different rules and strategies with different probabilities may be detected much better.

Second, the developmental phenomenon of phase transition in itself does not imply a particular shape of the frequency distribution of X_+. Thus, a statistical model that assumes a mix of binomial distributions may be too restrictive in several studies of phase transition. Whether this is true for a particular ability is an empirical question, but if the BMM is rejected as an explanatory model other, more flexible models may be called for.

Third, the assumption of a binomial distribution for X_+ implies that within a class the J tasks have the same difficulty level. Between classes these task difficulties may vary, so that classes may be ordered by means of these difficulties. This ordering is the same for each task. Equal task difficulty within a class is known as task-homogeneity (Formann, 2001, 2003). Task-homogeneity may be realistic for the water-level task (although this may also be questioned; see Thomas & Hettmansperger, 2001), but unrealistic for other task types and, again, more flexible models may be useful.

Finally, the binomial mixture distribution implicitly assumes that a particular distribution of number-correct scores X_+ was produced by just one strategy. For this reason, Thomas et al. (1999, p. 1025) noted that the BMM "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, accepting a particular "distorted" outcome of data analysis—a possibly incomplete representation of the strategies used—only because the statistical method used cannot reveal the correct outcome seems odd. We argue next that the assumption that a particular number-correct score was produced by just one strategy is unrealistic in many practical situations (see also Formann, 2003).

Number-Correct Score Versus Individual Task Scores

Several researchers who used the BMM for studying the existence of phases in cognitive developmental abilities assume that a particular distribution of the number-correct score X_+ or only the mode of this distribution corresponds to a developmental phase (Hosenfeld et al., 1997; Lohaus & Kessler, 1996; Lohaus, Kessler, Thomas, & Gediga, 1994; Thomas, 1989, 1994; Thomas & Lohaus, 1993; Thomas et al., 1999; Thomas & Turner, 1991). This approach focusses

on the number-correct score, X_+, as the statistic of interest, but ignores task scores, X_j $(j = 1, \ldots, J)$, as sources of information about strategy use.

It is of interest to note that the number-correct score is a summary of a pattern of task scores, and that knowledge of X_+ alone cannot reconstruct the original pattern. Let $\mathbf{X} = (X_1, \ldots, X_J)$ be the vector of the J task score variables and let $\mathbf{x} = (x_1, \ldots, x_J)$ be a realization of $\mathbf{X}$. For example, for $J = 4$ let $\mathbf{x} = (1001)$ be a possible pattern of task scores, with $X_+ = 2$; then, only knowing that $X_+ = 2$ cannot reproduce vector $\mathbf{x}$. Thus, a pattern of task scores contains more information than a summary number-correct score X_+.

Other researchers focussed on the pattern in $\mathbf{X}$, and assumed that the use of a particular strategy implies a correct answer to task j with a probability that is typical of this strategy (e.g., Bouwmeester et al., 2004; Jansen & Van der Maas, 1997, 2002; Raijmakers et al., 2004; Van Maanen, Been, & Sijtsma, 1989). For example, the use of Strategy S_1 may imply a probability of 0.9 of having task j correct, and the use of Strategy S_2 a probability of 0.2. Then for Strategy S_1 a score of 1 on task j is the most likely outcome and for Strategy S_2 a score of 0. When applied to each of the J tasks, this approach implies that Strategy S_1 is characterized by a most likely vector of J task scores, and Strategy S_2 by a different most likely vector. Obviously, such distinctive features for strategy use may be lost when the summary score X_+ is the unit of analysis rather than the pattern of task scores $\mathbf{X}$.

The LCM has the pattern of task scores $\mathbf{X}$ as the unit of analysis, and identifies strategy groups by means of these patterns rather than the aggregate score X_+. Not only is there more relevant information about strategy use in these patterns and their probability structure, but the LCM also allows tasks to have varying difficulty within a class. This greater flexibility of the LCM is expected to lead to better fit results for our transitive reasoning data.

The Latent Class Model

The LCM is a mixture model (Lazarsfeld & Henry, 1968; see also Hagenaars & McCutcheon, 2002; McCutcheon, 1987), which allows heterogeneity in both individual performance and task difficulty (Formann, 2003). Classes have prevalence or class probabilities π_u ($\sum_{u=1}^{c} \pi_u = 1$). Each class has class-specific probabilities, $\theta_{1|u}, \ldots, \theta_{J|u}$, for correctly solving the tasks. For vector $\mathbf{X}$ and its realization $\mathbf{x}$ the LCM is defined as:

$$f(\mathbf{X} = \mathbf{x}) = \sum_{u=1}^{c} \pi_u \prod_{j=1}^{J} \theta_{j|u}^{x_j} (1 - \theta_{j|u})^{1-x_j}. \tag{2}$$

Thus, the LCM models the J task scores in vector $\mathbf{X}$ whereas the BMM models the summary number-correct score X_+ based on vector $\mathbf{X}$. For sample size N

and individuals indexed i $(i = 1, \ldots, N)$, the likelihood L of the LCM is

$$L = \prod_{i=1}^{N} f(\mathbf{X}_i = \mathbf{x}_i). \qquad (3)$$

Thus, each unique task-score pattern is seen to contribute to L.

Note that if we would restrict the LCM task-probabilities to be equal for all J tasks, that is, for $j = 1, \ldots, J$, let $\theta_{j|u} = \theta_u$, then the product on the right-hand side in Equation 2 reduces to $\theta_u^{x_+}(1 - \theta_u)^{J-x_+}$, so that

$$f(\mathbf{X} = \mathbf{x}) = \sum_{u=1}^{c} \pi_u \theta_u^{x_+}(1 - \theta_u)^{J-x_+}; \qquad (4)$$

also see Equation (1), right-hand side. This probability only depends on number-correct score X_+, not on the pattern of task scores in $\mathbf{X}$. Thus, $f(\mathbf{X} = \mathbf{x})$ has the same value for all patterns of task scores, $\mathbf{X}$, that contain the same number of 1 scores. For given $X_+ = x_+$ there are $\binom{J}{x_+}$ such patterns. Take the sum of these $\binom{J}{x_+}$ patterns across both sides of Equation (4). This yields $P(X_+ = x_+|J, \boldsymbol{\theta})$ on the left-hand side and $\binom{J}{x_+}$ times the probability on the right-hand side in Equation (4). The result is the BMM; see Equation (1). Let n_{x_+} be the number of individuals that have a number-correct score of x_+; then the likelihood L' of the BMM can be written

$$L' = \prod_{i=1}^{N} f(X_{+i} = x_{+i}|J, \boldsymbol{\theta}) = \prod_{x_+=0}^{J} n_{x_+} \sum_{u=1}^{c} \pi_u Bin(X_+ = x_+|J, \theta_u). \quad (5)$$

Note that this likelihood depends only on number-correct score X_+. For actual maximum likelihood estimation one uses the log-likelihood functions. Obviously, the log-likelihood functions of the BMM and the LCM with equal response probabilities across tasks (or replications) are equal, except for a constant that does not depend on the model parameters. Thus, for parameter estimation it does not matter which log-likelihood function is used: both procedures are equally stable. When data are sparse it is an advantage of the BMM that the log-likelihood is computed over the collapsed contingency table.

The two approaches are different with respect to goodness-of-fit testing. For the BMM the goodness-of-fit test is usually performed on the collapsed table (i.e., the table with totals only). This is a check whether a BMM with c classes can reproduce the observed distribution of the total number of successes. However, this goodness-of-fit test does not evaluate whether the BMM can describe the distribution of the full response patterns; that is, whether the simultaneous

assumption of c classes and homogeneous success probabilities across tasks holds. For that we need to use the restricted LCM approach.

A problem with the LCM is that the frequency table is sparse when there are more than a few tasks (in this research, $J = 15$). Because of these sparse frequency tables the asymptotic p-values associated with the χ^2 statistics often cannot be trusted. A safe strategy is to test models relative to other models, and this is what has been done in this study. Conditional on the number of latent classes, we may compare nested LCMs with an unrestricted LCM using a likelihood-ratio test, which yields a powerful test for the BMM "homogeneous success probabilities" assumption. Also see Dayton (1998) for a comparison of the BMM and the LCM.

So far the LCM has not been applied often to study phases in cognitive development (for exceptions, see Formann, 2001, 2003; Jansen & Van der Maas, 2001, 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 phases 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.

Information processing theory posits the development of transitive reasoning to be a continuous process of forming internal representations and remembering them. The Rasch model (e.g., Glas & Verhelst, 1995) was fitted to the data in an effort to investigate this possibility. A fitting model would lend credibility to information processing theory in the sense that a continuous latent variable would explain much of individual differences and leave less room for discrete latent classes to be interpreted as distinctive phases in development.

HYPOTHESES

In this study, we investigated four hypotheses. Prior to investigating these hypotheses, artifactual phases due to the test's distribution of task difficulty was ruled out as an alternative explanation for developmental phases. This was done by evaluating whether subsets of the proportions of correct solutions for the tasks piled up so that the distribution of task difficulties might have caused "phases" (i.e., method bias). The four hypotheses were:

1. *The LCM leads to a better explanation of phases than the BMM*. We compared BMM results with LCM results to assess how much fit was lost when restricting the item parameters to be equal within classes. Unlike the BMM, the LCM is able to account for the relationship between task difficulty and strategy use, and this enhances the opportunities for a better explanation.

2. *The development of transitive reasoning is characterized by different phases*. Different phases are reflected by the existence of two or more different probability distributions for strategy use. This was investigated by fitting several LCMs to the data to determine presence of phases and if that was found, how many phases had to be distinguished.
3. *In one phase verbatim trace information is used predominantly to solve the tasks and in another phase gist trace information is used predominantly. In this latter phase children perform well, in particular, on tasks in which the ordering of the objects is obvious. In a third phase, children perform well on all tasks*. We derived this hypothesis from fuzzy trace theory. Latent classes were interpreted by means of verbal-explanation data to determine whether fuzzy trace theory was suited for interpreting the phases. Alternatively, information processing theory predicts continuous development without distinct phases. This was investigated by means of the Rasch model.
4. *Strategy groups give a clearer description of phases than age groups*. We investigated the relationship between class membership and age expressed in months by including age as a covariate in the LCM analysis. Within age groups individual differences are expected in performance on transitive reasoning tasks, and clear-cut age periods characterized by particular performance cannot be distinguished.

METHOD

Sample

The pooled sample consisted of 615 children from Grade 2 through Grade 6 of six elementary schools in the Netherlands. Children were from middle class social-economic status families. Table 1 gives the number of children in six age groups, and in each age group the mean and standard deviation of age.

Material

Transitive reasoning ability was investigated by means of a computerized test containing 16 transitive reasoning tasks (Bouwmeester & Aalbers, 2002). A transitive reasoning task has the following structure. First, a pair of objects, say, sticks A and B, is shown to the child on a computer screen, and the child is asked to indicate and memorize which one is longer. Then, the pair (B, C) is shown and the same question is asked. Based on this "premise" information the child is then asked to infer the relationship between sticks A and C. A correct explanation gives evidence of transitive reasoning ability. The tasks differed on

TABLE 1
Number of Children (N), Mean Age (M) and
Standard Deviation (SD) per Age Group

Age Group[a]	N	M	SD
≤96	73	91.78	3.06
97–108	83	103.02	3.14
109–120	126	114.45	3.31
121–132	108	126.70	3.08
133–144	116	138.70	3.01
≥145	59	149.46	3.46

[a]Number of months.

three task characteristics (which are are summarized in Table 2). Bouwmeester and Sijtsma (2004) and Bouwmeester et al. (2004) (see also Verweij, Sijtsma, & Koops, 1999) showed that in particular the task characteristics "format" and "presentation form" influenced the task's difficulty level. The task characteristics had 4, 2, and 2 levels, defining $4 \times 2 \times 2 = 16$ tasks. See Figure 1 for examples of the tasks.

Procedure

The transitive reasoning test was an individual test administered in a quiet room in the school building. Before a 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

TABLE 2
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

1.format: $Y_A > Y_B > Y_C$, presentation form: successive, content: physical. The blue stick is longer than the red stick, the red stick is longer than the yellow stick. Therefore, the blue stick is longer than the yellow stick

2.format: $Y_A = Y_B = Y_C = Y_D$, presentation form: successive, content: physical. The orange stick has the same length as the yellow stick. The yellow stick has the same length as the red stick. Therefore, the orange stick has the same length as the red stick.

3.format: $Y_A < Y_B < Y_C < Y_D < Y_E$, presentation form: simultaneous, content: physical. The red stick is longer than the purple stick. The purple stick is longer than the blue stick. Therefore, the red stick is longer than the blue stick

4.format: $Y_A = Y_B < Y_C = Y_D$, presentation form: simultaneous, content: physical. The green stick has the same length as the purple stick. The purple stick is longer than the blue stick. Therefore, the green stick is longer than the blue stick.

5.format: $Y_A > Y_B > Y_C$, presentation form: simultaneous, content: verbal. The lion is older than the camel, the camel is older than the hippo. Therefore, the lion is older than the hippo.

6.format: $Y_A > Y_B > Y_C > Y_D > Y_E$, presentation form: simultaneous, content: verbal. The cow is older than the zebra, the zebra is older than the horse. Therefore, the horse is older than the cow.

7.format: $Y_A = Y_B = Y_C = Y_D$, presentation form: successive, content: verbal. The duck has the same age as the cock. The cock has the same age as the parrot. Therefore, the parrot and the duck have the same age.

8.format: $Y_A = Y_B > Y_C = Y_D$, presentation form: successive, content: verbal. The dog is older than the sheep. The sheep has the same age as the cat. Therefore the dog is older than the cat.
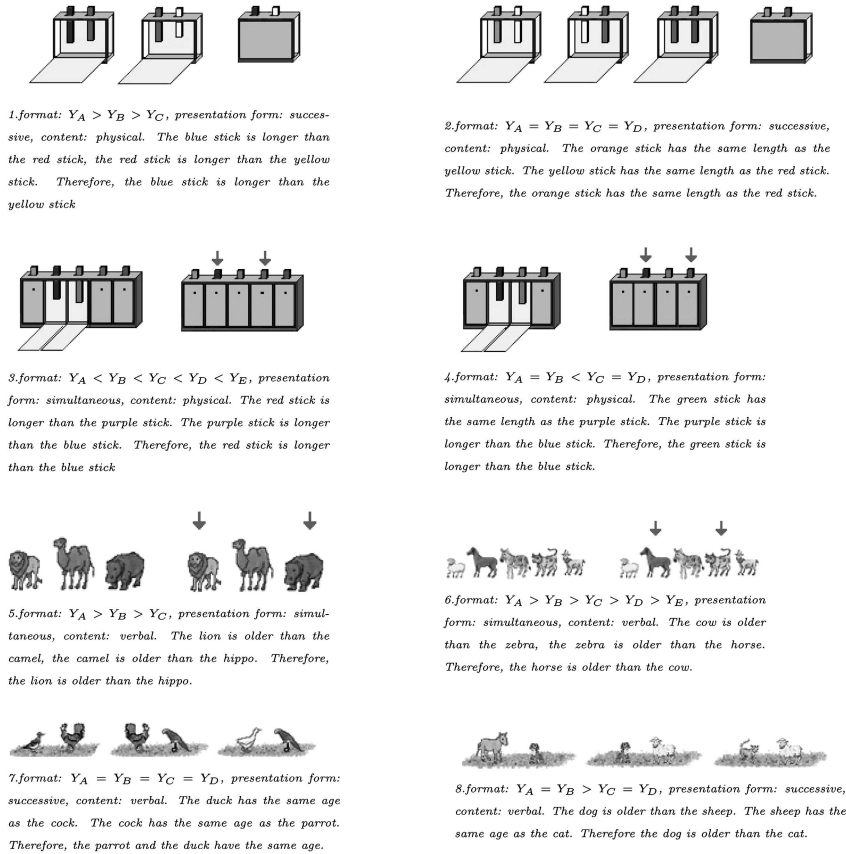
FIGURE 1 Eight examples of tasks used in the transitive reasoning test.

an hour, depending on the age of the child. For more details see Bouwmeester and Sijtsma (2004).

Task Scoring

For each task the response was automatically recorded by the computer. A verbal explanation of the response 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 considered incorrect. Previous research (Bouwmeester & Sijtsma, 2004) showed that the correct/incorrect-explanation scores reflected primarily transitive reasoning ability, whereas the

more "common" correct/incorrect scores also reflected irrelevant skills and abilities. Thus, the more valid correct/incorrect-explanation data served as input for data analysis.

Categorization of Verbal Explanations

The correct/incorrect-explanation scores were a dichotomization of an original categorical variable having 13 categories for different explanations. This original variable can be seen as the variable reflecting actual use of strategies and served to interpret the latent classes after it was recoded into four categories: (1) children used all the premise information in their explanation (literal premise information), or children gave 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.

RESULTS

Prior Analysis: Excluding Method Bias

The proportions correct (Table 3) showed much variation without obvious clustering, meaning that no distinct subsets of task difficulties could be identified which might explain the existence of phases. Thus, when phases are found in

TABLE 3
Proportion Correct of 15 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

subsequent analysis these will be attributed to a developmental process. Because one task (with the task characteristics: successive presentation, mixed format, and physical content) was incorrectly answered by 99% of the children, it was not used for further analysis.

Preliminary Information on BMM and LCM Analysis

The program Latent Gold 3.0 (Vermunt & Magidson, 2003) was used to estimate the parameters of the BMM and the LCM and to evaluate the fit of several models. The BMM was estimated as an LCM with equality restrictions on the task parameters within latent class u: $\theta_{1|u} = \theta_{2|u} = ... = \theta_{J|u} = \theta_u$. Then the same model can be estimated as when using the number-correct score (as outlined previously). The advantage was that the BMM and the LCM could be compared directly. For the LCM, $c - 1$ parameters were estimated for the latent class probabilities and $c \times 15$ parameters for the task parameters within latent classes, resulting in $(c - 1) + c \times 15$ parameters in total. For the BMM, $c - 1$ parameters were estimated for the latent class probabilities and c parameters for the tasks (i.e., one for each class), resulting in $(c - 1) + c$ parameters in total.

The likelihood-ratio chi-squared statistic L^2 gives an indication of model fit, and the Bayesian Information Criterion, abbreviated as BIC [defined as $-2 \times loglikelihood + \#parameters \times ln(N)$], serves as a selection criterion within the family of models fitted to the same data set. The BIC weights the fit and the parsimony [$\#parameters \times ln(N)$] of a model: The lower the BIC, the better the model in terms of parsimony and fit.

Investigating Hypotheses

Preliminary results.   The Rasch model was estimated and its fit evaluated by means of program Rasch Scaling Program (RSP; Glas & Ellis, 1994). This program provides fine-grained diagnostic information about model (mis-)fit to a degree that is not available from other software for the Rasch model (e.g., De Koning, Sijtsma, & Hamers, 2002; Glas & Verhelst, 1995). It was found that, in particular, local independence was violated ($\chi^2 = 1671$ and $df = 520$; based on a test by Van den Wollenberg, 1982). This means that conditioning on the sufficient statistic X_+ for the model's latent trait is not enough to explain the association between the J items; thus, some form of multidimensionality drove the responses to the items (e.g., see Sijtsma & Junker, 2006, for an extensive discussion of violation of the local independence assumption in item response models).

Two extensions of the Rasch model were evaluated, the two-class and the three-class mixed Rasch models (e.g., Rost & Von Davier, 1995). Like the BMM, these extensions account for a composite population and like the LCM they do

this allowing the items to vary in difficulty. Unlike the LCM, they assume equal discrimination power for all items, which again does not seem necessary *a priori* in many applications to cognitive data. For the Rasch model, Latent Gold 3.0 provided $L^2 = 3350.49$ and $BIC = 10140.81$; for the two-class mixed Rasch model, $L^2 = 1933.55$ and $BIC = 8826.62$; and for the three-class mixed Rasch model, $L^2 = 1719.21$ and $BIC = 8715.03$. Thus the mixed Rasch models had better fit than the Rasch model. The misfitting Rasch model renders the information processing theory interpretation of transitive reasoning unlikely, while the better fit of the mixed models lends support to the fuzzy trace theory interpretation in *Hypothesis 3*. This was further investigated by means of the BMM and the LCM.

Model comparison and detection of phases.   Table 4 shows the results of the model fit for the one-class through six-class models for both the LCM and the BMM. For both models, the decrease in L^2 was largest in going from the one-class model to the two-class model. The fit did not increase substantially when the number of classes increased from three to six. For both the LCM and the BMM, the BIC values show that the 3-class model was the most likely candidate for further interpretation given the trade-off between parsimony and fit. The L^2 and BIC values of the BMMs were much higher than those of the LCMs for all models, indicating that restricting the item parameters deteriorates the fit and masks possible phases 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. These results support *Hypothesis 1*.

The marginal probabilities (i.e., the π_us) for the three classes were .25, .39, and .36. For the LCM, for each of the three classes Figure 2 shows the success probabilities, $\theta_{j|u}$, for the 15 tasks. To facilitate readability of the plot, the order of the tasks is in accordance with their overall difficulty level, P_j (i.e.,

TABLE 4
LCM and BMM Fit Statistics

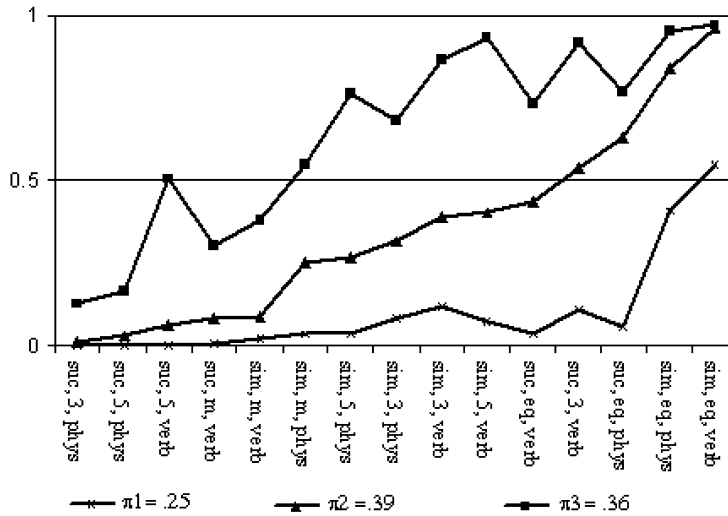|	LCM				BMM		
# Classes	L^2	BIC	# Par	# Classes	L^2	BIC	# Par
1	3350.491	10140.814	15	1	5680.532	12380.953	1
2	1933.549	8826.618	31	2	4849.405	11562.669	3
3	1719.218	8715.033	47	3	4785.039	11511.145	5
4	1627.415	8725.976	63	4	4777.327	11516.277	7
5	1553.946	8755.253	79	5	4772.144	11523.937	9
6	1504.750	8808.803	95	6	4772.086	11536.722	11

FIGURE 2   Marginal success probabilities ($\theta_{j|u}$) for tasks in each class. Notation along horizontal axis: sim: simultaneous; suc: successive; 3: $Y_A > Y_B > Y_C$; 5: $Y_A > Y_B > Y_C > Y_D > Y_E$; eq: $Y_A = Y_B = Y_C = Y_D$; mix: $Y_A = Y_B > Y_C = Y_D$; verb: verbal; phys: physical. Example: 'suc, 3, phys' means 'successive presentation of premises, 3 unequal objects, objects are compared on a physical property' (here: length).

proportions correct; see Table 3). The plot shows that allowing the tasks to differ in difficulty level resulted in highly varying success probabilities within latent classes. The first latent class differs clearly from classes 2 and 3. In this first latent class, except for the task denoted *"sim, eq, verb"* the success probabilities for the other tasks are smaller than .5, suggesting a low performance phase. In the second latent class, the probabilities are higher for several of the easier "equality" tasks (format $Y_A = Y_B = Y_C = Y_D$) and "inequality" tasks (format $Y_A > Y_B > Y_C$). In the third latent class, the probabilities are higher for most tasks except those which have a physical content and were successively presented. For these tasks it was very difficult to recognize the ordering. This result lends support to *Hypothesis 2*. Notice that consistent ordering results as in Figure 2 do not provide proof of unidimensionality of measurement; see, for example, Sijtsma and Meijer (2007).

Relationship with verbal explanation variables.   Figure 3 shows the distribution of the four explanation categories in each of the three latent classes. The first class is characterized by a relatively high percentage of explanations using external and visual information (21%) and absence of explanation (27%). Sometimes premise information was used, but often incorrectly. Children mostly
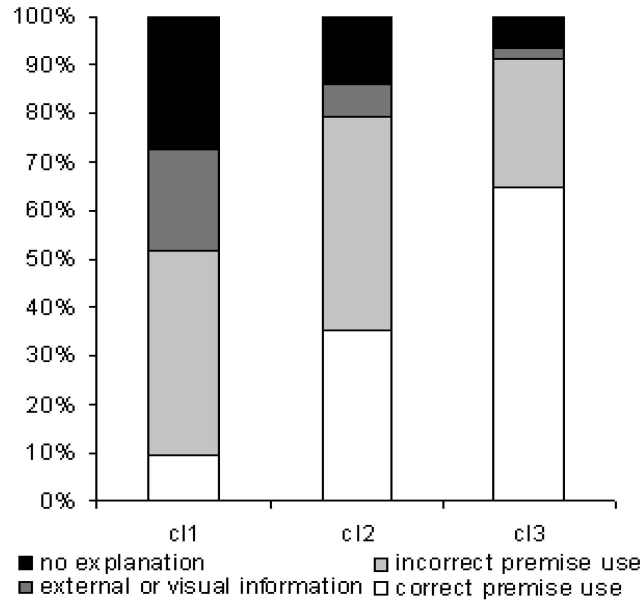
FIGURE 3   Distribution of the explanation categories for three classes.

used verbatim-trace information, but this information did not lead to a correct inference of the transitive relationship.

Children in the second latent class used the premise information, but more often incorrectly (44%) than correctly (35%). Moreover, the percentage of no explanation (14%) was higher in the second class than in the third class. Often use was made of gist-trace information but not for difficult tasks.

In the third latent class, premise information was used correctly in most cases (65%). Thus, either the literal premise information or the ordering of the premisses were used to infer a transitive relationship. For some tasks, children used the premise information incorrectly or they used only part of this information (incorrect premise use). The percentages of the categories "external/visual information" (2%) and "no explanation" (6%) were small, meaning that from the perspective of fuzzy trace theory children mostly used gist-trace information to solve the tasks and rarely verbatim trace information. These results support *Hypothesis 3* to a large extent.

Relationship with age.   A three-class LCM including age as covariate was compared with another three-class model without age as covariate. The task parameters ($\theta_{j|u}$) of both models did not differ significantly. This means that the

TABLE 5
Relationship between Age and Class Membership
Expressed by Means of Latent Class Probabilities π_u

Age in Months	Class 1	Class 2	Class 3
83–109	.448	.159	.063
110–214	.265	.213	.135
215–317	.158	.230	.196
318–423	.084	.201	.291
424–525	.045	.198	.315

relationship between the latent classes and the task effects was not affected by age. Therefore, it was inappropriate to use fixed age groups to describe development. The parameter estimates for age were -0.042 (95% confidence interval: -0.052–-0.032) for the first class; 0.007 (95% confidence interval: -0.001–(0.022) for the second class; and 0.035 (95% confidence interval: 0.026–0.044) for the third class. The negative parameter estimate for the first class indicates a negative relationship between age and class membership probability. The positive parameter estimate for the third class indicates a positive relationship. Because the 95% confidence interval for the second class contains the value 0, there is no significant relationship between age and class membership.

Table 5 shows the average probabilities of membership in one of the three classes for five age groups of equal size (we did not use the grades to define groups because of children who repeated a grade). The results support *Hypothesis 4* because the three latent classes did not form fixed age groups.

DISCUSSION

Often the term abruptness is used to indicate discontinuity in development, and to emphasize that the change curve, which shows performance on a particular cognitive task as a function of age, is expected to be jumpy in particular intervals. For example, Flavell (1970) and Brainerd (1993) agreed that a change curve as in Figure 4b is abrupt, showing discontinuity, and that a curve as in Figure 4a reflects continuous change (but they disagreed on 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 needed to decide on discontinuity is arbitrary. Third, development in particular aspects of behavior may be related to age but
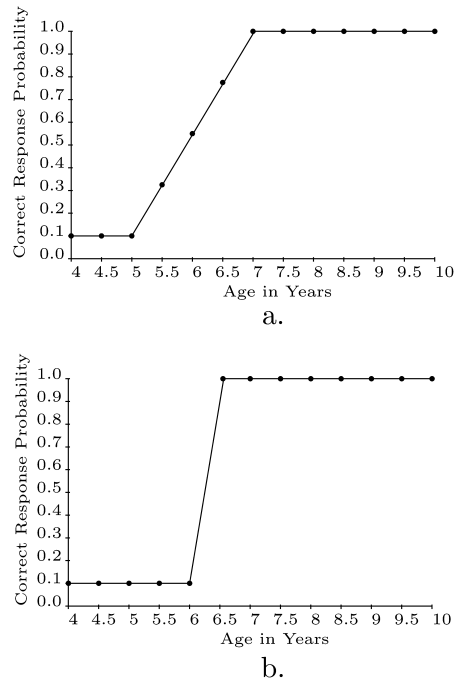
FIGURE 4   Continuous (a) and discontinuous (b) change curve, based on Brainerd (1993).

other variables than age presumably have a more direct, causal relationship to behavioral changes found with age (Wohlwill, 1973, p. 26). Finally, considerable individual differences in rate of development prevent chronological age to be a useful variable for studying behavioral development. That is, one four-year old child may attain a particular level at some behavioral dimension that another child may not reach until the age of six (Wohlwill, 1973, p. 26).

Taking these problems into account, we decided not to study discontinuity as abruptness in a change curve but to look for different phases in the form of multiple modes which reflect probability distributions of strategies. The results indicate that development can be described by three classes of low-ability, intermediate-ability and high-ability children. In each of these classes, children differ with respect to the probability of using a particular strategy to solve the tasks. In the low-ability class, children perform poorly on all transitive reasoning tasks. In the intermediate-ability class, children perform well on tasks in which the ordering of the objects is obvious, and worse on tasks in which this ordering is not obvious. In the high-ability class, children perform well on most transitive reasoning tasks. The results also showed that the performance on some

tasks was moderate, even in the high-ability class. These tasks had a physical content and premises were successively presented. Note that one task with these characteristics was not included in the analysis because 608 of the 615 children gave an incorrect response. Apparently, the combination of physical content and successive presentation made the recognition of the ordering extremely difficult.

The fit statistics (L^2, BIC) of the LCM help to select a most appropriate model from a set of potentially useful models. The decision whether the latent classes of this model indicate the existence of phases depends heavily on the interpretation of the classes in terms of underlying substantive theory. Thus, like BMM results LCM results should be interpreted with great care. Siegler and Chen (2002) emphasized that an LCM analysis has many advantages for distinguishing latent groups, but that the definition, use and interpretation of fit statistics also rely on arbitrary conventions. They argued that, when possible, results of other assessment methods for distinguishing classes should be compared with LCM results.

This study showed that children are characterized by a most likely phase and, as a result, smaller probabilities of being in other phases. A relationship between phases and fixed age periods was not detected. This result agrees with Wohlwill (1973, pp. 25–27, and chap. 9), who recommended to use other variables than chronological age for describing behavioral change. According to his approach fixed age-groups have no meaning because it is assumed that each child's development moves on its own, differential pace.

The latent classes could be interpreted well by means of the explanation children gave after they had answered the task. Bouwmeester et al. (2004) used a latent class regression model to investigate the relationships between these explanations, and the influence of the task characteristics on performance. They showed that task characteristics, which determined task difficulty, had an important influence on strategy use and that this influence varied between latent classes. In the present study, the BMM was too restrictive because it could not account for varying task difficulty level. Thus, ignoring varying task difficulty is not appropriate when studying phases in the development of transitive reasoning.

Researchers who investigate phases in the development of a particular cognitive ability are advised to first fit an unrestricted LCM to the data. Next, when such an LCM fits, for reasons of parsimony the BMM may be fitted when equal task difficulty within a class is expected to be realistic. Additionally, item response models with continuous latent variables may be fitted to find out whether a continuous dimension may explain individual differences in performance.

REFERENCES

Bouwmeester, S., & Aalbers, T. (2002). Tranred. 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., & Vermunt, J. K. (2004). Latent class regression analysis to describe cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology*, *1*, 67–86.

Brainerd, C. J. (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. (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.

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

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

Chapman, M. (1988). *Constructive evolution: Origins and development of Piaget's thought.* Cambridge, UK: Cambridge University Press.

Dayton, C. M. (1998). *Latent class scaling analysis.* Thousand Oaks, CA: Sage.

De Koning, E., Sijtsma, K., & Hamers, J. H. M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. *Applied Psychological Measurement*, *26*, 302–320.

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.

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, NJ: 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.

Glas, C. A. W., & Ellis, J. L. (1994). *Rasch scaling program.* Groningen, The Netherlands: iecProGAMMA.

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

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

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

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. (2001). Evidence for the phase transition from rule I to rule II on the balance scale task. *Developmental Review*, *21*, 450–494.

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.

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

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.

McCutcheon, A. L. (1987). *Latent class analysis.* Thousand Oaks, CA: Sage.

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

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.

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

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

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

Rost, J., & Von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp. 257–268). New York: Springer-Verlag.

Siegler, R. S., & Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology*, *81*, 446–457.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, *33*, 75–102.

Sijtsma, K., & Meijer, R. R. (2007). Nonparametric item response theory and related topics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, vol. 26: Psychometrics* (pp. 719–746). Amsterdam: Elsevier, North Holland.

Thom, R., & Fowler, D. H. (1975). *Structural stability and morphogenesis.* Reading, MA: Benjamin/Cummings.

Thomas, H. (1989). A binomial mixture model for classification performance: A commentary on Waxman, Chambers, Yntema and Gelman. *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., 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 den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika*, *47*, 123–140.

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: Erlbaum.

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. H., & 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.

Vermunt, J. K., & Magidson, J. (2003). *Latent Gold 3.0.* Belmont, MA: 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. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development*, *23*, 241–264.

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