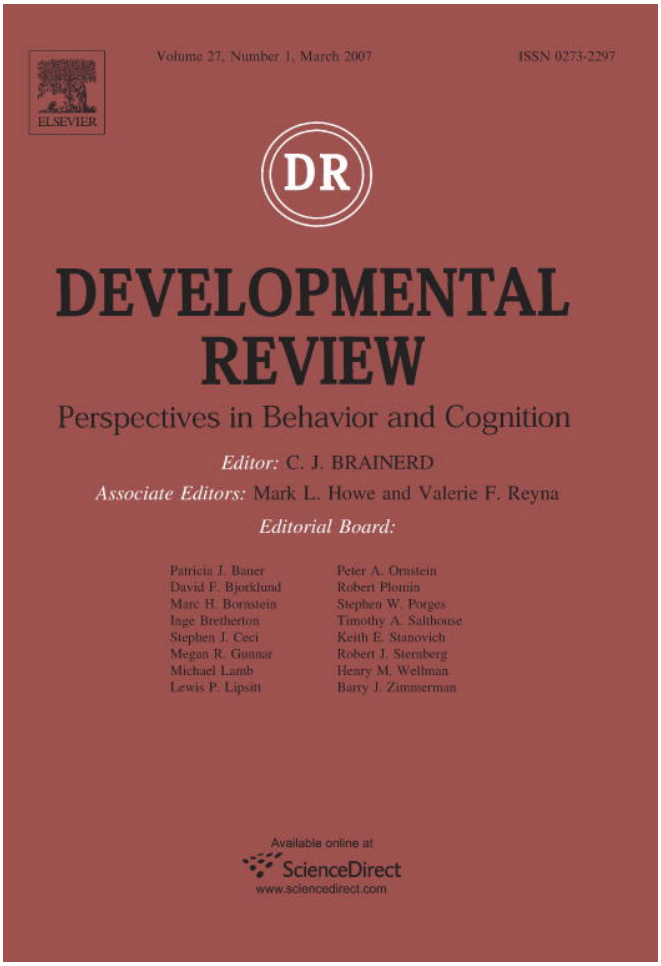


Available online at www.sciencedirect.com
ScienceDirect
Developmental Review 27 (2007) 127–149
DEVELOPMENTAL REVIEW
www.elsevier.com/locate/dr

Review

Classes in the balance: Latent class analysis and the balance scale task

Jan Boom *, Jan ter Laak

Department of Developmental Psychology, Universiteit Utrecht, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands

Received 26 May 2006
Available online 10 August 2006

Abstract

Latent class analysis (LCA) has been successfully applied to tasks measuring higher cognitive functioning, suggesting the existence of distinct strategies used in such tasks. With LCA it became possible to classify post hoc. This important step forward in modeling and analyzing cognitive strategies is relevant to the overlapping waves model for strategy development [Siegler, R. S. (1996). Emerging minds: The process of change in children's thinking. New York: Oxford University Press.]. However, so far, developmental trends were not part of the statistical model. Moreover, the theoretical importance of the fact that a few distinct classes were found was weakened because not all these classes represented pure strategies. To address these two issues, we model the development in class membership by incorporating age and working memory as covariates in the LCA model. Previous findings that the classes are well demarcated are replicated. A developmental sequence is supported by a strong effect of age on class membership and a moderate effect of working memory. Classification itself is hardly affected by the covariates: the problem of difficult to characterize classes remains. Nevertheless, classes describe large proportions of children's responses, classes are robust and fit a developmental trend, and some classes represent mixed rule use. In the discussion, the theoretical status of the overlapping waves model is clarified.

Keywords: Balance scale task; Working memory; Digit span; Latent class analysis; Logistic regression; Overlapping waves model; Cognitive development

* Corresponding author.
E-mail address: J.Boom@FSS.UU.NL (J. Boom).

0273-2297/$ - see front matter
doi:10.1016/j.dr.2006.06.001

Theoretical issues

Latent Class Analysis (LCA) elicited a revival of research with the balance scale task. LCA is a statistical technique for grouping participants according to their response patterns, based on similarity among these patterns (McCutcheon, 1987). The balance scale task which is used for studying higher cognitive functions, was introduced by Inhelder and Piaget in 1958 and became well-known in the 1970s and early 1980s thanks to the version coined by Siegler (1976, 1981). This version was specifically meant to elicit rule-governed response patterns in proportionality reasoning. Given the use of a certain strategy or rule by the participant, specific items are expected to be answered correct and others incorrect. Using a different strategy will result in a different pattern of correct and incorrect.

The first application of LCA to the balance scale task was original and promising (Jansen & Van der Maas, 1997). For the first time it was possible to classify children without the need for a predefined ideal pattern of correct and incorrect (as will be explained below). Application of LCA constituted an important step forward in modeling and analyzing the development of cognitive strategies. In studies that since used LCA for the balance scale task it was concluded that the task elicits discrete classes of response patterns for proportionality reasoning (Boom, Hoijtink, & Kunnen, 2001; Jansen & Van der Maas, 1997, 2002).

However, it is time to address and model the *development* of these classes. On the negative side, at least two serious theoretical problems have surfaced in the LCA studies on the balance scale task, on the positive side, the availability of new software allows more sophisticated models, e.g., more items can be analyzed at once and regressions of covariates can be integrated in the statistical model (Vermunt & Magidson, 2000, 2003b). The first problem is that Boom et al. (2001) found that some of the LCA classes did not map to the theoretically expected rules coined by Siegler. Apparently, some classes, while representing a distinct group, do not represent a group that uses one and the same rule or strategy. At least it became clear that these children do not use a strategy known from the original work of Siegler, nor one of the alternative strategies proposed by others since. We expect that with recent improvements in LCA and new models, problems regarding the interpretation of these classes can be solved. In particular, we will assess whether integrating the factors age and working memory in the model as covariates will facilitate interpretation of the classes. The expectation that working memory is relevant here is based on decades of Neo–Piagetian studies which suggested that, in addition to age, working memory is an important source of influence on cognitive performance (Case, 1992; Pascual-Leone, 1970). The second theoretical problem concerns the overlapping waves model for strategy development. This model, promoted recently by Siegler (1996) and others, seems a promising candidate for the kind of developmental model we are trying to build. However, on closer inspection a mix-up of an individual longitudinal model and a group cross-sectional model seem to have plagued the overlapping waves model. We will clarify what can and cannot be concluded from data of the kind we have, which is -as in most studies- cross sectional and not longitudinal in kind.

Our general aim is to model developmental trends for responses to the balance scale task and to clarify the theoretical status of classes in LCA. With these, we hope to overcome problems with difficult to characterize classes that are found by using LCA. In the following we first describe the balance scale task and discuss the operationalizations of the concepts of rules and classes, second, we briefly review some studies on working memory development. Third, we introduce our statistical modeling of latent group membership and of the

relationships between the balance scale task, age, and working memory. Fourth, we address the overlapping waves model. Finally, we formulate our aims in more detail.

The balance scale task: Rules or classes

The balance scale task and procedure by Siegler (1976, 1981) are still frequently used (Boom et al., 2001; Chletsos, 1986; Jansen & Van der Maas, 1997). The task involves showing a picture of a balance scale to the participating children (see Fig. 1). While the beam is fixed, a number of identical weights are placed on each side at regularly spaced distances from the fulcrum. The child is asked to predict which side will tip, if any. Based on previous research and rational task analysis, Siegler (1981) identified four rules for solving these items. Participants who use Rule l consider only the most salient aspect i.e. the number of weights on each arm. Participants who use Rule 2 consider the distance from the fulcrum when the number of weights is equal on both arms, but otherwise react similar to Rule 1 users. Participants who use Rule 3 consider the influence of the number of weights and the distance from the fulcrum in their predictions, but if the larger number of weights is on one arm, while the greater distance is on the other, random predictions result. Participants who use Rule 4 apply the torque rule, i.e., multiplying for each side the number of weights by the distance from the fulcrum. They predict that the balance will tip down on the side of the largest product.

Later, other rules were proposed by other researchers. For example, noticing the units of distance and number of weights on one side, children add these values instead of multiplying them as in Rule 4. This strategy became known as the addition rule. The buggy rule is based on the idea that displacing weights one hook to the left or to the right can be compensated by adding or subtracting one weight. The compensation rule requires computing the difference between weights left and right and comparing them to the difference in distance left and right. Application of any of these rules leads to the same

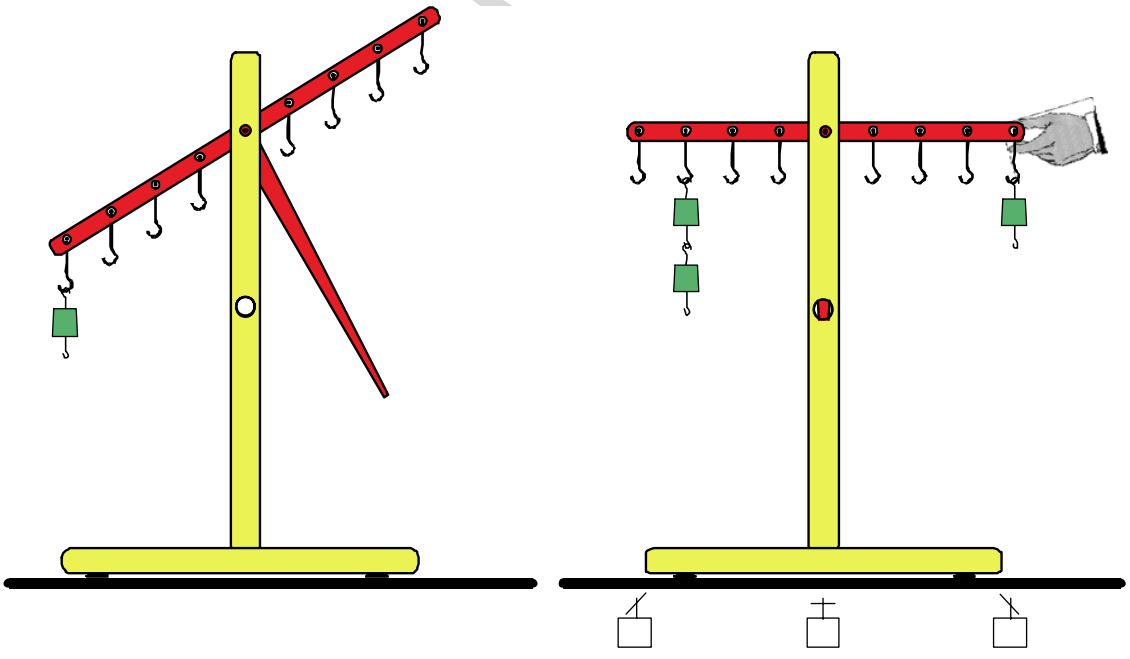


Fig. 1. Balance scale pictures: each picture was a full-page size. The left figure is one of the three examples presented first. Right is a conflict-weight item.

predictions. Since these rules cannot be distinguished by inspection of answer patterns, and we cannot know which variety a child has used, we will use the general term 'compensation rule' (Halford, Andrews, Dalton, Boag, & Zielinski, 2002; Normandeau, Larivee, Roulin, & Longeot, 1989; Van der Maas & Jansen, 2003).

Complementary to his original four rules, Siegler constructed six types of items to assess rule use. In simple-balance items, an equal number of weights are placed equidistant from the fulcrum on both sides. In simple-weight items, the distance is equal on both sides but the number of weights differs. In simple-distance items, the number of weights is the same on both sides but the distances are different. In conflict items, more weight is on one side with greater distance on the other, so that the side with more weight falls (conflict-weight item), or the side with the greater distance falls (conflict-distance item), or the balance scale beam remains horizontal (conflict-balance item). Despite efforts at standardization (Chletsos & De Lisi, 1991), balance scale items varied from study to study. Nevertheless, in studies using LCA, the choice of items is mostly limited to these six item-types or subsets of them (Boom et al., 2001; Jansen & Van der Maas, 2001, 2002; Van der Maas & Jansen, 2003), although all three conflict-item-types can be further subdivided in items that are or are not sensitive to the compensation rule (see Table 1).

Table 1
Item construction for Balance scale task: Pres-Seq represents order of presentation for 2003 cohort

Item characteristics					Number of weights on hook left				Number of weights on hook right			
Pres-Seq	Item-type	Torque	Side down	Mirrored	4	3	2	1	1	2	3	4
2	D	−2	Left	0	1					1		
6	D	−2	Left	1			2		2			
7	D	−2	Left	0		2				2		
10	D	−6	Left	1	3					3		
16	D	−2	Left	1	2						2	
9	CW	−2	Left	0			3					1
15	CW	−2	Left	1		2						1
20	CW	−1	Left	0		3						2
11	CWc	−4	Left	1			4					1
12	CWc	−1	Left	0				4			1	
3	CD	−2	Left	0			3		4			
13	CD	−2	Left	1		2			4			
1	CDc	−1	Left	1		1			2			
5	CDc	−4	Left	0	2				4			
18	CDc	−1	Left	1	1				3			
8	CB	0	Balance	1				4		2		
19	CB	0	Balance	0			2					1
4	CBc	0	Balance	0				3			1	
14	CBc	0	Balance	0			3				2	
17	CBc	0	Balance	1		4						3

Items are grouped by Item-type: D = distance, CW = conflict-weight, CD = conflict-distance, CB = conflict-balance, c = correct for compensation rule. Torque is product right minus product left. Mirrored: 1 if left and right were reversed for the actual item. E.g. the seventh item has two weights on the 3rd hook on the left and one weight on the 4th hook to the right (as in Fig. 1 left).

In the classical design of Siegler, rule-use is assumed to be assessable because the rules interact with item-type i.e., each rule generates a specific pattern of responses over item-types. In Siegler's (1981) Rule Assessment Methodology (RAM), scoring criteria operationalize the closest likeness to these *a priori* defined patterns of expected responses in terms of left-down/balance/right-down per item-type. The participant is assumed to have used the rule that generates the closest fitting pattern. LCA differs from this classical assessment because it allows classification without expected patterns. It must be admitted, however, that item selection is based on theoretically expected patterns. LCA searches for similar response patterns and groups them together in a predefined number of classes characterized by the probabilities of each of the possible responses (left, balance, right) to each of the items. These probabilities can be used to interpret which, if any, rule is used by the children who belong to such a class. LCA provides a statistical measurement model, does not need arbitrary cutoff rules and offers several fit measures. LCA and its advantages compared to the RAM are also discussed in Jansen and Van der Maas (2002).

We use the term *rule* to refer to a relatively simple strategy. A strategy describes the thought process of an individual that consistently applies an algorithm to a task and refers to a logic underlying the response pattern of an individual. This logic has to be recognizable for the researcher. In the original balance scale task analysis by Siegler the rules were referred to as 'decision rules', exemplified by a decision tree. He also noted "that the rule for generating correct solutions, once known, is trivially easy to execute, but inducing the rule in the first place is quite difficult" (Siegler, 1976; p. 483). The use of a specific rule can only be assessed if we have a response pattern hypothesized for that rule (note that no verbal explanations were requested from the participants) and compare such a predefined pattern with an observed response pattern.

We use the term *class* to refer to empirically derived and probabilistically defined collections of response patterns classified by LCA. Although each class is be characterized by a particular profile (exemplified by graphs as in Fig. 3) for the response pattern, no predefined pattern is required for classification itself.

Although LCA is clearly superior in detecting different response patterns compared to RAM, some unexpected results emerged. In Boom et al. (2001) no one-to-one correspondence between rules and classes was found. Some classes were elusive and difficult to characterize in terms of rules, while other classes clearly represented a rule. Clear results are almost always obtained for Rule 1, which emerges as a distinct class for elementary school children, and for a class with predominantly correct responses (suggesting Rule 4 use), which is observed in older samples. However, in the study of Boom et al. (2001), for some other classes, the deviations of the expected patterns were so large that interpretation for these classes was difficult.

The same problems can also be found in other studies using LCA. Jansen and Van der Maas (2002) detected problematic classes too and interpreted the discrepancies as inconsistencies or as the result of mixed rule use. They might have underestimated discrepancies due to computational limitations that required them to use equality constraints over items from the same type and to limit their analysis to much smaller subsets of items (Jansen & Van der Maas, 1997, 2001, 2002) than were used by Boom et al. (2001).

Distinguishing rules and classes brings to front the contrast between a top down approach and a bottom up approach. The top down approach begins with predefined rational strategies, which we seek to confirm in the data. The bottom up approach begins with empirical regularities in the data, and subsequently seeks to attach rational strategies

to them. The Balance Scale Task and the currently available LCA methodology offer a unique opportunity to see whether both approaches converge. Since classes are more recent, and less well understood and discussed than strategies, we henceforth focus on the LCA classes. Rules for the balance scale task have been discussed extensively in the eighties (e.g. Ferreti & Butterfield, 1986; Normandeau et al., 1989; Wilkening & Anderson, 1982).

Is development in proportionality reasoning related to age and working memory?

We expect that considering age and working memory capacity can solve problems involving the difficult to characterize classes. Developmental or age trends have frequently been found for the balance scale task (e.g. Boom et al., 2001; Jansen & Van der Maas, 2002). Using LCA, this would imply that the chances of belonging to a certain class (class membership) vary with age. However, given considerable individual variation, age will account for only part of the variation in classification. It is worthwhile to search for other factors related to class membership. Neo–Piagetians invoked resource constraints, especially working memory, to explain part of the variability in the age at which a certain level of thinking is attained (Case, 1992; Pascual-Leone, 1970). One of the surprisingly few studies relevant here (but see a review by Fry & Hale, 2000) is the monograph by Demetriou, Christou, Spanoudis, and Platsidou (2002). They concluded that the storage component of working memory, as measured by forward digit-span tasks, explained little of the variance in performance on the cognitive tasks. The executive aspect of working memory, in contrast, showed substantial correlations with performance. They measured this executive aspect of working memory with two complex memory tasks. We will measure the executive aspect of working memory with a backward digit-span task. This task is also believed to require complex processing (Gathercole, 1998). Note, that we will only evaluate the effect of working memory on this general level of classification; more detailed analysis involving analysis of working memory demand per item or rule are beyond the aims of this present study.

Statistical background: Modeling the effect of covariates on class membership

First, we address the statistical model for classification based on responses to the balance scale task itself. Next, we extend the statistical model to include the effect of covariates.

LCA can be used to help determining *the number of classes* needed to account for the differences among the response patterns in the data in an empirical, inductive way. The division into classes by LCA is based on the assumption of local independence (Haberman, 1974). LCA characterizes each class by its *class-specific probabilities*. These are the probabilities of giving each of the possible responses to the each of the items if a child is a member of the class. In addition, LCA estimates the *class weights* i.e., the proportion of children belonging to each class.

The number of latent classes needed to obtain an adequate explanation of the variability in the response patterns is traditionally determined using likelihood-ratio statistics. These statistics compare the observed frequency of each response pattern with the frequency predicted by a latent class model with a specific number of classes. If the summed differences between the observed and predicted frequencies are small, the model contains

enough classes to account for the variability in the response-patterns. The number of latent classes is increased (from one to fifteen in our case) to find the best fitting models. However, with sparse frequency tables the asymptotic *p*-value associated with the likelihood-ratio chi-squared statistic cannot be trusted. With only few items the number of possible patterns is likely to be lower than the number of participants (e.g. four items scores as correct/incorrect gives 16 possible patterns), however with 20 three-valued items we have over 3 billion possible patterns which results in a very sparse frequency table. Therefore, following Jansen and Van der Maas (2002), we used a (non-naïve) parametric bootstrap procedure which gives more reliable *p*-values (Van der Heijden, t Hart, & Dessens, 1997; Vermunt & Magidson, 2003b) in case of sparse frequency tables.

We used the popular Bayesian Information Criterion (BIC) to compare models that show an acceptable fit to the data. This criterion is based on the log-likelihood ($\log \mathcal{L}$) to compare different models (Schwarz, 1978). This criterion is based on a combination of the fit of the model and the number of parameters used and is calculated as $-2\log \mathcal{L} + (\log N)\ npar$, where $\mathcal{L}$ is the likelihood, *npar* is the number of parameters in the model, log is the natural logarithm, and N is the number of participants. The smaller the value of this criterion, the smaller the distance between the model at hand and the true model. Like Jansen and Van der Maas (2002) we prefer this Bayesian Information Criterion based on Log Likelihood (BIC-LL) over the Akaike's Information Criterion (AIC) since it favors more parsimonious models than the AIC measure (Akaike, 1974).

Our use of LCA diverges from the approaches in previous studies (Jansen & Van der Maas, 1997, 2002) because we make no use of equality restrictions. Neither did we use inequality constraints as in Laudy, Boom, and Hoijtink (2004) because our approach is exploratory and not confirmatory. Moreover, we used a large number of items in the LCA whereas Jansen and Van der Maas (1997, 2002) used LCA only for subsets or combinations of items (through equality restrictions) to avoid sparse frequency tables. Such sparseness is known to cause two additional problems for LCA, apart from the problems for the likelihood-ratio chi-squared statistic as discussed above (Raftery, 1986). First, the iterative search process for the optimal solution might stop prematurely for a single parameter if the estimate becomes zero or one. To eliminate the possibility of obtaining such boundary solutions Bayesian priors are used in Latent GOLD (Vermunt & Magidson, 2000). Second, using many items increases the likelihood of obtaining non-optimal local solutions for the overall model, therefore, we performed all analyzes with much larger than default sets of different random starting values and with higher than default number of iterations to be performed per start set (Vermunt & Magidson, 2003a).

Previously, age trends have been reported by cross-tabulating age groups with proportions for use of each rule. Plotting these proportions reveals age trends (Jansen & Van der Maas, 2002; Siegler, 1996; chapter 4). Unfortunately, in this way, no formal criteria are used to evaluate the magnitude and significance of the relationships. Although age is not a cause of class membership in a psychological and developmental sense, in a statistical sense age can be treated as a covariate that has an *effect* on class membership. Using normal *linear regression* to evaluate the effect of age on class membership would be problematic because 'class' is a categorical variable with a limited number of levels. The use of dummy variables for each category is not helpful because it breaks down the overall picture, and the estimated proportion class use will generally be biased (Menard, 2002). An alternative is the use of multinomial logistic regression.

Logistic regression models allow investigating effects of explanatory variables on categorical responses. A fitting model provides smoothed estimates of response probabilities (Agresti, 2002; Menard, 2002). Multi-category responses can be analyzed with a generalization of the logistic regression model: the multi-category (or polytomous) logit model. The relationship with continuous variables like age and digit-span, and categorical ones, like gender, can be assessed, and interactions can be estimated. The parameters can be transformed to proportion class use (see Eq. (2) below) resulting in a set of curves depicting the likelihood of belonging to each class (class weights) given a realized value on the covariate. The model can be visualized in smoothed curves that change with e.g., increasing age, or in smoothed surfaces when two covariates are used. This smoothed visualization might assist in interpretation of the classes.

Overlapping waves model

The combination of the LCA with a multinomial logistic regression model allows us to address development. We model the ratio between class proportions. This implies that when class membership for one class rises, class membership for other classes must fall. This results in a depiction like that of an overlapping waves model. Therefore our model is -at least visually- similar to the *overlapping waves* model proposed by Siegler (1996) and the complex developmental model proposed by Rest (1979). Siegler, like Rest, offered an overlapping waves model to accommodate findings of gradual change in strategy or stage use.

On the one hand, our model is formal, precise, can be tested, and is extended to two dimensions, on the other hand, our version of the overlapping waves model is more limited than Siegler's version because we did not pretend to model individual trajectories. However, precisely in so far as an overlapping waves model pretends to represent individual development it remains an heuristic or metaphorical model and not a formal, precise, and testable statistical model, because deriving a formal overlapping waves model from longitudinal data is difficult for several reasons. First, collecting longitudinal data is demanding. Second, taking into account large individual differences in timing of development make it difficult to generalize patterns. Finally, for tasks like the balance scale task (and task tapping higher cognitive functioning in general), repeated assessment will always imply some degree of repeated experience with the task and this will influence development. If so, the resulting overlapping waves model would reflect learning in addition to development and not pure development (see Boom, 2002). Unfortunately, an individual overlapping waves model cannot be inferred from cross-sectional group data either. Whereas summing or averaging of linear functions always results in a linear function, this is not necessarily the case for S-shaped curves (or waves). The logistic model is not dynamically consistent (Keats, 1983). This means that no conclusions can be drawn from the shape of the average trajectory concerning the shape of the individual trajectory (Siegler, 1996; Singer & Willett, 2003).

Our logistic regression model is conceptually and mathematically related to the logistic *growth simulations* advocated by van Geert (see Case et al., 1996; Demetriou et al., 2002; van Geert, 1994). We might conceive class weights in terms of a population growth model with competing growers that each represents the proportion strategy use in the population, given a certain level on the age and on the working memory variable. We assumed that the basic form of a logistic curve adequately captures such growth while the straight line of a default linear model cannot (Agresti, 2002; Eckstein, 2000).

Aims

Our specific aims are:

1. To assess the effect of working memory capacity on balance scale behavior predictions. In particular, to assess whether working memory capacity, or interactions between working memory capacity and age, can be seen as indicators of developmental trends in addition to age alone. Note, that we focus on class membership instead of on rule use.
2. To model, visualize, and estimate the precise magnitude and significance of the effects of age and working memory on class membership by means of a statistical model.
3. To improve classification by including and modeling age and working memory as covariates in the classification of responses to this task. Is it possible to (a) form classes more easily by obtaining better fitting models, or (b) to classify more respondents or assign them with more certainty to one of the classes, or to (c) obtain better interpretable classes (see next point).
4. To facilitate interpretation of the class profiles, in particular, to see if the problem that some of the classes found in previous analysis did not represent pure rules can be alleviated by considering points 1 to 3. Will new classes emerge as result of introducing covariates? Will classes remain as in earlier studies but with new interpretations easier to associate with rules? Will more information on the developmental order facilitate interpretation?
5. To address the overlapping waves model: The idea of overlapping waves of strategy use is based on the changes in strategy use along the developmental continuum. What are the implications of developmental trends for class membership (as in point 2 above) for such an overlapping waves model?

An empirical study of balance scale task answer patterns

We illustrate our approach in a sample of two cohorts of randomly selected Dutch children from 5 to 14 years. The final sample comprised 400 boys (mean age 9.9 *SD* 2.1) and 565 girls (mean age 9.7 *SD* 2.1). This age difference is not significant. For Cohort 2003 (427 children), a fixed sequence of presentation of the items was used; for Cohort 2004 (538 children), items were presented in random order. Children came from very diverse backgrounds from the urban and rural region in and around a mayor city in the Netherlands. However, children with higher socio-economic status were slightly over-represented. Given the nature of our questions, diversity and variance are important, but representativity of the Dutch population is not required.

Balance scale task

A 20 item paper and pencil test with each item on a separate A4 sheet is presented to the children. Each item depicts a balance scale with weights hanging down, designed by Wolters, Fischer, and Zuidema (1987). The hooks on which weights can be added are at regular distances from the fulcrum (Fig. 1). The maximum distance used is 4 units and the maximum number of weights is 4 in order to limit the influence of arithmetic ability. Five items of each of the following four item-types are presented to the children:

Distance, Conflict-Weight, Conflict-Distance, and Conflict-Balance (Table 1). Balance and Weight item-types are not included since these are expected to be answered correctly by nearly all children in our age range (see Boom et al., 2001). In addition, we made a distinction, for the conflict items, between items for which application of the compensation rule would lead to a correct prediction, and those that would lead to an incorrect response (Table 1). A hand is depicted holding the balance steady.

Backward digit-span task

Random digit strings of a certain length are presented orally to the child in a tempo of one digit per second. The children are required to repeat back verbally to the experimenter each string directly after hearing it, in reverse order. The answer is recorded verbatim. Starting with a length of two digits, when both strings are correctly reproduced in reverse order, the length is increased by one digit. If only one of the two strings of equal length is correctly reproduced a third trial is inserted. After two failures at the same string length testing is discontinued. Digit-span is the string-length of longest string that is correctly repeated two times. The score is increased by one-half if the next (longer) string is repeated correctly once. Digit-span tasks are part of intelligence tests like the WISC-R (Wechsler, 1974). Test-retest reliability coefficients range from .66 to .89 depending on interval length and age of subjects (Lezak, 1995). However, WISC norms pertain to a combination of forward and backward digit-spans. Recent information about age trends for backward digit-spans only is scarce. Morra (1994) found backward digit-span's increasing from 2.80 to 4.14 for an urban sample and from 2.19 to 3.57 for a rural sample for 6 to 10 year-olds (see Fig. 2).

Procedure

Participating children are required to indicate, by placing a mark on the drawing, which side would go down when the beam is released (see Fig. 1). The children did not receive

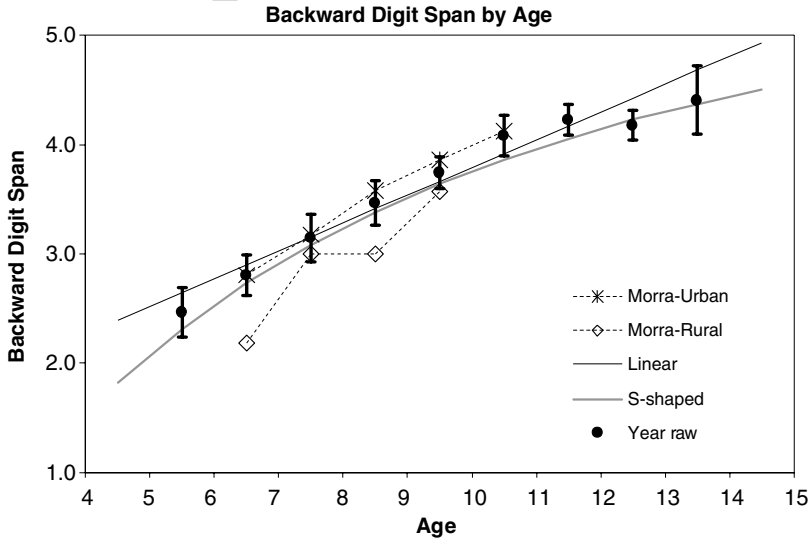


Fig. 2. Backward digit-span (length of longest string correctly repeated digits) by age. Year raw: Average with standard error of the mean based confidence intervals for the 5–95% range. Linear: linear regression line. S-shaped: best fitting nonlinear regression line. Morra-Urban and Morra-Rural are taken from Morra (1994) for comparison.

feedback during the assessment. The children were tested individually at home by as many trained students. Each student had to obtain a signed statement by parent of the child (with name, address, and phone number) given permission to use the data. The student experimenters were extensively prepared in small (max. 20), supervised groups for this specific assessment and followed a strict assessment protocol. Compared with a design with a few experimenters, this design decreases the chance of systematic biases due to differences in experimenters and it explains why the number of missing values was low. At the same time, it can increase the amount of random error in the data. The quality of our data is good simply because the task is easy to administer. The digit-span assessment is more susceptible to errors due to lack of experience by the students. To be able to evaluate the trustworthiness of this data, we compared our findings with trends for the same age-groups reviewed above (see Fig. 2).

Analysis

The data consisted of a matrix of 965 subjects × 20 items; and 3 child variables: Age[1], backward digit-span (BDS), and Gender. The response patterns consisted of 20 trichotomous nominal responses "left side down", "in balance", "right side down", so the number of possible patterns is 3^{20}. There were 708 different patterns for 974 children; 656 were unique. These data are appropriate to evaluate our models and approach. These patterns were analyzed by means of exploratory latent class analysis. A sequence of unrestricted models was explored using 1 to 15 latent classes. Covariates were: Age, BDS, gender, and new variables that represented interactions.

Latent GOLD uses logistic regression procedures combined with LCA. The logistic regression of age or working memory is used as a covariate in the LCA and both can contribute to the classification. Equation (1) can be used for modeling the basic probability density of observing, for example, a particular set of 20 y values given 2 covariates z_1 and z_2 with 5 latent classes denoted by x_1 to x_5 intervening.

$$\pi(y_1 \ldots y_{20}|z_1 z_2) = \sum_{i}^{5} \pi(x_i|z_1, z_2) \prod_{k=1}^{20} \pi(y_k|x_i) \qquad i = 1, 2, \ldots.., 5 \tag{1}$$

$$\pi_j(x) = \frac{\exp(\alpha_j + \beta_{j1} z_1 + \beta_{j2} z_2)}{\sum_{h=1}^{J-1} \exp(\alpha_h + \beta_{h1} z_1 + \beta_{h2} z_2)} \qquad j = 1, 2, \ldots.., (J-1) \tag{2}$$

Logistic regression uses the natural logarithm of the odds of using class $x_{1\ \text{to}\ j}$ versus a reference class x_{ref} in the regression equation, instead of the proportion class use directly (Agresti, 1996, 2002). Reversing this transformation, a smoothed visualization of the effects of the covariates can be obtained. The estimates for the intercept α and the added effect for two covariates without interaction z_1 and z_2 weighted by β_1 and β_2 can, for each

[1] Although the software we used (Latent Gold 3.0: Vermunt & Magidson, 2003a) can handle full precision of the Age variable (registered with precision of days), it is our experience, that using full precision (more than 500 categories) increases the sparseness of the frequency table so much that the bootstraps cannot be trusted anymore. Therefore, we broke up the Age variable in nine ordered categories in order to facilitate computations for the bootstrap procedure.

class j, be converted back to proportions of class use π (Agresti, 2002). Plotting π for various z_1 and z_2 in one graph depicts a set of J surfaces representing the estimated proportion use of a class as influenced by the two covariates (see Eq. (2)). The last surface J can be obtained by setting α and β to zero as it is the baseline category needed for reasons of identification (Agresti, 2002; p. 271).

We evaluate the contributions of the covariates by means of the logistic regression on three levels of detail. First, on the most general level, we consider the increase in fit in terms of the Bayesian Information Criterion based on Log Likelihood (BIC-LL). Models with covariates are compared to models without covariates. We also consider improvement in terms of classification statistics. In particular, we report the proportional reduction in error of the prediction to which class a participant belongs. Knowledge of the observed scores of a participant will reduce the error in prediction compared to the situation in which information for a participant is completely lacking. Knowledge of the values on the covariates is expected to reduce error in prediction even further. Second, on a more detailed level, we determine the effect of each covariate separately by means of a Wald statistic (Vermunt & Magidson, 2003b). This statistic allows to decide the inclusion or exclusion of a covariate or a term representing the interaction of covariates combined. All covariates and interaction terms are entered in the model one by one, and subsequently in all possible combinations. Third, on the most detailed level, once a covariate is included, we consider the standard errors for the parameters relating the covariate to the class separately for each class, to determine which classes in particular contribute to the effect. We take effects larger than 2 times the standard errors of measurement to be meaningful (Vermunt & Magidson, 2000).

Results

First, we compare the increase in working memory with age to trends reviewed in the introduction. Next, we present the LCA of responses to the balance scale task and address the optimal number of classes in terms of overall fit of the models and profiles for each class for the best fitting model. Finally, with these results in place, the stage is set to present results concerning our main hypotheses concerning the developmental trends of class use.

Working memory and age

Working memory increases with age. Increases per year for backward digit-span (BDS) capacity are visualized in Fig. 2. Differences between cohort 2003 and 2004 were not significant. Linear regression estimates a BDS increase of .247 units per year. The proportion explained variance (R^2) is .253 for BDS. A linear model is appropriate because the best alternative model resulted in only a slightly better fit to the data: R^2 is .293 for BDS for a non-linear S-shaped curve. S-shaped curves are theoretically more plausible, and confirm findings by Demetriou et al. (2002), but for our sample these regression curves diverge only marginally from the linear model. This implies that we can use BDS as covariate in the subsequent analysis.

Classification: Number of classes

We performed all analyzes with much larger than default sets of different random starting values. For up to six classes, this made no difference; for higher number of classes there were minor differences, but even in those cases, the default option would

have led to the same models. In Table 2, we report the Bayesian Information Criterion based on Log Likelihood (BIC-LL) for 1 to 15 classes. We use this fit index combined with theoretical considerations to choose the optimal number of classes. The five-class model with Age and the model with Age and BDS combined as covariates have the lowest BIC-LL values in Table 2. Since it makes sense theoretically to determine the number of classes first and then evaluate the contribution of the covariates, we have to consider which number of classes is appropriate overall. It appears that in each column, with or without using the covariates, the BIC-LL value for the five-class model is the lowest, so it gives the best description of the data according to the BIC-LL measures. In terms of the class-specific probabilities (discussed below and see Fig. 3), the five-class models perform well. With five classes, all classes have reasonable size and differ clearly from each other. The differences between the fit measures for the five-class and the six-class models are very small (in each column in Table 2). The six-class models offer only an additional small class without clear patterning and this situation is not improving with the seven- and the eight-class models.

The bootstrapped *p*-values for models with covariate(s) are all above *p*-values $> .05$, up to the 11-class models, except for the two-class model with Age as covariate (see next section). However, the bootstrapped *p*-values for models without covariate(s) are only satisfactory for the seven- and the eight-class models and not for the five-class model. Interestingly, adding the covariates has negligible effect on the parameter estimates for

Table 2
Comparison between models with different number of classes for four models: without covariates; with only Age; with Age and BDS; or with Age, BDS, and Age × BDS (A × B) as interaction term

	LCA with or without covariates							
Covariates	—	AGE	AGE	AGE	—	AGE	AGE	AGE
	—	—	BDS	BDS	—	—	BDS	BDS
	—	—	—	A × B	—	—	—	A × B
Classes	BIC-LL				*p*-value Bootstrapped			
1	31437	31437	31437	31437	0.000	0.000	0.004	0.006
2	26209	26009	25999	26006	0.000	0.000	0.086*	0.130*
3	25303	25088	25079	25092	0.006	0.070*	0.290*	0.312*
4	24585	24333	24329	24348	0.002	0.088*	0.262*	0.280*
5	24413	24158	24160	24184	0.016	0.066*	0.192*	0.274*
6	24445	24188	24182	24212	0.020	0.066*	0.212*	0.204*
7	24500	24249	24252	24288	0.084*	0.138*	0.190*	0.212*
8	24560	24316	24334	24380	0.062*	0.286*	0.182*	0.202*
9	24688	24458	24489	24528	0.044	0.078*	0.176*	0.156*
10	24835	24612	24647	24699	0.036	0.066*	0.124*	0.150*
11	24995	24783	24810	24881	0.026	0.058*	0.128*	0.114*
12	25172	24934	24976	25117	0.032	0.028	0.080*	0.100*
13	25346	25136	25160	25231	0.026	0.024	0.046	0.070*
14	25508	25259	25358	25414	0.008	0.016	0.038	0.040
15	25691	25473	25542	25617	0.000	0.008	0.042	0.038

BIC-Log Likelihood can be used for model comparison.
* indicates *p*-values > 0.05 = conventionally required for a model to be acceptable.
The number of parameters without covariates starts with 40, with 41 added for each additional class. With covariates, one parameter per additional class for each covariate, should be added.

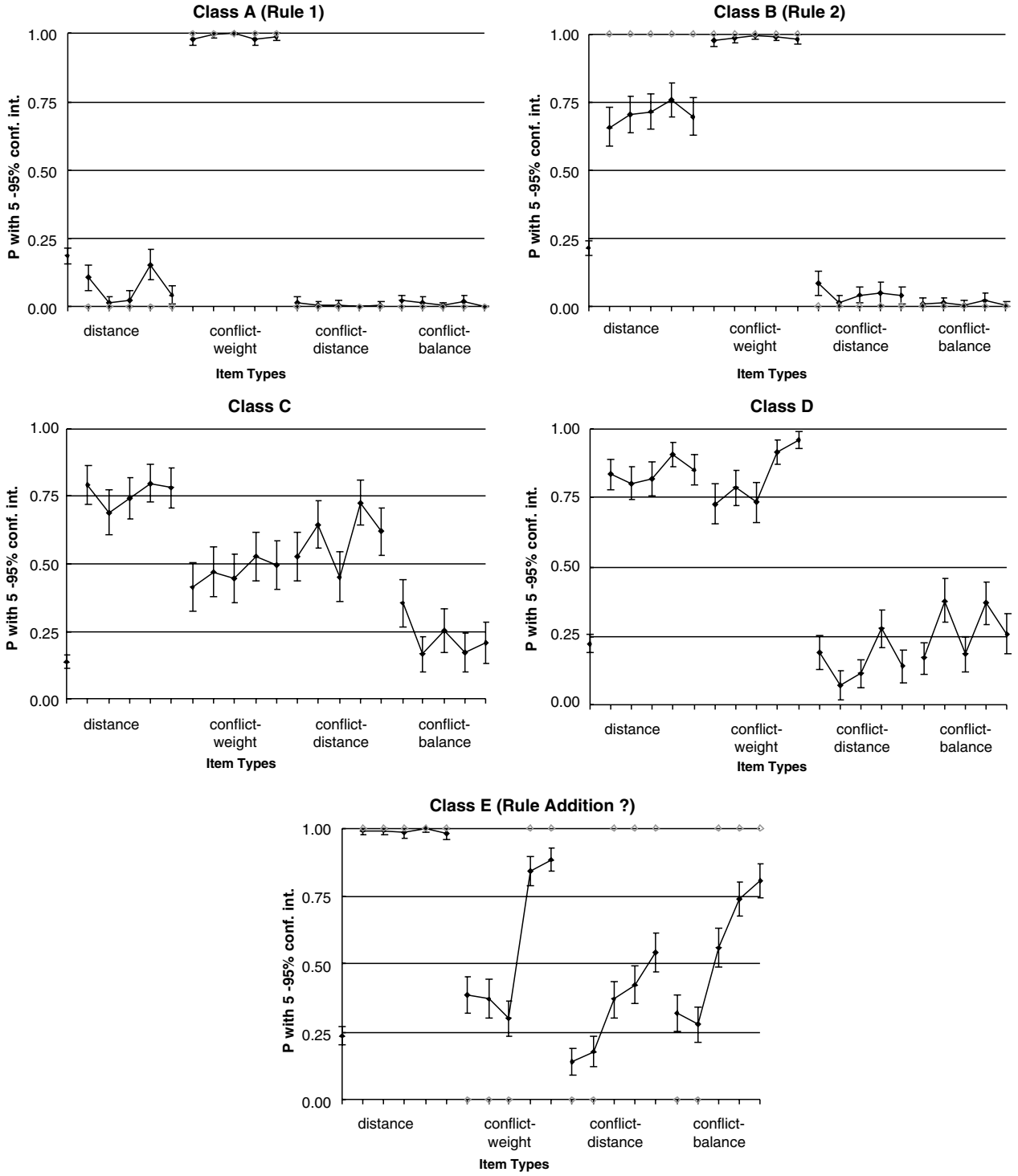


Fig. 3. Five-class model showing estimates of the class-specific probabilities for each item and corresponding intervals for the 5–95% range. Items are grouped by type (five each): d = distance, cw = conflict-weight, cd = conflict-distance, cb = conflict-balance. The marker on the y-axis indicates weights of each class. The gray markers in the background specify expected values (where applicable) for the Rules as explained in the text. Order of presentation is the presumed developmental order.

the items and thus the profiles to be presented shortly are almost unaffected. Moreover, classification is almost left unchanged by adding the covariates. Based on these considerations, and thanks to the inclusion of covariates, we choose the more parsimonious five-class models as the best models.

Classification: class profiles

We discuss these five-class models in terms of associated rules in as far as possible. In Fig. 3, the estimated proportion for giving the correct response for each item is depicted for the five-class model with Age as covariate but note that the other five-class models would have resulted in same pattern with differences to small to be visually perceptible in Fig. 3. The items are sorted by item-type in the same order as in Table 1. In Table 3 a selection of characteristics of the classes is presented.

Class A fits almost perfectly with expectations based on Rule 1 as defined by Siegler (see Fig. 3 and Table 3; the expected values for Rule 1 are indicated in grey in Fig. 3a). The confidence intervals are very small, on average .02. The class-specific probabilities are clearly separated, the items expected to be answered correctly (conflict-weight items) have an average probability of being correct of .99; and those expected to be answered incorrectly have an average probability of .03 for correct response. The average difference from expected for Rule 1 is .03. Class A contains around 19% of the children.

Class B shows a pattern of correct responding in which Rule 2 as defined by Siegler can be recognized for all except the distance items. The expected values for Rule 2 are indicated in Fig. 3b. The confidence intervals are small, on average .03. The class-specific probabilities are clearly separated with the items expected to be answered correctly (distance and conflict-weight items) having an average probability of being correct of .85 and those expected to be answered incorrectly having an average probability of .04 for correct response. Class B contains mainly Rule 2 thinkers but on the distance items more inter-individual responding is found as expected for this rule. The average difference from expected for Rule 2 is .07. Class B contains around 22% of the children.

Class C can be taken as random responding for all items. The confidence intervals are the largest of all classes (on average .08). The class-specific probabilities for correct responding are distributed over the whole range. Nevertheless, the rule that has the smallest deviation from this pattern is Siegler's Rule 3, the average difference being .14. Class C contains around 14% of the children.

Class D has a pattern of correct answers that resembles that of Rule 2 (see Fig. 3). This impression is misleading, however, because inspection of the probabilities for the incorrect response alternatives (not depicted in Fig. 3) reveals that in this class the balance response is given too often for the conflict items. On average, for these conflict items, the chance for answering balance is .22, whereas the expectation for Rule 2 is close to zero. The confidence intervals in general are modest, on average .06. The average difference from expected for Rule 2 is .19 and for Rule 3 this is .20. Class D contains around 22% of the children.

Table 3
Overview of class characteristics

LCA class	≈Rule	Average age (yrs)	Average BDS	Class weight (proportion)	Average SE of class prob.
A	R1	8.08	3.06	0.19	0.02
B	R2	9.11	3.56	0.21	0.03
C	—	9.81	3.74	0.14	0.08
D	R2-R3?	10.72	4.03	0.22	0.06
E	R3-RA?	11.01	4.17	0.24	0.05

For each class A–E (5-class model) the following characteristics are given: the associated Rule, the average age of children in the class, the class weight (the proportion of all children in the class), and the average confidence interval width (see Fig. 3) for the probability of giving the correct response.

Class E has a pattern of correct answers that resembles the pattern expected for the compensation rule but with some major deviations. Confidence intervals are modest, on average .05. The class-specific probabilities are clustered, with all distance items; conflict-weight item 4 and 5; conflict-distance item 3, 4, and 5; and conflict-balance items 3, 4, and 5 having an average probability of being correct of .78 (see Fig. 3e). The remaining items have an average probability of .28 for correct response. The expected values for the compensation rule are indicated in grey in Fig. 3e. The average difference from expected for the compensation rule is .18 and for Rule 3 this would be .17. Class E contains around 23% of the children.

Our data clearly show that classes and rules do not correspond on a one by one basis. Class A can be taken to represent Rule 1 because the average deviation is only .03. Class B can be taken to represent Rule 2 but here the average deviation is larger (.07). Visual inspection of Fig. 3 suggests that class D also represents Rule 2 but the average deviation of .19 is large. Visual inspection of the expected pattern for the compensation rule suggests it can be compared with class E but here the average deviation is again large (.18). Likewise, Rule 3 can be compared with classes C (.14), D (.20), and E (.17), but the deviations are large and it the definition of Rule 3 is too encompassing to be helpful. The absence of Rule 4 is to be expected for this age range. For better understanding of the classes, we now turn to the question how they are affected by the covariates.

Developmental trends for class membership

As argued in the theoretical introduction, improvements in task behavior related to increasing age and increasing working memory capacity might indicate developmental trends. We first evaluate the general effects of the continuous covariates Age and BDS. Table 2 contains the overall fit for models with and without a selection of covariate combinations for 1 to 15 classes. Model comparison based on BIC-LL, for all multiple-class models, reveals that the models with covariates included fit better than models without covariates. However, differences between the model with (1) Age alone, (2) both Age and BDS, and the model with (3) the Age × BDS interaction added are small. The proportional reduction in error for classification for the model without covariates is good (.92). This improves no further after adding covariates. In other words: the pattern classification is stable (which is desirable). The covariates in turn can be used to predict class membership: proportional reduction in error is .154 for Age alone and .163 for Age and BDS combined. Gender has no significant effects and will not be discussed further. Results for different numbers of classes varied slightly but the same covariates would have been selected as interesting in all instances. The models with other combinations of covariates (not reported in Table 2) had clearly lower BIC-LL values and are discussed no further.

Second, on a more detailed level of analysis, we compare the effects of the covariates with each other, restricting ourselves to the five-class models with covariates. It appears that adding Age has an extremely high Wald statistic (Wald = 188, p .001) indicating that there can be no doubt about the influence of Age on class membership in general. Adding BDS has a less impressive but still significant effect (Wald = 23.2, $p < .001$), and adding the interaction term between Age and BDS does not results in a significant effect.

Third, on still finer level of detail, and again focused on the five-class models with covariates, we assess which of the classes A to E is affected most by the covariates. The standard errors of the effects for each class separately suggest mixed significance: Age (only) has highly significant effects (z-values > 5) on all classes except class C. The

covariate BDS (entered after Age) has significant effect (z-values > 2) on classes A, D, and E. The interaction between Age and BDS has no significant effect for any class.

Fig. 4 clearly shows that Class A is dominant in the lower end of our age and BDS range. Class B use is at its heights in the lower end of our age range, but the use is dropping less with age than for class A (Fig. 4a). BDS effect is suggestive but not significant. Class C size is not affected by Age or BDS. Class D use peaks around 13 years and BDS effect is minimal (though significant). Class E use peaks at the highest Age and BDS levels in the sample. The ordering based on the average ages for each class (see Table 3) lead to the sequence A–E. Fig. 4 suggests the same developmental ordering but also suggests class C can be better left out.

Discussion

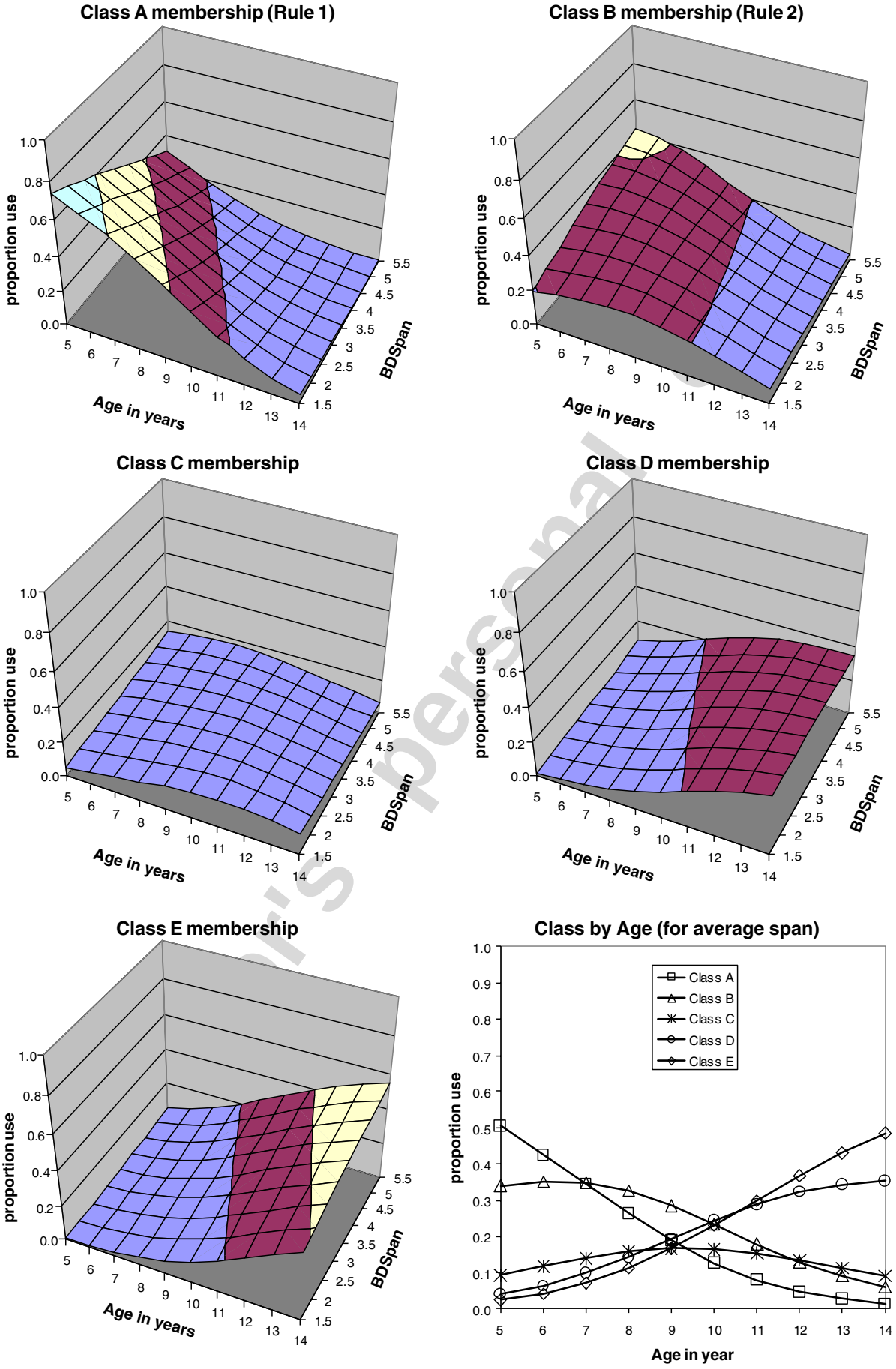
Previous findings that the classes are well demarcated and display a rough age trend were replicated. We fitted a two-dimensional, multinomial, logistic curves model with covariates to a large balance scale task data set. Below we discuss the implications of the findings for the theoretical issues raised above.

Working memory

The increase in average backward digit-span (BDS) with age was in accordance with published accounts (Demetriou et al., 2002; Morra, 1994). Consistent with recent literature on memory development (Gathercole, 1998), working memory explained some of the variance in class membership; effects on class membership could be attributed to backward digit-span (see below). Interaction between Age and BDS had no significant effects. However, this interaction was not far from being significant and we recommend considering the possible interaction effect again in future studies.

Modeling developmental trends

The parameter estimates of the logistic regression analysis part in the model, for each class, for age and working memory, characterize developmental trends that are visualized in Fig. 4a–e. The effect of age, given average working memory, is depicted in Fig. 4f. From these figures, we can derive a developmentally meaningful ordering of the classes. We assume that simple forms of thinking on this proportionality reasoning task will peak with lower age and working memory, whereas more complex forms of thinking will peak with higher age and working memory. Fig. 4a shows, as expected, that membership to this class A decreases with increasing age and increasing working memory. This makes sense because this class represents the least advanced rule: Rule 1 only considers one (the dominant) dimension of the task. Fig. 4b shows class B for which the membership is at its maximum for a combination of low age and high working memory although the effect of working memory is not significant. The use of this class, associated with Rule 2, drops at a slightly later age than the use of class A. This finding makes sense because Rule 2 is more advanced than Rule 1. Rule 2 is like Rule 1, but in addition implies the ability to switch to the subordinate dimension for non-conflict items. Fig. 4c shows that class C is not affected by age or working memory. Since this class has no clear response pattern, we feel confident to conclude that it represents a heterogeneous rest category. Therefore,



this class does not contribute the developmental ordering. Fig. 4d shows class D for which the membership increases slightly with increasing age and is weakly affected by working memory. This class was characterized as resembling both Rule 2 and Rule 3. The age and working memory trend is consistent with a designation between those two rules. Fig. 4e shows class E for which the membership increases with increasing age and working memory. This class represents the most advanced class in our sample in terms of age and working memory demands. This is consistent with a resemblance to the compensation rule and Rule 3. The compensation rule and Rule 3 are the most complex rules (in our data) since they require the simultaneous consideration of the two important dimensions of the task. We are confident that a relatively clear class representing Rule 4 would have shown up if the age range of participants would had been extended to higher ages (see Boom et al., 2001). We deliberately did not include high school participants in our study because in the Dutch educational system the mechanical principles underlying the balance task are often, but not always, part of the high school curriculum.

Classification

We hypothesized that we would be able to improve classification of responses to the balance scale task. To our surprise, we have to conclude that classification of the individual response patterns does not improve much by including the covariates. Unambiguous classification of participants to the classes, based on answer patterns without the covariates age and working memory, turns out to be reasonably good already, such that improvement is difficult to achieve: there was no improvement of the classification statistics. We also hypothesized that it was possible to estimate the magnitude and significance of the effects of age and working memory on class membership and in this we have succeeded. Age, and to a lesser degree working memory, are strong predictors of the use of classes (see Fig. 4 and in particular f), although their proportional reduction in error was not impressive. Note however, that the class-profiles themselves remain the same with or without these covariates. Apparently, classes are defined by the answer patterns and not much affected by the covariates. Almost all (99%) participants end up being assigned to the same class whether or not covariates are added to the model. Nevertheless, adding the covariates was useful because fewer classes were needed after adding the covariates, the fit of the models was slightly better, and in particular the bootstrap results were more positive. In sum, we found that these classes can be ordered in a developmental sequence based on age and working memory demand, but that classification statistics are hardly affected by these covariates.

Classes or rules?

The classes uncovered by LCA in this dataset do not align exactly with rules as introduced by Siegler in the 1980s. Nevertheless, these classes display a developmental ordering in terms of age and working memory demands. This ordering makes sense if we assume

Fig. 4. Five-class model. Effect of Age and Backward Digit-Span (BDS) on class A to E is shown in (a) to (e). The effect of Age, for average BDS, is shown in lower-right corner (f). The classes are as in Fig. 3. Please note, that standard errors or confidence intervals are not visualized to avoid cluttered graphs. The bands of white and gray shades are indicative of height (proportion use).

that classes D and E represent mixed rule use. For our developmentally most advanced class E, the chances for particular predictions to the items are as close to Rule 3 as to the compensation rule. However, note that this mix is a complex mix. First, we have explored the possibility that this class would have broken down in two new classes in models with more classes, but found no indication for such a state of affairs. The six-class model only added a small (5%) class without clear patterning and there was no sign of breaking up the classes designated here by mixed rule use. Second, for class D, the second most advanced class, the chances for particular predictions to the items are as close to Rule 2 as to the Rule 3, and for class E these chances are as close to the compensation rule as to the Rule 3. Third, pure rule use requires that the chances for correct as in Fig. 3 are either high (near one) or low (near zero). Therefore, efforts to invoke other rules to explain this task behavior are futile, as classes B to E do not comply with this basic requirement. Nevertheless, the patterns as shown in Fig. 3 are very stable. First, because these patterns as described in Fig. 4 are almost unaffected by the covariates. In addition, allowing for more classes does not alter the basic profile of the five classes already described. Finally, comparable results were found by Boom et al. (2001) and in datasets collected by other cohorts of our students with only slightly different item sets (van Vliet, Boom & Brand, submitted for publication). Perhaps more insight in these mixed classes would be obtainable with equality constraints added to the models. According to Jansen and Van der Maas (2002) more specific hypothesis can be tested in such a manner. It is a disadvantage of Latent Gold 3.0 that this is not possible.

Overlapping waves

We found strong support for our version of the overlapping waves model. Our version of the overlapping waves model is precise, testable, fits cross-sectional data, and has been extended to a two-dimensional variant that includes working memory. Our model is similar to the *overlapping waves* model as proposed by Siegler in that it pretends to model development, extends it in some respect, but is more limited because we do not pretend to depict individual development. Siegler (1996) pitted his own overlapping waves model against the traditional stage or staircase model. The traditional staircase model is clearly meant to be applicable to the individual level and a sequence of single strategies. The overlapping waves model is a probabilistic model, and therefore, by definition, refers to multiple strategies. Whether these multiple strategies must be located in an individual or whether they may represent different individuals is not clear. Anyhow, we do not consider our overlapping waves model to be the opposite of a stage model because a model for individual development cannot be inferred from cross-sectional group data if the basic relationships are non-linear as is clearly the case here. This means that no unambiguous conclusions can be derived from the shape of the average trajectory. We cannot infer from our data what the shape of the individual trajectories is because this cannot deduced from cross-sectional data (Keats, 1983; Singer & Willett, 2003). Whether individual curves are steeper up to the point where they approach the form of a staircase cannot be ruled out from group data.

Jansen and van der Maas encountered similar problems with non-pure strategy classes as we did and they proposed a restricted overlapping waves model, which combines rather abrupt transitions with gradual overlapping transitions, to accommodate these findings (Jansen & Van der Maas, 2001, 2002). However, their model, meant to describe the

idealized development of an individual, remains a hypothetical and not a formal statistical model based on real data. For support they cannot rely on the group based age trends, since, as pointed out above, there is no guarantee that individual developmental trends have the same shape as group trends. Moreover, their model is presented as referring to rules whereas our model sticks to the data and is modeling classes only. More troublesome is that they did not use all items in the LCA, which makes it possible that classes not representing any known rule may have remained undetected.

Conclusion

We were able to fit a two-dimensional, multinomial, logistic curves model, with covariates, to a large balance scale task data set. The model generalizes and formalizes the idea of overlapping waves to describe development in proportional reasoning, resulting in a typical rise and decline pattern for class membership dependent upon *age* and *working memory* capacity increases. This pattern is visualized in three-dimensional surface graphs in order to facilitate interpretation. Latent class analysis combined with logistic regression is a powerful new tool for developmental data.

We must conclude that, since including the covariates did not substantially alter the class profiles, the covariates do not throw much new light on the interpretation of the problematic classes. In fact, these results concerning the covariates only confirm the existence and robustness of these difficult to characterize classes. Complex classes, not representing known rule use for responses to the balance scale task, could not be explained away by considering the covariates age and working memory. In contrast: mixed rule use appears to be a robust phenomenon with a clear position in the developmental sequence. Unfortunately, therefore, the result of this investigation is that problems have become more prominent and less ignorable. Perhaps it is time to admit that the seemingly well established Rules 2 and 3 as coined by Siegler long ago cannot be used any longer to distinguish groups of children using different cognitive strategies. Much more individual variability must be reckoned with than the notion of simple rule use suggest. Interestingly, this conclusion is in line with Siegler's more recent ideas (Siegler, 1996).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.dr.2006.06.001.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: Wiley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Boom, J. (2002). Why longitudinal research is impossible for cognitive strategies and what to do instead. Paper presented at the 32nd Annual Meeting of The Jean Piaget Society, Philadelphia, PA.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: classes, strategies, or rules for the Balance Scale Task?. *Cognitive Development 16*, 717–735.

Case, R. (1992). *The mind's staircase*. Hillsdale, NJ: Erlbaum.

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., et al. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development*, 61 (Serial No. 246).

Chletsos, P. N. (1986). *A paper-and-pencil test replicating Siegler's rule assessment approach on Piaget's balance beam task. Instruction manual*. New York: Rutgers University.

Chletsos, P. N., & De Lisi, R. (1991). A microgenetic study of proportional reasoning using balance scale problems. *Journal of Applied Developmental Psychology, 12*, 307–330.

Demetriou, A., Christou, C., Spanoudis, G., & Platsidou, M. (2002). The development of mental processing efficiency, working memory, and thinking. *Monographs of the Society for Research in Child Development*, 67 (Serial No. 268).

Eckstein, S. G. (2000). Growth of cognitive abilities: dynamic models and scaling. *Developmental Review, 20*, 1–28.

Ferreti, R. P., & Butterfield, E. C. (1986). Are children's rule-assessment classifications invariant across instances of problem types? *Child Development, 57*, 1419–1428.

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology, 54*, 1–34.

Gathercole, S. E. (1998). The development of memory. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 39*, 3–27.

Haberman, S. J. (1974). Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. *Annals of Statistics, 2*, 911–924.

Halford, G. S., Andrews, G., Dalton, C., Boag, C., & Zielinski, T. (2002). Young children's performance on the balance scale: the influence of relational complexity. *Journal of Experimental Child Psychology, 81*, 417–445.

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.

Keats, J. A. (1983). Ability measures and theories of cognitive development. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick M. Lord* (pp. 81–101). Hillsdale, NY: Erlbaum.

Laudy, O., Boom, J., & Hoijtink, H. (2004). Bayesian computational methods for inequality constrained latent class analysis. In L. A. v. d. Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 63–82). Mahway, NJ: Erlbaum.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford: Oxford University Press.

McCutcheon, A. L. (1987). *Latent class analysis*. Beverly Hills, CA: Sage.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks: Sage.

Morra, S. (1994). Issues in working memory: testing for M-capacity. *International Journal of Behavioral Development, 17*, 143–160.

Normandeau, S., Larivee, S., Roulin, J. L., & Longeot, F. (1989). The balance-scale dilemma: either the subject or the experimenter muddles through. *Journal of Genetic Psychology, 150*, 237–250.

Pascual-Leone, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica, 32*, 301–345.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review, 51*, 145–146.

Rest, J. R. (1979). *Development in judging moral issues*. Minneapolis: University of Minnesota Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, 46 (Serial No. 189).

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press.

Van der Heijden, P., t Hart, H., & Dessens, D. (1997). A parametric bootstrap procedure to perform statistical tests in latent class analysis of antisocial behavior. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class analysis in the social sciences*. Munster: Waxman.

Van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology, 85*(2), 141–177.

van Geert, P. (1994). *Dynamic systems development: Change between complexity and chaos*. Hempstead: Harvester Wheatsheaf.

van Vliet, E., Boom, J., & Brand, A.N. (subm). Stability of latent class analysis classes for the balance scale task.

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

Vermunt, J. K., & Magidson, J. (2003a). *Addendum to the latent GOLD user's guide: Upgrade manual for version 3*. Belmont MA: Statistical Innovations Inc..

Vermunt, J.K., & Magidson, J. (2003b). Technical appendix for latent GOLD 3.0: http://www.statisticalinnovations.com/.

Wechsler, D. (1974). *Wechsler intelligence scale for children-revised*. New York: Psychological Corporation.

Wilkening, F., & Anderson, N. H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin, 92*, 215–237.

Wolters, M., Fischer, C., & Zuidema, J. (1987). Balanced measurement of cognitive development: a discussion on methodological problems with the balance scale. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 9*, 114–120.