PSYCHOMETRIKA—VOL. 70, NO. 4, 651–669
DECEMBER 2005
DOI: 10.1007/s11336-002-1047-z

AN IRT MODEL WITH A PARAMETER-DRIVEN PROCESS FOR CHANGE

FRANK RIJMEN
KATHOLIEKE UNIVERSITEIT LEUVEN, BELGIUM

PAUL DE BOECK
KATHOLIEKE UNIVERSITEIT LEUVEN, BELGIUM

HAN L.J. VAN DER MAAS
UNIVERSITY OF AMSTERDAM, THE NETHERLANDS

An IRT model with a parameter-driven process for change is proposed. Quantitative differences between persons are taken into account by a continuous latent variable, as in common IRT models. In addition, qualitative interindividual differences and autodependencies are accounted for by assuming within-subject variability with respect to the parameters of the IRT model. In particular, the parameters of the IRT model are governed by an unobserved or "hidden"' homogeneous Markov process. The model includes the mixture linear logistic test model (Mislevy & Verhelst, 1990), the mixture Rasch model (Rost, 1990), and the Saltus model (Wilson, 1989) as specific instances. The model is applied to a longitudinal experiment on discontinuity in conservation acquisition (van der Maas, 1993).

Key words: latent markov model, item response theory, panel data.

In longitudinal data, observations made on different subjects tend to be more heterogeneous than observations within the same subject. As a result, repeated observations of the same individuals are not statistically independent. Traditional latent class analysis (LCA) and item response theory (IRT) models are models for categorical longitudinal data that take the between-subject variability into account by incorporating latent variables. In LCA, the latent variables are categorical, each value representing a latent class. In IRT, the latent variables are continuous. LCA is best suited for modeling qualitative individual differences. Each latent class then represents a qualitatively distinct subgroup. In contrast, the continuous latent variables of IRT models naturally take into account quantitative individual differences.

Common to both types of models is the assumption of conditional (local) independence: the assumption that the responses on the items are independent of each other conditional on the latent variables. That is, all dependencies are assumed to stem from between-subject heterogeneity, which is taken into account by the latent variables of the model. However, another common finding in longitudinal data is that observations made closer in time are more homogeneous than observations made further apart. These autodependencies form a second source of dependencies in the data violating the assumption of conditional independence. Depending on the hypothesized nature of the autodependencies, LCA and IRT models have been extended in several ways to account for this extra source of dependencies. In this respect, Eid and Hoffmann (1998) distinguish between two kinds of models for longitudinal data. In *process-oriented models*, changes over time in response behavior (and hence autodependencies) are related to an underlying time-related process. An example of such a process is learning during the test as an effect of solving earlier

Frank Rijmen was supported by the Fund for Scientific Research Flanders (FWO), the GOA/2000/02 granted by the Katholieke Universiteit Leuven to Paul De Boeck and Iven Van Mechelen, and the PDM/02/067 granted by the Katholieke Universiteit Leuven to Paul De Boeck.

Requests for reprints should be set to frank.rijmen@psy.kuleuven.be.

© 2005 The Psychometric Society

items. In *variability models* (e.g., Steyer, Ferring, & Schmitt, 1992) on the other hand, differences over time are not explained in terms of alterations (e.g., learning, development, etc.) within persons, but as occasion-specific fluctuations around a (person-specific) constant level. Hybrid models for categorical longitudinal data, allowing for both variability and change, have been proposed by Eid and Hoffmann (1998) and by Eid and Langeheine (1999; 2003).

Process-oriented models can further be categorized into models incorporating observation- vs. parameter-driven processes (Cox & Snell, 1996). In models with an observation-driven process, the probability of a response is (partly) determined by the previously given responses. In models incorporating a parameter-driven process on the other hand, the response probabilities are (partly) determined by an unobserved process over time, taking place on the parameters. In the latter case, autodependencies are a consequence of dependencies between the parameters.

Within the context of LCA, models that incorporate a parameter-driven process were described by Collins and Wugalter (1992), Hagenaars (1998), Humphreys (1998), Langeheine and van de Pol (1990; 1994), Poulsen (1990), and Thomas and Hettmansperger (2001). Pioneering work was done by Wiggins (1973). With the model of Humphreys as a notable exception, these models only incorporate categorical latent variables and hence are less suited to model quantitative interindividual differences.

Within the domain of IRT, the dynamic Rasch model of Verhelst and Glas (1993) is an IRT model in which the response probabilities depend on past responses, and hence, autodependencies are accounted for by an observation-driven process. Furthermore, Böckenholt (1999) extended the latent trait model of Samejima (1969) for ordinal data with an observation-driven process. Contrary to the process-oriented latent class models, these models only incorporate continuous latent variables and hence fail in accounting for qualitative individual differences.

The model for binary longitudinal data that we present in the next section can either be seen as an extension of a latent class model with a parameter-driven process for autodependencies by taking into account quantitative individual differences, or as an extension of an IRT model by incorporating a parameter-driven process for change to account for qualitative inter- and intra-individual differences. More specifically, the responses are modeled through a linear logistic test model (LLTM) (Fischer, 1973), the parameters of which are governed by an unobserved, latent, or "hidden" homogeneous Markov process. Quantitative individual differences are taken into account by the continuous latent variable of the LLTM. Qualitative interindividual differences are attributed to the fact that persons can be in a different state of the Markov process at any given moment in time. Autodependencies arise in the model because it is more likely that a person will be in the same state for measurements that are taken close together in time. After presenting the model, relations with existing LCA and IRT models are discussed.

The model is applied to a longitudinal experiment on discontinuity in conservation acquisition (van der Maas, 1993). In the application, the stages of nonconservation and conservation correspond to the two unobserved states of a hidden Markov model. Within each stage, the responses are modeled through a LLTM. The transition from the former to the latter stage is reflected by a change in the parameters of the LLTM.

1. The Model

According to the model, there are three determinants of the probability that a person n ($n = 1, \ldots, N$) obtains a score of one on an item i ($i = 1, \ldots, I$):

1. The unobserved *state* ($s = 1, \ldots, S$) the person is visiting while solving the item. States correspond to the classes of latent class analysis. However, unlike latent class analysis, persons can switch states across time. The state a person n is visiting while solving item

i is denoted by the qualitative latent variable ζ_{ni} ($\zeta_{ni} = 1, \ldots, S$). As a consequence, the vector $\boldsymbol{\zeta}_n = (\zeta_{n1}, \ldots, \zeta_{ni}, \ldots, \zeta_{nI})'$ denotes the chain or trajectory through the state-space of person n. The state exerts its influence indirectly, both through the state-specific effects of the item characteristics and through the state-specific person-effect, as explained in determinants 2 and 3.

2. The *characteristics of the item*, such as, for example, the kind of elementary arithmetic operations to be carried out, and the number of unknowns in an algebraic problem. We denote the values of item i on a set of P item characteristics by the P-dimensional item covariate vector $\boldsymbol{x}_i$. The $I \times P$ matrix $\boldsymbol{X}$, with as the ith row $\boldsymbol{x}_i'$, is called the design matrix. Because item characteristics may play a different role in different states, the weights of the item characteristics are allowed to be different across states, and are denoted by the P-dimensional parameter vectors $\boldsymbol{\beta}_s$.
3. The *person*. Apart from differences in the trajectories through the state-space, persons may differ with respect to an overall tendency for obtaining a score of one. The person-effect is modeled through the continuous latent variable θ_{ns}. As the subscript s indicates, the person-effect is allowed to be state-specific.

Now, conditional on the unobserved state s a person n is visiting while solving item i, and conditional on the state-specific continuous latent variable θ_{ns}, the responses Y_{ni} are assumed to be independent Bernoulli realizations, where the logit of the probability of "success" is modeled as

$$\log\left(\frac{\Pr(Y_{ni}=1|\zeta_{ni}=s,\theta_{ns},\boldsymbol{x}_i,\boldsymbol{\beta}_s)}{1-\Pr(Y_{ni}=1|\zeta_{ni}=s,\theta_{ns},\boldsymbol{x}_i,\boldsymbol{\beta}_s)}\right)=\theta_{ns}+\boldsymbol{x}_i'\boldsymbol{\beta}_s. \tag{1}$$

This means that, within a latent state, the responses are modeled through a LLTM (Fischer, 1973) or, equivalently, through a logistic regression model with a person-specific intercept θ_{ns} (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). Assuming that θ_{ns} is a normally distributed conditional on state s, the marginal probability of a response pattern $\boldsymbol{y}_{ns}$ on the subset I_{ns} of items that were responded to while visiting state s equals

$$\Pr(\boldsymbol{y}_{ns}|\boldsymbol{x},\boldsymbol{\beta}_s)=\int\left\{\prod_{i:y_{ni}\in \mathrm{I}_{ns}}\frac{\exp[y_{ni}(\theta_{ns}+\boldsymbol{x}_i'\boldsymbol{\beta}_s)]}{1+\exp(\theta_{ns}+\boldsymbol{x}_i'\boldsymbol{\beta}_s)}\right\}N(\theta_{ns}|0,\sigma^2)d\theta_{ns}. \tag{2}$$

The constraint $E(\theta_{ns})=0$ is not a real constraint but a matter of convention. The mean of θ_{ns} is treated as a fixed effect by including the intercept in $\boldsymbol{x}_i$, the item covariate vector. θ_{ns} then represents the person-specific deviation from the mean intercept within state s. Finally, marginalizing over all K possible trajectories $\boldsymbol{\zeta}_n$ of a person through the state-space, the probability of a response pattern $\boldsymbol{y}_n$ is

$$\Pr(\boldsymbol{y}_n|\boldsymbol{X},\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_s)=\sum_{k=1}^{K}\Pr(\boldsymbol{\zeta}_n=k|\boldsymbol{\pi}_0,\boldsymbol{A})\left[\prod_{s=1}^{S}\Pr\left(\boldsymbol{y}_{ns}\left|\boldsymbol{X},\boldsymbol{\beta}_s\right.\right)\right], \tag{3}$$

where we define $\Pr(\boldsymbol{y}_{ns}|\boldsymbol{X},\boldsymbol{\beta}_s)\equiv 1$ if $I_{ns}=\phi$.

The chain $\boldsymbol{\zeta}_n$ is assumed to follow a homogeneous first-order Markov model with transition matrix A and initial distribution $\boldsymbol{\pi}_0$,

$$\boldsymbol{\pi}_t=\boldsymbol{\pi}_{t-1}\boldsymbol{A}, \tag{4}$$

where the index t, $t=1,\ldots,T$, refers to the time point or occasion a transition between classes is allowed, and $\boldsymbol{\pi}_t$ is the marginal probability distribution of the states at time point t. It has no person index because we assume that the same Markov model applies to all persons. The transition matrix $\boldsymbol{A}$ contains the probabilities of switching from one state to another. It bears

no time index because we assume a homogeneous Markov model, a Markov model for which the transition probabilities are independent of time point t. The Markov model is defined on the unobserved parameters of the LLTM and is therefore called a latent or hidden Markov model.

The occasions when a transition between states is allowed can refer to "real" time points, for example, when a test is taken repeatedly with regularly spaced time intervals. However, one could also consider the individual items of a test as occasions. In the latter case, the index t equals the item index i.

Changes across time that are equal for all persons can be incorporated into the model by defining the appropriate item covariate(s). For example, one could include the position of an item in a test as an item predictor to account for a (state-specific) learning effect that is constant across persons. Or, when the same test is administered repeatedly, dummy item covariates coding for the test occasion could be included to account for (again, state-specific but person-invariant) changes between test occasions.

The presented model is easily extended in three respects. First, some of the logistic regression weights (contained in the β-vectors) can be restricted to be the same across (some of the) states by using a common parameter for them. Second, the design matrix $\boldsymbol{X}$ can be made state-specific. This means that states not only mutually differ in terms of the effects item characteristics have on the probability of success, but also in terms in which item characteristics do play a role at all. Finally, the variances of the continuous latent variables θ_{ns} can be made state-specific. See the expectation-maximization-algorithm (EM-algorithm) described in the Appendix for more details.

2. Estimation and Hypothesis Testing

Treating the values of the persons on the continuous latent variables θ_{ns} and the unobserved Markov chains $\boldsymbol{\zeta}_n$ of the persons as missing data, an EM-algorithm (Dempster, Laird, & Rubin, 1977) can be used to find the stationary points of the log-likelihood function. A detailed description is given in the Appendix. This also describes how the observed information matrix and standard can be obtained within the EM-framework (McLachlan & Krishnan, 1997). It is common to use the observed information matrix as an approximation of the expected information matrix (McLachlan & Peel, 2000, p. 42). To avoid ending up with a local maximum solution, it is advisable to run the EM-algorithm several times with different starting values.

A necessary condition for the identifiability of the model that can be checked prior to any parameter estimation is that the design matrix $\boldsymbol{X}$ should be of full column rank. After the estimation is carried out, the expected information matrix conveys information on the identifiability of the model. If the latter is positive definite at a stationary point of the log-likelihood function (all eigenvalues larger than zero), the point corresponds to a (local) maximum and the model is locally identified (Formann, 1985, 1992; Goodman, 1974; McHugh, 1956).

The global goodness-of-fit of the model can be assessed by a likelihood-ratio, Pearson chi-squared, or other test from the power-divergence family of tests for categorical (dependent) variables (Read & Cressie, 1988). However, the number of possible response patterns will be very large in comparison to the sample size whenever the test contains more than a few items (e.g., for 10 dichotomous items, the number of possible response patterns equals 1024). In such cases of sparseness, one cannot rely on the asymptotic result that the statistics from the power-divergence family follow a chi-squared distribution. Instead, parametric bootstrap procedures have been proposed to approximate the reference distribution of the test statistics (Aitkin, Anderson, & Hinde, 1981; Collins, Fidler, Wugalter, & Long, 1993; Langeheine, Pannekoek, & van de Pol, 1996; von Davier, 1997). Based on the results of a simulation study, von Davier concluded that the Pearson and Read–Cressie chi-squared statistics are superior to the likelihood-ratio and Freeman–Tukey statistics in a bootstrap procedure when the data are sparse. It is unclear, however, to what extent

these results can be generalized to other models than those investigated in the simulation study. Furthermore, the parametric bootstrap technique can become very time-consuming, because it involves a reestimation of the parameters for each simulated dataset.

A hypothesis with respect to a subset of the parameters can be tested by comparing the model that is restricted according to the hypothesis with the more general alternative model to which those restrictions do not apply. In this regard, one can rely on the likelihood-ratio test, as well as on the score and Wald tests, which are based on an approximation of the log-likelihood function by a second-order Taylor expansion (Fahrmeir and Tutz, 2001). Under regularity conditions, all three statistics are asymptotically chi-squared distributed with, as degrees of freedom, the difference in the number of parameters between the unrestricted and restricted models. An important case in which regularity conditions are not met is when the restricted model is a boundary case of the unrestricted model. This occurs when testing for the need of a continuous latent variable within the states ($H_0 : \sigma^2 = 0$) or when comparing a model with S states to a model with $S + 1$ states. In the former case, the asymptotic distribution of the likelihood-ratio test is a mixture of a chi-squared distribution with, respectively, zero and one degree of freedom, both mixture components having an equal weight of 0.5 (Snijders & Bosker, 1999; Verbeke, 1997). The chi-squared distribution with zero degrees of freedom is a distribution with point mass one at zero, so that the appropriate p-value is obtained by halving the "classical" p-value that is obtained when testing a hypothesis for a single parameter. For the second case, for which the restricted model is on the boundary of the unrestricted model, when comparing a model with S states to a model with $S + 1$ states, matters are even more complicated because the restricted model corresponds to a nonidentifiable subset of the parameter space of the unrestricted model. Not much is known about the reference distribution of the likelihood-ratio statistic in this case, except for some specific cases (McLachlan & Peel, 2000).

Again, parametric bootstrap procedures could be used to simulate the appropriate reference distribution. However, the computational effort is even larger now than when testing for the global goodness-of-fit, because both the restricted and unrestricted models have to be reestimated for each simulated dataset.

No statistical tests are available for comparing nonnested models. In the absence of sound statistical tests, one frequently relies upon information criteria such as the AIC (Akaike, 1974) and BIC (Schwarz, 1978).

3. Related Models

When $\boldsymbol{A}$ is the identity matrix, no transitions occur, and hence there is no within-subject variability with respect to the parameters of the model. In this case, the model reduces to the mixture LLTM of Mislevy and Verhelst (1990), or to a specific instance of the latter, such as the mixture-Rasch (Rost, 1990) and the Saltus models (Mislevy & Wilson, 1996; Wilson, 1989). Specifically, if $\boldsymbol{A}$ is the identity matrix, there are only S possible sequences of states: always in state $1, \ldots$, always in state S. Hence, $K = S$, I_{ns} is either empty or contains all items, and $\Pr\left(\boldsymbol{\zeta}_n = k \,|\boldsymbol{\pi}_0, \boldsymbol{A}\right) = \pi_{0s}$, the initial probability of being in state s. Equations (3) can then be simplified to

$$\Pr(\boldsymbol{y}_n|\boldsymbol{X}, \boldsymbol{\beta}_1, \ \ldots \ , \boldsymbol{\beta}_s) = \sum_{s=1}^{S} \pi_{0s} \Pr(\boldsymbol{y}_n|\boldsymbol{X}, \boldsymbol{\beta}_s),$$

which is the marginal probability of a response pattern according to the mixture LLTM. In the mixture-Rasch model, $\boldsymbol{X}$ is an identity matrix. The Saltus model is a mixture-Rasch model with restrictions on the item parameters (the elements of $\boldsymbol{\beta}_s$) across classes.

If no continuous latent variables within the states are assumed ($\sigma^2 = 0$), a restricted latent Markov model is obtained (Collins & Wugalter, 1992; Langeheine and van de Pol, 1994; Wiggins, 1973). The latent Markov model is restricted in the sense that the transition matrix is time homogeneous, and that the state-specific response probabilities are modeled through a logistic regression. Both kinds of restrictions stem from the fact that our model is intended for cases where there are more than a few measurement occasions and/or more than one or a few items for each occasion, unlike the cases usually considered in the context of (unrestricted) latent Markov models. Without restrictions, the number of parameters would become very large with an increasing number of items and/or measurement occasions.

Humphreys (1998) also proposed a latent Markov model that incorporates continuous latent variables (random effects) to capture unobserved heterogeneity. Humphreys' model differs in several respects from the model that is proposed in this paper. First, Humphreys introduces three kinds of random effects, in order to account for heterogeneity in, respectively, the initial probability of belonging to a state, the transition probabilities between classes over time, and the class-specific response probabilities. We only take the heterogeneity in class-specific response probabilities into account. Second, Humphreys introduces separate, possibly correlated, random effects for each time point whereas, in our model, the continuous latent variable within a state is the same over time points. To overcome the computational burden for such high-dimensional models, Humphreys assumes a log-log link function for the class-specific response probabilities and a log-Gamma distribution for the random effects. This way, the integrals over the random effects can be solved analytically. A drawback of the log-Gamma distribution for the random effects is that only positive correlations between random effects can be taken into account. Third, in Humphreys' model, the response probabilities are modeled as a function of observed person characteristics and unobserved random effects, but not as a function of characteristics of the items. In contrast, in the model we propose, the response probabilities are restricted through a logistic regression model with item characteristics as predictors. Because the response probabilities are not further restricted in Humphreys' model, it is, like the latent Markov model, less suited for cases where there are more than a few measurement occasions and more than one or a few items for each measurement occasion.

3.1. Application: Longitudinal Experiment on Discontinuity in Conservation Acquisition

The data come from a study of van der Maas (1993; van der Maas & Molenaar, 1995) on discontinuity in conservation acquisition. First, we shortly describe the study.

4. Description of the Study

4.1. Theoretical Background

The study was conducted to assess the existence of stages in cognitive development, as put forward in the stage theory of Piaget (1960). The occurrence of transitions (i.e., abrupt changes), in behavior is an immediate implication of the existence of stages. Van der Maas and Molenaar (1992; 1995) described eight empirical criteria or "flags" for detecting transitions, all of them derived from catastrophe theory (Gilmore, 1981; Thom, 1975). Two main flags are (bi)modality in behavior across persons, reflecting the stages the persons are in, and a sudden jump in the behavior of individual persons, reflecting the transition from one stage to the other. However, both flags together do not differentiate a transition between qualitatively different stages from a continuous acceleration or step-wise growth, that is, a shift to an underlying quantitative dimension (for an extended discussion of the differences between models for discontinuous changes, see Jansen

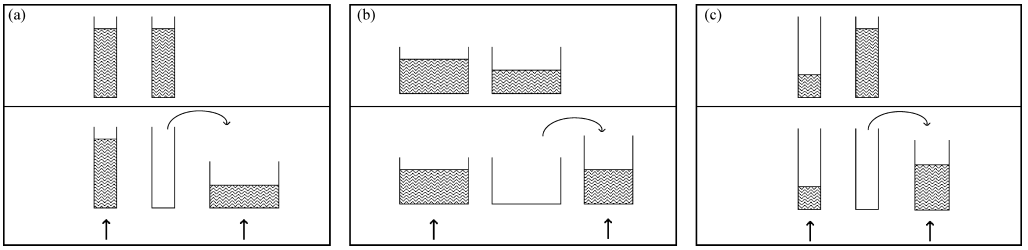


FIGURE 1.
Items of the liquid conservation test. (a) Standard equality item; the correct answer is "equal." (b) Standard inequality item; the correct answer is "left more." (c) Guessing item; the correct answer is "right more." Standard items discriminate between conservers and nonconservers. Guessing items between guessers, and conservers and nonconservers. The top two glasses form the initial situation, the bottom three the final situation.

and van der Maas, 2001). Instead of looking at other flags to make the distinction, we will explain how the distinction can be made based on the model presented in this paper.

4.2. Computer Test of Conservation of Continuous Quantity

Conservation is the invariance of a characteristic despite transformation of the object or collection of objects possessing this characteristic (Field, 1987). For example, liquid conservation refers to the fact that pouring liquid from one glass into another glass of a different shape does not change the quantity of liquid.

In the computer test of conservation of continuous quantity (van der Maas, 1993), children are presented with an initial situation (a two-dimensional display of two similar glasses filled with a liquid), a transformation (pouring one of the glasses in an empty glass of different dimensions), and the resulting final situation. Children have to judge both the initial and final situations. They can choose between three response alternatives: the left glass contains more liquid, the glasses contain the same quantity of liquid, or the right glass contains more liquid. Different types of items are part of the test to differentiate between conservers and nonconservers, and between different conserver and nonconserver strategies. We only make use of the responses to the final situation of the items belonging to the first part of the test, which contained, in the study, three equality standard items, three inequality standard items, and two guessing items.

The standard items discriminate conservers from nonconservers and guessers. An example of both equality and inequality standard items is shown in Figure 1(a) and (b). In a standard equality item (Figure 1(a)), the two similar glasses of the initial situation contain the same quantity of liquid. Conservers are aware that the transformation (pouring one glass into another of a different shape) does not change the quantity, and are expected to give the correct answer that both glasses of the final situation contain the same quantity of liquid. Nonconservers, in contrast, are expected to focus on the height, resulting in the incorrect answer that the glass with the highest level contains more liquid (the left glass in Figure 1(a)). For standard inequality items (Figure 1(b)), the glasses do not contain the same quantity, but the level of the glasses in the final situation is of equal height, so that nonconservers are expected to answer incorrectly that both glasses contain the same quantity of liquid. Again, conservers, being aware of the conservation principle, are expected to give the correct answer that (in Figure 1(b)) the left glass contains more liquid. Guessers will score at chance levels on both types of equality items. Guessing items, see Figure 1(c), should be solved correctly by conservers as well as by nonconservers. The height is not a misleading cue in this case: the glass with the highest level contains more liquid (the right glass in Figure 1(c). Again, guessers will score at chance level.

We call this test a test of conservation of continuous quantity and not a liquid conservation test because, as an anonymous reviewer pointed out, we cannot be sure that children interpret this task as a liquid conservation, rather than an area conservation, task in view of the two-dimensional display of the stimuli. Whether interpreted as a liquid conservation task or as an area conservation task bears no consequences as to which responses are to be expected from conservers and nonconservers, however.

4.3. Method

One hundred and one children, between 6.2 and 10.6 years of age at the beginning of the experiment (the mean age was 7.81, with a standard deviation of .91), took the test repeatedly. There were 11 test sessions, with a time lag of 2 to 4 weeks. There was a large number of missing sessions, 265 out of 1111.

4.4. Results and Discussion

We report only on the results that are relevant for the analysis we will present. Fitting mixtures of binomial probability distributions on the sum scores, separate for each session, van der Maas (1993) found that a two-component mixture provided the best fit at all sessions. The estimated probabilities of success of the two components were fairly constant across sessions. For one component, all estimated probabilities of success lay around .3. For the other component, all estimated probabilities of success were close to 1. Hence, the first component can be interpreted as a group of nonconservers, because they are expected to solve only the two guessing items but not the six standard items, and the second component as a group of conservers solving all the items. The proportion of subjects in the conservation mode shifted from .16 in the first session to .41 in the last session.

Based on the responses given to the individual items, the children were, for each session, classified according to the conserver or nonconserver strategy used. See van der Maas (1993) and van der Maas and Molenaar (1995) for a thorough description of the classification scheme. Forty-two children (proportion = .42) were classified as being nonconserver all the time, 20 children (proportion = .20) as being conserver from the beginning, 31 children (proportion = .31) were classified as transitional (starting as nonconserver, becoming conserver later on), and finally 8 children (proportion = .08) were classified in a residual group. Hence, according to these classification results, the proportion of conservers shifted from .20 (.22 when the 8 children of the residual category were not taken into account) in the first session toward .51 (.55) in the last session.

In order to differentiate between a stage transition and a continuous acceleration or stepwise growth, van der Maas (1993) assessed some other flags of transition. We do not report on the latter, but instead we will show how this issue can be addressed using the model proposed in this paper.

4.5. Reanalysis of the Conservation Data

The general model we presented in the previous sections was tailored to the application as follows. Because van der Maas (1993) found that a two-component mixture provided the best fit at all sessions, the hidden Markov model contained two states. The transition matrix was defined as $\boldsymbol{A} = \begin{bmatrix} a_{11} & 1 - a_{11} \\ 0 & 1 \end{bmatrix}$ reflecting the expectation derived from the stage theory that only transitions from "lower" to "higher" stages occur. Hence, the second state is an *absorbing* state (Ross, 1996): once it is entered, it is never left. The model assumption of a homogeneous Markov chain is reasonable, given that the time periods between successive sessions are comparable in

magnitude. Each session was considered as an occasion on which a transition was possible. Hence, there are 12 possible chains: being nonconserver all the time, being conserver all the time, or making a transition from nonconserver to conserver after session $1, \ldots, 10$.

Initially, three different models were estimated. In the first model (M_8), there were eight dummy item covariates, one for each of the eight items of a session. There were 11 test sessions, resulting in $8 \times 11 = 88$ measurements or "virtual" items in total. Hence, the design matrix was an 88 by 8 matrix consisting of 11 stacked 8 by 8 identity matrices. The assumption underlying the first model is that the item difficulties remained unchanged across sessions within a state. Because van der Maas (1993) found that the proportion correct was almost invariable within the two components of the mixture, the latter assumption seemed plausible. The other two models were constructed by further restricting the item difficulties. In the second model (M_3), items of the same type were restricted to have the same difficulty, resulting in an 88 by 3 design matrix consisting of 11 stacked 8 by 3 matrices of the following form:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$

where the first two rows refer to the guessing items, the next three to the standard equality items, and the last three to the standard inequality items. Hence, the item difficulty of a guessing item within state s is defined as β_{s1}, the item difficulty of a standard equality item as $\beta_{s1} + \beta_{s2}$, and the item difficulty of a standard inequality item as $\beta_{s1} + \beta_{s3}$. In the third model (M_2), the item difficulties were still further restricted by assuming a common parameter for the standard equality and standard inequality items, resulting in an 88 by 2 design matrix consisting of 11 stacked 8 by 2 matrices of the following form:

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Hence, the item difficulty of a guessing item within state s is defined as β_{s1}, and the item difficulty of a standard item as $\beta_{s1} + \beta_{s2}$.

The models were estimated with an implementation of the EM-algorithm described in the Appendix. The program was written in Matlab and can be obtained from the first author. In this, we ignored that some of the sessions were missing. However, if the dropout mechanism is random, a valid analysis can be obtained through a likelihood-based analysis that ignores the dropout mechanism, provided the parameters describing the measurement process are functionally independent of the parameters describing the dropout mechanism (Little & Rubin, 1987). A random dropout mechanism seemed to be a reasonable assumption in the present case.

To avoid ending up with a local maximum solution, the following estimation procedure was followed, for the three models just described and other models described below. First, a model without latent variables θ_{ns} (without individual differences within the states) was estimated

10 times, starting from different random starting values for the parameters. Second, the parameter estimates of the simplified models served as starting values for the models that included continuous latent variables. The starting value for the common variance σ^2 was one for the latter. All models could be estimated in reasonable time, the 10 runs for the most complex model taking about 20 minutes on a PC with a Pentium III processor and 256 Mb RAM.

For each of the three models without a continuous latent variable θ_{ns} within the states, the algorithm converged toward two different sets of parameter estimates. Because a solution with a lower log-likelihood corresponds to a local maximum solution, the solutions with the highest log-likelihood were selected, with log-likelihoods of -2406.8 (reached 7 times out of the 10 runs), -2411.1 (reached 7 times out of the 10 runs), and -2422.9 (reached 8 times out of the 10 runs), for, respectively, M_8, M_3, and M_2.

For the models that did include continuous latent variables, the log-likelihood amounted to -2187.9, -2194.5, and -2207.6, for M_8, M_3, and M_2, respectively. Because M_2 is nested within M_3, and M_3 within M_8, likelihood-ratio tests can be used for model comparison. Comparing M_3 with M_8, M_3 is not rejected, $LR = 13.2$, $df = 10$, $p = .21$. Comparing M_2 with M_3, M_2 is rejected, $LR = 26.2$, $df = 2$, $p < .001$. Based on these results, M_3 was selected.

The analogue of M_3 without continuous latent variables is nested within M_3, with $\sigma^2 = 0$. By comparing both models, the need for a continuous latent variable within each state can be tested. As explained in the section on estimation and hypothesis testing, the appropriate p-value for the likelihood-ratio test for this boundary case is obtained by halving the p-value obtained from a chi-squared distribution with one degree of freedom. The model without continuous latent variables was rejected, $LR = 433.2$, $p < .001$.

According to the hypothesis of a continuous acceleration or stepwise growth, the states should differ in a quantitative respect only, whereas according to stage theory, the two states should differ qualitatively. Whether the two states differed in a qualitative or merely a quantitative way can be tested as follows. If the states differ in a quantitative respect only, the regression weight of the intercept may be different across states ($\beta_{11} \neq \beta_{21}$), but the other regression weights should be equal across states ($\beta_{12} = \beta_{22}$ and $\beta_{13} = \beta_{23}$) in M_3. In contrast, if the two states are qualitatively different, all regression weights can be state-specific. The model defined similar to M_3, but with the two equality restrictions imposed on the regression parameters, resulted in a log-likelihood of -2349.5, and was rejected when compared with M_3, $LR = 310.0$, $df = 2$, $p < .001$. Hence, the difference between the two states is not merely quantitative. By consequence, the hypothesis of a continuous acceleration or stepwise growth is rejected.

To assess the stage theory, the parameter estimates and standard errors of M_3 are presented in Table 1. The state-specific probabilities of success, marginalized over θ_{ns}, were estimated by plugging in the maximum likelihood estimates for $\boldsymbol{\beta}_s$ and σ^2, and are presented in Table 2.

If the two states correspond to the stages of nonconservation and conservation, the probability of success for the guessing items should be high in both states. With respect to the standard equality items, the probability of success should be below the chance level of one-third for the nonconservers, but high for the conservers. The results are in line with these expectations (see Table 2). The probabilities of success for the guessing items amounted to .95 for state 1 and .86 for state 2. The probability of success for the standard items was below chance level for state 1, .10 and .18 for the equality and inequality items, respectively. For state 2, in contrast, the probabilities for the standard items are high, .69 for both the equality and the inequality items. The difference between standard equality and standard inequality items, found earlier by comparing M_2 with M_3, was due to a difference between both types for state 1. When restricting the difficulties for both types to be equal for state 2 only ($\beta_{12} \neq \beta_{13}$ and $\beta_{22} = \beta_{23}$; note that in this case the design matrix becomes state-specific, being equal to the design matrix of M_3 for state 1, and equal to the design matrix of M_2 for state 2), the log-likelihood amounted to –2194.6, so that

TABLE 1.
Parameter estimates and standard errors for the model with state-specific parameters for the guessing, standard equality, and standard inequality items (M_3).

	Estimate	Standard error
β_{11}	4.04	.29
β_{12}	−7.17	.22
β_{13}	−6.31	.19
β_{21}	2.64	.17
β_{22}	−1.43	.16
β_{23}	−1.46	.17
a_{11}	.91	.01
π_{01}	.79	.05
σ^2	2.95	.58

the model was not rejected when compared with M_3, $LR = 0.2$, $df = 1$, $p = .65$. The regression weight for the standard items amounted to −1.44 in state 2, with a standard error of .16. The other parameters were almost exactly the same as in M_3, the largest difference being smaller than .01.

The fact that standard inequality items were somewhat easier in state 1 than the standard equality items, can be explained by the fact that the inequality items can be solved even if one is not a conserver. The heights of the two glasses being the same, see Figure 1(b), the glass that contains more fluid is also wider, so that focusing on the width of the glasses also leads to the correct response, without being able to understand the conservation principle. However, because the probability is still low, apparently most of the nonconservers focused on the misleading height cue, which results in incorrect responses for both inequality and equality items.

The probability of success for the guessing items was higher for state 1 (nonconservers) than for state 2 (conservers). The difference was significant, the model defined as M_3, but with the additional restriction of equal probabilities for the guessing items across states ($\beta_{11} = \beta_{21}$), resulted in a log-likelihood of −2198.6, and was rejected when compared with M_3 without the restriction, $LR = 8.2$, $df = 1$, $p = .004$. This somewhat unexpected result can be explained by assuming that focusing on the height, the strategy used by nonconservers and leading to the correct response for the guessing items, is simpler and hence less prone to errors than applying the conservation principle. The latter might be more prone to errors because it was a newly acquired principle for at least the children that became conservers during the course of the longitudinal experiment. It is interesting that a similar result was also found by Rijmen and De Boeck (2003) in the field of deductive reasoning.

TABLE 2.
Marginal state-specific probabilities of success for the guessing, standard equality, and standard inequality items (for M_3).

	$P(Y_{ni} = 1\|\zeta_{ni} = 1, x_i, \boldsymbol{\beta}_s.)$	$P(Y_{ni} = 1\|\zeta_{ni} = 2, x_i, \boldsymbol{\beta}_s.)$
Guessing items	.95	.86
Standard equality items	.10	.69
Standard inequality items	.18	.69

The variance of the continuous latent variable amounted to 2.95. Hence, within a state, there was still a lot of heterogeneity between children. This can be attributed to individual differences in general intelligence, motivation, or more specific variables as, for example, computer skills.

At the first session, the *prior* probability (to be contrasted with the *posterior* probability, given the responses) of being in the nonconservation stage amounted to .79 ($\pi_{01} = .79$), and hence the prior probability of being in the conservation stage amounted to .21 ($1 - \pi_{01} = .21$). Evaluating equation (4) at the maximum likelihood estimates for π_{01} and a_{11}, it can be computed how the prior probabilities of being in the conservation stage gradually increased over time according to the model. At session 11, the prior probability of being in the conservation stage, which equals one minus the prior probability of being nonconserver all the time ($1 - \pi_{01}a_{11}^{10}$), amounted to .70. In addition, children were allocated to the Markov chain with the largest posterior probability (see the Appendix, equation (8), for how to compute these posterior probabilities). Thirty-three children (proportion = .33) were classified as being nonconserver all the time, 25 children (proportion = .25) as being conserver from the beginning, and 43 children (proportion = .43) were classified as transitional children. Hence, according to these classification results, the proportion of conservers shifted from .25 toward .68. Considering either the prior probabilities computed from equation (4) or the classification results based on the posterior chain probabilities, the transition from the nonconservation stage toward the conservation stage over time occurs somewhat sooner than the shift found by van der Maas (1993) reported earlier.

The cross-tabulation of the van der Maas (1993) strategy-based classification of children and the classification of children based on their posterior chain probabilities is given in Table 3. As can be seen in the table, the agreement between the two classification schemes was very high.

For each of the 12 possible chains, the mean of the observed sum score at the 11 test sessions was computed on all children allocated to that particular chain based on the maximum posterior chain probability criterion. The evolution of these mean scores over sessions is plotted in Figure 2. The number of children allocated to each chain is also given. The latter is an upper limit for the number of scores the means are computed on, because some of the sessions were missing. The mean sum score is high for all sessions for the group of children that are conservers from the beginning, and low for the group of children that stay in the nonconservation stage all the time (just above two, the number of guessing items that should be easy for nonconservers as well). Furthermore, for most of the chains in which a transition is made (T1–T10), a substantial jump appears in the mean sum score after the session that precedes the transition. The three chains for which this jump does not occur (T5, T7, and T8) contain only a few children: one, three, and four, respectively. The evolution of the mean sum scores over sessions for the different chains

TABLE 3.
Cross-tabulation of the van der Maas (1993) strategy-based classification of children and the classification of children based on their posterior chain probabilities.

	Strategy classification, van der Maas (1993)				
	Nonconservers all sessions	Transitional children	Conservers from the begining	Residual children	Total
Maximum posterior chain probability					
Nonconservers all sessions	33	0	0	0	33
Transitional children	7	30	2	4	43
Conservers from the beginning	2	1	18	4	25
Total	42	31	20	8	101

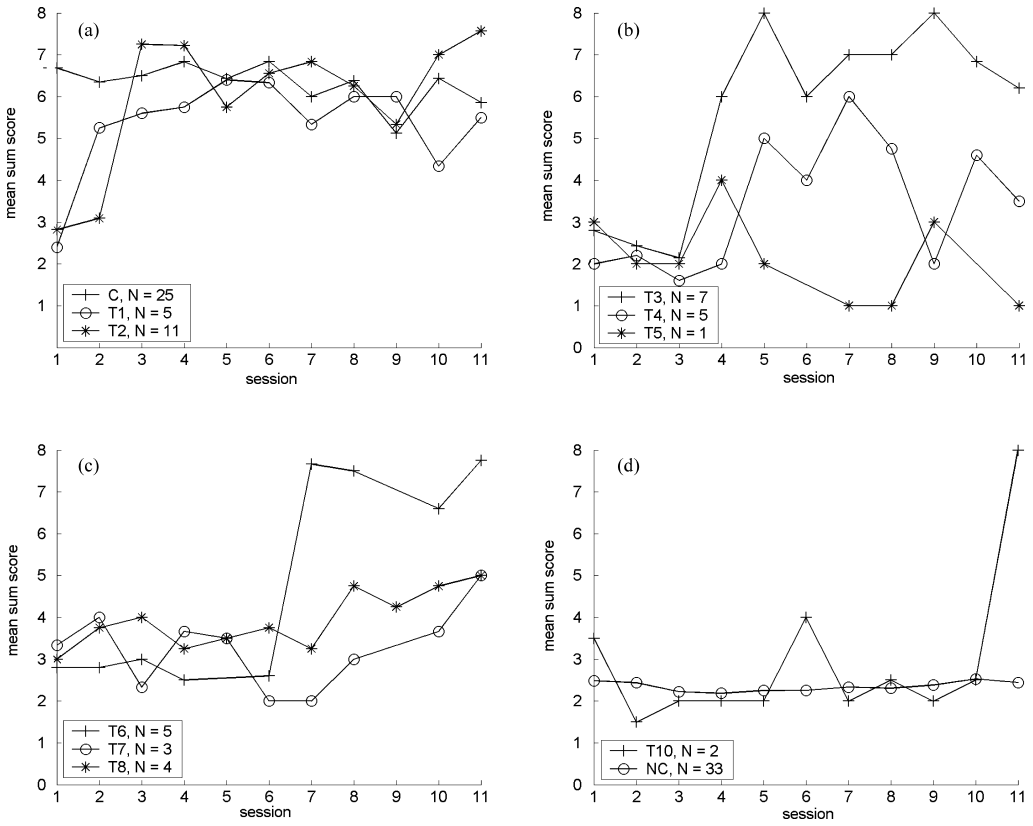


FIGURE 2.
Evolution of the mean sum score over the 11 sessions across all children allocated to the chain based on the posterior chain probabilities. (a) Conservers from the beginning (C), transition from the nonconservation stage to the conservation stage after session 1 (T1), and after session 2 (T2). (b) Transition after session 3 (T3), after session 4 (T4), and after session 5 (T5). (c) Transition after session 6 (T6), after session 7 (T7), and after session 8 (T8). (d) Transition after session 10, and nonconservers for all sessions (NC). *N* represents the number of children allocated to a chain, and is an upper limit for the number of scores the means are computed on. *N* equals the number of scores if the session was not missing for any of the children pertaining to the chain. When there is no marker for a particular session, that session was missing for all children allocated to that chain. There are no results for the chain in which a transition is made after session 9 (T9) because nobody was allocated to this chain.

shows that the allocation of the individual children to a particular chain, based on the parameter estimates of the model, was quite accurate.

The validity of the model was further assessed by comparing the mean of the observed to the expected sum score at each test session. Figure 3 displays the expected sum scores, the means of the observed sum scores, and the simultaneous 95% confidence intervals thereof. The confidence intervals were adjusted for multiple testing by means of the Bonferroni procedure. The top panel displays the sum scores marginalized over states, whereas the bottom panel shows the evolution of the sum scores over sessions separately for each state. For the latter, at each test session, the means of the observed sum scores were computed on all children pertaining to the state at that particular session according to the chain allocation based on the posterior chain probabilities. The expected sum scores were within the limits of the confidence intervals, except for the marginal mean sum score at session 9. Session 9 was also the session in which missingness was most severe: only 43 out of 101 children participated in that session. If the missingness was not completely at random, this may explain the discrepancy at session 9. Indeed, when, for each person, the missing values were replaced by their expectations, given the parameter estimates

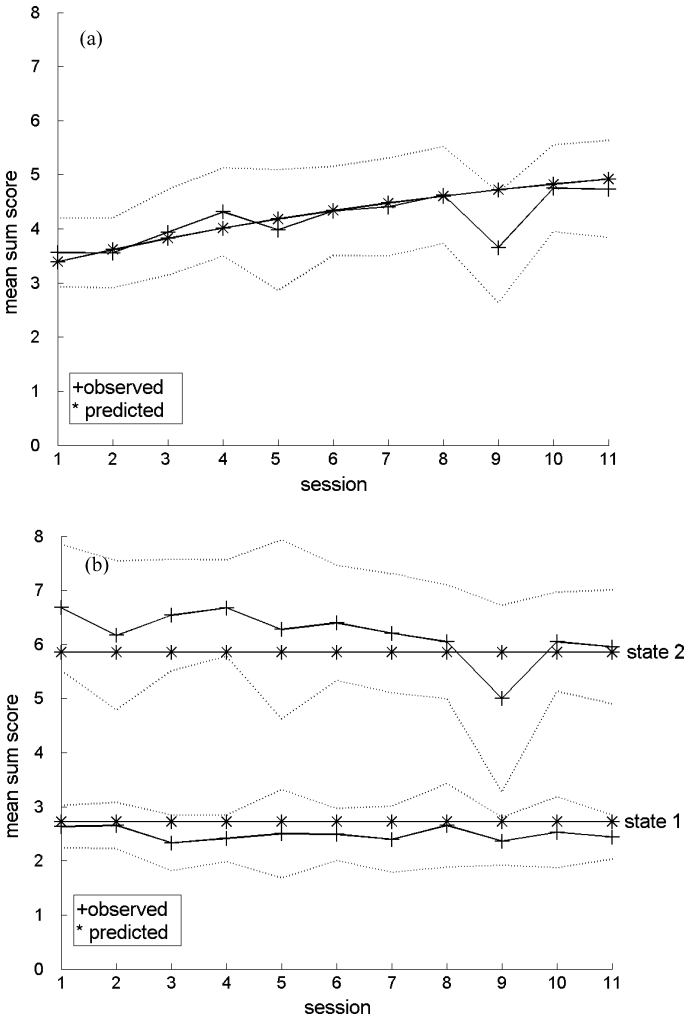


FIGURE 3.
Evolution of the predicted and the mean of the observed sum score over the 11 sessions. The dotted lines show the simultaneous 95% confidence intervals of the mean of the observed sum scores: (a) marginal over states; (b) conditional on the states.

and the observed data for that person, the expected and mean observed sum scores were in close correspondence at all sessions.

An important remaining issue concerns testing the absorbing state and first-order Markov assumptions of our model. This can be achieved by comparing the fit of the model to the fit of the more general models with, respectively, no absorbing state and a second-order Markov process. Estimating these more general models was beyond the scope of the software we developed, however. Nevertheless, the absorbing state assumption naturally follows from the stage theory of cognitive development (Piaget, 1960), where development is seen as an irreversible process characterized by transitions from lower to higher stages. Note that, one of the two states being an absorbing state, incorporating a second-order Markov process would not lead to a more general model in the present application because all information concerning the state history is redundant given the previous state: If your previous state is the state of nonconservers, you must have been in the state of nonconservers all the time; and if your previous state is the state of conservers, you

cannot leave that state, regardless of your state history. In addition, our choice for a Markovian model structure as such can be grounded in Piagetian theory with its finite number of qualitatively differently behaving stages, and abrupt transitions between them. The Markovian model structure is less suited when changes over time are continuous, for example, in a learning task where learning occurs gradually.

5. General Discussion

We presented a model for longitudinal binary data that takes into account dependencies caused by between-subject variability as well as autodependencies caused by within-subject variability. The proposed model is a linear logistic test model (Fischer, 1973), the parameters of which are determined by a homogeneous first-order Markov process. Each set of parameters corresponds to a state of the Markov process. The latent states of the hidden Markov process correspond to the latent classes from LCA.

If no continuous latent variables are incorporated, the model reduces to a latent Markov model (Collins & Wugalter, 1992; Langeheine & van de Pol, 1994; Wiggins, 1973). If no transitions between states are possible, that is, the transition matrix of the Markov process is an identity matrix, the model reduces to the mixture-LLTM of Mislevy and Verhelst (1990), or to a specific instance of the latter, such as the mixture-Rasch (Rost, 1990) and the Saltus model (Mislevy & Wilson, 1996; Wilson, 1989).

The model in its current formulation does not take into account autodependencies that are caused by an observation-driven process. An example of an observation-driven process is when learning occurs after each item that is solved correctly. Such a process can be taken into account by adding the responses of a participant on previous items, or some function of the previous responses such as the number of items solved correctly, as person-by-item covariates for the item considered next by that person (Verhelst & Glas, 1993). Including person-by-item covariates would of course mean that the design matrix becomes person-specific.

Given the wide range of models that are available for the analysis of categorical longitudinal data, it is not at all a trivial affair to decide upon which model to choose for a particular application. However, every model is built upon some basic assumptions regarding the nature of the stability and change of the data-generating process. These assumptions can be used as heuristics guiding the choice for a particular model. A first consideration is whether differences over time within a person are to be expected. If not, one can rely on models that solely take into account between-subject variability: Latent class models if the interindividual differences are all of a qualitative nature, IRT models if only quantitative interindividual differences play a role, and a hybrid model such as the mixture LLTM if both qualitative and quantitative interindividual differences are to be expected. Second, if differences within a person over time are to be expected, or are found in preliminary analyses, it is important to reflect on the sources of these intraindividual differences. If these are merely a consequence of occasion-specific fluctuations, a variability model is suited. On the other hand, if the intraindividual differences are related to an underlying time-related process, a process-oriented model is more appropriate. In the latter case, an additional choice to be made is between models incorporating an observation- vs. parameter-driven process for change.

After the question regarding the nature of the intraindividual differences is answered, a final issue is whether or not the mechanism, that is responsible for intraindividual differences over time, is also sufficient to explain the interindividual differences. In latent state-trait models, the answer is no. The latent state variables only explain the fluctuations within a person over time. Latent traits are included in the model to account for stable interindividual differences over time. In latent Markov models, the answer is yes. The Markov process that accounts for

the autodependencies in the data also accounts for all variability between persons. In the latent mixed Markov model on the other hand, additional interindividual variability is accounted for by assuming that the population consists of qualitatively distinct subpopulations that are each characterized by their own latent Markov model. In the model we propose, it is assumed that the additional interindividual variability is caused by quantitative interindividual differences within states. Likewise, a continuous latent variable is included within each state. Also in the models proposed by Humphreys (1998) and Böckenholt (1999), and in the dynamic Rasch model, a continuous latent variable is included to account for additional between-subject variability.

6. Appendix: Estimation

The algorithm presented here is an extension of the EM algorithm (Dempster, Laird, & Rubin, 1977) proposed by Mislevy and Verhelst (1990) for their mixture-LLTM. The complete data consist of the observed data, the unobserved values of the persons on the continuous latent variables θ_{ns}, and the unobserved Markov chains ζ_n of the persons.

Collecting the complete data for all persons into the complete data matrix $\boldsymbol{Y}_c$ with rows $\boldsymbol{y}'_{nc}$, and collecting all parameters $\sigma^2, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S, \boldsymbol{A}$, and $\boldsymbol{\pi}_0$ into the parameter vector $\boldsymbol{\psi}$, the complete data log-likelihood can be written as

$$\ell_c(\boldsymbol{\psi}|\boldsymbol{Y}_c) = \sum_{n=1}^{N} \log \Pr(\boldsymbol{\zeta}_n|\boldsymbol{\pi}_0, \boldsymbol{A}) + \sum_{n=1}^{N}\sum_{s=1}^{S} \log[\Pr(\boldsymbol{y}_{ns}|\theta_{ns}, \boldsymbol{X}, \boldsymbol{\beta}_s)] + \sum_{n=1}^{N}\sum_{s=1}^{S} \log[N(\theta_{ns}|0, \sigma^2)]. \tag{5}$$

The expected complete data log-likelihood, given the observed data and the provisional parameter estimates is

$$\begin{aligned} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(j)}) &= E_{\boldsymbol{\psi}^{(j)}}[\ell_c(\boldsymbol{\psi}|\boldsymbol{Y}_c)|\boldsymbol{Y}] \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} \Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}) \log[\Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{\pi}_0, \boldsymbol{A})] \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K} \Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}) \sum_{s=1}^{S} \int \log[\Pr(\boldsymbol{y}_{ns}|\theta_{ns}, \boldsymbol{X}, \boldsymbol{\beta}_s)] f\left(\theta_{ns}|\boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) d\theta_{ns} \\ &+ \sum_{n=1}^{N}\sum_{k=1}^{K} \Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}) \sum_{s=1}^{S} \int \log[\mathrm{N}(\theta_{ns}|0, \sigma^2)] f\left(\theta_{ns}|\boldsymbol{\zeta}_n = k, \mathbf{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) d\theta_{ns}, \end{aligned} \tag{6}$$

where the posterior densities $f(\theta_{ns}|\boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{\psi}^{(j)})$ and the posterior probabilities $\Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{y}_n, \boldsymbol{\psi}^{(j)})$ are computed using Bayes' theorem,

$$f(\theta_{ns}|\boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}) = \frac{\Pr(\mathbf{y}_{ns}|\theta_{ns}, \boldsymbol{X}, \boldsymbol{\beta}_s) N(\theta_{ns}|0, \sigma^2)}{\Pr(\boldsymbol{y}_{ns}|, \boldsymbol{X}, \boldsymbol{\beta}_s)}, \tag{7}$$

and

$$\Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}) = \frac{\Pr(\mathbf{y}_n|\boldsymbol{\zeta}_n = k, \boldsymbol{X}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S) \Pr(\boldsymbol{\zeta}_n = k|\boldsymbol{\pi}_0, \boldsymbol{A})}{\Pr(\mathbf{y}_n|\boldsymbol{X}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_S)}. \tag{8}$$

If a particular state s is never visited according to a chain k, $\boldsymbol{y}_{ns}$ is empty. In that case, the posterior density of θ_{ns} equals the prior, $f(\theta_{ns}|\boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{\psi}^{(j)}) \sim N(\theta_{ns}|0, \sigma^2)$.

Differentiating $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(j)})$ with respect to the parameters results in the following score equations:

$$\frac{\partial Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(j)})}{\partial \beta_{sp}} = \sum_{n=1}^{N}\sum_{k=1}^{K} \Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right)$$

$$\times \left\{ \sum_{i:y_i \in \mathrm{I}_{ns}} y_{ni} x_{ip} - \int \left\{ \sum_{i:y_i \in \mathrm{I}_{ns}} \frac{\exp\left(\theta_{ns} + \boldsymbol{x}_i' \boldsymbol{\beta}_s\right) x_{ip}}{1 + \exp\left(\theta_{ns} + \boldsymbol{x}_i' \boldsymbol{\beta}_s\right)} \right\} f\left(\theta_{ns} \middle| \boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) d\theta_{ns} \right\} \quad (9)$$

for $s = 1, \ldots, S$; $p = 1, \ldots, P$;

$$\frac{\partial Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(j)})}{\partial \sigma^2} = -\frac{SN}{2\sigma^2} + \sum_{n=1}^{N}\sum_{k=1}^{K} \Pr(\boldsymbol{\zeta}_n = k | \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)})$$

$$\times \sum_{s=1}^{S} \int \frac{\theta_{ns}^2}{2\sigma^4} f(\theta_{ns} | \boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)})\, d\theta_{ns}, \quad (10)$$

$$\frac{\partial Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(j)})}{\partial \pi_{0s}} = \sum_{n=1}^{N} \left\{ \sum_{k:\zeta_{n1}=s} \frac{\Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right)}{\pi_{0s}} - \sum_{k:\zeta_{n1}=S} \frac{\Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right)}{\pi_{0S}} \right\} \quad (11)$$

for $s = 1, \ldots, S-1$, where the second term between the curly brackets results from the fact that $\pi_{0S} = 1 - \sum_{s=1}^{S-1} \pi_{0s}$; and

$$\frac{\partial Q\left(\boldsymbol{\psi} \middle| \boldsymbol{\psi}^{(j)}\right)}{\partial a_{ss'}} = \sum_{n=1}^{N}\sum_{k=1}^{K} \left\{ \frac{\Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) m_{ss'}^k}{a_{ss'}} - \frac{\Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) m_{sS}^k}{a_{sS}} \right\} \quad (12)$$

for $s = 1, \ldots, S$; $s' = 1, \ldots, S-1$; where $a_{ss'}$ is the element of $\boldsymbol{A}$ of row s and column s', the second term between the curly brackets results from the fact that $a_{sS} = 1 - \sum_{s'=1}^{S-1} a_{ss'}$, and $m_{ss'}^k$ is the number of transitions from state s to state s' in chain k.

By equating the partial derivatives to zero, closed-form solutions can be found for σ^2 and for the elements of $\boldsymbol{A}$ and $\boldsymbol{\pi}_0$ that maximize $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(j)})$, serving as provisional parameter estimates for the E-step of the next iteration:

$$\sigma^{2(j+1)} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} \Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) \sum_{s=1}^{S} \int \theta_{ns}^2 f\left(\theta_{ns} \middle| \boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) d\theta_{ns}}{NS}, \quad (13)$$

$$\pi_{0s}^{(j+1)} = \frac{\sum_{n=1}^{N} \sum_{k:\zeta_{n1}=s} \Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right)}{N} \text{ for } s = 1, \ldots, S, \quad (14)$$

and

$$a_{ss'}^{j+1} = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} \Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) m_{ss'}^k}{\sum_{n=1}^{N}\sum_{k=1}^{K} \Pr\left(\boldsymbol{\zeta}_n = k \middle| \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right) \sum_{s'=1}^{S} m_{ss'}^k} \text{ for } s = 1, \ldots, S; s' = 1, \ldots, S. \quad (15)$$

For the logistic regression weights β_{sp} no closed-form solution exist. Instead, the parameter estimates are updated using one Newton–Raphson iteration. Hence, the algorithm we implemented is actually a Generalized EM algorithm (Dempster et al., 1977). The second derivatives needed

for the Newton–Raphson algorithm are

$$\frac{\partial^2 Q\left(\boldsymbol{\psi}\,|\,\boldsymbol{\psi}^{(j)}\right)}{\partial\beta_{sp}\partial\beta_{sp'}} = -\sum_{n=1}^{N}\sum_{k=1}^{K}\Pr\left(\boldsymbol{\zeta}_n = k\,|\boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right)\int\left\{\sum_{i:y_i\in\mathbf{I}_{ns}}\frac{\exp(\theta_{ns}+\mathbf{x}_i'\boldsymbol{\beta}_s)x_{ip}x_{ip'}}{[1+\exp(\theta_{ns}+\boldsymbol{x}_i'\boldsymbol{\beta}_s)]^2}\right\}$$
$$\times f\left(\theta_{ns}\big|\boldsymbol{\zeta}_n = k, \boldsymbol{y}_n, \boldsymbol{X}, \boldsymbol{\psi}^{(j)}\right)d\theta_{ns}. \tag{16}$$

The integrals over the θ_{ns} are approximated using Gauss–Hermite quadrature (Fahrmeir & Tutz, 2001).

After convergence, the observed information matrix is approximated by the empirical observation matrix:

$$\boldsymbol{I}(\hat{\boldsymbol{\psi}}|\boldsymbol{Y}) \approx \sum_{n=1}^{N}\boldsymbol{s}(\boldsymbol{y}_n|\hat{\boldsymbol{\psi}})\boldsymbol{s}'(\boldsymbol{y}_n|\hat{\boldsymbol{\psi}}), \tag{17}$$

where $\boldsymbol{s}(\boldsymbol{y}_n|\boldsymbol{\psi}) = \partial\ell_n(\boldsymbol{\psi}|\,\boldsymbol{y}_n)/\partial\boldsymbol{\psi}$, the score function based on the single observation $\boldsymbol{y}_n$. The latter can be computed using the complete data loglikelihood as follows (McLachlan & Krishnan, 1997):

$$\boldsymbol{s}(\boldsymbol{y}_n|\boldsymbol{\psi}) = E_{\boldsymbol{\psi}}[\partial\ell_{cn}(\boldsymbol{\psi}|\boldsymbol{y}_{nc})/\partial\boldsymbol{\psi}\,|\boldsymbol{y}_n] \tag{18}$$

Standard errors are computed as the square root of the diagonal elements of the inverse of the observed information matrix.

The three straightforward extensions discussed in the model section result in the following changes: First, if some of the logistic regression weights β_{sp} are restricted to be the same across (some of the) states, the expressions between curly brackets in equations (9) and (16) for those common parameters have to be summed over the respective states. Second, if the design matrix $\boldsymbol{X}$ is made state-specific, $\boldsymbol{X}$ is replaced by $\boldsymbol{X}_s$ (and thus $\boldsymbol{x}_i$ by $\boldsymbol{x}_{is}$) in all expressions. Third, if the variance is state-specific, there is a score equation for each variance, equal to equation (10) but without the summation over the states, and the new estimates are computed using equation (13) with N in the numerator instead of NS.

References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A*, *144*, 419–461.

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

Böckenholt, U. (1999). Measuring change: Mixed Markov models for ordinal panel data. *British Journal of Mathematical and Statistical Psychology, 52*, 125–136.

Collins, L.M., Fidler, P.L., Wugalter, S.E., & Long, S.E. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28*, 375–389.

Collins, L.M., & Wugalter, S.E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*, 131–157.

Cox, D.R., & Snell, E.J. (1996). *Analysis of binary data* (2nd ed.). London: Chapman & Hall.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Eid, M., & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics, 23*, 193–215.

Eid, M., & Langeheine, R. (1999). The measurement of consistency and occasion specificity with latent class models: A new model and its applications to the measurement of affect. *Psychological Methods, 4*, 100–116.

Eid, M., & Langeheine, R. (2003). Separating stable from variable individuals in longitudinal studies by mixture distribution models. *Measurement, 1*, 197–206.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer-Verlag.

Field, D. (1987). A review of preschool conservation training: An analysis of analyses. *Developmental Review, 7*, 210–251.

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

Formann, A.K. (1985). Constrained latent class models: Theory and applications. *The British Journal of Mathematical and Statistical Psychology, 38*, 87–111.

Formann, A.K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 418*, 476–486.
Gilmore, R. (1981). *Catastrophe theory for scientists and engineers.* New York: Wiley.
Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.
Hagenaars, J.A. (1998). Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociological Methods and Research, 26*, 436–486.
Humphreys, K. (1998). The latent Markov chain with multivariate random effects. *Sociological Methods and Research, 26*, 269–299.
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.
Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research, 24*, 492–516.
Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modelling in discrete space and discrete time. *Sociological Methods and Research, 18*, 416–441.
Langeheine, R., & van de Pol, F. (1994). Discrete-time mixed Markov latent class models. In A. Dale, & R.B. Davies (Eds.), *Analyzing social and political change: A casebook of methods.* London: Sage.
Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: Wiley.
McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika, 21*, 331–347.
McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: Wiley.
McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.
Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different persons employ different solution strategies. *Psychometrika, 55*, 195–215.
Mislevy, R.J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61*, 41–71.
Piaget, J. (1960). The general problems of the psychobiological development of the child. In J.M. Tanner, & B. Inhelder (Eds.), *Discussions on child development* (Vol. 4). London: Tavistock.
Poulsen, C.S. (1990). Mixed Markov and latent Markov modeling applied to brand choice data. *International Journal of Marketing, 7*, 5–19.
Read, T., & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer Verlag.
Rijmen, F., & De Boeck, P. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research, 67*, 219–231.
Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for IRT models. *Psychological Methods, 8*, 185–205.
Ross, S. (1996). *Stochastic processes* (2nd ed.). New York: Wiley.
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.
Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6*, 461–464.
Snijders, T., & Bosker, R. (1999). *An introduction to basic and advanced multilevel modeling*. London: Sage.
Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 2*, 79–98.
Thom, R. (1975). *Sructural stability and morphogenesis*. Reading, MA: Benjamin.
Thomas, H., & Hettmansperger, T.P. (2001). Modeling change in cognitive understanding with finite mixtures. *Applied Statistics, 50*, 435–448.
van der Maas, H. (1993). *Catastrophe analysis of stagewise cognitive development: Model, method and applications.* Unpublished doctoral dissertation. University of Amsterdam.
van der Maas, H., & Molenaar, P.C.M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99*, 395–417.
van der Maas, H.L.J., & Molenaar, P.C.M. (1995). Catastrophe analysis of discontinuous development. In A.A von Eye, & C.C. Clogg (Eds.), *Categorical variables in developmental research methods of analysis*. San Diego, CA: Academic Press.
Verbeke, G. (1997). Linear mixed models for longitudinal data. In G. Verbeke, & G. Molenberghs (Eds.), *Linear mixed models in practice*. New York: Springer-Verlag.
Verhelst, N.D., & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*, 395–415.
von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research* [On-line serial], 2, 29–48. Available http://www.mpr-online.de/.
Wiggins, L.M. (1973). *Panel analysis: Latent probability models for attitude and behavior processes*. Amsterdam: Elsevier.
Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Manuscript received 17 DEC 2002
Final version received 8 SEP 2004