State Mastery Learning: Dynamic Models for Longitudinal Data

Rolf Langeheine, Institute for Science Education at the University of Kiel
Elsbeth Stern, Max Planck Institute of Psychological Research
Frank van de Pol, Netherlands Central Bureau of Statistics

Macready & Dayton (1980) showed that state mastery models are handled optimally within the general latent class framework for data from a single time point. An extension of this idea is presented here for longitudinal data obtained from repeated measurements across time. The static approach is extended using multiple-indicator Markov chain models. The approach presented here emphasizes the dynamic aspects of the process of change, such as growth, decay, and stability. The general approach is presented, and models with purely categorical and ordered categorical states and several extensions of these models are discussed. Problems of estimation, identification, assessment of model fit, and hypothesis testing associated with these models also are discussed. The applicability of these models is demonstrated using data from a longitudinal study on solving arithmetic word problems. The advantages and disadvantages of using the approach presented here are discussed. *Index terms: arithmetic word problems, dynamic latent class models, latent class models, longitudinal categorical data, Markov models, state mastery models.*

Meskauskas (1976) showed that most mastery learning models can be grouped into two general classes. If a continuous trait is used to make statements about mastery, mastery is viewed as an interval on a test score scale. However, state mastery models assume that examinees can be grouped into a small number of classes of a categorical latent variable. In the simplest case, two states (masters and nonmasters) describe a given dataset, but multistate models exist as well.

In a review on the nature and use of state mastery models, Macready & Dayton (1980) showed that this class of models may be handled optimally within the general latent class framework (e.g., Goodman, 1974; Lazarsfeld & Henry, 1968). This paper extends single time latent class models to situations in which measurements are repeated across time. In these situations, the interest is in making statements about change in terms of growth, stability, or decay. This problem can be handled by extending the single-indicator latent Markov model (e.g., Langeheine & Van de Pol, 1990; Van de Pol & Langeheine, 1990) to a multiple-indicator model (Langeheine & Van de Pol, 1993; see also Langeheine, 1991). This is similar, in a sense, to latent transition analysis (Collins & Wugalter, 1992) and to latent class models that allow for latent change, such as a model presented by Macready & Dayton (1994) for longitudinal assessment of trait acquisition.

From Static to Dynamic Models

The General Approach

For simplicity, a hypothetical example will be used to demonstrate the approach. Assume that four dichotomous item indicators—A, B, C, and D—with response categories + (item answered correctly) and – (item answered incorrectly) have been given to a sample of examinees at a single time point. Further, assume that the hypothesis of a two-state mastery model is in accordance with the data (see Figure 1).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 3, September 1994, pp. 277–291

0146-6216/94/030277-15$2.00

Figure 1
A Hypothetical Example of a Two-Class Mastery Learning Model at $t = 1$ (Left Panel) and $t = 2$ (Right Panel): Values of ρ for Correct (+) and Incorrect (–) Responses to Four Items (A, B, C, D) at Two Points in Time ($t = 1$, $t = 2$) (The Empty Boxes Represent the Dynamic Data That Are Not Used in the $t = 1$ and $t = 2$ Cross-Sectional Analyses)

		Item	A	B	C	D
t=1	δ_1=.6	-	.75	.80	.86	.92
		+	.25	.20	.14	.08
	δ_2=.4	-	.05	.10	.15	.20
		+	.95	.90	.85	.80

t=2

	δ_1=.3		δ_2=.7	
Item	-	+	-	+
A	.75	.25	.05	.95
B	.80	.20	.10	.90
C	.86	.14	.15	.85
D	.92	.08	.20	.80

The general S-state model for a situation like this is given by

$$P_{ijkl} = \sum_{s=1}^{S} \delta_s^X \rho_{i\,s}^{AX} \rho_{j\,s}^{BX} \rho_{k\,s}^{CX} \rho_{l\,s}^{DX}. \tag{1}$$

where

P_{ijkl} is the proportion in cell $ijkl$,
δ_s^X is the proportion of state/class s of the latent categorical variable X, and
$\rho_{i\,s}^{AX}$ is the conditional probability for a response in category i of Item A, given membership in state s (and likewise for Items B, C, and D; i.e., j, k, and l refer to the categories of Items B, C, and D).

The model shown in Equation 1 will be referred to as Model 1. Local independence of items is assumed to hold within classes that are mutually exclusive and exhaustive. Thus, the mastery state (.4 at $t = 1$) has high probabilities for answering an item correctly and the nonmastery state (.6 at $t = 1$) has low probabilities for answering an item correctly.

Now assume that repeated measurements from a second point in time ($t = 2$) are available and that for these data, again, an acceptable fit was obtained for the two-state solution. As Figure 1 shows, conditional response probabilities remained stable across time. A comparison of the state proportions at $t = 1$ and $t = 2$ reveals that considerable change occurred between occasions. The proportion of masters increased from .4 at $t = 1$ to .7 at $t = 2$. However, separate analyses of the data from each time point do not indicate the states into which members of the two states at $t = 1$ fall at $t = 2$.

To gain insight into the dynamics of the process across time it is necessary to perform a simultaneous analysis of the $2^4 \times 2^4$ cross-classification of both time points. In this manner, probabilities then are obtained that quantify the transition from a given state at $t = 1$ to some state at $t = 2$. For a situation like this

($T = 2$ points in time, four items) the model is given by

$$P_{ijkl\,i'j'k'l'}^{1111\,2222} = \sum_{x=1}^{X} \sum_{y=1}^{Y} \delta_x^1 \rho_{i\,x}^1 \rho_{j\,x}^1 \rho_{k\,x}^1 \rho_{l\,x}^1 \tau_{y\,x}^{21} \rho_{i'\,y}^2 \rho_{j'\,y}^2 \rho_{k'\,y}^2 \rho_{l'\,y}^2, \tag{2}$$

where

superscripts refer to time points,
subscripts i, j, k, and l refer to the categories of the four items at $t = 1$,
subscripts i', j', k', and l' refer to the categories of the items at $t = 2$,
$P_{ijkl\,i'j'k'l'}^{1111\,2222}$ is the model expected proportion for a cell indexed by i, j, k, l, i', j', k', and l' at $t = 1$ and $t = 2$ [Note that this is a cell of the joint $2^4 \times 2^4$ table (if binary items are answered as in the hypothetical example)],
δ_x^1 is the proportion of latent state x at $t = 1$,
$\rho_{i\,x}^1$ is the conditional probability for a response in category i of Item A (given membership in state x at $t = 1$),
$\tau_{y\,x}^{21}$ is the probability of a transition from state x at $t = 1$ to state y at $t = 2$, and
$\rho_{i'\,y}^2$ is the conditional probability for a response in category i' of Item A at $t = 2$ (given membership in state y).

The model shown in Equation 2 will be referred to as Model 2. In most cases, the number of categories (states) of the latent variables is assumed to be equal to the number of categories of the manifest indicators (i.e., the items).

Thus, the ρs map the manifest indicators onto the latent variables and reflect how well a latent variable is measured by a specific indicator. Measurement of, for example, $x \times i$, is perfect if all $\rho_{i\,x}^1 = 1$ for $i = x$ and 0 otherwise. In order to make the model identifiable, all (indexed) sets of parameters must sum to unity, e.g.,

$$\Sigma_x \delta_x^1 = \Sigma_i \rho_{i\,x}^1 = \Sigma_y \tau_{y\,x}^{21} = 1.0. \tag{3}$$

Of course, this also holds for sets of parameters in Model 1. Extension to more than two occasions is straightforward.

Thus, Model 2 is a latent class model with latent change, which is a special case of the more general latent Markov model (Langeheine & Van de Pol, 1990). However, the classical latent Markov model applies to the situation of repeated measurements of a single item (indicator) across occasions. Therefore, Langeheine & Van de Pol (1993) extended the classical latent Markov model to multiple indicators measured across time (e.g., Model 2 above).

Note that Model 2 gives the equation for the hypothetical example in Figure 1. This is a special case of the general situation in which K items are measured at T occasions. In most cases, the number of classes considered for some specific model will be assumed to be equal across time (i.e., $Y = X$). There may be situations, however, in which some substantive theory requires the number of classes to differ across time.

Categorical and Ordered Categorical Latent Variables

An implicit assumption has been made that the latent variables are categorical. With multiple indicators it is possible to conceive of latent variables as being ordered by forcing items to follow some probabilistic version of a Guttman-type scale at each point in time using latent distance models (Lazarsfeld & Henry, 1968; see also Clogg & Sawyer, 1981; Dayton & Macready, 1980; Langeheine, 1988). These models are appropriate if manifest items have been constructed a priori to form a hierarchy, with item difficulties increasing as in the example in Figure 1 and an example presented below.

For the above example, the assumption of ordered categorical variables would imply specifying five instead of two latent states at $t = 1$ and $t = 2$. The response probabilities are shown in Table 1 where an L

denotes a low probability for solving an item, and an H denotes a high probability for solving the item.

Note, however, that in general a model such as that depicted in Table 1 would neither be identified (there would be too many free parameters to be estimated) nor would ordered states (which require nonintersecting item profiles of the conditional response probabilities across items) be guaranteed. Latent distance models solve both of these problems simultaneously by putting certain forms of equality constraints on the conditional response probabilities (see references above). For example, in the "item specific error rate model" (Clogg & Sawyer, 1981; Dayton & Macready, 1980; Langeheine, 1988; Lazarsfeld & Henry, 1968) each item is characterized by a single parameter $p(\mathrm{H})$ with $p(\mathrm{L}) = 1 - p(\mathrm{H})$ across states.

Table 1
Response Probabilities in a Five-State Latent Distance Model With Four Manifest Items

	Items			
State	A	B	C	D
1	L	L	L	L
2	H	L	L	L
3	H	H	L	L
4	H	H	H	L
5	H	H	H	H

Models assuming ordered categorical latent variables are especially relevant in the context of learning or, in a more general sense, development assumed to pass through a sequence of qualitatively different stages in which the stages correspond to the states in Table 1. Examples are provided in Collins & Wugalter (1992) and Langeheine (1991).

Extensions of the Models

Three issues concerning these models were addressed (at least in part) by Macready & Dayton (1980):

1. The models are not restricted to manifest dichotomous items but are extended easily to items having more than two categories.
2. Whereas population heterogeneity is an implicit characteristic of latent class models, a Markov model postulating a single (latent) chain of movements across time ("chain" refers to a discrete time process) assumes population homogeneity with respect to the parameters generating the dynamics. However, modeling unobserved heterogeneity is possible by extending the single chain model to a mixture of several (e.g., S) chains:

$$P^{1111\,2222}_{ijkl\,i'j'k'l'} = \sum_{s=1}^{S}\sum_{x=1}^{X}\sum_{y=1}^{Y} \pi_s \delta^{1}_{x:s} \rho^{1}_{i:xs} \rho^{1}_{j:xs} \rho^{1}_{k:xs} \rho^{1}_{l:xs} \tau^{21}_{y:xs} \rho^{2}_{i':ys} \rho^{2}_{j':ys} \rho^{2}_{k':ys} \rho^{2}_{l':ys}. \quad (4)$$

The model presented in Equation 4 will be referred to as Model 3. This is the latent mixed Markov model of Langeheine & Van de Pol (1990) relevant for the case of $K = 4$ items measured at $T = 2$ occasions. Note that Equation 4 is very similar to Equation 2 (Model 2), but an additional term appears in Equation 4. π_s is the proportion of chain s with all other sets of parameters considered conditional on chain membership; it also is called the mixing proportion of a mixture model.

A special case of Model 3 (Equation 4) is the "partially latent mover–stayer model" (Langeheine & Van de Pol, 1990) with two chains. Members of the mover-chain follow an ordinary latent Markov chain (i.e., with response error). Stayers respond in a deterministic fashion and stay with their initial state across time with probability 1.0.

3. If there is some a priori reason to assume population heterogeneity (e.g., that males and females or an experimental and a control group may differ in their dynamics), the data should not be pooled; rather, a simultaneous multiple-group analysis should be performed. For Markov chains, the extension is presented in Van de Pol & Langeheine (1990). In the present context, this leads to Model 4:

$$P^{1111\,2222}_{h,ijkl\,i'j'k'l'} = \gamma_h \sum_{s=1}^{S}\sum_{x=1}^{X}\sum_{y=1}^{Y} \pi_{s\,h} \delta^{1}_{x:sh} \rho^{1}_{i:xsh} \rho^{1}_{j:xsh} \rho^{1}_{k:xsh} \rho^{1}_{l:xsh} \tau^{21}_{y:xsh} \rho^{2}_{i':ysh} \rho^{2}_{j':ysh} \rho^{2}_{k':ysh} \rho^{2}_{l':ysh}, \quad (5)$$

where γ_h is the proportion of subpopulation h defined by some additional external discrete variable with h categories, and all other sets of parameters are considered conditional on h. Macready & Dayton (1980) called Model 4 a covariate model.

Model 4 is similar to the Collins & Wugalter (1992) model, which they termed a "latent transition analysis," but there are two differences. First, it is assumed in Model 4 that subpopulation membership is measured without error; however, in the Collins & Wugalter model, the probability of having a specific value on the indicator of subpopulation membership is estimated. Second, contrary to the Collins & Wugalter model, in Model 4 unobserved heterogeneity is modeled within observed heterogeneity (there may be S latent chains within each subpopulation h).

Although transition probabilities are an integral part in both the model of Collins & Wugalter (1992) and Models 2–4, Macready & Dayton (1994) used a modified classical latent class approach to make statements about transitions from one point in time to the next. In the case of two points in time only, their model is identical to Model 4 with the exception that they do not consider unobserved heterogeneity within subpopulations.

Estimation, Identification, and Assessment of Model Fit

Because the multiple-indicator Markov model is embedded in a single-indicator latent Markov model, parameter estimation is similar (for details, see Van de Pol & de Leeuw, 1986; Van de Pol & Langeheine, 1990). Parameter estimation involves setting up a single-indicator latent Markov model with $T' = K \times T$ points in time, where K is the number of items and T is the number of time points under study, and nonstationary transition matrices. To ensure that the indicators for each time point t measure the same construct, all transition matrices, $\mathbf{T}^{t,t+1}$, for each time point are restricted to the identity matrix. In the above example (Figure 1, Model 2), $T' = 4 \times 2 = 8$ with $\mathbf{T}^{21} = \mathbf{T}^{32} = \mathbf{T}^{43} = \mathbf{I}$ and $\mathbf{T}^{65} = \mathbf{T}^{76} = \mathbf{T}^{87} = \mathbf{I}$. The unconstrained matrix $\mathbf{T}^{54}$ gives the transition probabilities from $t = 1$ to $t = 2$ that are of interest (denoted by $\tau^{21}_{y,x}$ in Equation 2; for details see Langeheine & Van de Pol, 1993). Consequently, existing software for single-indicator models (e.g., PANMARK; Van de Pol, Langeheine, & de Jong, 1991) may be used to fit multiple-indicator models.

Whereas the single-indicator latent Markov model is not identified unless conditional response probabilities are restricted to be equal across (at least some points in) time, this problem does not occur in one-chain multiple-indicator models. However, for more than one chain, identifiability depends on the number of time points under study and the number and the nature of the chains. PANMARK offers several checks for model identification.

As with other models that use contingency table analysis, χ^2 statistics are used to assess model fit. The likelihood ratio χ^2 (G^2) is preferred over the Pearson χ^2 (X^2), because it compares nested models using conditional likelihood ratio tests. However, multiple-indicator models often have the problem of sparse data, unless sample size is extremely large. This will (at least in most cases) cause no problems in estimating the parameters of a given model, but many of the expected frequencies will be small (even less than 1). Consequently, the resulting statistic may have a distribution that is badly approximated by the χ^2 distribution. Below, three statistics from the "family of power divergence statistics" (see Read & Cressie, 1988) are reported with $\lambda \to 0$ (which is equal to G^2), $\lambda = 2/3$, and $\lambda = 1$ (which is equal to X^2). The reason for includ-

ing the $\lambda = 2/3$ statistic is because Read and Cressie argued that for a minimum expected frequency not smaller than 1 the resulting statistic has a distribution that is generally well approximated by the χ^2 distribution. They stated that this statistic will still be a good choice when many expected frequencies are below 1. However, this problem has not been fully resolved. One solution to the dilemma would be to generate the reference distribution using monte carlo studies (see, e.g., Aitkin, Anderson, & Hinde, 1981). Unfortunately, this is tedious and costly, especially if many models are fit to a given dataset.

Because exact or at least approximate exact tests are not available, another possibility is to rely on descriptive fit indexes such as the AIC (Akaike, 1974) or BIC (Schwarz, 1978). Neither the AIC nor the BIC are restricted to comparisons of nested models. The BIC is preferred because it has been shown to be a consistent and asymptotically optimal estimator of model dimensionality in contingency table analysis (Raftery, 1986), especially in Markov chain analysis (Katz, 1981). These studies showed that the BIC selects the correct model with probability 1.0 as N (sample size) goes to infinity, even when non-nested models are compared. However, the BIC is not without problems. One problem is the question of whether the model that really generated the result in essentially the same BIC. Another problem concerns whether the model that really generated the data is among those fit. The answer to this second problem is twofold. The BIC can identify the "best" model from among those considered. Substantive theory can restrict the range of these models, but at the risk of missing a better fitting model.

Hypothesis Testing

Macready & Dayton (1980) mentioned various specific state mastery learning models and showed how these are related to Model 1 for a single time point. Of course, all of these models may be considered when measurements are available from more than one occasion.

The multiple-indicator latent Markov model offers great flexibility in testing hypotheses about change, all of which fit in with the general 2×2 classification shown in Table 2. If examinees are allowed to change (cells G and H), the model becomes a latent Markov model with items (indicators) restricted to be time-homogeneous or to be free across time. If data are available from more than two points in time, transition stationarity may be assumed and matrices of transition probabilities may be constrained to be equal across time (time-homogeneous). As with structural equation models, the initial latent distribution (vector of δs) together with matrices of transition probabilities give the structural part of the model. Many hypotheses may result by putting additional constraints on the transition probabilities.

Table 2
A 2×2 Classification of Multiple-Indicator Latent Markov Models

Change in Examinee	Change in indicator: No	Change in indicator: Yes
No	E	F
Yes	G	H

For example, if no decay is hypothesized all entries in the lower triangle of the transition probability matrices are fixed to 0. If one-state progress only [called slow growth by Collins & Wugalter (1992)] is hypothesized, additional elements in the upper triangle of the transition probability matrices are fixed to 0. The measurement part of the model (i.e., the relationship between the latent variables and the manifest indicators) is given by the conditional response probabilities. Apart from constraining the measurement model to be equal across time, various additional specific measurement models can be hypothesized if the latent variables are considered to be ordered categorical (see above).

If all transition probability matrices are restricted to the identity matrix, no change is allowed for examinees. The latent Markov model thus reduces to a latent class model (cells E and F of Table 2) with or without items allowed to change over time. The latter case is the most restrictive model with no change at all.

Note that for two points in time, Models 2–4 also may be formalized as special cases of a latent class model by combining latent variables X and Y into a single variable with as many classes as there are nonempty cells in the transition matrix of a given model (see, e.g., Goodman, 1974; Hagenaars, 1990). For more than two occasions, however, it becomes impossible to handle most situations using a classical latent class approach such as Goodman's (1974) or Haberman's (1979) formalization. A typical example is the assumption of stationarity of transition probabilities.

Applications to Solving Arithmetic Word Problems

Items and Problem-Solving Models

Table 3 shows the test items used in this study. Arithmetic problems require different competencies for their solution. Item 1 in Table 3 can be solved using an action-based counting procedure because the information given in the text can be processed successively. In contrast, Item 2, although it requires the same mathematical operation as Item 1, requires the transformation into an equation, such as $x + 3 = 7$, because the first sentence cannot be represented in an action-based counting procedure. Item 3 requires advanced mathematical competencies because two equations must be combined. Development in performance on Items 1–3 depends on the growth in one dimension (i.e., growth in mathematical competencies). To solve Item 4, first the number of rabbits that Tom has must be determined. Then this information is used to answer the question. Item 5 involves a similar process. This item is especially difficult because two relational statements are involved.

Table 3
Items Used in The Study

Item	Problem/Item Wording
1	Peter had 7 sweets. Then he gave 2 sweets to Susan. How many sweets does Peter have now?
2	Peter had some sweets. Then Jack gave 3 additional sweets to Peter. Now Peter has 7 sweets. How many sweets did Peter have in the beginning?
3	Jack and Beth have 6 apples altogether. Jack has 2 apples. Ken and Ina have 9 apples altogether. Ken has 5 apples. How many apples do Beth and Ina have altogether?
4	John has 7 rabbits. He has 4 more rabbits than Tom. How many rabbits do John and Tom have altogether?
5	Joyce has 7 marbles. She has 2 more marbles than Tom has. Oliver has 3 more marbles than Tom. How many marbles does Oliver have?

Items 3, 4, and 5 require an advanced level of abstract mathematical part-whole representation. In addition, Items 4 and 5 require an advanced understanding of quantitative relations. These items may therefore be considered indicators of a latent categorical variable termed "flexibility in dealing with complex prob-

lems." Because these items were designed a priori to have approximately equal difficulty, a simple two-state mastery model was assumed. With respect to transitions, it was assumed that progress only (i.e., no decay) took place across time because such complex problems can only be understood and solved if examinees already understand quantitative part-whole relations at a high level and use mathematical strategies. The model for Items 3, 4, and 5 was called Model X.

Items 1, 2, and 3, on the other hand, clearly differ in their requirement of mathematical strategies and thus form a hierarchy. These items, therefore, were considered indicators of an ordered categorical latent variable termed "availability of mathematical strategies." For these three items, a four-state Guttman-type latent distance model was assumed (Model Y), again with progress across time only.

Data

The data were collected in a longitudinal study by the Max Planck Institute of Psychological Research in Munich, Germany (Renkl & Stern, 1994). 1,453 children from 54 elementary school classrooms were administered several group tests. The data analyzed here were the arithmetic word problems shown in Table 3 presented on three occasions: the beginning of second grade (mean age 7 years, 9 months), the end of second grade (mean age 8 years, 3 months), and the beginning of third grade (mean age 8 years, 9 months). At each measurement point, the children were presented with 12 to 20 different arithmetic word problems. Only the five problems discussed here were presented at all three measurement points. Only examinees who participated in all three measurements were included in the analysis. The actual sample size thus was reduced to $N = 965$.

An item was scored correct if the correct answer and an appropriate mathematical solution was presented. For the complex items (Items 3, 4, and 5), it was sufficient to provide only one step of the solution. Thus, five pass/fail items were measured repeatedly at three occasions.

Results

In principle, the performance of both Model X and Model Y could be evaluated using three indicators each measured at three points in time. However, at $t = 1$, Items 4 and 5 had extremely low solution rates that increased considerably from $t = 1$ to $t = 2$ and became similar to those of Item 3. In testing Model X, therefore, it was decided to use data from $t = 2$ and $t = 3$ only. Although both Model X and Model Y were considered reasonable models for describing the data, other models might have described the data equally well or better. Therefore, a series of models were fit to the data.

Model X and Related Models for Items 3, 4, and 5

The contingency table resulting from the measurement of three binary items at two points in time is of size $2^{T'} = 2^6 = 64$ $(T' = K \times T = 3 \times 2)$. Models A through D were variants of Model 2 above for 2×3 items. Model F extended Model 2 to a multiple-chain model (i.e., Model 3 with $S = 2$ chains). All of these models differed with respect to the assumptions made about the conditional response probabilities and the form of the transition matrix from $t = 2$ to $t = 3$. Results of various model tests are given in Table 4. None of the models given in Table 4 fit on purely statistical grounds.

In Model A, the most restrictive model, it was assumed that neither items nor examinees changed across time—the conditional response probabilities were restricted to be equal across time, and the latent transition matrix was equal to the identity matrix. This is indicated in Table 4 by – signs under the columns labeled "item change" and "examinee change." In Model B, items were allowed to change across time (e.g., to become easier). In Model X, time-homogeneous item probabilities were assumed and latent change without decay was allowed. To be sure that the latter assumption was consistent with the data, no constraints were imposed on the transition matrix of Model C.

Table 4
Change in Items and/or Examinees (– = No, + = Yes), the Form of the Latent Transition Matrix **T** (**I** = Identity Matrix, r = Entries in the Transition Matrix Restricted to 0, and f = Full Unrestricted Transition Matrix) for Models A, B, X, C, D, and F and Items 3, 4, and 5, and Fit Statistics

Model and Number of Classes	Item Change	Examinee Change	Latent Transition Matrix	*df*	Fit Statistic G^2	Fit Statistic X^2	Fit Statistic BIC
A, 2	–	–	**I**	56	199.4	205.6	7,102
B, 2	+	–	**I**	50	99	100.7	7,044
X, 2	–	+	r	55	118.8	108.4	7,028
C, 2	–	+	f	54	118.8	108.4	7,035
D, 2	+	+	f	48	86.8	85.9	7,045
F, 2 × 2	–	+,–	r, **I**	53	89.4	77.8	7,013

Estimated parameter values of Model C converged to those of Model X. That is, the data supported the assumption that $\tau_{21} = 0$ for Model X. In Model D, it was asumed that both items and examinees changed across time.

Because there were no problems with sparse data and both G^2 and X^2 led to the same conclusion, the Read-Cressie statistic was not provided. Therefore, none of the models considered thus far fit in purely statistical terms. Note, however, that BIC favored Model X. Also, in comparison with Model A (no change at all), only one degree of freedom was lost and a considerable improvement in fit was obtained.

Because of the unsatisfactory fit of Models A–D, several extensions of Model 2 were considered that allowed for unobserved heterogeneity. A model that has been a viable candidate in previous research is the "partially latent mover–stayer" model (Langeheine & Van de Pol, 1990). This model, denoted Model F in Table 4, extended the single-chain latent Markov model to a two-chain model (i.e., Model 3). In the present case, the mover chain was assumed to be equal in definition to that of Model X, whereas stayers were assumed to be measured without error (i.e., response probabilities were fixed at 0.0 or 1.0) and, by definition, to stay in their respective state with probability 1.0 across time. Therefore, compared to Model X, two additional parameters (one chain proportion and one state proportion at $t = 2$) had to be estimated. Although Model F led to some improvement in fit over Model X, the fit was still unsatisfactory in absolute terms. However, the BIC fared slightly better for Model F than for Model X.

Table 5 provides the estimated parameter values for Model X and Model F. For Model X, the proportion of masters (Class 2, characterized by high probabilities in passing all three items) at $t = 2$ was .46. Class 1 contained nonmasters with low probabilities for correct responses on all items. Out of the 54% who were nonmasters, 22% made progress to the mastery class from $t = 2$ to $t = 3$. Writing the initial probabilities in row-vector $\boldsymbol{\delta}^2$ and transition probabilities from $t = 2$ to $t = 3$ of

$$\mathbf{T}^{32} = \begin{vmatrix} \tau_{11} & \tau_{21} \\ \tau_{12} & \tau_{22} \end{vmatrix}, \tag{6}$$

the latent distribution at $t = 3$ is given by $\boldsymbol{\delta}^3 = \boldsymbol{\delta}^2 \mathbf{T}^{32}$ or more generally

$$\boldsymbol{\delta}^t = \boldsymbol{\delta}^{t-1} \mathbf{T}^{t, t-1}. \tag{7}$$

Thus, $\boldsymbol{\delta}^3 = (.42, .58)$, indicating a progress of 12%.

According to Model F, 86% were movers with initial proportions of masters and nonmasters approximately equal to those in Model X. Due to the fact that perfect masters (Class 2 of the stayers) and pure nonmasters (Class 1 of the stayers) were allocated to a separate chain, measurement precision was somewhat lower in the mover chain compared to Model X. Consequently, turnover was estimated to be somewhat

Table 5
Estimated Parameter Values for Models X and F: Chain Proportion (π), Intitial Distribution (Class and δ), Probability of Correct Response to Items 3, 4, and 5 (ρ) and Transition Probabilities (τ)

Model and π	Initial Distribution: Class	Initial Distribution: δ	Item	ρ	Transition Probability τ^{32}	
Model X, $\pi = 1$	1	.54	3	.19	.78	.22
			4	.10		
			5	.14		
	2	.46	3	.77	0	1
			4	.69		
			5	.75		
Model F, $\pi_1 = .86$	1	.55	3	.26	.69	.31
			4	.14		
			5	.20		
	2	.45	3	.74	0	1
			4	.66		
			5	.73		
Model F, $\pi_2 = .14$	1	.65	3	0	1	0
			4	0		
			5	0		
	2	.35	3	1	0	1
			4	1		
			5	1		

Note. Conditional response probabilities were assumed to be time-homogeneous. Parameter values of 0 and 1 were fixed by definition.

higher for the movers, resulting in a marginal latent distribution of $\boldsymbol{\delta}^3 = (.38, .62)$.

Note that Model F also may be formalized as a one-chain, four-state, latent Markov model. The transition matrix is shown in Table 6 (cf. Langeheine & Van de Pol, 1990), where Class 1 was pure nonmasters, Class 2 was nonmasters, Class 3 was masters, and Class 4 was perfect masters. These four classes, characterized by conditional response probabilities (see Table 5), clearly reveal an order from pure nonmastery to perfect mastery. Of course, standard errors of parameter estimates (not reported) were somewhat higher for Model F (between .03 and .09) than for Model X (between .02 and .03).

Table 6
One-Chain, Four-State Markov Model Transition Matrix (1 = Pure Nonmasters, 2 = Nonmasters, 3 = Masters, and 4 = Perfect Masters)

State at $t = 2$	State at $t = 3$: 1	2	3	4
1	1	0	0	0
2	0	.69	.31	0
3	0	0	1	0
4	0	0	0	1

With respect to the total sample, for Model F there was a small proportion of examinees classified as permanent pure nonmasters ($.14 \times .65 = .09$) and an even smaller proportion classified as perfect masters throughout time ($.14 \times .35 = .05$). The large majority of examinees (86%) were classified as either nonmasters or masters at $t = 2$, with some progress of the nonmasters across time ($.86 \times .55 \times .31 = .15$ in proportion).

Model Y and Related Models for Items 1, 2, and 3

For Model Y, the contingency table was of size 512 (three binary items measured at three occasions). With a sample size of $N = 965$, there was a problem of sparseness because there were 363 zero cells in the table. All models considered below were variants of Model 2 above; however, they were adjusted for having 3×3 items.

Models I through IVb in Table 7 were two-state models, thus a simple two-state mastery model was assumed. Models Ya through Vb were four-state latent distance models (see Table 1), and item-specific error rates were assumed; that is, each item was characterized by a single parameter across classes that, in addition, was assumed to be time-homogeneous.

Table 7
Change in Items and/or Examinees (– = No, + = Yes), the Form of the Latent Transition Matrix **T** (**I** = Identity Matrix, r = Entries in the Lower Triangular are Restricted to 0, f = Full Matrix, Hom = Time-Homogeneous, Het = Time-Heterogeneous) for Model Y and Rival Models for Items 1, 2, and 3, and Fit Statistics

Model and Number of States	Item Change	Examinee Change	Latent Transition **T**	*df*	Fit Statistic: $\lambda \to 0$ G^2	$\lambda = 2/3$	$\lambda = 1$ X^2	BIC
I, 2	–	–	**I**	504	584	631	738	8,343
IIa, 2	–	+	f, Hom	502	409	505	730	8,181
IIb, 2	–	+	f, Het	500	402	487	686	8,188
III, 2	+	–	**I**	492	325	340	549	8,166
IVa, 2	+	+	f, Hom	490	305	368	514	8,160
IVb, 2	+	+	f, Het	488	304	367	509	8,174
Ya, 4	–	+	r, Hom	499	469	619	958	8,262
Yb, 4	–	+	r, Het	493	456	592	879	8,291
Va, 4	–	+	f, Hom	493	311	329	386	8,146
Vb, 4	–	+	f, Het	481	295	317	376	8,212

Model I was the "no change at all" model and was considered a baseline model for other two-state models. Models IIa and IIb allowed for examinee-specific change. Transitions for Model IIa were assumed to be time-homogeneous, and for Model IIb they were assumed to be time-heterogeneous. Model III was characterized by item-specific change only. Models IVa and IVb allowed for both item- and examinee-specific change across time, again with time homogeneous (Model IVa) versus heterogeneous (Model IVb) transitions.

For all of the four-state latent distance models (Models Ya, Yb, Va, and Vb), an ordered latent variable was assumed at each point in time. In this case, only models with examinee-specific change were included (models with both item- and examinee-specific change improved the fit only marginally). Model Y was the model hypothesized to be adequate for these data, that is, a model without decay. Model V relaxed this assumption by allowing for decay.

Table 7 shows that the power divergence statistics provided the following results:

1. Model I was rejected by all three statistics [G^2, Read-Cressie ($\lambda = 2/3$), and X^2]. For Models IIa and IIb, there was a discrepancy; X^2 led to rejection and the other two statistics favored accepting. Note that there was also a discrepancy for Models III, IVa, and IVb (X^2 was larger than the other two statistics), although all three statistics would lead to acceptance on statistical grounds alone. For the hypothesized Model Y, there was also disagreement of the statistics. Models Ya and Yb were accepted according to

G^2, but were rejected by both the Read-Cressie statistic ($\lambda = 2/3$) and X^2. For Models Va and Vb, the statistics agreed.

2. Allowing for time-heterogeneous versus time-homogeneous transitions led to marginal improvement in fit only. In all of these cases, the BIC favored the more parsimonious model.
3. Models that allowed both item and examinee specific change did no better than models that allowed for one or the other.

Thus, the hypothesized Model Ya was rejected and replaced by Model Va, which has the lowest BIC in Table 7. Table 8 contains estimated parameter values for Model Va. The model was both parsimonious (18 nonredundant estimated parameters) and well identified, with low to moderate standard errors (not reported) ranging from .01 to .09.

Table 8
Estimated Parameter Values for Model Va: Initial Distribution (δ), Probability of Correct Response to Items 1, 2, and 3 (ρ), and Transition Probabilities (τ)

					Transition Probabilities (τ)							
		ρ			t to $t+1$				$t=1$ to $t=3$			
		Item			Class				Class			
Class	δ	1	2	3	1	2	3	4	1	2	3	4
1	.05	.04	.10	.11	.20	.40	.36	.04	.08	.27	.41	.24
2	.23	.96	.10	.11	.08	.36	.51	.05	.06	.23	.40	.31
3	.50	.96	.90	.11	.02	.14	.36	.48	.03	.11	.28	.58
4	.22	.96	.90	.89	.00	.01	.15	.84	.01	.03	.19	.77

Note. Conditional response probabilities and transition probabilities were assumed to be time-homogeneous.

The estimated conditional response probabilities revealed an order to the four classes. Class 1, estimated to contain 5% of the examinees at $t = 1$, had low probabilities for passing all items (see Table 8). 23% (Class 2) had a high probability of solving the easiest item (Item 1), but low probabilities of solving the more difficult items (Items 2 and 3). 50% of the examinees (Class 3) had low probabilities of solving the most difficult item (Item 3) only. The remainder of the examinees (22%, Class 4) had high probabilities of solving all of the items correctly. Note that only three item probabilities were estimated, because of the restriction that $p(\mathrm{L}) = 1 - p(\mathrm{H})$ for each item across classes and because of assuming time-homogeneous item probabilities.

Latent transition probabilities are given in Table 8 for both t to $t+1$ (assumed to be constant over time) and $t = 1$ to $t = 3$. For t to $t+1$, examinees made considerable progress from one point in time to the next. This held especially for those in Class 1 (80% progress) but also for Classes 2 (56% progress) and 3 (48% progress). However, there was some non-negligible decay for Classes 2 (8%), 3 (16%), and 4 (16%) because Model Va (with decay) fit better than Model Ya (without decay): G^2 difference was $469 - 311 = 158$, with degrees of freedom $= 499 - 493 = 6$. The $t = 1$ to $t = 3$ transition probabilities show what happened over the entire time span. Of those initially having low probabilities of solving all items (Class 1), 27% did well on Item 1, 41% passed both Items 1 and 2, and 24% passed all three items at $t = 3$. Total progress in Class 2 was estimated to be 71% and 58% in Class 3. Most of the decay was to the adjacent lower state. Comparing the initial distribution and the marginal distribution at $t = 3$, $\boldsymbol{\delta}^3 = (.03, .13, .30, .54)$ showed net change across time. Classes 1, 2, and 3 decreased, whereas Class 4 increased in size. Total proportions of decay, stability, and progress were .13, .37, and .50, respectively, from $t = 1$ to $t = 3$.

Thus, there was overwhelming evidence for progress, but at the same time decay was observed in a small but significant proportion of examinees. By focusing on individual learning over time, results from quite different areas reveal that learning is not only characterized by a steady increase of performance, but rather by temporal relapses to less appropriate solution strategies. Thus, only models considering decay are appropriate for describing longitudinal change.

Conclusions

The methodology presented here makes available flexible extensions from one-time state mastery models to dynamic models. These methods have much in common with the work of Collins, Cliff, & Dent (1988), Collins & Cliff (1990), Collins & Wugalter (1992) and Macready & Dayton (1994).

Collins et al. (1988) and Collins & Cliff (1990) extended the deterministic Guttman simplex to a method for measuring dynamic constructs in longitudinal panel studies. Collins & Wugalter (1992) extended this work to what they called "latent transition analysis," in which the dynamics on the latent rather than the manifest level are the focus of interest. Their general model is similar to Model 4 developed here. The model used in their empirical data example is identical to Model 2 above. Thus far, however, there have been no applications to data from more than two occasions. Note that, for two points in time only, a (latent) multiple-indicator Markov model may be formalized as a classical latent class model with as many classes as there are nonempty entries in the matrix of transition probabilities. The approach presented here is also flexible with respect to other extensions, for example, modeling unobserved population heterogeneity using multiple chains.

Macready & Dayton (1994) also used a classical latent class formalization for longitudinal assessment of trait acquisition. They showed how various hypotheses about migration may be tested using latent class models for several items measured at three points in time and, in addition, with several groups. However, the simultaneous group analysis approach of Clogg & Goodman (1984, 1985) has several limitations in this context; for example, (1) unconstrained migration implies the assumption of a rather complicated second-order latent Markov model; (2) for more than two points in time, the classical latent class approach does not allow for specifying models with time-homogeneous transition probabilities; (3) it is rather difficult and laborious to fit these models using Clogg's (1977) MLLSA program. The approach presented here and the program PANMARK (Van de Pol et al., 1991) can easily handle these and a wide variety of other cases.

Of course, the approach presented here is not without limitations. For Model X (and several other models), multiple-indicator Markov models often resulted in unacceptable fit. To address this problem, structural equation modeling methods have included effects for correlated errors in, for example, LISREL (Jöreskog & Sörbom, 1984) models. In the Markov chain context, this means extending a model by direct effects between indicators over time. Such effects cannot be handled because the parameters estimated in the approach presented here are the probabilities. In principle, such models may be fit using Haberman's (1979, 1988) software for latent class analysis. Practically, however, it is unrealistic to do so because of at least the following two problems. First, Haberman's programs require user-supplied design matrices that become rather large very quickly. For Model Yb, for example, this matrix would be of size $32{,}768 \times 18$. Second, it is not possible to fit models with time-homogeneous transition probabilities.

As shown here, models assuming ordered categorical latent variables may be fit by specifying that the measurement part of the model follows some latent distance model. However, latent distance models imply severe constraints on the conditional response probabilities in that item response functions follow step functions (cf. Langeheine, 1988). Ordered classes, defined by nonintersecting item profiles also may be obtained by applying less severe inequality constraints (Croon, 1990). Currently, the latter cannot be handled.

Another problem is missing data due to attrition. In principle, the approach presented here allows the determination of whether data are (completely) missing at random or not missing at random. For the data analyzed above, only examinees with valid observations at all occasions were retained. In fact, however, the data were probably not missing at random because some teachers refused to participate in the study at certain points in time. Consequently, there was a dropout of full classes.

Despite these limitations, the methodology presented here appears well suited to describe progress in mathematical competencies at different levels of development. Model X described progress in advanced mathematical skills in dealing with complex problems. Models Y and V described change from action-based

strategies to building equations developed from abstract problem models based on part-whole knowledge. Of course, the methodology is not restricted to this field. For applications in other settings see, for example, Graham, Collins, Wugalter, Chung, & Hansen (1991) on substance use prevention; Rendtel, Langeheine, & Berntsen (1992) on poverty research; and Van de Pol & Langeheine (1992) on labor market data. In contrast to mean-change approaches [cf., e.g., the review of Spada & McGaw (1985) on linear logistic test models for measuring change], the models considered here focus on groups or classes of examinees (defined by item profiles, the ρs) and how they change. These models are therefore well suited for measuring qualitative, structural change.

References

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

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

Clogg, C. C. (1977). *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users* (Working Paper 1977-09). University Park PA: Population Issues Research Center.

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79,* 762–771.

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several populations. In N. B. Tuma (Ed.), *Sociological methodology 1985* (pp. 81–110). San Francisco: Jossey-Bass.

Clogg, C. C., & Sawyer, D. O. (1981). A comparison of alternative models for analyzing the scalability of response patterns. In S. Leinhardt (Ed.), *Sociological methodology 1981* (pp. 240–280). San Francisco: Jossey-Bass.

Collins, L. M., & Cliff, N. (1990). Using the longitudinal Guttman simplex as a basis for measuring growth. *Psychological Bulletin, 108,* 128–134.

Collins, L. M., Cliff, N., & Dent, C. W. (1988). The longitudinal Guttman simplex: A new methodology for measurement of dynamic constructs in longitudinal panel studies. *Applied Psychological Measurement, 12,* 217–230.

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

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43,* 171–192.

Dayton, C. M., & Macready, G. B. (1980). A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika, 45,* 343–356.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology, 79,* 1179–1259.

Graham, J. W., Collins, L. M., Wugalter, S. E., Chung, N. K., & Hansen, W. B. (1991). Modeling transitions in latent stage-sequential processes: A substance use prevention example. *Journal of Consulting and Clinical Psychology, 59,* 48–57.

Haberman, S. J. (1979). *Analysis of qualitative data: New developments* (Vol. 2). New York: Academic Press.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. In C. C. Clogg (Ed.), *Sociological methodology 1988* (pp. 193–211). Washington DC: American Sociological Association.

Hagenaars, J. A. (1990). *Categorical longitudinal data. Log-linear panel, trend, and cohort analysis.* Newbury Park CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide.* Chicago: National Educational Resources.

Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics, 23,* 243–249.

Langeheine, R. (1988). New developments in latent class theory. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 77–108). New York: Plenum.

Langeheine, R. (1991). Latente Markov-Modelle zur Evaluation von Stufentheorien der Entwicklung [Latent Markov models for evaluating stage theories of development]. *Empirische Pädagogik, 5,* 169–189.

Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research, 18,* 416–441.

Langeheine, R., & van de Pol, F. (1993). Multiple indicator Markov models. In R. Steyer, K. F. Wender, & K. F. Widaman (Eds.), *Psychometric methodology. Proceedings of the 7th European Meeting of the Psychometric Society in Trier* (pp. 248–252). Stuttgart: Fischer.

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

Macready, G. B., & Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4,* 493–516.

Macready, G. B., & Dayton, C. M. (1994). Latent class models for longitudinal assessment of trait acquisition. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 245–273). Thousand Oaks CA: Sage.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46,* 133–158.

Raftery, A. E. (1986). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, 48,* (Series B), 249–250.

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer.

Rendtel, U., Langeheine, R., & Berntsen, R. (1992, June). *Household income: Self-assessed or computed from components? The estimation of poverty-dynamics using different household income measures.* Paper presented at the International Conference on Social Science Methodology, Trento, Italy.

Renkl, A., & Stern, E. (1994). Die Bedeutung von kognitiven Eingangsvoraussetzungen und Lernaufgaben für das Lösen von einfachen und komplexen Textaufgaben [The significance of cognitive prerequisites and learning opportunities at school for solving simple and complex mathematical word problems]. *Zeitschrift für Pädagogische Psychologie, 8,* 27–39.

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

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 169–191). Orlando FL: Academic Press.

van de Pol, F., & de Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research, 15,* 118–141.

van de Pol, F., & Langeheine, R. (1990). Mixed Markov latent class models. In C. C. Clogg (Ed.), *Sociological methodology 1990* (pp. 213–247). Oxford: Blackwell.

van de Pol, F., & Langeheine, R. (1992, June). *Analysing measurement error in quasi-experimental data. An application of latent class models to labor market data.* Paper presented at the International Conference on Social Science Methodology, Trento, Italy.

van de Pol, F., Langeheine, R., & de Jong, W. (1991). *PANMARK user manual. PANel analysis using MARKov chains. Version 2.2* (3rd ed.). Voorburg: Netherlands Central Bureau of Statistics.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

Author's Address

Send requests for reprints or further information to Rolf Langeheine, IPN, the University of Kiel, Olshausenstrasse 62, D-24098 Kiel, Germany. Internet: npn27@rz.uni-kiel.d400.de.