Fachbereich Erziehungswissenschaft und Psychologie

der Freien Universität Berlin

Modeling latent change in categorical variables

Dissertation

zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr. phil.)

vorgelegt von

Dipl.-Psych. Claudia Crayen

Berlin, 2015

Erstgutachter:

(First Advisor)

Prof. Dr. Michael Eid

Freie Universität Berlin

Zweitgutachter:

(Second Advisor)

Prof. Dr. Jeroen Vermunt

Tilburg University

Datum der Disputation:

(Date of defense)

16.07.2015

Acknowledgments

First of all, I'd like to thank my adivsors Michael Eid and Jeroen Vermunt. Michael, thank you for your constant support and trust and patience. Jeroen, thank you for your enlightening emails in dark moments. I feel very fortunate to be in contact with two such great minds and open hearts.

I'd also like to thank the fellows transitioning through the methodology group at FU Berlin for making that particular state so very pleasant (in approximate order of appearance): Tanja Lischetzke, Fridtjof Nussbeck, Christian Geiser, Maike Luhmann, Natalie Schüz, Luna Beck, Irina Kumschick, Martin Wertenbruch, Tobias Koch, Georg Hosoya, Jana Mahlke, Tanja Kutscher, Martin Schultze, Fenne große Deters, Jana Holtmann, and Johannes Bohn, with recurring guest appearances by Leona Aiken, Steve West, Hugo Carreteira-Dios, and Christopher Beam.

Sophie, Jana, Fadi and Christian, thanks for helping out with formatting on very short notice.

I'd like to thank my parents for their high hopes, low control and for knowing when to step in.

Last but not least, I love and thank my husband and son for being more important than academics without being academic about it.

Contents

Summary

This thesis aims at extending statistical models for social and behavioral science data that allow capturing change over time in categorical variables and at making them more accessible for applied researchers. Change is considered on the level of latent variables that are corrected for measurement error. The concept of change is briefly summarized, pointing out aspects that are relevant in choosing an appropriate longitudinal statistical model. One important aspect is the timescale of the change process under investigation. Nesselroade (1991) established the distinction between long-term *change* and short-term *fluctuation*. Examples for both phenomena are covered here. In study one, a longitudinal structural equation model for multitrait-multimethod data (Geiser, 2009) is extended to multiple groups and categorical indicators. In an application to a data set with parent and teacher ratings for 659 young children, the treatment effect of an intervention program is estimated as the group difference in mean change. In study two, it is illustrated how interindividual differences in intraindividual mood fluctuation patterns can be identified by applying mixture latent Markov models (Vermunt, Tran, Magidson, 2008). Data from an ambulatory assessment study ($N = 164$ students with up to 56 repeated measurement occasions) are considered. The model was extended to fit the nested structure of measurement occasions within days (Vermunt, 2009). Two latent classes that differ with regard to their mood fluctuation pattern are identified and related to self-report measures of mood regulation competencies. In contrast to study one, both, the manifest indicators and the latent variable (states) are categorical in nature. In study three, the model obtained in study two is extended to account for varying time intervals between measurement occasions, incorporating continuous-time parameters (Böckenholt, 2005). A simulation study is conducted to explore parameter recovery qualities with small sample sizes for the continuous-time mixture latent Markov model and compare them to the discrete-time model. Advantages and limitations of the models are discussed.

Chapter 1

Introduction

The aim of this thesis is the extension of longitudinal latent variable models for social and behavioral science data. Emphasis is also put on making these models more accessible for applied researchers. Change is considered on the level of latent variables that are corrected for measurement error. The concept of change is briefly summarized, pointing out aspects that are relevant in choosing an appropriate longitudinal statistical model. In the main body, three studies are combined. In each, a latent variable model has been adapted and applied to match the specific change process measured by categorical variables. At the end of this chapter, I will place the models and processes of the three studies within the framework provided.

1.1 Change and variability

For a long time, emphasis in personality research was on stable interindividual differences in *traits* and their ability to predict behavior. In this context, trait *change* describes a long-term shift in such an (otherwise stable) trait. Trait change is mostly expected in developmental settings (growth and decline) or after an intervention (e.g., a therapy). Situational variability, on the other hand, was background noise to the assessment of traits. It was only with latent state trait theory (LST; Steyer, Ferring, & Schmitt, 1992; Steyer, Schmitt, & Eid, 1999) that both stable and situational aspects of personality were integrated into a measurement theory, quantifying the relative size of reliably measured trait and state specific variance. Nesselroade

(1991, 2001) considers a similar concept from a developmental perspective. In his taxonomy, intraindividual *change* is defined as enduring and developmental, while intraindividual *variability* is mostly reversible and manifests on a shorter timescale. Based on the ideas of LST theory, it has become quite popular to assess stable {and variable aspects of personality constructs (e.g., life satisfaction, Lucas Donnellan, 2007). The measurement of intraindividual variability with ambulatory assessment techniques is also on the rise (e.g., Mehl & Connor, 2012). Interestingly, with interindividual differences in intraindividual variability commonly studied, variability itself has become some trait-like attribute and is integrated into personality theory (e.g., Fleeson, 2004, 2007). Another advantage of ambulatory assessment is the multitude of recorded information of situational influences that can be used to predict behavior. Therefore, this data acquisition technique often serves to identify preconditions of maladaptive behavior in health and clinical psychology (e.g., Tyler, Jones, Black, Carter, & Barrowclough, 2015).

1.2 Longitudinal data

Very generally speaking, longitudinal data sets can be described by their number of time points, the number of subjects, the number of variables per measurement occasion, and the spacing of the measurement occasions. Number and spacing of occasions are dependent on the timescale of the process of interest. With long-term *change* slowly unfolding, spacing is usually weeks to years and the number of measurement occasions is at least two. Large scale panel studies may have many time points spaced about a year apart, which allows testing assumptions about the form of the mean change process. The variability between subjects in the spacing is usually small in relation to the timescale of the process. Studies assessing short-term *variability* will have fewer subjects, but more measurement occasions. Both types of studies, intervention/panel studies with many subjects and ambulatory assessment studies with many measurement occasions, are limited with regard to the variable set that can be administered. Because of high costs and little space, scales are often short and items have few categories. In this case, treating the data as categorical has several advantages: Measurement invariance can be tested on the level of the item categories and no assumptions for continuous scales have to be met.

1.3 Basic latent variable models for categorical longitudinal data

Because questionnaires or ratings are never perfectly reliable, only multiple indicator models that take occasion specific measurement error into account are considered here. With manifest categorical indicators, longitudinal latent variable models can be classified according to the nature of the latent occasion-specific variable (continuous or categorical) and according to the number of time points they might cover. For a simple intervention study, there are only two measurement occasions.

Usually, models with continuous latent variables (factors) are based on manifest continuous indicators. However, categorical indicators can also be linked to a continuous latent variable on each of two measurement occasions. Here, a continuous latent variable underlying the available categorical variable is assumed. The categories are the result of coarse measurement of the latent dimension. The manifest categorical variables can be linked to a continuous latent variable via parameters that define points on the continuum that separate two neighboring categories, so-called threshold parameters. Latent change models developed for continuous data can be modified to include these threshold parameters. The modifications have some important implications that I will address in study 1. In such a model, a latent change factor can be formulated (Steyer, Partchev, & Shanahan, 2000), making interindividual differences in intraindividual change accessible. There may be no mean change across the sample, but one could still include predictors for differential change. With few measurement occasions, latent change models can be extended to include more than one latent change factor in which change is expressed relative to a reference occasion. Latent change models do not impose any restrictions on the form of the change process. When the form of the change process (or trajectory) is assumed to be, for example, linear, latent growth curve models (LGM; Duncan, Duncan, & Strycker, 2006) can be applied. Similarly to the latent change model, interindividual differences in the change process are included. In a data situation where many closely spaced measurement occasions are available, the change process will most likely not be linear. Instead, Latent-State-Trait models (Eid & Langeheine, 1999) can be used. Here, a latent trait captures the stable variance and variability is expressed in occasion-specific residual factors. If dependencies across consecutive occasions persist, an autoregressive

structure can be imposed.

In models with categorical latent variables (latent classes), the categories of the manifest variables are seen as distinct and qualitative different (albeit ordered) states. They are linked to the states of the latent class variable by response probabilities. Measurement error is considered because the response probability of a certain manifest category given the corresponding latent category does not have to equal one. Change from one measurement occasion to the other can be modeled by means of transition probabilities between the latent categories. This form of a dynamic latent class is called latent transition analysis (LTA). It is also feasible with more than two measurement occasions. When the number of occasions becomes large, transitions from one measurement occasion to the next are usually restricted to be equal across occasions and the latent Markov model is obtained. An overview over longitudinal latent class models gives Eid (2007).

These are the very basic latent variable models for categorical longitudinal data. They have been extended in many ways. Multilevel extensions take group dependencies within the data into account. Mixture extensions allow for population heterogeneity. Multitrait-multimethod models add a method dimension to each state-occasion unit.

1.3.1 Measurement invariance

Measurement invariance concerns the equivalence of parameters in the measurement part of the model across measurement occasions and/or groups. It is an important issue in longitudinal modeling, because it secures that psychometrically, the measured occasion-specific states hold the same meaning (same properties of measure). Measured change is in these cases interpretable as change in the construct itself, not in the measure. For example, a measure may change its meaning between two measurement occasion of a developmental study, because children mature and react differently to the same questions. In ANOVA, measurement invariance is usually asssumed, while in latent variable models, it can be tested. In the case of categorical indicators, the parameters that measurement invariance relates to are different from the continuous case (e.g., thresholds or logits). The timescale of change is relevant here, too, because the shorter the intervals between measurement occasions, the less likely it is that the properties of the measure

change. In study 1, measurement invariance across occasions and groups is established for ordinal indicators. In study 2, measurement invariance for multinomial indicators is established across time, but differences between unobserved subgroups exist.

1.3.2 Indicator specificity

Another phenomenon in longitudinal modeling concerns the autocorrelation of measures. When indicators of longitudinal multiple indicator models are repeatedly measured, they tend to exhibit autocorrelation, i.e., systematic variance that is not shared with the other indicator(s) of the measurement model. In cross-sectional models, this proportion of variance would just be counted as measurement error. We will see in study 1, how this can be dealt with accordingly. In study 2, the model itself contains an autoregressive structure. Here, it could be tested whether higher order relationships (not only to the direct neighbour) exist.

1.4 Featured studies

From the basic latent variable models for categorical indicators, two from the opposite corner are featured in the studies: In the first study, a latent change model for two measurement occasions is be applied to an intervention study data set with young children. Timescale and number of measurement points are straight-forward: Two measurement occasions, before and after the implementation of the intervention. Because the data set exhibits a multitrait-multimethod (MTMM) data structure with two ratings on two traits, a correlated-state-correlated-methods-1 model (Geiser, 2009) is extended to multiple groups (intervention and control group) and ordered categorical indicators. The MTMM structure allows to assess the convergent validity of true (latent) change. Particular attention is paid to the assessment of measurement invariance in the case of multiple occasions and multiple groups and ordered categorical indicators.

For the second study, an ambulatory assessment study with 164 students was conducted to learn more about mood fluctuation patterns and mood regulation competencies. Here, there are up to 56 measurement occasions per subject and the mood fluctuation process is assumed to be best represented by a stationary first order Markov process on the level of latent state variables. However, two complications arise: The measurement occasions are nested in days (8 occasions

per day), and the assumption of a homogenous fluctuation process for all subjects in the sample is relaxed, yielding thereby a hierarchical mixture latent Markov model (Rijmen, Vansteelandt, & de Boeck, 2008; Vermunt, 2010) that had not been applied outside of the illustrating example of Rijmen et al. (2008).

In the third study, a shortcoming of the hierarchical mixture latent Markov model in the application to AA data is addressed, the need to account for interindividually varying time intervals between measurement occasions. This is done by extending the hierarchical mixture latent Markov model to incorporate continuous-time transition parameters (e.g., Böckenholt, 2005). To learn more about the performance of this model and differences to the discrete-time model that was applied in study two, a Monte-Carlo simulation study is conducted with a focus on small sample sizes. Guidelines for data requirements are given.

1.5 References

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues and Applications.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Eid, M. (2007). Latent-class models for analyzing variability and change. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (pp. 591-607). Oxford: Oxford University Press.

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

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology, 80*, 1011

Fleeson, W. (2004). Moving personality beyond the person-situation debate the challenge and the opportunity of within-person variability *Current Directions in Psychological Science, 13*, 83-87

Geiser, C. (2009). *Multitrait-multimethod-multioccasion modeling.* Munich, Germany: AVM.

Lucas, R. E. & Donnellan, M. B. (2007). How stable is happiness? Using the STARTS model to estimate the stability of life satisfaction. *Journal of Research in Personality, 41*, 1091-1098.

Mehl, M. R. & Connor, T. S. (2012). *Handbook of Research Methods for Studying Daily Life.* New York: Guildford Press.

Nesselroade, J. R. (1991). The warp and the woof of the developmental fabric. In R. M. Downs, L. S. Liben, & D. S. Palermo (Eds.), Visions of aesthetics, the environment, and development: The legacy of Joachim F. Wohlwill (pp. 213-240). Hillsdale, NJ: Erlbaum.

Nesselroade, J. R. (2001). Intraindividual variability in the development within and between individuals. *European Psychologist, 6,* 187-193.

Steyer, R., Ferring, D. & Schmitt, M. J. (1992). *On the definition of states and traits.* Trier: Universität, Fachbereich I - Psychologie.

Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*, 389-408.

Steyer, R., Partchev, I., & Shanahan, M. (2000). Modeling true intra-individual change in structural equation models: The case of poverty and children's psychosocial adjustment. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 109-126). Hillsdale, NJ: Erlbaum.

Tyler, E., Jones, S., Black, N., Carter. L.-A., & Barrowclough, C. (2015). The relationship between bipolar disorder and cannabis use in daily life: An experience sampling study. *PLoS ONE, 10,*(3): e0118916. doi:10.1371/journal.pone.0118916.

Chapter 2

Evaluating Interventions with Multimethod Data: A Structural Equation Modeling Approach

Crayen, C., Geiser, C., Scheithauer, H., & Eid, M. (2011). Evaluating Interventions with Multimethod Data: A Structural Equation Modeling Approach. *Structural Equation Modeling, 18*, 497-524. doi: 10.1080/10705511.2011.607068

Abstract

In many intervention and evaluation studies, outcome variables are assessed using a multimethod approach comparing multiple groups over time. In this article, we show how evaluation data obtained from a complex multitrait-multimethod-multioccasion-multigroup design can be analyzed with structural equation models. In particular, we show how the SEM approach can be used to (1) handle ordinal items as indicators, (2) test measurement invariance, and (3) test the means of the latent variables to examine treatment effects. We present an application to data from an evaluation study of an early childhood prevention program. 659 children in intervention and control groups were rated by their parents and teachers on prosocial behavior and relational aggression before and after the program implementation. No mean change in relational aggression was found in either group, whereas an increase in prosocial behavior was found in both groups. Advantages and limitations of the proposed approach are highlighted.

Keywords: Multitrait-multimethod analysis, longitudinal confirmatory factor analysis, multiple group analysis, measurement invariance, preschool intervention

2.1 Introduction

The use of multiple methods to ensure the valid measurement of a construct found its way into common practice after Campbell and Fiske's (1959) seminal paper introduced the multitrait-multimethod (MTMM) approach. In MTMM studies, several constructs are assessed by two or more methods (e.g., ratings of depression and anxiety by therapist and patient or by mother and child). Multimethod data not only provide information about convergent and discriminant validity, but also allow capturing a construct in greater complexity, as each method contributes specific aspects and facets of a construct (Eid & Diener, 2006).

Multimethod assessment has also grown popular in evaluation research (e.g., Kochanska, Barry, Jimenez, Hollatz, & Woodard, 2009; McDowell & Parke, 2009). In those domains, multimethod measurement seems particularly advisable, given the wide scope of decisions based on study results (evaluation) and the uncertainty that is inherent in, for example, the validity of children's self-reports in many areas of psychological research. Recent applications of the MTMM approach in evaluation research include questionnaires completed by students and teachers to evaluate an education project (Stone, 2006), parent warmth rated by children and their parents (Kwok, Haine, Sandler, Ayers, Wolchik, & Tein, 2005), and aggression in early childhood rated by expert observers and teachers (Ostrov & Crick, 2007). In addition to multiple methods, evaluation studies often include multiple measurement occasions (e.g., pre-post design), as well as multiple groups (e.g., treatment- and control groups).

The analysis of data obtained from multimethod intervention and evaluation studies is the focus of this article. Data generated by such a complex design are extensive. Researchers may be tempted to break down the data set, and to simply compare observed mean values across groups using the repeated measures ANOVA or MANOVA. However, as we discuss in detail below, these approaches do not only neglect significant aspects of multimethod evaluation data, but are also limited in testing important assumptions. For example, the ANOVA and MANOVA methods do not allow testing specific hypotheses regarding the convergent and discriminant validity of different methods. Furthermore, these methods do not allow researchers to test important assumptions, such as the assumption of measurement invariance across groups and time, and they are restricted to metrical dependent variables.

Our aim is to demonstrate how structural equation models (SEM) can be used to deal with a number of important issues that typically arise in multimethod evaluation studies and that are not adequately addressed with conventional methods of data analysis. We will first provide an overview of specific issues that arise in multimethod evaluation studies and explain why conventional data analytic strategies are limited in resolving these issues. We then demonstrate how each of these problems can be dealt with by applying SEM. Later, we illustrate our SEM approach to multimethod evaluation data in an empirical application. In this study, the effects of an intervention program aimed at reducing antisocial behavior in preschool children were assessed. Finally, we discuss advantages and limitations of our approach compared to conventional data analytic strategies.

2.2 Challenges in Multimethod Evaluation Studies

2.2.1 Measurement Error

Measurement error is ubiquitous in social science data. Measurement instruments such as questionnaires or tests are never perfectly reliable. In order to obtain unbiased estimates of convergent and discriminant validity as well as associations between variables, measurement error needs to be taken into account. ANOVA and MANOVA do not explicitly address the issue of measurement error in the outcome variables, as these methods focus on observed rather than latent variables. In the framework of SEM and confirmatory factor analysis (CFA), measurement error is explicitly modeled by using multiple indicators per construct to separate the reliable variance from error variance in the manifest (observed) variables. The analysis of correlations is then carried out on the level of latent variables that only contain the reliably measured "true score" variance and are thus corrected for measurement error. SEM/CFA is therefore appropriate to deal with the problem of measurement error.

2.2.2 Multiple Constructs

The evaluation of a treatment or intervention program often involves multiple outcome variables. For example, in an intervention study, one might not only measure depressive symptoms but

also include measures of anxiety. This is beneficial for several reasons. On the one hand, the intervention could show effects beyond the target construct (e.g., depression). In this sense, multiple outcome variables allow for a broader view of the field of interest and provide the possibility to detect changes in adjacent domains. Furthermore, including multiple constructs allows for an assessment of discriminant validity by studying correlations among different outcome variables. Although MANOVA allows analyzing multiple outcome variables simultaneously, the SEM approach is more flexible in the testing of specific hypotheses with respect to convergent and discriminant validity.

2.2.3 Multiple Methods

The constructs of interest might not be fully captured by a single method. Therefore, evaluation studies are often designed to contain more than one measure of each outcome variable, following Campbell and Fiske's (1959) MTMM approach. This holds especially true for evaluation studies in a developmental or clinical context. Here, for example, self-reports might be considered insufficient and external reports are often used as an additional source of information (Achenbach, McConaughy, & Howell, 1987; Jensen et al., 1999). In the depression example, this could, for example, mean that both the patient and the therapist rate the patient's depression and anxiety. To analyze cross-sectional MTMM data, a number of CFA models have been developed in the past decades (for an overview, see Eid, Lischetzke, & Nussbeck, 2006; Marsh & Grayson, 1995; Widaman, 1985). An important feature of MTMM longitudinal data is that convergent validity of change can be analyzed and models have been developed for this purpose (Geiser, Eid, Nussbeck, Courvoisier, & Cole, 2010a, 2010b). In this article, we generalize MTMM models of change to multimethod evaluation studies including multiple groups.

2.2.4 Longitudinal Design

Most evaluation studies include at least two measurement occasions, such as in a standard pre-post design. Multiple measurement occasions (MO) are inevitable for assessing change and determining the effect of an intervention program. SEM has proven to be a flexible tool for analyzing longitudinal data, especially because — in contrast to traditional methods — SEM

allows analyzing change at the latent level (i.e., change scores corrected for measurement error) and testing important underlying assumptions such as the assumption of measurement invariance (Widaman & Reise, 1997).

Furthermore, as Burns and Haynes (2006) put it,"a single source (parent) at a single time point provides little information about the time course of the particular problem" (p. 417). Even though data of this kind are reported frequently (e.g., Biesanz & West, 2004; Burns, Walsh, & Gomez, 2003; Corwyn, 2000), CFA models especially concerned with this MTMM-MO structure have only recently been developed (Geiser, 2009; Geiser et al., 2010b; Grimm, Pianta, & Konold, 2009; LaGrange & Cole, 2008) and to our knowledge have not yet been applied to multimethod evaluation studies involving multiple groups.

2.2.5 Indicator Specificity

In longitudinal designs, a complication frequently arises when the same measures (e.g., items of a questionnaire) are repeatedly administered. Responses to item A at the first time point might be more strongly correlated to responses to that very same item at the second time point than responses to a similar (but not identical) item B. When such heterogeneous items are used as indicators in CFA, their inhomogeneity will become apparent in shared (indicator-specific) variance over time. To avoid misspecification and bias in parameter estimates, these so-called indicator-specific effects need to be taken into account. The SEM framework offers a number of different approaches for handling indicator-specific effects in longitudinal data (e.g., Eid, Schneider, & Schwenkmezger, 1999; Marsh & Grayson, 1994; Raffalovich & Bornstedt, 1987; Sörbom, 1975).

2.2.6 Measurement Invariance

The core interest in the analysis of evaluation data is to compare scores of two or more groups. Comparisons across groups (e.g., with regard to means) are based on the assumption that measurement of the constructs is comparable across groups and across time. When conducting mean comparisons by using conventional methods such as ANOVA or MANOVA, one makes the implicit assumption that measurement invariance holds across groups or across time. However, in

ANOVA and MANOVA, this assumption is not testable. When using the SEM/CFA framework, one can formally test this assumption by constraining the parameters of the measurement model to be invariant across time and groups (Cheung & Rensvold, 2002; Meredith, 1993). Only if measurement invariance holds to a sufficient degree is it tenable to draw conclusions from latent mean differences (Widaman & Reise, 1997). Depending on the degree of similarity, four levels of measurement invariance are typically distinguished for continuous measures (Millsap & Meredith, 2007; Widaman & Reise, 1997). At the lowest level (so-called *configural* invariance), only the number of latent variables and their loading patterns (allocation of indicators) are equal across groups and time. *Weak* invariance requires that the loadings themselves are equal across groups and time. In addition to invariant loadings, *strong* factorial invariance requires the intercepts of the manifest (observed) variables to be equal. *Strict* (or full) factorial invariance holds if the unique (sometimes called error or residual) variances of the indicators are also equal. Hence, in complex multimethod evaluation designs, measurement invariance needs to be tested both across measurement occasions and across groups.

2.2.7 Item-Level Data

Because long questionnaires are time consuming and costly in large studies, multimethod evaluation studies often use short scales, with few items for each construct. Item parcels of this type are prone to heterogeneity and violations of distributional assumptions and might not satisfy the requirement of a truly continuous scale. Furthermore, the use of item parcels as indicators in SEM has been criticized (e.g., Bandalos, 2002; Little, Cunningham, Shahar, & Widaman, 2002). If only a few items are available, it is often preferable to analyze item-level data instead of constructing parcels. This has the additional advantage that measurement invariance can be tested on the level of the actual item response process.

If dichotomous or ordered categorical (ordinal) items with few response categories are used as indicators in a structural equation model, specific measurement models for ordinal variables and appropriate estimators for such kind of data should be used (DiStefano, 2002). In SEM for dichotomous and ordinal variables, the categories of the items are thought to reflect divisions of an underlying latent continuous response variable (see later discussion). Consequently, the

definition of measurement invariance given in the previous section has to be slightly modified when data are analyzed at the level of single items that are ordinal rather than continuous (Millsap & Tein, 2004). The categories are linked to the underlying continuous response variable by means of threshold parameters. These thresholds need to be held constant to establish strong invariance in the ordinal case.

We outlined the key features of a typical longitudinal MTMM evaluation data set with ordinal indicators. To our knowledge, a model that takes all of these features into account has not been discussed in the current literature. In the next section, we present an MTMM-MO for the multiple group case and ordinal variables.

2.3 The CSC($M-1$) change model for ordinal indicators

The model presented here is an extension of the Correlated State-Correlated Method Minus One [CSC($M-1$)] model developed by Geiser (2009; Geiser et al., 2010b). The CSC($M-1$) model is itself an extension of the CTC($M-1$) model (Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Eid et al., 2008) to multiple occasions of measurement and the measurement of change. This study expands on the previous model by (a) formulating a model for ordinal outcomes, and (b) describing an application to data from an evaluation study with multiple groups. In this section, we briefly review the CSC($M-1$) latent change model and then discuss its extension to ordinal indicators and multiple groups. The basic idea of the CSC($M-1$) model is to choose one method as the reference method that all other methods are contrasted against (Geiser, Eid, & Nussbeck, 2008). The reference method could, for example, be a long-established "gold standard" in the field or a method of particular interest or relevance to the study. In the case of multiple raters as methods, the self-report, if available, is often selected as the reference method, because it represents an internal perspective that differs structurally from external perspectives. The reports of other raters (e.g., parent, therapist, or teacher) would then be the nonreference methods that are contrasted against the self-report. In such a model, systematic deviations of the external ratings from self-perception become apparent.

A CSC($M-1$) change model for one trait, three indicators, two methods, and two occasions of measurement is depicted in Figure 2.1. In Figure 2.1, Y_{ikl} denotes the ith indicator measured

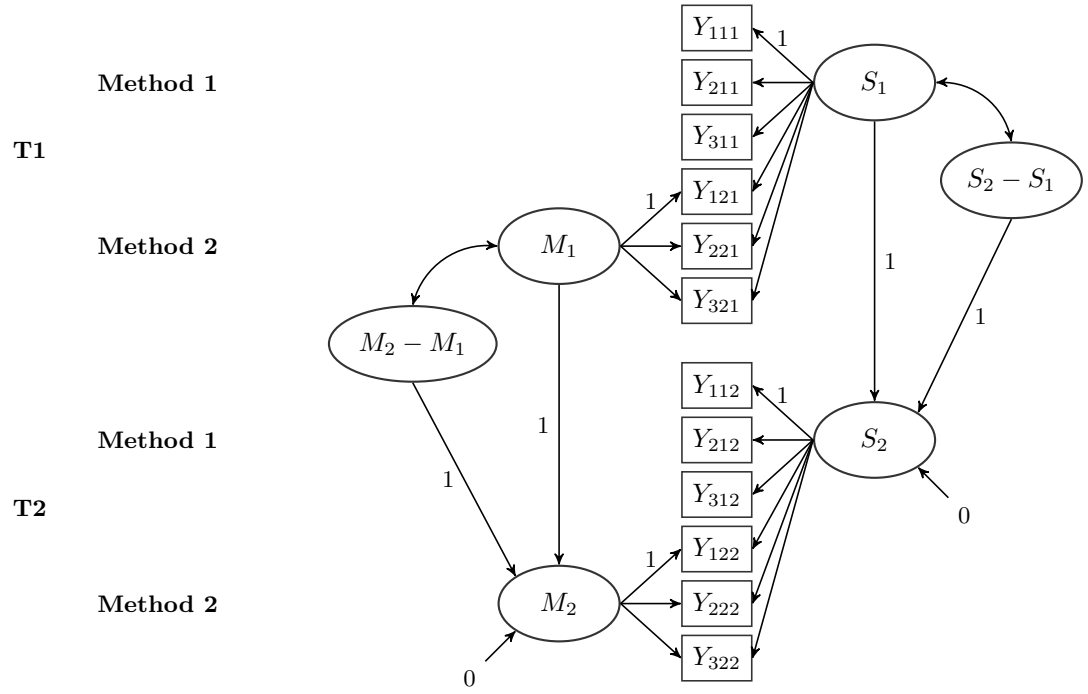


Figure 2.1: CSC($M - 1$) change model for one trait without indicator-specific factors (Geiser, 2009; Geiser et al., 2010b). Method 1 is selected as the reference method. Method 2 is contrasted against the reference method. The method factors capture the method-specific deviation of the nonreference method from the reference method. Indicator Y_{ikl} represents the ith item of method k at time point l. For the sake of clarity, error variables E_{ikl} are omitted and loading parameters are only shown for the first indicator. T1 = time point 1; T2 = time point 2.

by method k on occasion l. Because we are only looking at a single trait for now, subscript j is omitted. In this example, the first method ($k = 1$) serves as the reference method. All indicators Y_{i1l} pertaining to this method load only on the corresponding occasion-specific reference state factors S_l. Therefore, the reference state factors S_l represent the common occasion-specific factors of the indicators belonging to the reference method. The indicators of the second method also load on this reference state factor, as well as on an occasion-specific method factor M_l. The method factors account for systematic residual variance in the nonreference indicators that is not shared with the indicators pertaining to the reference method. In other words, the method factors contain the effect of a particular method in measuring a construct, compared to the reference method. As a residual factor, the method factor has a mean of zero and is uncorrelated with all state factors belonging to the same construct.

For the purpose of directly investigating interindividual differences in true intraindividual change, latent difference variables are included in the model to measure change directly (McArdle & Hamagami, 2001; Steyer, Eid, & Schwenkmezger, 1997; Steyer, Partchev, & Shanahan, 2000). The rationale behind latent difference modeling is a simple decomposition of the latent state factor S_2 into the initial state factor S_1 and a latent difference factor $(S_2 - S_1)$. The factor $(S_2 - S_1)$ represents latent change from measurement occasion 1 to measurement occasion 2:

$$S_2 = S_1 + (S_2 - S_1). \tag{2.1}$$

A latent change factor for the method factors can be defined in the same manner:

$$M_2 = M_1 + (M_2 - M_1). \tag{2.2}$$

The CSC($M-1$) latent change model allows us to look at the way change is measured by different methods. The latent difference factor $(S_2 - S_1)$ represents individual differences in change as measured by the reference method. The latent difference factor $(M_2 - M_1)$ represents individual differences in method change. That is, this factor captures the *deviation* of the latent change scores of the nonreference method from change predicted by the reference method. Method difference factors can therefore be used to study the question of whether (and why) different methods diverge in the assessment of change. A prerequisite for this interpretation of change factors is strong measurement invariance over time (see later).

2.3.1 Indicator-Specific Factors

As mentioned earlier, when using multiple indicators in longitudinal studies, the same indicators often share specific variance with themselves over time (i.e., indicators are often more highly correlated with themselves across time than with the other indicators of the same construct; Jöreskog, 1979; Raffalovich & Bohrnstedt, 1987; Sörbom, 1975). To account for these indicator-specific effects over time, the CSC($M-1$) model can be extended to include indicator-specific factors (Geiser, 2009; Geiser et al., 2010b). A CSC($M-1$) change model with indicator-specific factors is depicted in Figure 2.2. In contrast to Figure 2.1, there are four additional indicator-

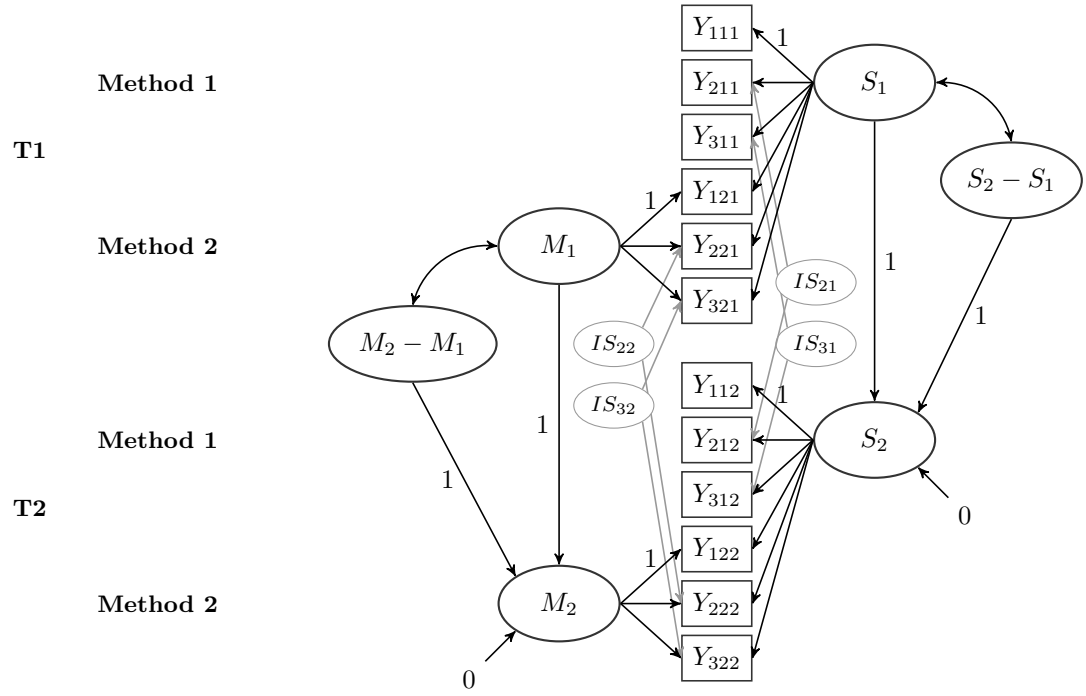


Figure 2.2: CSC($M-1$) change model for one trait with indicator-specific factors. The indicator-specific (but occasion-unspecific) factors IS_{ik} represent item-specific variance that the indicators Y_{ikl} and $Y_{ikl'}$ share over time.

specific factors IS_{ik}. Transferring Eid's (2000) approach of a reference method to the indicator level, indicator-specific factors are defined for all indicators except the first one ($i = 1$), which serve as the *reference indicators*. The indicator-specific factor represents that part of an indicator that is not shared with the reference indicator and is therefore unique to this specific indicator. By this definition, the reference state factor is the true-score variable of the reference indicator of the reference method (Geiser, 2009). This approach is similar to dummy-coding in regression analysis, in which one reference category has to be chosen and all other categories are compared to this reference category. In our model, we have to choose a reference category for the methods as well as for the indicators. Introducing as many indicator-specific factors as indicators considered and as many method factors as methods considered would not be reasonable and can cause problems of underidentification and improper parameter estimates (e.g., negative factor variances; Eid, 2000; Kenny & Kashy, 1992; Marsh & Bailey, 1991).

2.3.2 Ordinal Indicators

In a longitudinal MTMM data set, the indicators Y_{ijkl} (the ith measure of trait j measured by method k on occasion l) might not be measured on a metrical scale. This is often the case if item-level data are analyzed and the items are used as indicators of latent variables. In the case of item-level data, indicators are often measured on an ordinal rather than a metrical scale, usually with only a small number of categories. By definition, ordinal variables cannot be normally distributed, an assumption that is required for maximum likelihood (ML) estimation, the default estimator in CFA. If multivariate normality is violated, ML estimation might not be efficient and an inflated Type 1 error rate could lead to the rejection of too many proper models based on the χ^2 statistic (Curran, West, & Finch, 1996). Moreover, to test measurement invariance over time, it is necessary to consider appropriate measurement models that take the ordinal character of the measures into account (Lubke & Muthén, 2004). Ordinal variables require different methods of parameter estimation, such as the weighted least squares (WLS) estimator (Bollen, 1989; B. O. Muthén, du Toit, & Spisic, 1997; Satorra, 1989, 1992). When estimating large models, simulation studies suggest that the robust WLS mean and variance adjusted (WLSMV) estimator should be preferred (Beauducel & Herzberg, 2006; Flora & Curran, 2004; B. O. Muthén et al., 1997; Nussbeck, Eid, & Lischetzke, 2006).

For ordinal outcomes, WLS approaches are based on polychoric correlations. These are obtained by assuming that latent response variates (LRVs) Y^*_{ijkl} underlie the observed ordinal variables Y_{ijkl}. The LRVs are assumed to be unobserved metrical variables that are multivariate normally distributed (Jöreskog & Moustaki, 2001; Takane & De Leeuw, 1987) and the categorical values of Y_{ijkl} are thought to coarsely represent the underlying continuous distribution of Y^*_{ijkl}. The observed ordinal variables Y_{ijkl} and the LRVs Y^*_{ijkl} are linked to each other by so-called thresholds that mark the points on the latent continuum at which values start being attributed to the next highest category of the observed item. Formally, this relation is given by (Eid, 1996; B. O. Muthén, 1983; Nussbeck et al., 2006; Takane & De Leeuw, 1987):

$$Y_{ijkl} = s, \text{ if } \tau_{ijkls} < y^* \leq \tau_{ijkls+1} \tag{2.3}$$

with the threshold parameters τ_{ijkls} determining the categories $s \in \{0, ..., c\}$, where $\tau_{ijkl0} = -\infty$

and $\tau_{ijklc} = +\infty$. The CSC($M-1$) model with indicator-specific factors for the LRVs of the ordinal indicators is given by:

$$Y^*_{ijkl} = \begin{cases} \lambda^S_{1j1l} S_{jl} + E^*_{1j1l}, \text{ for } i,k = 1 \\ \lambda^S_{ij1l} S_{jl} + \lambda^{IS}_{ij1l} IS_{ij1} + E^*_{ij1l}, \text{ for } i \neq 1, k = 1 \\ \lambda^S_{1jkl} S_{jl} + \lambda^M_{1jkl} M_{jkl} + E^*_{1jkl}, \text{ for } i = 1, k \neq 1 \\ \lambda^S_{ijkl} S_{jl} + \lambda^M_{ijkl} M_{jkl} + \lambda^{IS}_{ijkl} IS_{ijk} + E^*_{ijkl}, \text{ for } i \neq 1, k \neq 1. \end{cases} \tag{2.4}$$

Here, λ^S_{ijkl} are the factor loadings on a reference state factor S_{jl}, λ^M_{ijkl} are the loadings on a residual method factor M_{jkl}, and λ^{IS}_{ijkl} are the loadings on an indicator-specific (but not occasion-specific) factor IS_{ijk}. Without loss of generality, the first method ($k = 1$) has been selected as the reference method. As residual factors, all method factors M_{jkl} have a mean of zero and are uncorrelated with the reference state factors pertaining to the same construct j. The indicator-specific factors are residual factors with a mean of zero and are uncorrelated with the reference state and method factors pertaining to the same construct and method (same indexes j and k; see Geiser, 2009). The error variables E^*_{ijkl} have a mean of zero and are uncorrelated with all latent factors and all error variables.

One way of standardizing the LRVs and identifying the model is to assume that (a) all latent variable means (factors and LRVs) equal zero, and that (b) the residual variances E^*_{ijkl} of the LRVs equal one (Eid & Hoffmann, 1998; B. O. Muthén & Asparouhov, 2002). Additionally, at least two thresholds per marker indicator have to be set equal across time. In the case of ordinal indicators, strong measurement invariance across time requires equal threshold parameters; that is, $\tau_{ijkls} = \tau_{ijkl's}$ with $l \neq l'$.

Note that there are four types of indicators that differ with regard to their variance decomposition: reference indicators of the reference method that only load on their respective reference state factors S_{jl}; the remaining indicators of the reference method that additionally load on an indicator-specific factor; reference indicators of nonreference methods that load on a reference state factor and a method factor but not on an indicator-specific factor; and nonreference indicators of nonreference methods, that load on a reference state, a method, and an indicator-specific

factor.

The CSC($M-1$) change model for ordinal indicators can be used to determine the reliability, convergent validity, and method specificity of each indicator's LRV. To be able to express these indexes in terms of variance components, strong measurement invariance (equal loadings and thresholds) has to hold across time. For this purpose, we make the additional assumption that all state factor loadings, method factor loadings, and indicator-specific factor loadings are time-invariant (i.e., $\lambda^{S}_{ijkl} = \lambda^{S}_{ijkl'} = \lambda^{S}_{ijk}$, $\lambda^{M}_{ijkl} = \lambda^{M}_{ijkl'} = \lambda^{M}_{ijk}$, $\lambda^{IS}_{ijkl} = \lambda^{IS}_{ijkl'} = \lambda^{IS}_{ijk}$, for all $l, l' = 1, ...L$). Note that this assumption can be empirically tested by comparing the fit of a model with constrained loadings against a model with unconstrained loadings. Given Equation 2.4 and the preceding equality constraints, the variance of an LRV at a given time point can be decomposed into:

$$Var(Y^{*}_{ijkl}) = \begin{cases} (\lambda^{S}_{1j1})^2 Var(S_{jl}) + Var(E^{*}_{1j1l}), \text{ for } i,k = 1 \\ (\lambda^{S}_{ij1})^2 Var(S_{jl}) + (\lambda^{IS}_{ij1})^2 Var(IS_{ij1}) + Var(E^{*}_{ij1l}), \text{ for } i \neq 1, k = 1 \\ (\lambda^{S}_{1jk})^2 Var(S_{jl}) + (\lambda^{M}_{1jk})^2 Var(M_{jkl}) + Var(E^{*}_{1jkl}), \text{ for } i = 1, k \neq 1 \\ (\lambda^{S}_{ijk})^2 Var(S_{jl}) + (\lambda^{M}_{ijk})^2 Var(M_{jkl}) + (\lambda^{IS}_{ijk})^2 Var(IS_{ijk}) + Var(E^{*}_{ijkl}), \\ \text{for } i \neq 1, k \neq 1. \end{cases} \tag{2.5}$$

Likewise, using the change version of the CSC($M-1$) model, the variance of an observed *difference score* can be decomposed into:

$$Var(Y^{*}_{ijkl'} - Y^{*}_{ijkl}) = \begin{cases} (\lambda^{S}_{1j1})^2 Var(S_{jl'} - S_{jl}) & + Var(E^{*}_{1j1l}), \text{ for } i,k = 1 \\ (\lambda^{S}_{ij1})^2 Var(S_{jl'} - S_{jl}) & + Var(E^{*}_{ij1l}), \text{ for } i \neq 1, k = 1 \\ (\lambda^{S}_{1jk})^2 Var(S_{jl'} - S_{jl}) & + (\lambda^{M}_{1jk})^2 Var(M_{jkl'} - M_{jkl}) \\ & + Var(E^{*}_{1jkl}), \text{ for } i = 1, k \neq 1 \\ (\lambda^{S}_{ijk})^2 Var(S_{jl'} - S_{jl}) & + (\lambda^{M}_{ijk})^2 Var(M_{jkl'} - M_{jkl}) \\ & + Var(E^{*}_{ijkl}), \text{ for } i \neq 1, k \neq 1. \end{cases} \tag{2.6}$$

Note that the indicator-specific factors do not contribute to the variance of difference scores, because assuming the indicator-specific factor loadings to be time-invariant, they cancel out (see Geiser, 2009). Given Equation 2.6, it is possible to define coefficients of *convergent validity* (*consistency*), *method specificity*, and *reliability* for each LRV difference score. The consistency coefficient $CO(Y^*_{ijkl'} - Y^*_{ijkl})$ captures that part of the variance of a difference score that is determined by change in the reference factor. Thus the consistency coefficient is a measure of convergent validity of change with respect to the reference method:

$$CO(Y^*_{ijkl'} - Y^*_{ijkl}) = \frac{(\lambda^S_{ijk})^2 Var(S_{jl'} - S_{jl})}{Var(Y^*_{ijkl'} - Y^*_{ijkl})}. \tag{2.7}$$

The proportion of variance of a difference score that is specific to a particular nonreference method is given by the method specificity coefficient $MS(Y^*_{ijkl'} - Y^*_{ijkl})$:

$$MS(Y^*_{ijkl'} - Y^*_{ijkl}) = \frac{(\lambda^M_{ijk})^2 Var(M_{jkl'} - M_{jkl})}{Var(Y^*_{ijkl'} - Y^*_{ijkl})}. \tag{2.8}$$

Consistency and method specificity add up to reliability:

$$\begin{aligned} Rel(Y^*_{ijkl'} - Y^*_{ijkl}) &= 1 - \frac{Var(E^*_{ijkl'}) + Var(E^*_{ijkl})}{Var(Y^*_{ijkl'} - Y^*_{ijkl})} \\ &= CO(Y^*_{ijkl'} - Y^*_{ijkl}) + MS(Y^*_{ijkl'} - Y^*_{ijkl}). \end{aligned} \tag{2.9}$$

The reliability coefficient indicates that part of the variance of a change score that is not due to measurement error.

2.3.3 Multiple Groups

Multimethod evaluation studies often feature multiple groups (e.g., intervention vs. control group). The CSC($M-1$) change model for ordinal indicators is extended to a multiple group model by adding the additional subscript g to allow for group-specific parameters (i.e., differences

between groups):

$$Y^*_{gijkl} = \begin{cases} \lambda^S_{g1j1l} S_{gjl} + E^*_{g1j1l}, \text{ for } i,k = 1 \\ \lambda^S_{gij1l} S_{gjl} + \lambda^{IS}_{gij1l} IS_{gij1} + E^*_{gij1l}, \text{ for } i \neq 1, k = 1 \\ \lambda^S_{g1jkl} S_{gjl} + \lambda^M_{g1jkl} M_{gjkl} + E^*_{g1jkl}, \text{ for } i = 1, k \neq 1 \\ \lambda^S_{gijkl} S_{gjl} + \lambda^M_{gijkl} M_{gjkl} + \lambda^{IS}_{gijkl} IS_{gijk} + E^*_{gijkl}, \text{ for } i \neq 1, k \neq 1, \end{cases} \tag{2.10}$$

where Y^*_{gijkl} is the LRV of item i in group g measuring trait j with method k at Time l . As Millsap and Tein (2004) showed for the general cross-sectional multiple group case, one possibility to identify this model is to set the following constraints in one of the groups (e.g., $g = 1$): All latent variable means are set to 0, and the residual variances of the LRVs are set to unity. In addition, two thresholds of the first indicator for each latent variable are constrained to be equal across groups. As usual, one loading per factor is fixed to 1. Compared to this least restrictive case, strict measurement invariance across time and groups now implies (a) equal thresholds across time and groups: $\tau_{gijkls} = \tau_{g'ijkl's}$ with $g \neq g'$ and $l \neq l'$, (b) equal loadings across time and groups, and (c) equal residual variances of the LRVs, $Var(E^*_{gijkl}) = Var(E^*_{g'ijkl'}) = 1$ with $g \neq g'$ and $l \neq l'$. The means of the latent variables can then be estimated in each group separately. Moreover, hypotheses concerning differences in latent mean change between groups can be tested.

Presented next is an application of the CSC($M-1$) change model for ordinal indicators and multiple groups using data from an evaluation study of an early childhood prevention program.

2.4 Empirical application

Data from the Augsburg Longitudinal Study for the Evaluation of the Prevention Program Papilio (ALEPP) as analyzed. Detailed information about the study can be found, for example, in Mayer, Heim, and Scheithauer (2007). In short, this study examined the effectiveness of an early childhood prevention program implemented in nurseries. The Papilio program aimed at increasing emotional competence in young children and reducing the short- and long-term risk of behavior problems. The project was developed and evaluated by the beta Institute

in Augsburg, Germany, a nonprofit institution for applied health management and sociomedical research in collaboration with the Universities of Bremen, Augsburg, and Freie Universität Berlin (all in Germany). The evaluation study started in 2003 with a representative randomized sample of nurseries from Augsburg and Augsburg County, a middle-size town located in the south of Germany. In addition to the training of nursery teachers (e.g., providing information concerning child development and on how to implement the interventions), the modularized program Papilio involves three child-oriented interventions.

1. On one fixed day per week, the toys are "on holiday" and put away. The children are encouraged to think of ways to play interactive games instead.

2. The "Box Puppet Story" is an interactive tale dealing with four elf-like characters representing the basic emotions of sadness, fear, anger, and joy. The story is supported by pictures, audio material, and the puppets. In this way, children learn about their own and others' perception of emotions, as well as ways to regulate their emotions.

3. In a modified version of the good behavior game, positive behavior in line with agreed rules is rewarded.

2.4.1 Sample and Measures

A representative sample of 50 groups in 25 nurseries from Augsburg (town) and Augsburg County was assigned randomly to a waiting control group (CG) or an intervention group (IG). In total, the sample used in this analysis consisted of $N = 659$ children aged 3 to 6, $n = 342$ of which were in the CG. The nested structure of the data (children in groups) was controlled for in the analysis by correcting the standard errors and fit statistics (Asparouhov & Muthén, 2005). Data collection took place on a baseline occasion prior to program implementation (T1) and at follow-up 12 months after the beginning of the intervention (T2). On both occasions, nursery teachers and mothers rated the children on a variety of scales, among them the German version of the Strengths and Difficulties Questionnaire (SDQ-Deu; Klasen, Woerner, Rothenberger, & Goodman, 2003). The SDQ is a commonly used instrument in clinical and developmental research and available for multiple raters and age groups. It consists of five subscales with five items each: (a) emotional symptoms, (b) conduct problems, (c) hyperactivity and inattention, (d) peer

relationship problems, and (e) prosocial behavior. The items are ordered categorical, that is, answered using a 3-point scale ranging from 0 (*not true*) to 2 (*certainly true*). In addition, a measure of relational aggression was included, a subscale of the Preschool Social Behaviour Scale (PSBS-T; Crick, Casas, & Mosher, 1997). Relational aggression is a subtle form of aggressive behavior that is revealed in such actions as harming peers by excluding them from the peer group (Crick & Grotpeter, 1995). The relational aggression scale consists of six items that are answered using a 5-point scale ranging from 1 (*never or almost never true of this child*) to 5 (*always or almost always true of this child*). In our application, we focus on two constructs: prosocial behavior (PB) as measured by the PB subscale of the SDQ and relational aggression (RA) as measured by the RA scale of the PSBS. At the level of our multitrait-multimethod-multioccasion measurement model for PB and RA, we are interested in quantifying convergent validity, method specificity, and reliability of parent and teacher ratings for assessing change over time. Furthermore, we want to investigate to what extent measurement invariance holds across groups and time. Of key interest is the latent mean structure: The children in the IG are expected to show a more pronounced increase in PB on average than children in the CG. To rule out the possibility that an increase in PB is associated with a shift toward more subtle ways of expressing aggression, RA is controlled for and not expected to increase in either group. This simultaneous examination of two outcomes and their change in relation to each other is a particular strength of the proposed approach.

2.4.2 Statistical Analysis

For illustrative purposes and to simplify the model, we picked three out of the five items (see Table 2.1 for wording) of the PB scale. The items were answered by both teachers and parents on T1 and T2, adding up to 12 observed ordered categorical variables. For RA, three of the six items (Table 2.1) were selected as indicators. Of the original five categories, the upper three (*often* to *always*) were pooled together to avoid empty categories in any of the groups. This resulted in three items with three categories, answered on the two occasions by both raters. For each trait-method-occasion unit (TMOU; e.g., teacher rating of PB at T1) three items served as indicators. The use of multiple indicators per trait-method unit in SEM of MTMM data is

Table 2.1: Item Wording and Indicator Labels in the Empirical Application

		Teacher Rating		*Parent Rating*	
Scale	*Wording*	*T1*	*T2*	*T1*	*T2*
PB	Often offers to help others (parents, teachers, other children)	tPB11	tPB21	pPB11	pPB21
	Helpful if someone is hurt, upset, or feeling ill	tPB12	tPB22	pPB12	pPB22
	Considerate of other people's feelings	tPB13	tPB23	pPB13	pPB23
RA	Tells a friend that he or she won't play with that peer or be that peer's friend unless he or she does what the child asks	tRA11	tRA21	pRA11	pRA21
	Tells a peer he or she won't be invited to their birthday party unless he or she does what the child wants	tRA12	tRA22	pRA12	pRA22
	Verbally threatens to keep a peer out of the play group if the peer doesn't do what the child asks	tRA13	tRA23	pRA13	pRA23

Note. PB = Prosocial Behavior subscale of the Strengths and Difficulties Questionnaire; RA = Relational Aggression subscale of the Preschool Social Behavior Scale; teacher and parent rating = naming of the indicators in the model; the first letter refers to the rater (t = teacher, p = parent), the two capital letters refer to the scale; and the first number refers to the occasion of measurement (1 or 2), while the second number indicates the position of the indicator (1st to 3rd).

generally recommended (Marsh, 1993; Marsh & Hocevar, 1988). With multiple indicator models, possible flaws in the psychometric properties of a scale can be detected on the item level and occasion-specific as well as construct-specific method effects can be estimated (Eid et al., 2003; Geiser 2009). In this illustrative application, teachers were chosen as the reference method.

All models were analyzed using the software M*plus*, Version 5.2 (L. K. Muthén & Muthén, 1998-2007). Observed variables were specified as ordered categorical. The WLS means and variance adjusted estimator (WLSMV; B. O. Muthén et al., 1997) was employed. We chose the theta parameterization that is available in M*plus* for multiple group analysis with ordinal indicators, because it allows distinguishing between loadings, residual variances, and factor variances as sources of (non)invariance in multiple group analysis (B. O. Muthén & Asparouhov, 2002).

Because the χ^2 value obtained from the WLSMV estimator cannot be used for conventional χ^2 difference testing, we used the adjusted procedure described in Asparouhov and Muthén (2006) to statistically compare nested models. The M*plus* input specifications for the final model described here can be found in the Appendix.

2.4.3 Model Specification

The final model is shown as a path diagram in Figure 2.3. With regard to the specification of indicator-specific factors, this model differs slightly from the model in Figure 2.2. As discussed in detail later, the reason is that indicator-specific effects were present only for some items, so that some of the indicator-specific factors could be dropped. Beginning with PB at T1, a reference ("teacher" factor was defined, measured by a total of six items (three teacher items plus three parent items). For the three parent items, an additional residual method ("parent") factor was specified. As a residual factor, this parent factor was not allowed to covary with the teacher factor. The parent factor thus represents the specific variance that the parent items share beyond what they all have in common with the teacher items. In the same way, factors were specified for PB on T2 and for RA on T1 and T2, yielding four teacher reference state factors and four parent residual factors. Covariances between reference state factors and residual factors belonging to the same construct were fixed at 0.

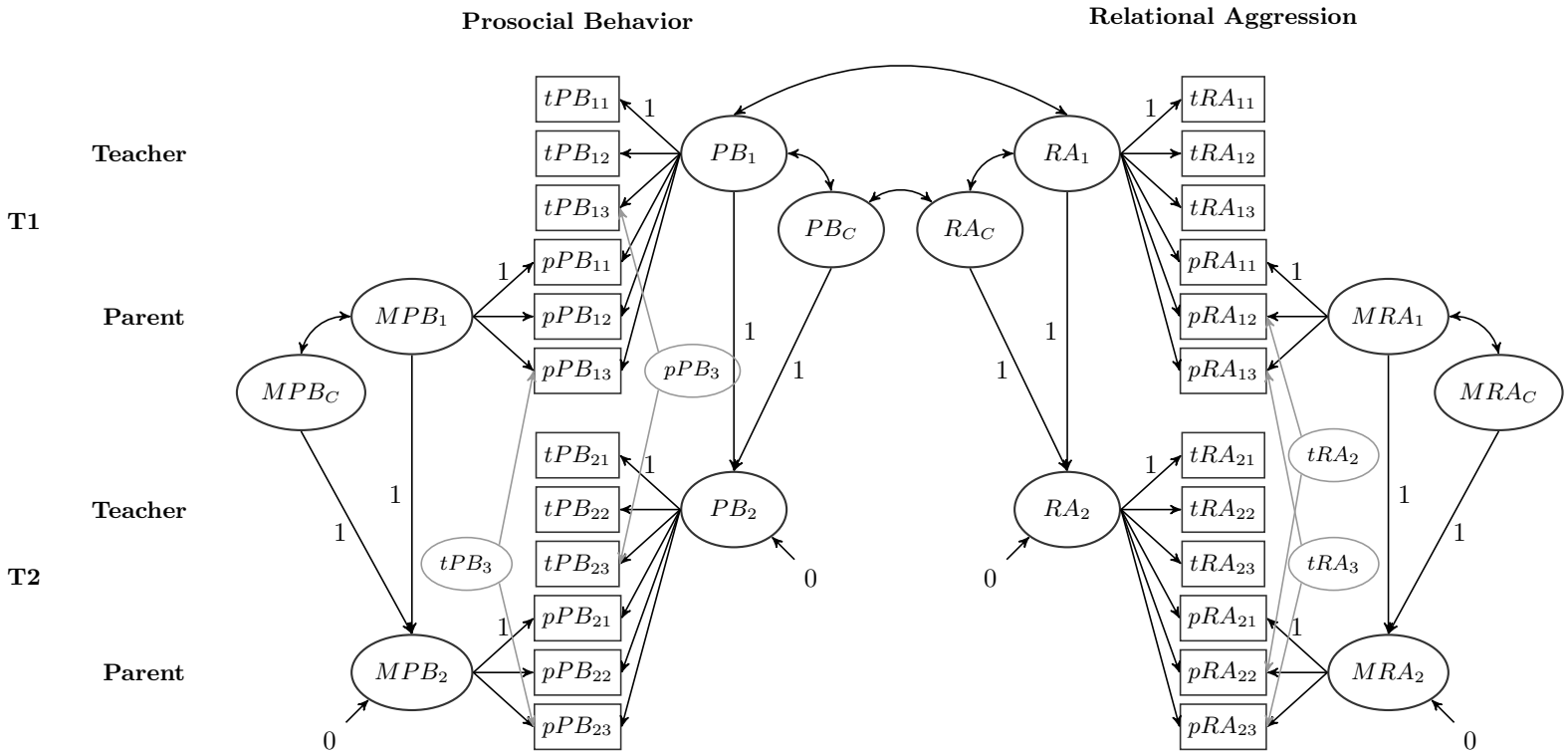


Figure 2.3: The CSC($M-1$) change model as used in the presented application. Teachers are selected as reference method. Parents are contrasted against teachers. PB_1 = Prosocial Behavior reference state factor at time point 1 (T1); RA_1 = Relational Aggression reference state factor at T1; PB_C = Prosocial Behavior change factor ($PB_2 - PB_1$), and so on. Four indicator-specific factors are included in the model to account for shared item variance over time.

2.4.4 Change Factors

To analyze latent change, we included latent change factors in the model. Starting again with the teacher reference state factor for PB, a new factor was introduced that represents latent change in PB from T1 to T2 as measured by teacher ratings. Change factors were included in the same way for RA and the parent method factors.

Indicator-specific factors

Preliminary analyses revealed that in this particular application, only four indicator-specific factors were needed. For PB, an indicator-specific factor had to be included for the teacher and the parent rating of Item 3, respectively. This was necessary to account for the variance that the third indicator shared with itself over time. By looking at the item wording in Table 2.1, we can clearly see that this particular item represents a different facet of PB compared to the two other items pertaining to this scale: The first two items measure helpfulness, whereas the third item aims at consideration. Hence, the first two PA items proved to be homogeneous, but differed significantly from the third item. Therefore, we included just one indicator-specific factor per method for PB (see Figure 2.3). This factor mirrored the deviation of the consideration item from the two helpfulness items. For RA, indicator-specific factors were needed for both nonmarker indicators pertaining to the parent ratings. Indicator-specific effects were not significant for the teacher indicators of RA, so these factors were dropped and do not appear in Figure 2.3.

Measurement invariance

To systematically test for measurement invariance across time and groups, we compared four versions of the CSC($M-1$) model. First, we specified an unrestrictive model as a baseline model that assumed only configural invariance (Model A). To identify Model A, the following specification was chosen: The first group (CG) served as reference group. In the reference group, (a) the means of all latent variables were fixed to 0 and (b) the residual variances of the LRVs underlying the observed categorical variables were fixed to 1. In addition, the thresholds were held equal across groups. Loadings of the marker indicators were fixed to 1 to identify the metric of the latent factors in both groups. All other loadings, residual variances, and latent means

were free in the nonreference group (IG) and allowed to differ across time and groups.

In the next step, we specified a second Model B with strict measurement invariance. Here, all loadings, residual variances, and thresholds were set equal across time and groups. Moreover, the teacher factor means on T1 were also set equal across groups to reflect the same initial level in PB and RA in both groups. Changes in PB were freely estimated and allowed to differ across groups. This setting is reasonable from a theoretical point of view, because some natural change in children's PB can occur even without an intervention. In line with theoretical expectations, mean change in RA was fixed to 0 in both groups.

Mean change

To examine the potential differences between groups in mean PB change more carefully, we formulated two models with additional restrictions that tested specific hypotheses about group differences in mean PB change. The first hypothesis was that the same amount of change in PB occurred in both groups. That is, there might have been some change in teacher's ratings of children's PB on T2 compared to T1. However, this change might not be due to the intervention program but in fact to maturation processes that occurred in both groups. Therefore, we imposed the additional constraint of equal mean change across groups in Model C. To test the even stronger hypothesis of no mean change in either group, the PB change factor means were fixed to 0 in an even more restrictive Model D. To determine which model should be retained, appropriate χ^2 difference tests were performed comparing each model to the less restrictive model, respectively (B to A, C to B, and D to C). Results of the difference tests and global fit statistics for all four models can be found in Table 2.2.

2.5 Results

Compared to the baseline Model A, the strict measurement invariance Model B did not fit the data worse. In addition, Model B showed a good absolute fit according to χ^2 test, the comparative fit index, and the root mean square error of approximation. Hence, the assumption of strict measurement invariance across time and groups was not rejected. Model C with the additional assumption of equal mean change across groups did not fit worse than Model B and still

Table 2.2: Goodness-of-Fit Measures

Model	χ^2	*df*	*p*	$\Delta\chi^2$	Δdf	*p*	*CFI*	*RMSEA*
A. Baseline, minimal restrictions	48.37	35	.07	—	—	—	.991	.034
B. Strict invariance, free change	44.53	37	.18	20.8	24	.65	.995	.025
C. Strict invariance, equal change	43.59	36	.18	2.03	1	.15	.995	.025
D. Strict invariance, zero change	48.76	37	.09	45.1	1	.00	.992	.031

Note. $N = 659$. Strict invariance = invariant loadings, intercepts, and error variances for all indicators across time and groups. The χ^2 difference values and *df* for the weighted least squares mean and variance adjusted testing procedure do not necessarily equal the differences in parameters (B. O. Muthén et al., 1997). CFI = comparative fit index; RMSEA = root mean square error of approximation.

fit the data well globally. In spite of an acceptable overall fit, Model D fit the data significantly worse than Model C, $\Delta\chi^2(1) = 45.1, p < .01$, indicating that the assumption of no mean change lead to a significant decrease in model fit. In sum, the results indicate that significant mean change over time occurred in PB. However, contrary to our hypothesis, the statistical model comparisons indicated that mean change in PB was the same in the intervention and control groups. We therefore decided to retain Model C. In this model, strict measurement invariance (equal thresholds, loadings, and residual variances) holds across time and groups, and all latent means are assumed to be equal across groups. Detailed outcomes for Model C are discussed in the following.

2.5.1 Measurement Model

The parameter estimates for the measurement part of our final model (Model C) are provided in Tables 2.3 (PB) and 2.4 (RA). The low loadings of the parent items on the teacher reference state factor (as well as the high loadings of the same indicators on the parent method factors) indicate low convergent validity (and high method specificity) of teacher and parent ratings. The variance components for the observed difference scores are in concert with this finding (see Table 2.5). Reliabilities are higher for RA than for PB, and higher for teacher ratings compared to

Table 2.3: Estimated Factor Loadings for Prosocial Behavior

		Reference State Factor (PB) Loading			*Method Factor (MPB) Loading*			*Indicator-Specific Factor (pPB3) Loading*			*Indicator-Specific Factor (tPB3) Loading*		
Group	*Indicator*	*Estimate*	*SE*	*Stand. Est.*	*Estimate*	*SE*	*Stand. Est.*	*Estimate*	*SE*	*Stand. Est.*	*Estimate*	*SE*	*Stand. Est.*
CG	tPB11	1.00	—	.86	—	—	—	—	—	—	—	—	—
	tPB12	1.02	0.13	.86	—	—	—	—	—	—	—	—	—
	tPB13	0.92	0.11	.74	—	—	—	1.00	—	.46	—	—	—
	pPB11	0.18	0.03	.23	1.00	—	.56	—	—	—	—	—	—
	pPB12	0.18	0.03	.21	1.30	0.13	.66	—	—	—	—	—	—
	pPB13	0.17	0.04	.19	1.24	0.13	.57	—	—	—	1.00	—	.47
	tPB21	1.00	—	.88	—	—	—	—	—	—	—	—	—
	tPB22	1.02	0.13	.88	—	—	—	—	—	—	—	—	—
	tPB23	0.92	0.11	.78	—	—	—	1.00	—	.43	—	—	—
	pPB21	0.18	0.03	.24	1.00	—	.63	—	—	—	—	—	—
	pPB22	0.18	0.03	.22	1.30	0.13	.72	—	—	—	—	—	—
	pPB23	0.17	0.04	.19	1.24	0.13	.63	—	—	—	1.00	—	.44
IG	tPB11	1.00	—	.87	—	—	—	—	—	—	—	—	—
	tPB12	1.02	0.13	.88	—	—	—	—	—	—	—	—	—
	tPB13	0.92	0.11	.75	—	—	—	1.00	—	.48	—	—	—
	pPB11	0.18	0.03	.25	1.00	—	.57	—	—	—	—	—	—
	pPB12	0.18	0.03	.22	1.30	0.13	.67	—	—	—	—	—	—
	pPB13	0.17	0.04	.18	1.24	0.13	.52	—	—	—	1.00	—	.60
	tPB21	1.00	—	.88	—	—	—	—	—	—	—	—	—
	tPB22	1.02	0.13	.89	—	—	—	—	—	—	—	—	—
	tPB23	0.92	0.11	.77	—	—	—	1.00	—	.47	—	—	—
	pPB21	0.18	0.03	.25	1.00	—	.60	—	—	—	—	—	—
	pPB22	0.18	0.03	.23	1.30	0.13	.70	—	—	—	—	—	—
	pPB23	0.17	0.04	.18	1.24	0.13	.55	—	—	—	1.00	—	.59

Note. CG = control group; IG = intervention group; Stand. Est. = Standardized Estimate. Refer to Table 2.1 for the naming of indicators. Estimates without standard errors were fixed at 1. Measurement invariance holds across time and groups; that is, the unstandardized estimates and standard errors are the same for, for example, tPB12 and tPB22 in both groups.

Table 2.4: Estimated Factor Loadings for Relational Aggression

		Reference State Factor (RA) Loading			*Method Factor (MRA) Loading*			*Indicator-Specific Factor (pRA2) Loading*			*Indicator-Specific Factor (pRA3) Loading*		
Group	*Indicator*	*Estimate*	*SE*	*Stand. Est.*	*Estimate*	*SE*	*Stand. Est.*	*Estimate*	*SE*	*Stand. Est.*	*Estimate*	*SE*	*Stand. Est.*
CG	tRA11	1.00	—	.97	—	—	—	—	—	—	—	—	—
	tRA12	0.47	0.08	.87	—	—	—	—	—	—	—	—	—
	tRA13	0.77	0.17	.95	—	—	—	—	—	—	—	—	—
	pRA11	0.04	0.02	.07	1.00	—	.85	—	—	—	—	—	—
	pRA12	0.05	0.02	.13	0.50	0.13	.53	1.00	—	.53	—	—	—
	pRA13	0.06	0.02	.14	0.67	0.15	.66	—	—	—	1.00	—	.44
	tRA21	1.00	—	.96	—	—	—	—	—	—	—	—	—
	tRA22	0.47	0.08	.86	—	—	—	—	—	—	—	—	—
	tRA23	0.77	0.17	.94	—	—	—	—	—	—	—	—	—
	pRA21	0.04	0.02	.08	1.00	—	.82	—	—	—	—	—	—
	pRA22	0.05	0.02	.13	0.50	0.13	.48	1.00	—	.55	—	—	—
	pRA23	0.06	0.02	.14	0.67	0.15	.61	—	—	—	1.00	—	.46
IG	tRA11	1.00	—	.95	—	—	—	—	—	—	—	—	—
	tRA12	0.47	0.08	.82	—	—	—	—	—	—	—	—	—
	tRA13	0.77	0.17	.92	—	—	—	—	—	—	—	—	—
	pRA11	0.04	0.02	.06	1.00	—	.83	—	—	—	—	—	—
	pRA13	0.06	0.02	.12	0.67	0.15	.63	—	—	—	1.00	—	.45
	tRA21	1.00	—	.95	—	—	—	—	—	—	—	—	—
	tRA22	0.47	0.08	.82	—	—	—	—	—	—	—	—	—
	tRA23	0.77	0.17	.92	—	—	—	—	—	—	—	—	—
	pRA21	0.04	0.02	.06	—	—	.85	—	—	—	—	—	—
	pRA22	0.05	0.02	.11	0.50	0.13	.52	1.00	—	.54	—	—	—
	pRA23	0.06	0.02	.12	0.67	0.15	.66	—	—	—	1.00	—	.43

Note. CG = control group; IG = intervention group; Stand. Est. = Standardized Estimate. Refer to Table 2.1 for the naming of indicators. Estimates without standard errors were fixed at 1. Measurement invariance holds across time and groups; that is, the unstandardized estimates and standard errors are the same for, for example, tRA12 and tRA22 in both groups.

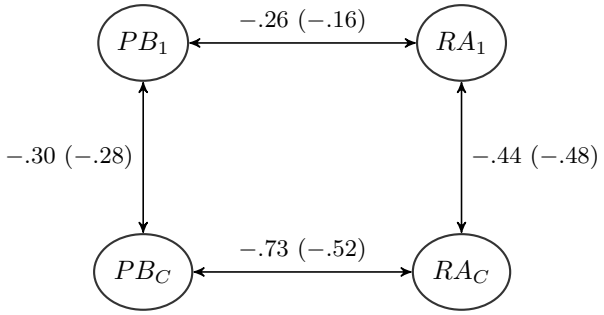


Figure 2.4: Correlations between reference state factors in the structural model. Estimates for the intervention group are followed by estimates for the control group in brackets. PB_1 = prosocial behavior at time point 1; PB_C = change in prosocial behavior from time point 1 to time point 2; RA_1 = relational aggression at time point 1; RA_C = Change in relational aggression from time point 1 to time point 2; All correlations are significant at the .05 level.

parent ratings. Particularly unreliable are the difference scores of parent ratings in assessing change in PB. Note, however, that these reliability estimates refer to difference scores of single items; it is expected that these show rather low reliability estimates.

2.5.2 Structural Model

In this paragraph, we focus on the key part of the structural model, the reference state factors at time point 1 and the reference change factors (see Figure 2.4). Both PB and RA are negatively related to their respective change factors. This indicates that the lower a child is rated at time point 1, the stronger is his or her increase over time. This finding is similar in both groups. Concerning discriminant validity, the constructs are weakly related at time point 1, indicating a high level of discriminant validity between PB and RA. The correlation is slightly higher in the IG ($r = -.26$) compared to the CG ($r = -.16$). The change factors are also negatively related, meaning that an increase in PB is associated with a decrease in RA. This is particularly true for the IG ($r = -.73$), which indicates a lack of discriminant validity in the assessment of change by the reference method (teachers).

2.5.3 Mean Change

The estimated mean change in PB was 0.65. Divided by the standard deviation of the change factor, this yields a medium-sized effect of .48 in the IG and .47 in the CG.

Table 2.5: Estimated Variance Components for Item Difference Scores

	Control Group			*Intervention Group*		
Difference Score	*CO*	*MS*	*Rel*	*CO*	*MS*	*Rel*
tPB21-tPB11	.49	—	.49	.48	—	.48
tPB22-tPB12	.50	—	.50	.49	—	.49
tPB23-tPB13	.45	—	.45	.44	—	.44
pPB21-pPB11	.03	.06	.08	.03	.07	.10
pPB22-pPB12	.03	.09	.12	.02	.12	.14
pPB23-pPB13	.03	.08	.11	.02	.11	.13
tRA21-tRA11	.84	—	.84	.79	—	.79
tRA22-tRA12	.54	—	.54	.46	—	.46
tRA23-tRA13	.76	—	.76	.69	—	.69
pRA21-pRA11	.00	.44	.44	.00	.51	.51
pRA22-pRA12	.01	.16	.17	.01	.20	.21
pRA23-pRA13	.01	.26	.27	.01	.32	.33

Note. CO = Consistency, MS = Method Specificity, Rel = Reliability. The variance components refer to the difference scores of the manifest (observed) variables. Rounding errors might prevent the consistency and method specificity to exactly add up to the reliability coefficient.

2.6 Discussion

In this article, we showed a comprehensive way to analyze multimethod evaluation data within the SEM framework. This approach has several key advantages compared to conventional statistical methods that are routinely used to analyze evaluation data (e.g., repeated measures ANOVA and MANOVA). The SEM approach allows for a much more flexible and comprehensive analysis of such data. ANOVA and MANOVA are special cases of SEM that make more restrictive assumptions that cannot be tested in these approaches. First of all, measurement error is not explicitly modeled in traditional approaches, and consequently, estimates of score reliabilities are not available and results could be biased. Second, although multiple outcome variables can be considered simultaneously in MANOVA, only a single indicator per construct (or an aggregate or sum score) is used. Therefore, indicator-specific effects cannot be tested and items that actually measure different facets of a construct might be aggregated and erroneously treated as unidimensional. In the SEM approach presented here, data can be analyzed at the item level and item homogeneity can be tested by comparing the fit of models with and without indicator-specific factors.

Furthermore, relationships between different constructs (i.e., discriminant validity) on the latent level cannot be directly assessed by MANOVA. In SEM, the mean and covariance structure of multiple latent constructs can be considered simultaneously, and correlations among constructs can be assessed. Furthermore, individual differences in change over time and convergent as well as discriminant validity of change scores can be assessed. In sum, SEM offers a much more fine-grained analysis of dimensionality issues and relationships between different constructs in a single model.

Another important assumption that is implicitly made in the (M)ANOVA framework (but not explicitly tested) is the assumption of measurement invariance across groups and time. That is, it is implicitly assumed that scores in different groups and at different time points are measurement-equivalent and therefore comparable. This assumption is often violated in practice, and there is no way to formally test this assumption in (M)ANOVA. Consequently, when applying (M)ANOVA in practice, it remains unclear whether mean comparisons are actually tenable. In contrast, SEM allows for a detailed assessment of measurement invariance by testing parameters of the

measurement model for invariance using, for example, chi-square difference tests. In the case reported here, we found that strict measurement invariance held both across groups and time so that we could safely investigate differences in latent variable means.

Another significant advantage of the SEM framework is that it easily allows one to use dichotomous or ordinal items as indicators for the constructs of interest. This is of great relevance, particularly for multimethod evaluation studies, as these studies are very costly in the first place. Therefore, metrical indicators formed by summing up multiple items of a scale might not be available for the simple reason that only a few items can be administered per construct, method, and time point. In this case, researchers must conduct their analyses on the item-level data that do not satisfy the condition of a metrical scale. However, (M)ANOVA assumes that the outcome variables are measured at least on an interval scale, and might lead to incorrect results if ordinal outcomes are treated as if they were continuous. In SEM, this problem is easily (and correctly) addressed by specifying a model for ordinal variables that assumes an underlying continuous response variable as discussed in this article.

2.7 Acknowledgements

We gratefully acknowledge Eva Fondel's and Martin Schultze's assistance in data preparation. The Papilio project was supported by the Bavarian Ministry for Environmental Issues, Public Health, and Consumer Protection, betapharm Pharmaceuticals, the BMW Group, and the Puppet Theatre "Augsburger Puppenkiste". We gratefully acknowledge the project staff, colleagues, and collaborators at the beta Institute and cooperating institutions for their hard work and valuable assistance during different phases of the evaluation study. We are especially grateful to the teachers, children, and parents at the participating preschools for their cooperation, without which this study would not have been possible.

2.8 References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101(2)*, 213-232.

Asparouhov, T., & Muthén, B.O. (2005). *Multivariate statistical modeling with survey data.* Retrieved from `http://www.fcsm.gov`

Asparouhov, T., & Muthén, B. O. (2006). *Robust chi square difference testing with mean and variance adjusted test statistics.* Retrieved from `http://www.statmodel.com`

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*, 78-102.

Barquero, B., Mayer, H., Heim, P., Scheithauer, H., Meir-Brenner, S., Koglin, U., Petermann, F., & Erhardt, H. (2007). Papilio: Ein Programm zur Primärprävention von Verhaltensproblemen, zur Förderung sozial-emotionaler Kompetenzen im Kindergarten und zur langfristigen Prävention von Sucht und Gewalt. [Papilio: A program for primary prevention of behavior disorders, for the promotion of emotional-social competence in kindergarten and for long-term prevention of addiction and violence.] In B. Röhrle (ed.), *Prävention und Gesundheitsförderung Bd. III Kinder und Jugendliche* (pp. 297-418). Tübingen, Germany: dgtv.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186-203.

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality, 72*, 845-876.

Bollen, K. A. (1989). *Structural equations with latent variables.* NY: Wiley.

Burns, G. L., & Haynes, S. N. (2006). Clinical psychology: Construct validation with multiple sources of information and multiple settings. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 401-418). Washington, DC: American Psychological Association.

Burns, G. L., Walsh, J. A., & Gomez, R. (2003). Convergent and discriminant validity of trait and source effects in ADHD-Inattention and Hyperactivity/Impulsivity measures across a 3-month interval. *Journal of Abnormal Child Psychology, 15*, 529-541.

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling, 9*, 233-255.

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality, 34*, 357-379.

Crick, N. R., & Grotpeter, J. K. (1995). Relational Aggression, Gender, and Social-Psychological Adjustment. *Child Development, 66*, 710-722.

Crick, N. R., Casas, J. F., & Mosher, M. (1997). Relational and overt aggression in preschool. *Developmental Psychology, 33(4)*, 579-588.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*, 327-346.

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research – Online, 1*, 65-85.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241-261.

Eid, M., & Diener, E. (2006). *Handbook of multimethod measurement in psychology.* Washington, DC: American Psychological Association.

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., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283-299). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple indicator CT-C($M-1$) model. *Psychological Methods, 8*, 38-60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., Lischetzke, T. (2008): Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*, 230-253.

Eid, M., Schneider, C., & Schwenkmezger, P. (1999). Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *European Journal of Personality, 13*, 283-306.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Geiser, C. (2009). *Multitrait-multimethod-multioccasion modeling.* Munich, Germany: AVM.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C($M-1$) model: A comment on Maydeu-Olivares & Coffman (2006). *Psychological Methods, 13*, 49-57.

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D., & Cole, D. (2010a). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multimethod change model to depression and anxiety in children. *Developmental Psychology, 46*, 29-45.

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D., & Cole, D. (2010b). Multitrait-multimethod change modeling. *Advances in Statistical Analysis.*

Grimm, K. J., Pianta, R. C., & Konold, T. (2009). Longitudinal multitrait-multimethod models for developmental research. *Multivariate Behavioral Research, 44*, 233-258.

Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E., & Lahey, B. B. (1999). Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? *Journal of the American Academy of Child & Adolescent Psychiatry, 38*, 1569-1579.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal-developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303-351). New York, NY: Academic Press.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*, 347-387.

Kenny, D. A. & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165-172.

Klasen, H., Woerner, W., Rothenberger, A. & Goodman, R. (2003). Die deutsche Fassung des Strengths and Difficulties Questionnaire (SDQ-Deu) - Übersicht und Bewertung erster Validierungs- und Normierungsbefunde. [The German version of the Strengths and Difficulties Questionnaire (SDQ-Deu): Overview and evaluation of initial findings on validity and standardization.] *Praxis der Kinderpsychologie und Kinderpsychiatrie, 52*, 491-502.

Kochanska, G., Barry, R. A., Jimenez, N. B., Hollatz, A. L., & Woodard, J. (2009). Guilt and effortful control: Two mechanisms that prevent disruptive developmental trajectories. *Journal of Personality and Social Psychology, 97*, 322-333.

Kwok, O., Haine, R. A., Sandler, I. N., Ayers, T. S., Wolchik, S. A., & Tein, J.-Y. (2005). Positive parenting as a mediator of the relations between parental psychological distress and mental health problems of parentally bereaved children. *Journal of Clinical Child and Adolescent Psychology, 34*, 260-271.

LaGrange, B., & Cole, D. A. (2008). An expansion of the trait-state-occasion model: Accounting for shared method variance. *Structural Equation Modeling, 15*, 241-271.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173.

Lubke, G., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*, 514-534.

Marsh, H. W. (1993). Multitrait-multimethod analyses: Inferring each trait/method combination with multiple indicators. *Applied Measurement in Education, 6*, 49-81.

Marsh, H. W. & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*, 47-70.

Marsh, H. W., & Grayson, D. A. (1994). Longitudinal confirmatory factor analysis: Common, time-specific, item-specific, and residual-error components of variance. *Structural Equation Modeling, 1*, 116-145.

Marsh, H. W., & Grayson, D. A. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 177-198). Thousands Oaks, CA: Sage.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*, 107-117.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In R. B. Cattell & J. Nesselroade (Eds.), *Handbook of multivariate experimental psychology* (pp. 561-614). New York: Plenum Press.

McDowell, D. J., & Parke, R. D. (2009, Jan.). Parental correlates of children's peer relations: An empirical test of a tripartite model. *Developmental Psychology, 45*, 224-235.

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods, 9*, 301-333.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Millsap, R. E., & Tein, J.-Y. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*, 479-515.

Millsap, R. E., & Meredith, W. (2007). Factorial Invariance: Historical Perspectives and New Problems. In: R. Cudeck & R. MacCallum (Eds.), *Factor Analysis at 100: Historical Developments and Future Directions* (pp. 130-152). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers.

Muthén, B. O. (1983). Latent variable structural equation moderling with categorical data. *Journal of Econometrics, 22*, 48-65.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple group and growth modeling in Mplus.* Retrieved from `http://www.statmodel.com`

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Retrieved from `http://www.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf`

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus User's Guide. Fifth Edition.* Los Angeles, CA: Muthén & Muthén.

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analyzing MTMM data with SEM for ordinal variables applying the WLSMV-estimator: What is the sample size needed for valid results? *British Journal of Mathematical and Statistical Psychology, 59*, 195-213.

Ostrov, J. M., & Crick, N. R. (2007). Forms and functions of aggression during early childhood: A short-term longitudinal study. *School Psychology Review, 36*, 22-43.

Raffalovich, L. E., & Bohrnstedt, G. W. (1987). Common, specific, and error variance components of factor models: Estimation with longitudinal data. *Sociological Methods & Research, 15*, 385-405.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrica, 54*, 131-151.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *American Sociological Association, 22*, 249-278.

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology, 28*, 138-151.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research – Online, 2*, 21-33.

Steyer, R., Partchev, I., & Shanahan, M. (2000). Modeling true intra-individual change in structural equation models: The case of poverty and children's psychosocial adjustment. In T. D. Little, K. U. Schnabel & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 109-126). Hillsdale, NJ: Erlbaum.

Stone, N. (2006). Evaluating interprofessional education: The tautological need for interdisciplinary approaches. *Journal of Interprofessional Care, 20*, 260-275.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

2.9 Appendix

Mplus input for estimating the CSC(M-1) change model with invariant thresholds, factor loadings, residual variances, and equal PB mean change across time and groups (Model C in the text).

```
TITLE: CSC(M-1) change model with multiple groups and ordinal indicators.
       2 constructs each measured by 6 items from 2 raters
       - PROSOCIAL BEHAVIOR (PB)
       - RELATIONAL AGGRESSION (RA)
       All lines beginning with an exclamation mark represent comments
!  Definition of the ASCII data file to be used
DATA: FILE = Papilio.dat;
VARIABLE:
!  Naming of the variables in the file
         NAMES = treat group
                 tPB11 tPB12 tPB13 pPB11 pPB12 pPB13
                 tPB21 tPB22 tPB23 pPB21 pPB22 pPB23
                 tRA11 tRA12 tRA13 pRA11 pRA12 pRA13
                 tRA21 tRA22 tRA23 pRA21 pRA22 pRA23;
!  Observed variables (items) to be used in the analysis
         USEVARIABLES = tPB11-tPB13 tPB21-tPB23 pPB11-pPB13 pPB21-pPB23
                        tRA11-tRA13 tRA21-tRA23 pRA11-pRA13 pRA21-pRA23;
!  Definition of observed variables as ordinal
         CATEGORICAL = tPB11-tPB13 tPB21-tPB23 pPB11-pPB13 pPB21-pPB23
                       tRA11-tRA13 tRA21-tRA23 pRA11-pRA13 pRA21-pRA23;
!  Missing value flag
         MISSING = all(-9);
!  Definition of the cluster variable
!  Children where nested within groups
         CLUSTER = group;
!  Definition of the grouping variable
         GROUPING = treat (0 = CG 1 = IG);
!  This type of analysis takes clustering of observations into account
ANALYSIS: TYPE = COMPLEX;
!  The WLSMV is the default estimator for this analysis
!  when at least one dependent categorical variable is involved
          ESTIMATOR = WLSMV;
!  This restores the default of Mplus 5.2
          SATTERTHWAITE=ON;
!  Change the default Delta parameterization to Theta
          PARAMETERIZATION = THETA;
!  Request difference test for WLSMV
          DIFFTEST B-free_change.dat;
!  Model specification
MODEL:
!  The first method k = 1 is selected as reference method
!  Definiton of Prosocial Behavior factor PB1 at time point
PB1 by tPB11
       tPB12 (1)
       tPB13 (2)
       pPB11 (3)
       pPB12 (4)
```

```
       pPB13 (5);
!  Definiton of a Prosocial Behavior factor PB2 at time point 2
!  Loadings are fixed to be the same as at time point 1
PB2 by tPB21
       tPB22 tPB23 (1-2)
       pPB21-pPB23 (3-5);
!  Definition of a Relational Aggression factor RA1 at time point 1
RA1 by tRA11
       tRA12 tRA13 (6-7)
       pRA11-pRA13 (8-10);
!  Definition of a Relational Aggression factor RA2 at time point 2
!  Loadings are fixed to be the same as at time point 1
RA2 by tRA21
       tRA22 tRA23 (6-7)
       pRA21-pRA23 (8-10);
!  Construct-specific method factor for the nonreference method
!  Prosocial Behavior Method factor MPB1 at time point 1
MPB1 by pPB11
       pPB12 pPB13 (11-12);
!  Prosocial Behavior Method factor MPB2 at time point 2
MPB2 by pPB21
       pPB22 pPB23 (11-12);
!  Relational Aggression Method factor MRA1 at time point 1
MRA1 by pRA11
       pRA12 pRA13 (13-14);
!  Relational Aggression Method factor MRA2 at time point 2
MRA2 by pRA21
       pRA22 pRA23 (13-14);
!  Item-specific factors
tPB3 by tPB13 tPB23@1;
pPB3 by pPB13 pPB23@1;
pRA2 by pRA12 pRA22@1;
pRA3 by pRA13 pRA23@1;
!  Definition of change factor PBC
PBC by;
PB2 on PB1@1 PBC@1;
PB2@0 [PB2@0];
!  Definition of change factor RAC
RAC by;
RA2 on RA1@1 RAC@1;
RA2@0 [RA2@0];
!  Definition of change factor MPBC
MPBC by;
MPB2 on MPB1@1 MPBC@1;
MPB2@0 [MPB2@0];
!  Definition of change factor MRAC
MRAC by;
MRA2 on MRA1@1 MRAC@1;
MRA2@0 [MRA2@0];
!  Restrictions in the structural model
PB1 with MPB1@0 MPB2@0 MPBC@0 RA2@0 MRA2@0 tPB3@0 pPB3@0;
PBC with MPB1@0 MPB2@0 MPBC@0 RA2@0 MRA2@0 tPB3@0 pPB3@0;
PB2 with MPB1@0 MPB2@0 MPBC@0 tPB3@0 pPB3@0 pRA2@0 pRA3@0
```

```
         MRA1@0 MRA2@0 MRAC@0 RA1@0 RA2@0 RAC@0;
RA1 with MRA1@0 MRA2@0 MRAC@0 MPB2@0 pRA2@0 pRA3@0;
RAC with MRA1@0 MRA2@0 MRAC@0 MPB2@0 pRA2@0 pRA3@0;
RA2 with MPB1@0 MPB2@0 MPBC@0 pRA2@0 pRA3@0 tPB3@0 pPB3@0
         MRA1@0 MRA2@0 MRAC@0;
MPB1 with MRA2@0 pPB3@0;
MPBC with MRA2@0 pPB3@0;
MPB2 with MRA1@0 MRA2@0 MRAC@0 pRA2@0 pRA3@0 tPB3@0 pPB3@0;
MRA1 with pRA2@0 pRA3@0;
MRAC with pRA2@0 pRA3@0;
MRA2 with pRA2@0 pRA3@0 tPB3@0 pPB3@0;
!Fix thresholds across time
!RA
[tRA11$1 tRA21$1] (15);
[tRA11$2 tRA21$2] (16);
[tRA12$1 tRA22$1] (17);
[tRA12$2 tRA22$2] (18);
[tRA13$1 tRA23$1] (19);
[tRA13$2 tRA23$2] (20);
[pRA11$1 pRA21$1] (21);
[pRA11$2 pRA21$2] (22);
[pRA12$1 pRA22$1] (23);
[pRA12$2 pRA22$2] (24);
[pRA13$1 pRA23$1] (25);
[pRA13$2 pRA23$2] (26);
!PB
[tPB11$1 tPB21$1] (27);
[tPB11$2 tPB21$2] (28);
[tPB12$1 tPB22$1] (29);
[tPB12$2 tPB22$2] (30);
[tPB13$1 tPB23$1] (31);
[tPB13$2 tPB23$2] (32);
[pPB11$1 pPB21$1] (33);
[pPB11$2 pPB21$2] (34);
[pPB12$1 pPB22$1] (35);
[pPB12$2 pPB22$2] (36);
[pPB13$1 pPB23$1] (37);
[pPB13$2 pPB23$2] (38);
!  Fix means of residual factors to 0
[MPB1@0 MPBC@0 tPB3@0 pPB3@0];
[MRA1@0 MRAC@0 pRA2@0 pRA3@0];
!  Fix latent factor means to 0 in both groups
[PB1@0 RA1@0 RAC@0];
!  Fix residual variances to 1
tPB11-pRA23@1;
!  Estimate means of Prosocial Behavior Change factor in both groups
MODEL CG:
[PBC] (99);
MODEL IG:
[PBC] (99);

!Request standardized solution
OUTPUT: STDYX;
```

```
!Save derivatives for further difference testing
SAVEDATA: DIFFTEST = C_equal_change.dat;
```

Chapter 3

Exploring Dynamics in Mood Regulation — Mixture Latent Markov Modeling of Ambulatory Assessment Data

This is a non-final version of an article published in final form in Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring Dynamics in Mood Regulation — Mixture Latent Markov Modeling of Ambulatory Assessment Data. *Psychosomatic Medicine, 74*, 366-376. doi: 10.1097/PSY.0b013e31825474cb

Abstract

Objective: To illustrate how fluctuation patterns in ambulatory assessment data with features such as few categorical items, measurement error, and heterogeneity in the change pattern can adequately be analyzed with mixture latent Markov models. The identification of fluctuation patterns can be of great value to psychosomatic research concerned with dysfunctional behavior or cognitions, such as addictive behavior or noncompliance. In our application, unobserved subgroups of individuals who differ with regard to their mood regulation processes, such as mood maintenance and mood repair, are identified. **Methods:** In an ambulatory assessment study, mood ratings were collected 56 times during 1 week from 164 students. The pleasant-unpleasant mood dimension was assessed by the two ordered categorical items unwell-well and bad-good. Mixture latent Markov models with different number of states, classes, and degrees of invariance were tested, and the best model according to information criteria was interpreted. **Results:** Two latent classes that differed in their mood regulation pattern during the day were identified. Mean classification probabilities were high ($> .88$) for this model. The larger class showed a tendency to stay in and return to a moderately pleasant mood state, whereas the smaller class was more likely to move to a very pleasant mood state and to stay there with a higher probability. **Conclusions:** Mixture latent Markov models are suitable to obtain information about interindividual differences in stability and change in ambulatory assessment data. Identified mood regulation patterns can serve as reference for typical mood fluctuation in healthy young adults. **Key words:** ambulatory assessment, experience sampling method, mood regulation, latent class analysis, hierarchical latent Markov model, mixture distribution.

3.1 Introduction

Affective states (e.g., pleasant-unpleasant mood, calm-tense mood), body states (e.g., blood pressure, sleep quality), cognitions (e.g., appraisals, self-esteem), and behaviors (e.g., treatment compliance, drinking behavior) typically fluctuate over time. Importantly, individuals may differ in the specific pattern of fluctuations they show (e.g., slow versus fast transitions between states), and these individual differences are of key interest to psychosomatic research. For instance, in health psychology, one might be interested in the patterns of instability of a specific type of health behavior, or in psychiatry, one might be interested in specific patterns of selfdestructive behavior over time. Away to explore these patterns and to gain insight into their circumstances is the repeated measurement of individuals' affective states, body states, cognitions, and behaviors via ambulatory assessment (AA). Ambulatory assessment studies provide intensive longitudinal data (e.g., several measurements a day across a period of 2 weeks) that permit researchers to analyze individual differences in patterns of change and stability (Walls, Höppner, & Goodwin, 2007). There are several statistical approaches appropriate for analyzing intensive longitudinal data. A summary of their main strengths and limitations can be found in Table 3.1.

Of particular importance to the selection of a statistical approach is the type of variable that is to be analyzed. In AA studies, states and behaviors of interest are often categorical in nature (e.g., compliant versus noncompliant behavior) or are assessed by only few items with a categorical response format (e.g., very bad mood, rather bad mood, rather good mood, very good mood). The key aim of this article was to explain and illustrate one particular approach to the analysis of intensive longitudinal data that is appropriate for categorical observed and categorical latent variables and that is able to separate variability due to occasion-specific influences from variability due to measurement error: mixture latent Markov (MLM) models (Vermunt, Tran, & Magidson, 2008). Originally, latent Markov models were developed for the analysis of panel data (Wiggins, 1973). Until recently, the application of MLM models required very large sample sizes and was restricted to few measurement occasions. Methodological developments by Vermunt et al. (2008) make it now possible to apply these models to the analysis of interindividual differences in intraindividual fluctuations in intensive longitudinal studies. Previous applications based on the new approach include models with 23 measurement occasions, but models for much

longer time series can be dealt with. Dias and colleagues (Dias, Vermunt, & Ramos, 2010) applied these methods to financial time series consisting of almost 2000 time points (days). In this article, we will show how MLM models can be applied to AA datawith many measurement occasions of many individuals. We will use the models to test hypotheses about the existence of subgroups differing in their pattern of mood fluctuations over time, which can be conceptualized as indicating different mood regulation competencies. These differences can have important consequences for subjective well-being and psychological health. Many forms of psychopathology (e.g., depression, phobias) may arise from, and be maintained by, unsuccessfully implemented mood regulation (Parkinson, Totterdell, Briner, & Reynolds, 1996).

The remainder of this section is organized as follows. First, we consider mood and mood regulation processes and review empirical findings on individual differences in mood regulation competencies. Second, we present the properties of the MLM model and demonstrate how the model can be applied to assess mood regulation patterns by AA data.

3.1.1 Mood and Mood Regulation

Mood states are diffuse and unfocused affective states which shape the background of our moment-to-moment experience (Frijda, 1994; Morris, 1989). Structural models of mood assume that mood states can be described by a few dimensions (e.g., Schimmack & Grob, 2000). The three-dimensional model of mood (Matthews, Jones, & Chamberlain, 1990; Steyer, Schwenkmezger, Notz, & Eid, 1994), for instance, includes wakefulness-tiredness, relaxation-tension, and pleasant-unpleasant mood as basic dimensions. In our application, we will focus on the pleasant-unpleasant dimension of mood.

Mood has both a stable and a variable aspect; that is, individuals have a characteristic (habitual) level of mood (also called set point), and their momentary mood fluctuates around this set point (Parkinson et al., 1996). Research has demonstrated that individuals differ both in their set point of mood and in their pattern of mood fluctuations (e.g., Eid & Diener, 1999; Eid, Notz, Steyer, & Schwenkmezger, 1994). This pattern of fluctuation is partly due to mood regulation behavior (Parkinson et al., 1996; Morris, 1989). Research has demonstrated that individuals differ considerably in their ability to effectively improve a negative mood or maintain

Table 3.1: Methodological Approaches to the Analysis of Ambulatory Assessment Data

Advantages	Limitations	Application
Time series analysis, frequency domain analysis (Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009; Shumway & Stoffer, 2011)		
- Large number of occasions - Singel-case estimates - Weekly or daily cycles	- Few individuals, no interindividual differences - Measurement error not considered - Continuous outcomes	- Effects of daily stress on well-being (Reicherts & Pihet, 2000) - Mood change frequency (Larsen, 1987)
Dynamic factor analysis (Nesselroade & Molenaar, 2004)		
- Structure of observed variables can be tested - Complex relations between variables	- Few individuals, no interindividual differences - Continuous outcomes	- Daily emotions after a romantic breakup (Sbarra, & Ferrer, 2006)
Multilevel analysis (Singer & Willett, 2003; Walls, Jung, & Schwartz, 2006)		
- Many individuals and occasions - Many different types of change processes - Intraindividual and interindividual differences	- Measurement error not considered - Measurement invariance assumed	- Intraindividual variability in positive and negative affect (Röcke, Li, & Smith, 2009)
Structural equation modeling (Eid, Courvoisier, & Lischetzke, 2012)		
- Latent variables free of measurement error - Unbiased estimates of stability and variability	- Usually restricted to continuous outcomes - No qualitative differences in change	- Mobile phone assessment of mood in daily life (Eid et al., 2012)

a positive mood (Josephson, Singer, & Salovey, 1996; Larsen, 2000; Showers & Kling, 1996).

To date, most research on mood regulation competencies has attempted to measure stable individual differences in negative mood repair and positive mood maintenance by self-report questionnaires (Bryant, 1989; Lischetzke & Eid, 2006; Salovey, Mayer, Goldman, Turvey, & Palfai, 1995). As an alternative, AA allows us to measure mood regulation competencies indirectly by drawing on information of individuals' mood course over a longer period. Mixture latent Markov models are well suited to investigate interindividual differences in the intraindividual course of mood as an indirect measure of mood regulation because these models take into account that the fluctuation process might differ between individuals and might depend on the specific state that is maintained or modified. For example, the information that over time, individuals have a high probability of changing their mood state has to be judged differently if this refers to a negative mood state or to a positive mood state.

3.1.2 Aim of the Study

The aim of the present work was to show how MLM models could be used to assess interindividual differences in mood regulation. We expected to find several latent classes of individuals who differ in their fluctuation pattern. According to current theories of mood regulation, we expected classes that differ in their ability to maintain their positive mood and to repair their negative mood. Classes with high mood maintenance should show a high probability to stay in a pleasant mood state. Classes with high mood repair should show a high probability to leave an unpleasant mood state.

3.1.3 The Mixture Latent Markov Model

In Markov models, stability and change are represented by transition probabilities, which describe the probabilities of staying in the same category over time or moving to another response category. The transition probabilities are estimated based on a set of measurement occasions. Applied to mood measurement, the model estimates the overall probability of being in a certain mood state given the state at the previous time point.

The MLM model is an extension of the simple Markov model and the latentMarkovmodel.

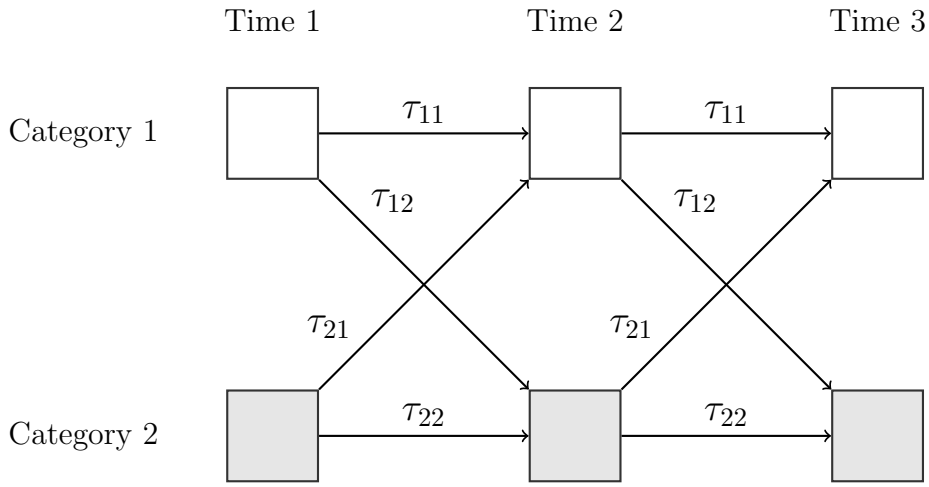


Figure 3.1: Simple Markov chain for a manifest response variable with two categories and three occasions of measurement. τ_{ij} is the transition probability from category i to j. Depicted is a time-homogeneous Markov process, because τ_{ij} are independent of the specific time point.

The simple Markov model describes relations among categories at different points in time in a so-called Markov chain. Only two kinds of parameters are needed to describe this process: the initial probabilities $Prob(Response_0)$, which contain information about the size of each category at the very first time point; and the transition probabilities $Prob(Response_t|Response_{t-1})$, which indicate the probability of a category given a certain previous category. In the Markov model with time-homogeneous transition probabilities, it is assumed that the transition probabilities are constant across different time points (time-homogeneous or stationary). To illustrate this point with a simple example, the probability of switching from a positive to a negative mood state between Time Point 1 and Time Point 2 (e.g., from morning to noon) is the same as between Time Point 2 and Time Point 3 (e.g., from noon to afternoon). The general structure of a simple Markov model with time-homogeneous transition probabilities for two observed categories is depicted in Figure 3.1. Here, the coefficient τ_{11} is the probability to stay in the first category, τ_{22} is the probability to stay in the second category, τ_{12} is the probability to move from the first to the second category, and τ_{21} is the probability to move from the second to the first category. In other words, the transition probabilities between the same category at different time points (τ_{11} and τ_{22}) describe stability. The ones between different categories (τ_{12} and τ_{21}) contain information about change.

The assumption of time-homogeneous transition probabilities can be tested by comparing a model with time-homogeneous transition probabilities with a model assuming time-heterogeneous transition probabilities. One should keep in mind that time-homogeneous transitions are only sensible if the time points are equidistant. Assuming the same influence on a current state by a state 1 hour ago or a state 5 hours ago is very restrictive. We will return to this point in the discussion. Another notable assumption in this model is the fact that a first-order Markov process is assumed. This means that the probability of a state someone is in on a certain occasion depends only on the previous occasion (and not, for example, on the occasion before that).

A disadvantage of simple Markov models is that it is unclear whether change is due to measurement error or to true change processes. Because most measures in psychological and clinical research are afflicted with measurement error, some of the observed change between categories may be attributable to measurement error instead of reflecting true change. To separate measurement error from true change, latent Markov models have been developed.

Latent Markov models are multiple indicator extensions of simple Markov models. At each measurement occasion, at least two indicators are linked to a "true" latent state variable by state-specific response probabilities. In the latent Markov model, the Markov process takes place on the level of errorfree latent categories (categorical latent state variables). Because the observed (manifest) indicators have to be linked to the latent state, the latent Markov model contains an additional type of parameter: the conditional response probabilities $Prob(Response_t|State_t)$. They describe how likely an observed category is, given a certain latent state at the same point in time. Even if, for example, someone is in a pleasant mood state, he or she might not respond with the according category *I feel well*. This would be reflected in a response probability $Prob(Response_t = well|State_t = pleasant)$ lower than 1. Whenever an observed category is linked to a corresponding latent category indicating the same state, deviations of the response probabilities from 1 indicate the influence of measurement error, and the response probabilities indicate reliability. There are as many response probabilities as combinations of observed categories and latent states. Whether these response probabilities are constant over time or not is a question of whether it is the same construct that is measured over time (Cheung & Rensvold, 2002; Meredith, 1993). In AA studies, time lags are usually too short to assume that the meaning of the latent states or the properties of the measurement instrument changes over

time. Nevertheless, the assumption of measurement invariance has to be tested.

The latent Markov model assumes that all individuals show the same fluctuation pattern. If distinct subgroups are observed, for example, via a patient/control group variable, a multigroup model with a chain for each group could be considered and differences in parameters between the chains could be formally tested. However, the groups that differ in their fluctuation pattern are often not identifiable by means of observed variables. To identify unobserved subgroups in latent Markov processes, MLM models have been defined. In MLM models, each subpopulation (latent chain or latent class) of individuals is characterized by a latent Markov model (Van de Pol & Langeheine, 1990). The aim of the analysis is to detect the number of latent chains (classes) that differ in their parameters (initial state probabilities, response probabilities, and state transition probabilities (Eid, 1996, 2007). As additional parameters, the MLM model contains the probabilities of belonging to a particular latent class $Prob(Class)$, which is also referred to as the size of the class. All other parameters are conditional on the latent class membership in the MLM model.

The MLM model can be extended by including covariates (Vermunt et al., 2008). These covariates can be either time constant, such as measures of stable personality traits or sex, or time varying, such as situational factors collected via AA (e.g., events, physiological measures). Covariates can be used to predict the different types of parameters in the model, for example, the transition probabilities. Individuals might be more likely to move to a more pleasant mood state in social situations compared with nonsocial situations. The effect of covariates could also differ between latent classes of individuals.

An attribute of AA data that requires special attention is the nesting of measurement occasions in days and the dependency between days. Not only is the time lag between the last signal at night and the first signal in the morning much longer than the time lags within the day, but the processes that operate at night might also be different. These transitions on different levels can be accounted for by treating the measurement occasions as nested within days. Such a hierarchical model was suggested by Rijmen and colleagues (Rijmen, Vansteelandt, & de Boeck, 2008) in their application of MLM models to AA data. They reported results for a study with 32 female patients and 63 signals during the course of a week, assessing emotional states. The structure of such a hierarchical MLM model is depicted in Figure 3.2. In our illustrated exam-

ple, there are two subgroups or latent day classes and three latent states for each measurement occasion. Two Markov processes are operating, one on each level: On the lower within-day level, transitions between latent states occur only during the day but not across days. On the upper between-day level of latent day classes, transitions between latent classes occur only across days but not during the day. The latent day classes for each day are obtained by separating individuals who differ in their pattern of fluctuations between latent states over the day.

In a simplified example, there may be two classes on each day: one with individuals who had a bad day (high initial probability and high stability of unpleasant mood state) and one with individuals who had a good day (high initial probability and high stability of pleasant mood state). Because of the transition between days on the upper level of the latent classes, a person can have a good day after a bad day and vice versa, independent of the person's last state on the previous night.

To sum up, the hierarchical MLM model contains the following parameters (disregarding covariates):

$Prob(Class_0)$ is the initial class probability, that is, the probability of belonging to a particular class at $d = 0$ (e.g., the first day of assessment; circled "A" in Fig. 3.2).

$Prob(Class_d|Class_{d-1})$ is the latent transition probability of being in a certain latent class given the latent class on the previous day (circled "B" in Fig. 3.2).

$Prob(State_{d0}|Class_d)$ is the initial state probability on the beginning of each day that depends on the latent class of the same day (circled "C" in Fig. 3.2).

$Prob(State_{dt}|State_{dt-1}, Class_d)$ is the latent transition probability on the latent state level, that is, the probability of being in a certain state given the previous state and the class on that day (circled "D" in Fig. 3.2).

$Prob(Response_{jdt}|State_{dt}, Class_d)$ is the response probability for the observed categories of indicator j, given the latent state on that particular time point and the latent class on that day (not depicted in Fig. 3.2).

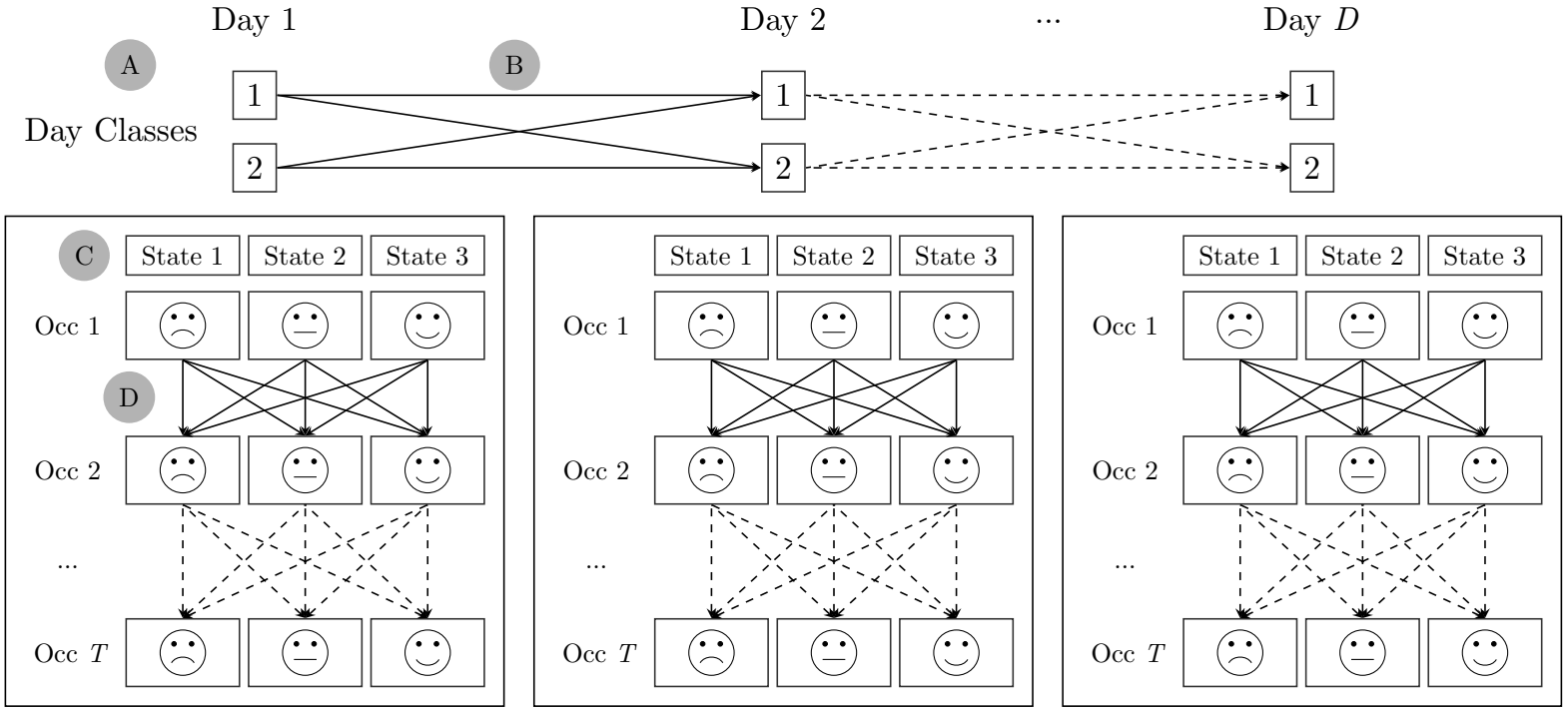


Figure 3.2: A hierarchical mixture latent Markov model with three latent states and two latent day classes. The measurement part of the model has been omitted. Occ = occasion within a day; D = number of days; T = number of occasions. Parameters denoted by gray-shaded letters are referred to in the text: A = initial day class probabilities; B = day class transition probabilities; C = initial state probabilities; D = state transition probabilities.

3.2 Application

The data analyzed here are a subset of data from a larger study on mood regulation processes that combined a laboratory session with a 14-day AA period. In the laboratory session, various personality variables were assessed via self-report. During the AA period, the focal construct of momentary mood was measured, as well as a number of additional variables. In this application, we will use data of the first week of the AA period only.

3.2.1 Participants

A total of 165 participants were recruited from the Freie Universität Berlin via a notice posted on campus. Criteria for inclusion were student status in a subject other than psychology and German as a native tongue. Data from one participant were excluded from the analyses because this person's average mood level across the AA period was exceptionally low (-5 SD). The final sample consisted of 164 students (88 women) with an M (SD) age of 23.7 (3.31) years (min = 18 years, max = 35 years). Students received 80 EUR in exchange for their participation, and an additional 20 EUR if at least 80% of the field signals were answered.

3.2.2 Procedure

Initial laboratory sessions were done in groups of one to six. Participants gave informed consent and completed a computerbased questionnaire assessing several personality dimensions. After the computer-based part, participants were given detailed instructions in the use of the handheld device and the ambulatory questionnaire. The AA period started for all participants on the Wednesday after the laboratory session, which was either the next or the next but 1 day. Data collection took place between late October 2009 and early May 2010, covering mostly the winter half year.

During the first week of the AA period, momentary mood was assessed eight times per day using signal-contingent time sampling. Participants were requested to respond on handheld devices (HP iPAQ rx 1950 Pocket PCs) when signaled by an alarm (software: Izybuilder, IzyData Ltd., Fribourg, Switzerland). The signal sounded pseudo-randomly within a 13-hour period during the day. Participants were able to choose the period according to their waking hours.

The delay between adjacent signals could vary between 60 and 180 minutes[1] (M [SD] = 100.24 [20.36] minutes, min = 62 minutes, max = 173 minutes). Responses had to be made within a 30-minute time window after the signal on the touch screen of the device using a stylus. If participants failed to respond within the 30-minute time window, the session was counted as missing. On average, the 164 participants responded to 51 (of 56) signals (M [SD] = 51.07 [6.05] signals, min = 19 signals, max = 56 signals). In total, there were 8374 nonmissing measurement occasions in the present analysis.

3.2.3 Measures

Momentary Mood

At each measurement occasion during the AA period, participants rated their momentary mood on an adapted short version of the Multidimensional Mood Questionnaire (MMQ) (Steyer et al., 1994; Steyer, Schwenkmezger, Notz, & Eid, 1997). Instead of the original monopolar mood items, a shorter bipolar version was used to fit the need for brief scales in an AA study (Piasecki, Hufford, Solhan, Trull, 2007). Several studies (Steyer et al., 1994; Steyer, Schwenkmezger, Notz, & Eid, 1997) have shown that the items belonging to the same scale of the MMQ but different poles are strongly negatively correlated when momentary mood is assessed, resulting in a common factor. Hence, building bipolar items on the basis of monopolar items of the same scale of the MMQ is acceptable. Four items assessed pleasant-unpleasant mood (happy-unhappy, content-discontent, good-bad, and well-unwell). Participants rated how they momentarily feel on a 4-point bipolar intensity scales (e.g., *very unhappy, rather unhappy, rather happy, very happy*). For the current analysis, we focused on the items well-unwell and good-bad to keep the model simple. Preliminary analysis of the response category frequencies showed that the lowest category (i.e., *very bad* and *very unwell*) was only chosen in approximately 1% of all occasions. We therefore decided to collapse the two lower categories together into one *unwell* and *bad* category, respectively. The following analyses are based on the recoded items with three categories.

[1] Two signals that violated this rule owing to device malfunctioning were excluded from the analysis.

Trait Mood Regulation

In the laboratory session, participants completed an 11-item scale measuring perceived effectiveness in mood regulation (Lischetzke & Eid, 2003). Six items assessed negative mood repair (e.g., "It is easy for me to improve my bad mood") and five items assessed positive mood maintenance (e.g., "It is easy for me to maintain my good mood for a long time"). The items were answered on 4-point frequency scales (ranging from *almost never* to *almost always*).

3.2.4 Data Analysis

Software

To estimate MLM models with many occasions and feasible sample sizes, the special forward-backward EM algorithm as described by Vermunt et al. (2008) has to be integrated in the software. For all analyses, the Latent GOLD 4.5 software package (Vermunt & Magidson, 2008) was used. Syntax for Latent GOLD and the corresponding code for the R system (R Development Core Team, 2010; Sturtz, Ligges, & Gelman, 2005) to run WinBUGS (Lunn, Thomas, Best,& Spiegelhalter, 2000) is provided in the online appendix (Supplemental Digital Content 1, `http://links.lww.com/PSYMED/A42`). Rijmen et al. (2008) provided the according functions for the Matlab environment. Functions that are able to estimate similar models in the R system may be found in the depmixS4 package (Visser & Speekenbrink, 2010).

Determining the Number of Latent States and Classes

Following a bottom-up strategy, we started building models for each of the 7 days separately to see which combination of number of latent states and number of latent classes would fit the observed data best. We expected the number of latent states to mirror the three observed states (i.e., item categories). Alternative models with two and four latent states were tested in addition. The number of latent states can deviate from the number of observed categories if the observed variables vary in their difficulties (Rost, 2002). The number of latent classes in the models we tested ranged from 1 to 4. The best fitting model was selected according to information criteria (Eid, Langeheine,& Diener, 2003; Read & Cressie, 1988). Other fit statistics that rely on the χ^2 distribution are not applicable here because of sparse contingency tables (Agresti & Yang,

1986). This problem occurs because, with many categories and time points, many combinations of the categories of the observed variables that would theoretically be possible do not appear in the data when the sample size is not extremely large. The distribution of fit statistics is no longer known and cannot be used for calculating valid p values. Information criteria depend on the fit of the model as well as its complexity. According to information criteria, the best-fitting model is the simplest model showing an adequate fit. This model can be found by comparing the information criteria of different models and selecting the model with the smallest value of the information criterion considered. There are many different information criteria that differ in how model complexity is penalized. For latent class models, the Bayesian Information Criterion (BIC) (Schwarz, 1978) has been shown to perform well (Nylund, Asparouhov, & Muthen, 2007; Li, Cohen, Kim, & Cho, 2009). For the special case of MLM models, there is some evidence (Dias, 2007) suggesting the use of the modified Akaike Information Criterion (AIC3; Bozdogan, 1987).

Testing Invariance

Next, we combined the single-day models into a model of the first week. Days were linked by a transition between latent classes at the beginning of the day.[2] We proceeded in several steps to test parameter invariance across days, starting from a baseline model, in which all parameters (class- and state-dependent response probabilities, state transition probabilities, initial state probabilities at the beginning of the day, and transition probabilities between classes) were allowed to differ between days. Subsequently, we imposed equality constraints across days on the response probabilities (Model B), the latent state transition probabilities (Model C), the initial state probabilities (Model D), and the day class transition probabilities (Model E). Finally, we analyzed a model without transitions between day classes (probabilities to stay in the same class constrained to 1; Model F).

[2]We tested whether an additional link allowing the last mood state of a day to influence the first state of the following day would improve the model, but it did not.

Table 3.2: Fit Measures for the Estimated Models

Model	LL	BIC	$AIC3$	n_{par}
A. Unrestricted baseline model	$-10,080$	$20,930$	$20,613$	151
B. Response probabilities restricted	$-10,097$	$20,902$	$20,610$	139
C. State transition probabilities restricted	$-10,151$	$20,643$	$20,502$	67
D. Initial state probabilities restricted	$-10,163$	$20,546$	$20,456$	43
E. Class transition probabilities restricted	$-10,166$	$20,500$	$20,430$	33
F. No class transition allowed	$-10,188$	$20,535$	$20,470$	31
G. Model E + covariates	$-10,152$	$20,483$	$20,410$	35

Note. LL=Loglikelihood, BIC=Bayesian Information Criterion, $AIC3$= modified Akaike Information Criterion, n_{par}=Number of parameters.

3.3 Results

The results for the single-day models showed that, in general, a model with three latent states and two latent classes can be adopted for each day. Transition probabilities were assumed to be time-invariant during the course of a day, and response probabilities were allowed to differ between latent classes. The single day models were combined into a single 7-day model. The BIC, the AIC3, and the number of parameters for each 7-day model tested are reported in Table 3.2. Because both information criteria were in agreement in our application, from here on, we will only refer to the AIC3. In the baseline model (Model A), all parameters (class- and state-dependent response probabilities, state transition probabilities, initial state probabilities at the beginning of the day, and transition probabilities between classes) were allowed to differ between days. It should be noted that this baseline model was not completely unrestrictive. Some equality assumptions had to be made to secure model identification. Model identification refers to the situation where it is possible to obtain unique estimates for all free parameters in the model (see Langeheine & Van de Pol, 2002, where identification of latent Markov models is discussed). One restriction in the baseline model concerned the response probabilities. They were restricted to be equal across measurement occasions within the same day and the same class in all models.

Next, we tested equality constraints. In Model B, we tested whether the measurement part of the model, the link between latent states and observed response categories, remained stable across days.We restricted the response probabilities to be equal across days (but not classes). Model B had a lower AIC3 than Model A did, implying that equal response probabilities could be assumed. The mood states did not change their meaning across days, but they were slightly different for the two classes. Next, the same procedure was applied to test homogeneity across days concerning the state transition probabilities. Model C contained these restrictions and yielded a lower AIC3 than Model B. In the following step, initial state probabilities were restricted to be equal across all days. The obtained Model D showed an even lower AIC3 than Model C did. The class transitions between days were set equal in Model E. Again, this restriction led to a lower AIC3. Because the stability of the classes was very high (the probability of staying in the same class across days was larger than 0.9), we tested whether it was necessary to even let people change classes between days or whether these classes could be seen as trait classes rather than day-specific classes linked by a Markov process. Compared with Model E with equal class transitions, Model F with no class transition allowed (the probabilities to stay in the same class were restricted to 1) had to be rejected based on a higher AIC3. We therefore kept Model E as our final model, which is described in more detail in the next paragraph.

In Model E, the mean classification probabilities were high for the latent states (0.94, 0.92, and 0.88) and the latent classes (0.96 and 0.93), showing that this model yielded a reliable classification of individuals. Mean classification probabilities were calculated in the following way. First, for each individual, the classification probabilities to belong to the different classes and states were calculated based on his or her response vector. Then, an individual was assigned to the latent state on each occasion and the latent class on each day for which his or her classification probability is maximum. Then, the mean of the assignment probabilities of all individuals belonging to the same class and state were calculated. The closer the mean probabilities are to 1, the better is the classification of individuals. From the perspective of psychological assessment, the classification of individuals to different types of mood regulation is a very important task. The high mean classification probabilities show that this assessment could be reliably done based on the model selected.

Two latent classes that differed with respect to their within-day mood fluctuation patterns

Table 3.3: Estimated Conditional Response Probabilities in Model E

	Item "Well"			Item "Good"		
	Unwell	Rather Well	Very Well	Unwell	Rather Well	Very Well
Class 1						
State 1	0.90	0.10	0.00	0.93	0.07	0.00
State 2	0.05	0.94	0.02	0.02	0.96	0.01
State 3	0.00	0.39	0.61	0.00	0.48	0.52
Class 2						
State 1	0.72	0.25	0.03	0.64	0.34	0.02
State 2	0.01	0.93	0.05	0.00	0.91	0.09
State 3	0.00	0.16	0.84	0.00	0.11	0.89

Note. Probabilities may not add up to one due to rounding error.

were identified. These two classes characterize the pattern of mood change within a single day. The transition probabilities did not differ between days but individuals were allowed to change classes between days. We named the larger of the two classes as Class 1 and the smaller one as Class 2. The size of Class 1 was 0.68 on the first day of our AA week and remained very stable across days (the probability to stay in this class between 2 days was 0.98). Accordingly, the smaller Class 2 had a size of 0.32 and was a little less stable (the probability to stay in this class between 2 days was 0.90). To determine the character of a class, it was crucial to first characterize the latent states in the class by looking at the response probabilities for the two different items *well* and *good*. For Model E, these can be found in Table 3.3.

Latent mood State 1 of Class 1 was characterized by a high probability of choosing the observed categories *unwell* (0.90) and *bad* (0.93), respectively. This state could be interpreted as "unpleasant mood". On the other hand, latent mood State 2 in Class 1 was associated with a very high probability of choosing the observed categories *rather well* (0.94) and *rather good* (0.96). We labeled this state "rather pleasant mood". For the last latent mood State 3 in Class 1, the probabilities of choosing the observed categories *very well* (0.61) and *very good* (0.52) were

not as high. There was still a considerable probability of choosing the middle categories *rather well* (0.39) and *rather good* (0.48). This means that individuals in this state were in a mood that is somewhat between a rather pleasant and a very pleasant mood. Because the probabilities were highest for the last item categories, we labeled this state "very pleasant mood".

Class 2 differed from Class 1 in the response probabilities given the latent states. In latent mood State 1, individuals in Class 2 had a lower probability of responding with the lowest observed categories *unwell* (0.72) and *bad* (0.64) than individuals in Class 1. Instead, there was a tendency toward the middle categories *rather well* (0.25) and *rather good* (0.34).

Compared with the "unpleasant mood" state in Class 1, this latent State 1 in Class 2 was between an unpleasant and a rather pleasant mood. Because the probabilities were highest for the first item categories, we named this state "unpleasant mood". Latent State 2 was very similar to the "rather pleasant mood" state in Class 1 and characterized by a very high probability of choosing the middle categories *rather well* (0.93) and *rather good* (0.91). In latent State 3, individuals in Class 2 had a clearly higher probability of choosing the highest observed categories *very well* (0.84) and *very good* (0.89) than individuals in Class 1. Keeping this difference in mind, this state in Class 2 was also labeled "very pleasant mood". In Class 2, the latent mood states seemed to reflect a higher basic mood level.

The initial state probabilities and state transition probabilities are depicted in Figure 3.3. In Class 1, the initial probability for the rather pleasant mood state was by far the highest (0.75). Looking at the transition probabilities, this mood state was very stable (0.82). The probabilities of changing one's rather pleasant mood state for the better (0.10) or the worse (0.09) were equally low. In this class, there was a high probability to start in a rather pleasant mood in the morning and to stay in this rather pleasant mood over the day. By comparison, the unpleasant and very pleasant mood states showed lower stabilities (0.53 and 0.52), and there was a tendency toward returning to the rather pleasant mood state. In sum, the rather pleasant mood state prevailed in Class 1.

We found a different pattern looking at Class 2. Here, people started off in a very pleasant mood (0.39) almost as often as in a rather pleasant mood (0.47). This class was more likely to remain in a very pleasant mood state (0.74) than to decline from there. In mood regulation terms, this number quantifies the extent of positive mood maintenance. Also, compared with

Class 1

Initial	*State* − 1	*State*		
		1	2	3
.07	1	.53	.44	.03
.75	2	.09	.82	.10
.18	3	.05	.43	.52

Class 2

Initial	*State* − 1	*State*		
		1	2	3
.14	1	.33	.43	.24
.47	2	.14	.57	.29
.39	3	.05	.21	.74

Figure 3.3: Estimated initial state probabilities and state transition probabilities in Model E. The upper part of the figure gives the estimates in Class 1, and the lower part of the figure gives the estimates for Class 2. The first column provides the initial probabilities for each state. For example, a member of Class 1 has a 75% chance of starting in State 2. The numbers and shaded circular areas in the grid represent the transition probabilities from the former state ($State - 1$, rows) to the current state ($State$, columns). For example, the probability for members of Class 1 to stay in State 2 is 0.82. Probabilities may not add up to 1 due to rounding error.

Class 1, Class 2 had a higher probability of entering the very pleasant mood state (0.24 and 0.29) coming from one of the other two mood states that were not as stable. Negative mood repair (the transition probabilities of moving to a better mood state if in an unpleasant mood state) added up to 0.67 compared with 0.47 in Class 1. Overall, this class had a higher mean mood level and showed a pattern of regulation toward an elevated mood state. This pattern may very well be representative for people who are exceptionally skilled in regulating their mood. They are highly able to repair their negative mood and to maintain their positive mood.

To further test this interpretation of mood regulation patterns, we included measures from the laboratory session as time-constant covariates into the model. Specifically, we regressed the class proportions on the first day on self-reported mood repair and mood maintenance by means of a binary logistic regression model (Model G in Table 2). Class 1 served as the reference category. The positive and significant regression weights for both predictors, mood repair (b = 1.02; standard error [SE] = 0.43, p = .017) and mood maintenance (b = 1.36; SE = 0.44, p = .002), reflect that individuals with high self-reported mood regulation competencies were much more likely to start in the class with the very positive regulation pattern than in the moderately positive class.

3.4 Discussion

In our empirical application, the MLM model allowed a reliable classification of individuals to different classes of mood fluctuation patterns. Moreover, it also allowed a reliable assignment to latent mood states within a day. We will review the substantive findings first and conclude with prospects of the modeling approach.

3.4.1 Individual Differences in Mood Regulation

The results revealed that there were only two classes (or patterns) of mood fluctuations in our sample.We found a class with pronounced abilities to repair negative and maintain positive mood and a class that was very stable in a moderately positive mood state. Individuals in this class were somewhat able to repair their negative mood and to maintain their very positive mood, and there seemed to be a high ability to maintain a moderately positive mood. These classes

appeared to be distinct both in their habitual mood level (or set point) and their fluctuation pattern. The smaller class with the higher habitual mood level exhibited a higher rate of overall fluctuation. This fluctuation was mainly a result of the improvement of mood states. Whether these mood fluctuations were due to mood regulation behavior or other influences (rhythmic processes due to biologic and social factors, activities and situations, positive and negative daily events) cannot be answered by the present analysis. We are able to see that people were successful in regulating their mood, but we do not know how they achieved this. Investigating this question would require incorporating information about the situation or on individuals' regulation behavior into the analysis. Ways in which the presented model can be extended to include this type of information are discussed below. An indication that mood regulation competencies do play a role in these different patterns comes from the self-reported trait measures of mood regulation. Higher reported competency is linked to the class with the higher habitual mood level. With this interpretation, one should keep in mind that individuals were allowed to change classes between days. Even if the assignment to the classes was very stable across days, there remains a day-to-day variation that cannot be accounted for by trait measures. It would be interesting to determine conditions of this day-to-day variability in future research.

It is also of interest that there was no class of individuals with a very high probability of staying in an unpleasant mood state (i.e., with low mood repair competence). This might be because we have analyzed a sample from a nonclinical population. In a clinical population, one might expect different classes, for example, a class with high probability to stay in a negative mood (depression) or a class characterized by unusual high variability of mood (borderline personality disorder). Depending on the symptoms considered, many other classes are conceivable. The classes we found may serve future clinical studies as an example of standard regulation patterns.

In addition, the results provide some insight into the relationship between the two mood regulation abilities under consideration: positive mood maintenance and negative mood repair. The fact that we did not find a class in which only one ability was present but not the other shows that they do not seem to occur independently from each other. One reason may lie in the concurrent acquisition of these competencies or in regulation strategies that can be valuable for positive mood maintenance as well as negative mood repair (e.g., social sharing). Mixture latent Markov models allow analyzing mood regulation in an indirect way on the basis of repeated

measurements. This has many advantages over the more traditional way of assessing mood regulation competencies by self-report questionnaires. The results might be less distorted by memory effects and are representative for individuals' daily life. Nevertheless, our results also demonstrate the convergent validity of our assessment method with traditional questionnaires because the mood regulation questionnaires predicted class membership.

3.4.2 The MLM model

There are at least three ways in which the MLM model applied here can be extended. First, the model can be modified to allow for between-day differences in the number and structure of the latent classes. In the application presented, we found homogeneous structures across the different days. However, this might not be the case for other constructs.

Second, time-varying covariates could be included. Mood fluctuations depend not only on mood regulation but also on situational influences. Time-varying covariates characterize the situations in which individuals are and could be used to investigate whether the latent classes differ in the way they react to these situational influences. There might be classes of individuals with high resilience that do not react strongly to negative events but also classes that might be very reactive to situational influences (Courvoisier, Eid, & Nussbeck, 2007). Mixture latent Markov models could be applied to measure resilience in an indirect way by separating latent subgroups that differ in the way situational factors influence behavior and feelings. If covariates are included, parameters are accordingly conditioned on the set of covariates. Hence, the meaning of the parameters can change when covariates are included. In which way the meaning of the parameter changes depends on the specific model considered. If covariates are observed variables, this does usually not affect the identifiability of the model parameters. However, if the covariates are latent variables, measurement models have to be specified for them as well.

Finally, we have assumed in our application that the transition probabilities between two states are the same for all individuals within a class. However, individuals might differ in the time lag between two occasions of measurement. If the time lag varies between individuals, the assumption of homogeneous transition probabilities might be inappropriate. The probability to stay in the same state might be higher for shorter time lags than for longer time lags. Although

there were individual differences in time lags in our study, the intraindividual distributions of the time lags were homogeneous. Therefore, the model seems to be appropriate in our application. If there are large interindividual differences in time lags, the model has to consider these differences. A way to adequately include interindividually varying time lags in MLM models is suggested by Vermunt (2010). Another option in this case would be to employ continuous time Markov models (Böckenholt, 2005) that do not assume equal or at least similarly spaced measurements within and between individuals, as opposed to discrete time Markov models (like the MLM model we applied here).

From a more general point of view, the application illustrates some properties of the model that are attractive for AA studies. In contrast to more traditionally used models for analyzing AA data such as classic multilevel analyses, the model has three major advantages. First, it separates change due to measurement error from true change. Second, it allows single categories (states) to differ in the process of change. The degree of state-specific stability and change can easily be modeled with MLM models, whereas it is much harder to model state-specific change processes with other statistical approaches. Third, the model permits population heterogeneity with respect to the change process. In contrast to multiple group analysis, the subpopulations do not have to be known but are rather a result of the analyses.

3.4.3 Recommendations

When is it appropriate and beneficial to employ MLM models? First of all, the data at hand should show characteristics that are suitable for Markov processes: Is it sensible to assume qualitatively distinct states (categories) for each measurement occasion? Can the switching process among these categories considered to be autoregressive? If this is the case, one can start to think in greater detail about the model. If the occasions exhibit a nested structure that is likely to affect the change process, it has to be accounted for. In our application, measurement occasions were nested in days, and we found the hierarchical approach to provide a suitable solution. If the measures used are likely to contain measurement error, researchers should include at least two indicators of the same construct so that a measurement model for the latent state can be included. With this information, a latent Markov model that adequately reflects the basic

structure of the data can be constructed.

The next step involves determining the number of latent states and latent classes, often the core question in such an analysis. The number of latent states is often expected to reflect the number of observed categories. As mentioned before, this number may increase if the observed variables vary in their difficulties, thereby capturing in-between states. Determining the number of latent classes may well be the hardest part, because as of yet, statistical fit criteria that ought to guide this decision are not well scrutinized. Theoretical considerations should be involved: How many classes can be expected? Do the classes of a particular solution make sense? Can the classes be separated well, as indicated by high mean classification probabilities? In the context of AA studies, one might even expect differences in the number or profile of the latent classes between, for example, weekdays and the weekend. In this case, one has to exercise reasonable care in arranging the data according to weekdays for all individuals in the study.

Once one has decided on the number of latent states and latent classes, one can go about to test specific assumptions, such as a homogeneous change process, by comparing models with different restrictions. In an additional step, one could include time-constant and/or time-varying covariates to gain insight into factors influencing the identified fluctuation patterns.

3.5 Acknowledgements

The authors thank Martin Schultze for his help in preparing the figures.

3.6 References

Agresti, A. & Yang, M. (1986). An empirical investigation of some effects of sparseness in contingency tables. *Comput Stat Data Anal, 5*, 9-21.

Böckenholt, U. (2005). A Latent Markov Model for the Analysis of Longitudinal Data Collected in Continuous Time: States, Durations, and Transitions. *Psych Methods, 10*, 65-83.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Bryant, F. B. (1989). A four-factor model of perceived control: Avoiding, coping, obtaining and savoring. *J Pers, 57*, 773-797.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Struct Equ Modeling, 9*, 233-255.

Courvoisier, D., Eid, M. & Nussbeck, F. W. (2007). Mixture distribution state-trait models. *Psych Methods, 12*, 80-104.

Dias, J. G. (2007). Model selection criteria for model-based clustering of categorical time series data: A monte-carlo study. In R. Decker & H. J. Lenz (Eds.), *Advances in data analysis. Proc. 30th Annual conference of the Gesellschaft für Klassifikation* (pp. 23-30). Berlin: Springer.

Dias, J. G., Vermunt, J. K. & Ramos, S. (2010). Mixture hidden Markov models in finance research. In A. Fink, B. Lausen, W. Seidel & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 451-459). Berlin: Springer.

Eid, M. (1996). Longitudinal Confirmatory Factor Analysis for Polytomous Item Responses: Model Definition and Model Selection on the Basis of Stochastic Measurement Theory. *Meth Psychol Res, 1*, 65-85.

Eid, M. (2007). Latent-Class Models for Analyzing Variability and Change. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford Handbook of Methods in Positive Psychology* (pp. 591-607). New York: Oxford University Press.

Eid, M., Courvoisier, D. & Lischetzke, T. (2012). Structural equation modeling. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 384-406). New York: Guilford.

Eid, M. & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *J Pers Soc Psychol, 76*, 662-676.

Eid, M., Langeheine, R. & Diener, E. (2003). Comparing typological structures across cultures by latent class analysis: A primer. *J Cross Cult Psychol, 34*, 195-210.

Eid, M., Notz, P., Steyer, R. & Schwenkmezger, P. (1994). Validating scales for the assessment of mood level and variability by latent state-trait analyses. *Pers Individ Dif, 16*, 63-76.

Frijda, N. H. (1994). Varieties of affect: Emotions and episodes, moods, sentiments. In P. Ekman & R. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 59-67). New York: Oxford University Press.

Josephson, B. R., Singer, J. A. & Salovey, P. (1996). Mood regulation and memory: Repairing sad moods with happy memories. *Cognit Emot, 10*, 437-444.

Langeheine, R. & Van de Pol, F. (2002). Latent Markov chains. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 304-341). Cambridge University Press.

Larsen, R. J. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *J Pers Soc Psychol, 52*, 1195-1204.

Larsen, R. J. (2000). Toward a science of mood regulation. *Psychol Inq, 11*, 129-141.

Li, F., Cohen, A. S., Kim, S.-H. & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Appl Psychol Meas, 33*, 353-373.

Lischetzke, T. & Eid, M. (2003). Is attention to feelings beneficial or detrimental to affective well-being? Mood regulation as a moderator variable. *Emotion, 3*, 361-377.

Lischetzke, T. & Eid, M. (2006). Why are extraverts happier than introverts? Exploring the role of mood regulation processes. *J Pers, 74*, 1127-1162.

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput, 10*, 325-337.

Matthews, G., Jones, D. M. & Chamberlain, A. G. (1990). Refining the measurement of mood: The UWIST Mood Adjective Checklist. *Br J Psychol, 81*, 17-42.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543

Molenaar, P. C. M., Sinclair, K. O., Rovine, M. J., Ram, N. & Corneal, S. E. (2009). Analyzing developmental processes on an individual level using non-stationary time series modeling. *Dev Psychol, 45*, 260-271.

Morris, W. N. (1989). *Mood: The frame of mind*. New York: Springer.

Nesselroade, J. R. & Molenaar, P. C. M. (2004). Applying dynamic factor analysis in behavioral and social science research. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 335-344). Thousand Oaks, CA: Sage.

Nylund, K. L., Asparouhov, T.& Muthen B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Struct Equ Modeling, 14*, 535-569.

Parkinson, B., Totterdell, P., Briner, R. B. & Reynolds, S. (1996). *Changing moods*. New York: Addison, Wesley, Longman.

Piasecki, T. M., Hufford, M. R., Solhan, M. & Trull, T. J. (2007). Assessing clients in their natural environments with electronic diaries: rationale, benefits, limitations, and barriers. *Psychol Assess, 19*, 25-43.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna. Retrieved from `http://www.R-project.org/`.

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

Reicherts, M. & Pihet, S. (2000). Job newcomers coping with stressful situations: A micro-analysis of adequate coping and well-being. *Swiss J Psychol, 59*, 303-316.

Rijmen, F., Vansteelandt, K. & de Boeck, P. (2008). Latent Class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*, 167-182.

Röcke, C., Li, S.-C. & Smith, J. (2009). Intraindividual Variability in Positive and Negative Affect Over 45 Days: Do Older Adults Fluctuate Less Than Young Adults? *Psychol Aging, 24*, 863-878.

Rost, J. (2002). Mixed and latent Markov models as item response models. *MPR Online, 7*, 53-72.

Salovey, P., Mayer, J. D., Goldman, S. L., Turvey, C. & Palfai, T. P. (1995). Emotional attention, clarity, and repair: Exploring emotional intelligence using the Trait Meta-Mood Scale. In J. W. Pennebaker (Ed.), *Emotion, disclosure, and health* (pp. 125-154). Washington, DC: American Psychological Association.

Sbarra, D. A. & Ferrer, E. (2006). The Structure and Process of Emotional Experience Following Nonmarital Relationship Dissolution: Dynamic Factor Analyses of Love, Anger, and Sadness. *Emotion, 6*, 224-238.

Schimmack, U. & Grob, A. (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *Eur J Pers, 14*, 325-345.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann Stat, 6*, 461-464.

Showers, C. J. &, Kling, K. C. (1996). Organization of self-knowledge: Implications for recovery from sad mood. *J Pers Soc Psychol, 70*, 578-590.

Shumway, R. H. & Stoffer, D. S. (2011). *Time Series Analysis and its Applications. With R examples.* New York: Springer.

Singer, J. & Willett, J. (2003). *Applied Longitudinal Data Analysis. Modeling Change and Event Occurrence.* Oxford: Oxford University Press.

Steyer, R., Schwenkmezger, P., Notz, P. & Eid, M. (1994). Testtheoretische Analysen des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF). *Diagnostica, 40*, 320-328.

Steyer, R., Schwenkmezger, P., Notz, P. & Eid, M. (1997). *MDBF – Mehrdimensionaler Befindlichkeitsfragebogen.* Göttingen: Hogrefe.

Sturtz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *J Stat Softw, 12*, 1-16.

Van de Pol, F. & Langeheine, R. (1990). Mixed Markov latent class models. *Sociol Methodol*, 213-247.

Vermunt, J. K. (2010). Longitudinal research using mixture models. In K. van Montfort, J. H. L. Oud & A. Satorra (Eds.), *Longitudinal Research with Latent Variables* (pp. 119-152). Berlin: Springer.

Vermunt, J. K. & Magidson, J. (2008). *LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., Tran, B. & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement and analysis* (pp. 373-385). Burlington, MA: Elsevier.

Visser, I. & Speekenbrink, M. (2010). depmixS4: An R Package for Hidden Markov Models. *J Stat Software, 36*, 1-21.

Walls, T. A., Höppner, B. B. & Goodwin, M. S. (2007). Statistical issues in intensive longitudinal data analysis. In A. Stone, S. Shiffman, A. Atienza & L. Nebelling (Eds.), *The Science of Real-time Data Capture* (pp. 338-360). New York: Oxford University Press.

Walls, T. A., Jung, H. & Schwartz, J. (2006). Multilevel Models and Intensive Longitudinal Data. In T.A. Walls & J. S. Schafer (Eds.), *Models for Intensive Longitudinal* (pp. 3-37). Data. New York: Oxford University Press.

Wiggins, L. M. (1973). *Panel analysis.* Amsterdam: Elsevier.

3.7 Appendix

3.7.1 Appendix A

Latent Gold syntax for estimating the unrestricted Baseline model

(Model A in the text), and the model with equality constraints on response probabilities, state transition probabilities, initial state probabilities, and class transition probabilities (Model E in the text). Lines beginning with // represent comments. The options and variables sections are the same for all models considered and stated only once. The equations section that differs between the two models is fully reproduced.

```
// options section with non-default lines only
options
   bayes categorical=1 variances=1 latent=1 poisson=1;
   missing includeall;
   output parameters=first standarderrors profile estimatedvalues;
variables
// identify each person by id variable
   caseid UserId;
// identify categorical dependent variables
   dependent well, good;
// identify independent (time structuring) variables
// first is a dummy variable that takes on the value 1 for the first
// measurement occasion of each day
// notfirst is a dummy variable that takes on the value 1 for every
// measurement occasion but the first of each day
   independent   day nominal, first nominal, notfirst nominal;
// specify names, type and number of categories for latent class variables
// "dynamic" indicates that latent variables follow a Markov process
// DClass is the latent day class variable, State the latent state class
   latent DClass nominal dynamic 2, State nominal dynamic 3;
```

Equation section for Unrestricted Baseline Model (Model A)

```
equations
// initial day class probabilities
   DClass[=0] <- 1;
// initial state probabilities dependent on day class
// parameter is given name "a" for later use in restrictions
   State[=0] <- (a) 1 | DClass[=0];
// measurement models with special error coding
// day classes may differ by constant that can differ by day
```

```
   well <- (~err) 1 | State + DClass | day;
   good <- (~err) 1 | State + DClass | day;
// Transition between day classes may occur on the beginning of each day
// but not during the day; transitions may differ between days.
   DClass <- (b~tra) first | DClass[-1] day + (-100~tra) notfirst | DClass[-1];
// Initial state probabilities may differ between day classes and days
// Transition between states dependent on day class and day
   State <- (c) first | DClass day + (~tra) notfirst | State[-1] DClass day;
//do not estimate transitions from day 0 to day 1 (there is no day 0)
   b[1,1] = -100;
   b[2,1] = -100;
// initial state probabilities on day 1 are estimated in two equations
// this sets the four parameters equal
   a = c;
```

Equation section for restricted model (Model E)

```
equations
// initial day class probabilities
   DClass[=0] <- 1;
// initial state probabilities dependent on day class
// parameter is given name "a" for later use in restrictions
   State[=0] <- (a) 1 | DClass[=0];
// measurement models with special error coding
// day classes may differ by constant
// that constant is no longer allowed to differ across days
   well <- (~err) 1 | State + DClass;
   good <- (~err) 1 | State + DClass;
// Transition between day classes may occur on the beginning of each day
// but not during the day;
// dependence on day removed, transitions may NOT differ between days.
    DClass <- (~tra) first | DClass[-1] + (-100~tra) notfirst | DClass[-1];
// Initial state probabilities may differ between day classes
// Transition between states dependent on day classes
// Dependence on day removed, hence restricted to be equal across days
   State <- (c) first | DClass + (~tra) notfirst | State[-1] DClass;
// initial state probabilities on day 1 are estimated in two equations
// this sets the four parameters equal
a = c;
```

3.7.2 Appendix B

To run this syntax, the free software environment R, the R Package 'R2WinBUGS' and WinBUGS itself need to be installed (see text for references). The code can be run in R and calls WinBUGS. Lines beginning with # represent comments.

```
# import data set (person-period format)
mydata <- read.table("C:/DataForR.csv", header=TRUE, sep=";")
# identify the case variable in the dataset
case <- mydata$case
# identify variable with day unit information
day <- mydata$day
# identify variable with occasion information
occasion <- mydata$occasion
# identify response variables
good <- mydata$good
well <- mydata$well
# define number of latent states (K) and latent day classes (L)
K <- 3
L <- 2
# define function with parameter names and starting values
inits <- function() {
list(
  beta0well = structure(.Data = c(0,-3,-3,-3,0,-3,-3,-3,0), .Dim=c(3,3)),
  beta0good = structure(.Data = c(0,-3,-3,-3,0,-3,-3,-3,0), .Dim=c(3,3)),
  beta1well = c(0, .5),
  beta1good = c(0, .5),
  pstateinit = structure(.Data = c(.3,.3,.4,.4,.3,.3) , .Dim=c(2,3)),
  pstatetran = structure(.Data = c(.8,.1,.1,.1,.8,.1,.1,.1,.8,
                                   .8,.1,.1,.1,.8,.1,.1,.1,.8),
                         .Dim=c(2,3,3)),
  pclassinit = c(.5, .5),
  pclasstran = structure(.Data = c(.9,.1,.1,.9), .Dim=c(2,2))
)
}
# create list with data vector names to be used in bugs()
alpha <- c(1,1,1,1,1)
data <- list ("case","day","occasion","good","well","K","L","alpha")
# run MCMC simulation
my.sim <- bugs(data,
               inits, parameters=c("pwell","pgood","pstateinit",
                                   "pstatetran","pclassinit", "pclasstran"),
  model.file="C:/bugsmodel.txt",
  n.chains=1, n.iter=1000, n.burnin=500, n.thin=1,
  bugs.directory="C:/WinBUGS14/",debug=TRUE)

# Notes:
# (1) bugs.directory is the directory where WinBUGS has been installed
# (2) contents of the model.file are stated below;
```

```
# Store these lines in a separate model file (here: bugsmodel.txt)
model {
# response model for well and good given class and state
for (i in 1: 8374) {
  well[i] ~ dcat(pwell[class[case[i],day[i]],state[case[i],day[i],occasion[i]],])
  good[i] ~ dcat(pgood[class[case[i],day[i]],state[case[i],day[i],occasion[i]],])
}
# within-day Markov model for state given class
for (i in 1:165) {
  for (j in 1:7) {
    state[i,j,1] ~ dcat(pstateinit[class[i,j],])
    for(k in 2:8) {
      state[i,j,k] ~ dcat(pstatetran[class[i,j],state[i,j,k-1],])
    }
  }
}
# between-day Markov model for class
for (i in 1:165) {
  class[i,1] ~ dcat(pclassinit[])
  for (j in 2:7) {
    class[i,j] ~ dcat(pclasstran[class[i,j-1],])
  }
}
# logit model for responses
for (l in 1:L) {
  for (k in 1:K) {
    for (m in 1:3) {
      ewell[l,k,m] <- exp(beta0well[k,m] + (m-1)*beta1well[l])
      egood[l,k,m] <- exp(beta0good[k,m] + (m-1)*beta1good[l])
    }
    swell[l,k] <- sum(ewell[l,k,1:3])
    sgood[l,k] <- sum(egood[l,k,1:3])
    for (m in 1:3) {
      pwell[l,k,m] <- ewell[l,k,m]/swell[l,k]
      pgood[l,k,m] <- egood[l,k,m]/sgood[l,k]
    }
  }
}
# priors for response logits
for (k in 1:K) {
  for (m in 1:3) {
    beta0well[k,m] ~ dnorm(0, 1.0E-6)
    beta0good[k,m] ~ dnorm(0, 1.0E-6)
  }
}
for (l in 1:L) {
  beta1well[l] ~ dnorm(0, 1.0E-6)
  beta1good[l] ~ dnorm(0, 1.0E-6)
}
# priors for initial state and state transition probabilities
for (l in 1:L) {
  pstateinit[l,1:K]  ~ ddirch(alpha[1:K])
  for (k in 1:K) {
```

```
    pstatetran[l,k,1:K] ~ ddirch(alpha[1:K])
  }
}
# priors for initial class and class transition probabilities
pclassinit[1:L] ~ ddirch(alpha[1:L])
for (l in 1:L) {
  pclasstran[l,1:L] ~ ddirch(alpha[1:L])
}
}
```

Chapter 4

A continuous-time mixture latent Markov model for ambulatory assessment data: Parameter recovery in small samples.

Crayen, C., Eid, M., Lischetzke, T., & Vermunt, J. K. (2015). A continuous-time mixture latent Markov model for ambulatory assessment data: Parameter recovery in small samples. *A modified version of this Manuscript was submitted for publication in the European Journal of Psychological Assessment.*

Abstract

Ambulatory assessment studies produce intensive longitudinal data with many timepoints and inter- and intraindividually varying time intervals. The mixture latent Markov model is combined with the continuous time approach for latent Markov models to arrive at a model that perfectly matches the AA data structure. It takes into account measurement error, transition between qualitative states, varying time intervals, the nested time-unit structure and heterogeneous subgroups. A simulation study with a focus on small samples (35, 50, 75, 100, 150) is conducted to learn more about minimal data requirements and differences in estimation performance between continuous time and discrete time models.

Keywords: Longitudinal categorical data, continuous time, ambulatory assessment, experience sampling method, mixture latent Markov model

4.1 Introduction

For the study of intraindividual dynamic processes, assessing individuals repeatedly in everyday settings (ambulatory assessment, AA) has become the golden standard (Stone & Shiffman, 1994; Bolger, Davis, & Rafaeli, 2003; Shiffman, Stone, & Hufford, 2008). While modern electronic devices allow a multitude of media content and physical data being recorded, subjects' self-reports are still of key interest to behavioral researchers. Presenting items of self-report questionnaires many times and on a small screen has consequences for the data structure, with the following typical characteristics:

- Reduced scales with few items
- Items with few response categories
- Many measurement occasions
- Nesting of measurement occasions (mostly within days)
- Chronological order of measurement occasions
- Varying time intervals between measurement occasions

Depending on the assumptions about the nature of the momentary state, there are several ways to analyze this type of data. For example, one could decide to aggregate the item scores on each occasion of measurement to obtain an approximately continuous outcome for longitudinal multilevel models (Singer & Willett, 2003; Walls, Jung, & Schwartz, 2006). Taking measurement error into account, one could also start with longitudinal structural equation models (Eid, Courvoisier, & Lischetzke, 2012). These often assume discrete-time data with time-lags that are constant across individuals. This situation is more often found in panel data but hardly in AA data. For the case of varying time intervals, continuous time structural equation models have been developed (Voelkle, Oud, Davidov, & Schmidt, 2012). These are, however, models for continuous (metrical) indicators, and the change process is also modeled based on continuous latent variables (factors). Also, the parameters describing the change process are assumed to be the same for all individuals. Models that feature categorical indicators and latent categorical variables are latent Markov models (van de Pol & de Leeuw, 1986). Here, the transitions between

latent states describe the change process. To allow for differences in this change process between unobserved subpopulations, latent Markov models have been extended to mixed Markov latent class models (van de Pol & Langeheine, 1990) and mixture latent Markov models (Vermunt, Tran, & Magidson, 2008). Mixed Markov models have been popular for the analysis of panel data with a limited number of measurement occasions, but for many measurement occasions like in AA data, estimation became possible through the use of efficient algorithms (Vermunt et al., 2008).

To our knowledge, there have been only two applications of the mixture latent Markov model to AA data (Rijmen, Vansteelandt, & de Boeck, 2008; Crayen, Eid, Lischetzke, Courvoisier, & Vermunt, 2012). While appropriately handling the nested-day structure in the data, both studies do not take into account varying time intervals and assume constant time intervals between measurement occasions. The potential that lies in the application of mixture latent Markov models to AA data has certainly not been realized. We suspect two reasons for this: (1) The model is still limited with regard to varying time intervals in AA data, and (2) the conditions under which the model can be successfully applied are not well examined. Therefore, the present study has two aims: (1) to expand the mixture latent Markov model to include time-varying intervals by incorporating the continuous-time approach to latent Markov models (Böckenholt, 2005); and (2) to gain insight into data requirements for the application of the model by performing a Monte-Carlo simulation study.

In the remainder of this section, we will introduce the motivating example, subsequently using it in the non-technical introduction to the hierarchical mixture latent Markov model, which we will eventually bring together with the continuous-time latent Markov approach. We conclude with a review of similar mixture simulation studies.

Our motivating example throughout the article will be the assessment of momentary mood, which is a popular outcome in AA studies, often in clinical populations (e.g., Huffziger et al., 2013; Bujarski et al., 2015). We will follow the design of a mood regulation study in a healthy student sample reported in Crayen et al. (2012). Here, the pleasant-unpleasant mood dimension (Matthews, Jones, & Chamberlain, 1990; Steyer, Schwenkmezger, Notz, & Eid, 1994), was repeatedly measured during the day by two items with three categories each over the period of one week. The process of alternating between mood states - the mood fluctuation pattern -

can be conveniently described by a Markov transition process. For ordered states like the mood ones (e.g., not good - fairly good - very good) such a transition table provides an added benefit compared to treating the outcome as continuous: The rate of change (the transition probabilities) can differ depending on the base level. While some mood variability measures have been shown to be associated with low psychological well-being (see Houben, van den Noortgate, and Kuppens, 2015, for a meta-analysis), information on the base level of fluctuation is often lost. For example, affective inertia in a fairly or very good state (stability) might indicate mood savoring competencies (Salovey, Mayer, Goldman, Turvey, & Palfai, 1995), while inertia in a bad mood more likely indicates low ability to repair one's mood. In the example application, two latent classes were found that differed with regard to their daily mood fluctuation pattern.

4.2 The Mixture Latent Markov Model

The model evaluated in this simulation study is a continuous-time hierarchical mixture latent Markov model. As the name suggests, it extends the mixture latent Markov models (MLM; van de Pol & Langeheine, 1990; Vermunt, Tran, & Magidson, 2008) to the continuous time approach to latent Markov models (CT-LM; Böckenholt, 2005). The hierarchical mixture latent Markov model for discrete-time data has been considered by Rijmen et al. (2008) and Vermunt (2010). This hierarchical model should not be confused with a multilevel mixed latent Markov model (MLMM; Yu, 2007; Bartolucci & Lupparelli, 2012; also described in Vermunt, 2010). In a MLMM model, the Markov process takes places within Level-1 units (e.g., students) which are nested in Level-2 units (e.g., schools). The mixture part can be located on both levels (e.g. schools are grouped into latent classes according to the typical transitions of their students). In our AA example, on the other hand, there is one Markov chain within each individual, but this chain is cut into within-day and between-day processes according to the nested time units. Nevertheless, we will refer to the within-day level as Level-1 and to the between-day level as Level-2.

We start our description of the discrete-time hierarchical mixture latent Markov model with the measurement part, progressing to the within-day structural part and finally to the between-day structure. The model's empirical base are responses y_{idtj} from subject i ($i = 1, ..., N$), measured at occasion t ($t = 0, ..., T$) on study day d ($d = 0, ..., D$) on a set of J manifest

Table 4.1: Example initial state and transition probabilities for $K = 3$ latent mood states. Shaded cells represent stability.

		SC_{dt}		
$P(SC_d)$	SC_{dt-1}	1	2	3
.18	1: Not good	.38	.55	.07
.65	2: Fairly good	.09	.75	.16
.17	3: Very good	.06	.42	.52

indicators. The number of categories of each indicator is M_j. For each measurement occasion dt, there is a latent state class variable SC_{dt} (*state* for its occasion-specific nature) with k ($k = 1, ..., K$) categories, which are related to the categories of the manifest indicators via conditional response probabilities $P(y_{idtj} = m_j \mid SC_{dt} = k)$. Unless restricted, there are $M_j * K$ conditional response probabilities per indicator, one for every combination of categories. The local independence assumption from standard LCA (Goodman, 1974) is made here, stating that there is no association between the indicators of the same occasion conditional on the common latent state. Conditional response probabilities lower than 1 therefore imply measurement error. The latent state categories can be labeled according to the association pattern eminent from the conditional response probabilities. The conditional response probabilities are usually restricted to be the same over time to preserve the interpretation of the latent states. For our mood example, we assume that the number and meaning of manifest categories directly carries over to the latent state categories (i.e., not good - fairly good - very good).

The within-day fluctuation process we are interested in is now regarded on the level of the occasion-specific latent state variable SC_{dt} (we consider only the within-day process for now but keep the index for the day). It is thought to follow a simple first-order stationary Markov process, which means that each state is only dependent on the previous one and that this dependency is the same across occasions. This process can be conveniently described by a set of K initial state probabilities $P(SC_{d0} = k)$ and (in the discrete-time case) a set of $K * K$ transition probabilities $P(SC_{dt} = k \mid SC_{dt-1} = k_{t-1})$. An example set of initial state and transition probabilities for the three mood states is given in Table 4.1.

Table 4.2: Initial state and transition probabilities for the example with $K = 3$ latent mood states and $L = 2$ latent day classes. Shaded cells represent stability.

			SC_{dt}		
DC	$P(SC_d)$	SC_{dt-1}	1	2	3
	.18	1: Not good	.42	.54	.04
1	.76	2: Fairly good	.10	.79	.10
	.06	3: Very good	.05	.56	.39
	.12	1: Not good	.22	.54	.24
2	.51	2: Fairly good	.14	.51	.35
	.37	3: Very good	.05	.29	.66

The pattern in Table 4.1 can be interpreted in the following way: At the beginning of the day ($t = 0$), the probability for a random person in the sample for being in a fairly good mood is .65. The probability for a person in a fairly good mood to transition to a very good mood for the next measurement occasion is .16. For a person in a very good mood the probability to stay there until the next measurement occasion is .52 and so on. Transition probabilities are often referred to by the position of their cell in the table, i.e., $p_{23} = .16 = P(SC_{dt} = 3 \mid SC_{dt-1} = 2)$.

As appealing as such a parsimonious description of a large longitudinal sample might be, it is of course very restrictive to assume that every person in the sample is well described by the same process. Instead, the typological approach used here assumes that there is more than one typical pattern, and that individuals can be grouped together according to their observed pattern (i.e., it is not necessary to assume individual patterns). Such a latent categorical variable that classifies individuals according to their within-day transition pattern is situated on the between-day level and will be called day class variable (DC) here. The number of categories ($l = 1, ., L$) has to be determined from theory or from the data in an exploratory manner. Because there are multiple (D) consecutive days, there is one DC for each day. In fact, it is really a day-specific *state*. In our example, there are two day-classes (two categories of the DC variable) that differ with regard to their typical initial mood and their mood fluctuation pattern (see Table 4.2).

Conditioning the probabilities introduced so far on the day class of the particular day yields

$$P(y_{idtj} = m_j \mid SC_{dt} = k, DC_d = l) \tag{4.1}$$

$$P(SC_{d0} = k \mid DC_d = l) \tag{4.2}$$

$$P(SC_{dt} = k \mid SC_{dt-1} = k_{t-1}, DC_d = l). \tag{4.3}$$

For the Markov process on the between-day level, there are also initial day class probabilities and transition probabilities:

$$P(DC_0 = l) \tag{4.4}$$

$$P(DC_d = l \mid DC_{d-1} = l_{d-1}) \tag{4.5}$$

Now we can model the probability of subject i's responses y_{idtj} as a combination of these five probability types:

$$P(y_{idtj}) = \sum_{DC_0=1}^{L} \sum_{DC_1=1}^{L} ... \sum_{DC_D=1}^{L} P(DC_0) \left[\prod_{d=1}^{D} P(DC_d \mid DC_{d-1})\right] \left[\prod_{d=0}^{D} P(y_{idtj} \mid DC_d)\right] \tag{4.6}$$

gives the between-day part, where the within-day part is

$$P(y_{idtj} \mid DC_d) = \sum_{SC_0=1}^{K} \sum_{SC_1=1}^{K} ... \sum_{SC_T=1}^{K} P(SC_{d0} \mid DC_d) \left[\prod_{t=1}^{T} P(SC_{dt} \mid SC_{t-1}, DC_d)\right] \left[\prod_{t=0}^{T} \prod_{j=1}^{J} P(y_{idtj} \mid SC_{dt} DC_d)\right]. \tag{4.7}$$

An additional (but not necessary) restriction in the model is that the first state of each day is independent from the last state of the previous day given the current day-class.

4.3 Continuous time

So far, we have treated the measurement occasions as if they were equally spaced (with a constant time interval). However, in AA studies, intervals between measurement occasions mostly vary within and between individuals. The set of transition probabilities from a discrete-time model is most correctly interpreted as being representative for a time interval equal to the overall mean interval. Still, the estimated transition process will not be precise, because the effect of the previous state is assumed to be constant. In the data, more closely spaced observations will have stronger associations (higher stability) than more distant observations. Another assumption in discrete-time models concerns the timing of transition: The can only occur when a state is observed. A time-continuous process with transitions occurring at any point in time is often more in line with psychological theory. The continuous-time approach has a long history in Markov model theory (Singer & Spilerman, 1976; Kalbfleisch & Lawless, 1985; Böckenholt, 2005). Böckenholt (2005) applied the continuous-time latent Markov model to AA data but focused on data from a single day and on covariates for the change process rather than on differential change. The main difference between the discrete time (DT in the following) and the continuous time (CT) approach is that instead of transition probabilities, transition intensities are defined (Coleman, 1981; Kalbfleisch & Lawless, 1985). Transition intensities (or rates) q_{ab} can be thought of as probability per time unit. For our example from Table 4.1, they can be written as

$$q_{ab} = \lim_{\Delta t \to 0} \frac{P(SC_{dt} = b \mid SC_{dt-\Delta t} = a)}{\Delta t} \qquad \text{and} \tag{4.8}$$

$$q_{ab} * \Delta t = P(SC_{dt} = b \mid SC_{dt-\Delta t} = a) \qquad \text{with } a \neq b \tag{4.9}$$

When two states are observed at nearly the same time point ($\Delta t \to 0$), it is highly unlikely to observe a transition from state a to state b (because with $a \neq b$, stability is not included). The longer the time interval, the more likely it becomes to observe a transition. The transition intensity is the parameter that defines this rate of change in transition probabilities. Just like the transition probabilities in each row of Table 4.1 sum to 1, the corresponding transition intensities sum to 0, thereby making it convenient to obtain the non-transition rate

$$q_{aa} = P(SC_{dt} = a \mid SC_{dt-\Delta t} = a) = -\sum_{a \neq b} q_{ab} * \Delta t \tag{4.10}$$

Note that while the specific transition probabilities are a function of the time interval, the transition rates themselves are constant over time (we have a stationary process). Just like the probabilities, they are assumed to be constant for all subjects (within the same day-level class). In order to obtain a set of transition probabilities from the intensities for a specific time interval, the following relationship is crucial:

$$P(\Delta t) = e^{Q\Delta t} \tag{4.11}$$

Where $P(\Delta t)$ is the transition probability matrix for the specific time interval Δt, $e^{Q\Delta t}$ is the matrix exponential and Q is the transition intensity matrix with entries

$$Q = \begin{pmatrix} -(q_{ab} + q_{ac}) & q_{ab} & q_{ac} \\ q_{ba} & -(q_{ba} + q_{bc}) & q_{bc} \\ q_{ca} & q_{cb} & -(q_{ca} + q_{cb}) \end{pmatrix} \tag{4.12}$$

To give an example, with this set of transition intensities

$$Q = \begin{pmatrix} -.81 & .75 & .06 \\ .17 & -.32 & .15 \\ .06 & .33 & -.39 \end{pmatrix} \tag{4.13}$$

We obtain the transition probabilities in Table 4.3 for $\Delta t = 1$ (e.g. one hour) and $\Delta t = 2$ (see R code in the Appendix for computation). As can be seen in Figure 4.1, stability decreases (and transition probabilities increase) exponentially with an increasing time interval. Note that even though it is possible to obtain transition probabilities for a wide range of values for the time interval, interpretation for values outside the range of the data set used to estimate the intensity

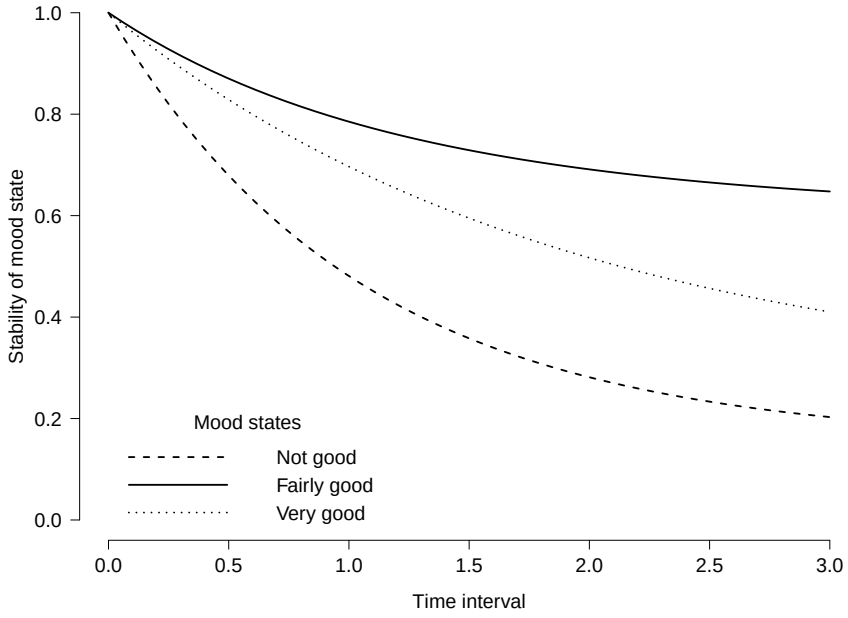


Figure 4.1: Stability of the mood states as a function of the time interval

parameters is poor practice.

This continuous-time hierarchical mixture latent Markov model perfectly matches the initially stated typical characteristics of AA data: Through its transition process on the level of categorical latent variables, the chronological order of the multivariate longitudinal categorical indicators is taken into account. Depending on the nesting of measures, two processes at different time units are defined. And finally, heterogeneity and varying time intervals are incorporated.

4.3.1 Related simulation studies

To our knowledge, no simulation study has been published for the hierarchical mixture latent Markov model. Fan (2014) conducted a simulation study on the accuracy of a mixed (manifest) Markov model in retrieving the correct number of clusters when assumptions are violated. Results were robust and more sensitive to the number of time points (50, 100 or 200) than to the number of subjects (50 or 100 in each of 3 to 6 classes). In this study, the minimum number of total data points added up to 2500 (50*50*3).

Table 4.3: Two sets of transition probabilities obtained from the transition intensities given in the text with $\Delta t = 1$ and $\Delta t = 2$.

		SC_{dt}		
Δt	$SC_{d(t-\Delta t)}$	1	2	3
1	1: Not good	.48	.45	.07
	2: Fairly good	.10	.79	.11
	3: Very good	.05	.25	.70
2	1: Not good	.28	.59	.13
	2: Fairly good	.14	.69	.17
	3: Very good	.09	.40	.52

Multilevel latent class models without longitudinal data structure have received some more attention. Most of the simulation studies conducted in this field are concerned with the performance of information criteria in determining the model with the correct number of classes, such as Lukociene and Vermunt (2010), more thoroughly Lukociene, Varriale, and Vermunt (2010), as well as Yu and Park (2014). Lukociene and Vermunt (2010) combined different numbers of lower-level units per higher-level unit (5, 10, 15, 20, 30) with 50 and 500 higher level units, which leads to a minimum total sample size of 250, the minimum number of total data points in the design of Lukociene, Varriale and Vermunt (2010) was 5*30 = 150. Neither study reported estimation problems.

Focused on differences between parametrizations of ML-LCA models, Finch and French (2014) looked at parameter estimation accuracy and the quality of classification, depending on the number of higher level units (25, 75, 150 ,200), lower level units (10, 20, 35), and number of indicators (5, 10, 15). The results for their non-parametric model might partly be applicable to our study, with parameter coverage on the lower level equal to the desired .95 and on the higher level around .9, on both levels unaffected by the number of units.

4.4 Goal of the present study

Given that few studies to date applied or simulated discrete-time hierarchical MLM models *or* continuous-time (non-mixture) latent Markov models, little is known about the general performance and data requirements of these models. Our goal was to explore the minimal data requirements needed for trustworthy results by means of a Monte Carlo simulation study. Monte Carlo simulations are a tool to evaluate the performance of statistical models under various conditions (Harwell, Stone, Hsu, & Kirisci, 1996). Multiple data sets are sampled based on a population model with known parameters. The data sets may differ with regard to sample size, or, in our case, number of time points. Each sampled data set is then analyzed along the lines of real empirical data and all results from one condition are aggregated to yield sampling distributions for the estimated parameters. These can then be compared to known population parameters to detect systematic deviations (bias). In a Monte Carlo simulation study, the sample and model characteristics are the independent factors and indicators of model performance are the dependent measures.

We were particularly interested in the comparison of discrete-time and continuous-time models, hypothesizing that the lack of use of CT models might stem from a more unstable performance. From an applied point of view, we were interested in how different ways of adding data points to an AA study (subjects, occasions, or days) could have a compensatory effect. It is, for example, usually cheaper to add measurement days for each subject than to recruit more subjects, or the number of days could be limited in some way (days spent in the hospital) but more measurement occasions were an option. Our general expectations were that additional empirical information would lead to more precise parameter estimation and that parameters on the between-day level would gain more from additional days than from additional measurement occasions within a day.

4.5 Method

4.5.1 Population model

A CT hierarchical mixture latent Markov model was refit to data reported in the motivating study (Crayen et al., 2012). There were three latent states mirroring the manifest categories and two latent day classes, the smaller one with a size of .33. With slight alterations[1], this served as the population model. All population parameters can be found in Table 4.4. The central assumptions in the model were invariance of (a) conditional response probabilities and (b) the transition process across measurement occasions and across days. The conditional response probabilities differed between Level-2 classes only in a shift parameter. The Level-2 transition process was also assumed to be invariant across days. Note that the data generating model was always a CT model, but that the bias measures for the DT model are based upon the converted parameters that result from constant time intervals equal to the mean time interval in the data.

4.5.2 Independent factors

Model type

We were interested in comparing the performance of models treating the within-day occasions as equally spaced (DT models) versus regarding the exact time interval (CT models). Because the population model was a CT model, adjustments were made for the DT model for a fair comparison: The DT model included four additional transition parameters on the lower level to allow for direct transitions between the two extreme categories of the latent state (which are allowed indirectly in CT models). These additional parameters (marked grey in Table 4.4) were not included in further analyses.

[1]To avoid an excessive amount of boundary values in small samples, some parameters associated with very low probabilities were fixed: Four conditional response probabilities linking the extreme manifest category to the opposite latent category, and transition intensities linking the lowest mood state to the highest mood state. Note that the latter four were the additional parameters in the DT model. In addition, two parameters that had an absolute value > 4 (p9 and p12) were lowered.

Empirical information

Three main aspects to decide upon in any ambulatory assessment study concern the number of subjects in the sample (N), the length of the study period (or number of days, D), and the number of measurement occasions within each day (T). In the present simulation study, those three aspects were varied systematically. Based on a literature review of ambulatory assessment studies that examined affective processes (e.g., Miller, Hedeker, Mermelstein, Berbaum, and Campbell, 2009; Miller, Vachon, & Cynam, 2009), five values for the sample size were chosen with an emphasis on small samples: 35, 50, 75, 100, and 150. For the number of days, a short period ($D = 3$), the very common period of one week ($D = 7$), and a longer period of $D = 10$ days was selected. The number of measurement occasions within a day (T) varied between 4, 6, and 8. The simulation design was fully crossed and resulted in a total of 90 conditions.

Type of parameter

The 25 parameters in the CT model were grouped into five types as follows (see Table 4.4):

1. Five parameters per item in the measurement part of the model (four conditional response probabilities plus one constant describing the difference between day-classes)

2. Four parameters represented the initial state within a day

3. Eight (common) Level-1 transition parameters for the within-day Markov chain

4. One (initial) class size parameter

5. Two Level-2 transition parameters for the transition between the day-classes

Parameter types 2, 3, and 5 are transition parameters for which the parametrization differs between CT and DT models (see Table 4.4). In the case of 2 and 5, however, time intervals in the data sets are constant (1 and 4.8, respectively). Dependent measures will be analyzed separately for each type of parameter.

Table 4.4: Parameters of the generating model can be found in the column CT. The same values are used to calculate bias measures for the CT models. Parameters for calculation of bias measures for the DT models differ and are partly dependent on number of measurement occasions. The additional L1 transition parameters 19, 21, 26, 28 for the DT model (shaded grey) were not analyzed.

		Parameter values			
		CT	DT		
Type and Index	Parametrization (DT)		$T=4$	$T=6$	$T=8$
Initial class size					
1	$\ln\left(\frac{P(DC_{d=0}=2)}{P(DC_{d=0}=1)}\right)$	-0.7			
Measurement part					
2	$\ln\left(\frac{P(Y_{1dt}=2\|SC_{dt}=1,DC_d=1)}{P(Y_{1dt}=1\|SC_{dt}=1,DC_d=1)}\right)$	-2.4			
3	$\ln\left(\frac{P(Y_{1dt}=1\|SC_{dt}=2,DC_d=1)}{P(Y_{1dt}=2\|SC_{dt}=2,DC_d=1)}\right)$	-2.9			
4	$\ln\left(\frac{P(Y_{1dt}=3\|SC_{dt}=2,DC_d=1)}{P(Y_{1dt}=2\|SC_{dt}=2,DC_d=1)}\right)$	-3.6			
5	$\ln\left(\frac{P(Y_{1dt}=2\|SC_{dt}=3,DC_d=1)}{P(Y_{1dt}=3\|SC_{dt}=3,DC_d=1)}\right)$	-0.6			
6	$\ln\left(\frac{\frac{P(Y_{1dt}=2\|SC_{dt}=1,DC_d=2)}{P(Y_{1dt}=1\|SC_{dt}=1,DC_d=2)}}{\frac{P(Y_{1dt}=2\|SC_{dt}=1,DC_d=1)}{P(Y_{1dt}=1\|SC_{dt}=1,DC_d=1)}}\right)$ etc.	1.2			
7	$\ln\left(\frac{P(Y_{2dt}=2\|SC_{dt}=1,DC_d=1)}{P(Y_{2dt}=1\|SC_{dt}=1,DC_d=1)}\right)$	-2.7			
8	$\ln\left(\frac{P(Y_{2dt}=1\|SC_{dt}=2,DC_d=1)}{P(Y_{2dt}=2\|SC_{dt}=2,DC_d=1)}\right)$	-3.6			
9	$\ln\left(\frac{P(Y_{2dt}=3\|SC_{dt}=2,DC_d=1)}{P(Y_{2dt}=2\|SC_{dt}=2,DC_d=1)}\right)$	-3.6			
10	$\ln\left(\frac{P(Y_{2dt}=2\|SC_{dt}=3,DC_d=1)}{P(Y_{2dt}=3\|SC_{dt}=3,DC_d=1)}\right)$	-0.3			
11	$\ln\left(\frac{\frac{P(Y_{2dt}=2\|SC_{dt}=1,DC_d=2)}{P(Y_{2dt}=1\|SC_{dt}=1,DC_d=2)}}{\frac{P(Y_{2dt}=2\|SC_{dt}=1,DC_d=1)}{P(Y_{2dt}=1\|SC_{dt}=1,DC_d=1)}}\right)$ etc.	1.9			
Level-2 transition					
12	$\ln\left(\frac{P(DC_d=2\|DC_{d=d-1}=1)}{P(DC_d=1\|DC_{d=d-1}=1)}\right)$	-2.2	-0.7		
13	$\ln\left(\frac{P(DC_d=1\|DC_{d=d-1}=2)}{P(DC_d=2\|DC_{d=d-1}=2)}\right)$	-2.3	-0.8		
Initial state					
14	$\ln\left(\frac{P(SC_{dt}=2\|SC_{dt=0}=1,DC_d=1)}{P(SC_{dt}=1\|SC_{dt=0}=1,DC_d=1)}\right)$	0.5	1.5		
15	$\ln\left(\frac{P(SC_{dt}=3\|SC_{dt=0}=1,DC_d=1)}{P(SC_{dt}=1\|SC_{dt=0}=1,DC_d=1)}\right)$	-2.1	-1.1		
16	$\ln\left(\frac{P(SC_{dt}=2\|SC_{dt=0}=1,DC_d=2)}{P(SC_{dt}=1\|SC_{dt=0}=1,DC_d=2)}\right)$	0.2	1.4		
17	$\ln\left(\frac{P(SC_{dt}=3\|SC_{dt=0}=1,DC_d=2)}{P(SC_{dt}=1\|SC_{dt=0}=1,DC_d=2)}\right)$	-0.1	1.1		
Level-1 transition					
18	$\ln\left(\frac{P(SC_{dt}=2\|SC_{dt=t-1}=1,DC_d=1)}{P(SC_{dt}=1\|SC_{dt=t-1}=1,DC_d=1)}\right)$	-0.4	0.8	0.3	-0.1
19	$\ln\left(\frac{P(SC_{dt}=3\|SC_{dt=t-1}=1,DC_d=1)}{P(SC_{dt}=1\|SC_{dt=t-1}=1,DC_d=1)}\right)$		-1.4	-2.2	-2.8
20	$\ln\left(\frac{P(SC_{dt}=2\|SC_{dt=t-1}=1,DC_d=2)}{P(SC_{dt}=1\|SC_{dt=t-1}=1,DC_d=2)}\right)$	0.5	1.1	0.9	0.7
21	$\ln\left(\frac{P(SC_{dt}=3\|SC_{dt=t-1}=1,DC_d=2)}{P(SC_{dt}=1\|SC_{dt=t-1}=1,DC_d=2)}\right)$		0.8	0.1	-0.5
22	$\ln\left(\frac{P(SC_{dt}=1\|SC_{dt=t-1}=2,DC_d=1)}{P(SC_{dt}=2\|SC_{dt=t-1}=2,DC_d=1)}\right)$	-2.0	-1.8	-2.0	-2.2
23	$\ln\left(\frac{P(SC_{dt}=3\|SC_{dt=t-1}=2,DC_d=1)}{P(SC_{dt}=3\|SC_{dt=t-1}=2,DC_d=1)}\right)$	-2.0	-1.9	-2.0	-2.2
24	$\ln\left(\frac{P(SC_{dt}=1\|SC_{dt=t-1}=2,DC_d=2)}{P(SC_{dt}=2\|SC_{dt=t-1}=2,DC_d=2)}\right)$	-0.9	-1.3	-1.3	-1.4
25	$\ln\left(\frac{P(SC_{dt}=3\|SC_{dt=t-1}=2,DC_d=2)}{P(SC_{dt}=2\|SC_{dt=t-1}=2,DC_d=2)}\right)$	-0.7	-0.1	-0.38	-0.6
26	$\ln\left(\frac{P(SC_{dt}=1\|SC_{dt=t-1}=3,DC_d=1)}{P(SC_{dt}=3\|SC_{dt=t-1}=3,DC_d=1)}\right)$		-1.3	-2.1	-2.7
27	$\ln\left(\frac{P(SC_{dt}=2\|SC_{dt=t-1}=3,DC_d=1)}{P(SC_{dt}=3\|SC_{dt=t-1}=3,DC_d=1)}\right)$	-0.3	0.9	0.37	0
28	$\ln\left(\frac{P(SC_{dt}=1\|SC_{dt=t-1}=3,DC_d=2)}{P(SC_{dt}=3\|SC_{dt=t-1}=3,DC_d=2)}\right)$		-2.1	-2.6	-2.9
29	$\ln\left(\frac{P(SC_{dt}=2\|SC_{dt=t-1}=3,DC_d=2)}{P(SC_{dt}=3\|SC_{dt=t-1}=2,DC_d=2)}\right)$	-0.9	-0.6	-0.8	-1.1

4.5.3 Dependent measures

Estimation problems, Information Criteria and Classification

The proportion of non-converged replications and replications with other estimation problems was recorded for each condition. There are several information criteria (IC) recommended for this type of longitudinal model, the BIC (Nylund, Muthén, & Asparouhov, 2007; Costa & de Angelis, 2010) and the AIC3 (Dias, 2007). Because sample sizes on both levels and the number of (common) parameters are equal for CT and DT models in this simulation, both IC would produce the same difference which is equal to the difference in the raw log-likelihood values. We will therefore only report the BIC. The mean classification probabilities for the response vectors based on the posterior class membership probabilities for the state-classes on Level-1 and the day-classes on Level-2 will be reported.

Root median squared error

For the evaluation of parameter estimation performance, the root median squared error ($RMdSE$) was calculated. It is the square root of the median of the squared differences between parameter estimate and population value:

$$RMdSE = \sqrt{Md_{(\hat{e}-e)^2}} \tag{4.14}$$

This median-based measure is less sensitive to few extreme estimates than measures commonly used in the context of factor analysis (mostly standardized deviations, see, for example, Bandalos, 2006). Extreme estimates occur more frequently when the logit parametrization is used.

Coverage and median width of confidence interval

The 95% CI was as usual constructed with upper and lower limits at 1.96 times the estimated standard error below resp. above the estimated parameter for each parameter in each replication. It was recorded whether the population parameter was covered by this CI or not, and the mean proportion across replications (and in our case: parameters of the same type) gives the *coverage*. This value should mirror the theoretical 95%. The median width of the confidence interval

($CIMd$) was used as a robust measure for the performance of standard error estimation.

$$CIMd = Md_{2*1.96*SE} \tag{4.15}$$

The standardized standard error bias measure often used as an outcome in factor analysis simulations derives its population value from the simulation estimates, which are themselves susceptible to bias in small samples.

4.5.4 Data generation and analysis

For each of the 45 data conditions, 500 data set templates were constructed by our own routine in R (R Core Team, 2014) which is provided in the supplementary material. Time points were randomly drawn from a uniform distribution within segments of a presumed observation day lasting from 8 am to 10 pm (840 minutes), with the additional condition of a minimum time interval of 30 minutes. The resulting time intervals for the different number of within-day measurement occasions were approximately normally distributed within their possible value range (values of 1 correspond to 100 minutes, see Table 4.5). Based on the time intervals and the parameters of the population model, the manifest responses on the two items with three categories were generated by Latent Gold 5 (Vermunt & Magidson, 2013). Within the same software, both, CT and DT models were applied to each of the 22,500 data sets with the population values serving as starting values, the number of iterations for the EM algorithm set to 300 and the number of iterations for the Newton-Raphson algorithm set to 200. The estimated parameters, standard errors and statistics were saved to single result files, which were further processed in R (R Core Team, 2014). Sample syntax is provided in the supplementary material.

For the calculation of the performance measures, non-converged replications and replications with other estimation problems (mainly rank deficiency) were excluded. In addition, single parameter estimates with an absolute value of 15 or larger and standard errors with values above 100 were counted as boundary values and, together with the corresponding standard error or parameter estimate, excluded from further analysis. Label switching was ruled out on both levels by inspecting the orientation of day-class specific parameters. Within each replication, measures were aggregated as the median across parameters of the same type.

Table 4.5: Distribution of time intervals for different number of measurement occasions T. One unit corresponds to 100 minutes.

	T		
	4	6	8
M	2.10	1.40	1.05
SD	0.84	0.55	0.40

While we put emphasis on the graphical representation of the median bias measures (Figures 4.4 to 4.13), we also report effect sizes on the means (Table 4.7). Separately for each bias measure and model type, a three-way factorial analysis of variance was performed within the general linear model framework, using zero-sum contrasts and log-transformed dependent measures. Because all effects are significant ($\alpha = .001$) with this large sample (number of replications), emphasis is put on the total $\hat{\eta}^2$ for each effect, which was calculated using the Type-III sum of squares provided by the car-package (Fox and Weisberg, 2011). Only those effects that reached Cohen's (1988) large effect size criterion of .14 are reported in the text.

4.6 Results

4.6.1 Estimation problems

Overall, about 5% of replications in the CT condition and 1% in the DT condition did not converge. As can be seen in Figure 4.2, almost all of the non-converged replications in the CT model were either (51%) in conditions with the combination of small sample sizes ($N > 75$) and the shortest study period ($D = 3$), or (another 44%) in the remaining conditions with few measurement occasions ($T = 4$). In converged replications, the most frequent estimation problem was rank deficiency, especially in DT models (41%). Because many of these cases in DT models were associated with standard errors of the (quite extreme) four additional DT parameters, the count depicted in Figure 4.2 takes only the parameters common to both models into account. This corrected count leads to similar numbers for both, DT (13%) and CT (12%) models. Estimation problems occurred very frequently for the shortest study period ($D = 3$) in combination with

Table 4.6: Number of valid replications per condition. Counts below 100 appear in boldface.

		Length of study period (D) and number of measurement occasions (T)								
		$D = 3$			$D = 7$			$D = 10$		
Model	N	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$
CT	35	**18**	**69**	283	329	337	497	285	499	488
	50	**82**	263	486	460	396	498	256	500	500
	75	232	391	478	490	400	497	452	498	500
	100	391	441	499	492	494	500	443	500	500
	150	459	427	499	496	500	500	496	500	500
DT	35	**16**	386	470	452	381	480	**75**	488	464
	50	**64**	387	487	486	471	493	198	498	492
	75	315	358	491	494	422	498	401	500	500
	100	425	393	492	500	496	500	424	500	500
	150	429	374	498	499	500	500	489	500	500

Table 4.7: Total $\hat{\eta}^2$ for bias measures by model and parameter type. Values greater than .13 appear in boldface.

			CT Model					DT Model				
Measure	Effect	Df_1	Msr	L1Tr	IStP	L2Tr	CS	Msr	L1Tr	IStP	L2Tr	CS
	N	4	**.15**	**.14**	**.23**	.13	.03	**.14**	.12	**.20**	**.15**	.03
	T	2	**.15**	**.16**	.01	.07	.07	**.16**	**.20**	.01	.10	.09
	D	2	.06	.08	.04	**.24**	.06	.07	**.14**	.02	**.18**	.06
$RMdSE$	$N \cdot T$	8	.02	.01	.08	.03	.03	.02	.01	.07	.04	.05
	$N \cdot D$	8	.06	.00	.01	.01	.01	.06	.01	.04	.01	.01
	$T \cdot D$	4	.06	.02	.11	.03	.10	.09	.02	**.21**	.04	.09
	$N \cdot T \cdot D$	16	.07	.01	.09	.09	.07	.06	.02	.03	.09	.06
	N	4	**.33**	**.23**	**.40**	**.29**	**.51**	**.31**	**.32**	**.31**	**.31**	**.45**
	T	2	**.30**	**.37**	.03	.04	**.17**	**.30**	**.27**	.01	.08	.13
	D	2	**.20**	**.26**	**.44**	**.45**	.01	**.19**	**.31**	**.43**	**.43**	.01
$CIMd$	$N \cdot T$	8	.01	.00	.01	.01	.02	.01	.00	.03	.01	.04
	$N \cdot D$	8	.01	.01	.00	.02	.02	.01	.00	.02	.01	.03
	$T \cdot D$	4	.03	.02	.02	.00	.02	.03	.02	.03	.02	.02
	$N \cdot T \cdot D$	16	.01	.01	.01	.02	.05	.02	.00	.04	.02	.07

Note. Msr = Measurement part parameters; L1Tr = Level-1 transition parameters; L2Tr = Level-2 transition parameters, IStP = Initial state parameters; CS = class size parameter. All effects were significant with $p < .001$. $Df_{2CT} = 18776$; $Df_{2DT} = 19241$. Dependent measures were log-transformed before analysis.

few measurement occasions ($T = 4$) and small sample sizes ($N < 100$). Problems occurred also for small sample sizes and few measurement occasions for the longest study period ($D = 10$). Because non-converged replications and replications with estimation problems were excluded from further analysis, the cell counts decreased dramatically for some conditions. The number of valid replications per condition can be found in Table 4.6. We decided to keep all conditions in the analysis, even if, for example, fewer than 100 of the 500 original replications are valid (which applies to six cells). In return, we will point out where caution in interpreting the results is advised based on reduced reliability of measures or the selection of well-behaved replications.

4.6.2 Information Criteria

The CT model would be preferred over the DT model in 97.6 % of the cases in which both model types were successfully applied to the data set (17,550 of 22,500 data sets). The remaining

proportion (DT model preferred over CT model) can mainly be found in conditions with small sample sizes ($N < 100$) or short study periods ($D = 3$).

4.6.3 Classification

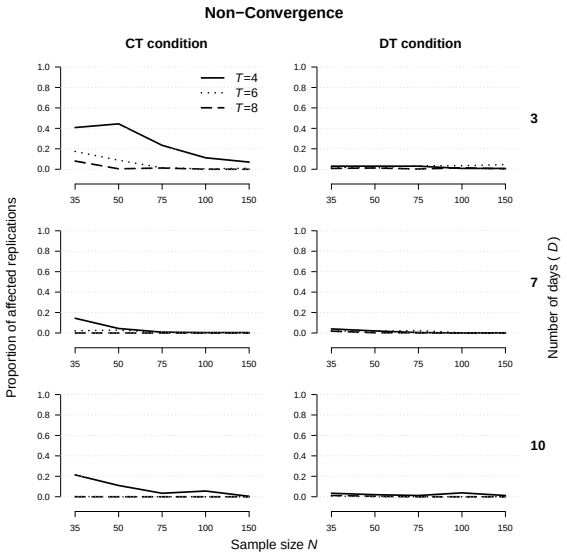
There was little variability in the mean classification probabilities for the latent state classes on Level-1 ($M = .94$, $Min = .91$, $Max = .96$; see Figure 4.3a). For $D = 3$, classification was best for $T = 6$. For the day-classes on Level-2, classification probabilities were slightly lower ($M = .89$, $Min = .84$, $Max = .93$), with more measurement occasions leading to higher classification probabilities (see Figure 4.3b).

4.6.4 Bias

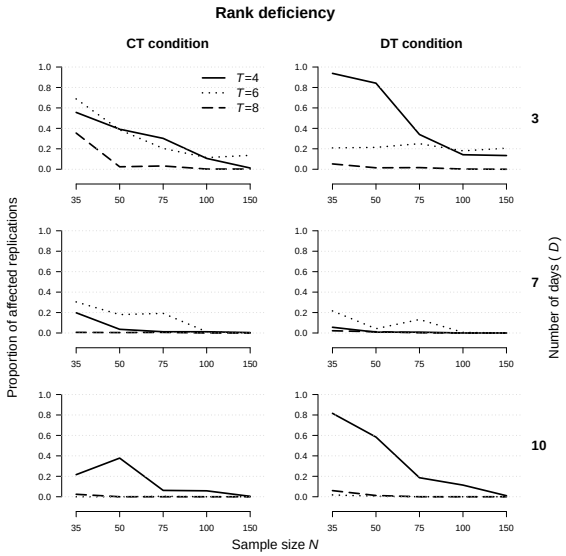
In this section, we will review the RMdSE and the CIMd by type of parameter. The corresponding Figures are 4.4 to 4.13. The ANOVA effect sizes reported can be found in Table 4.7. We start with the parameters on the lower level.

Measurement model parameters. Overall, the pattern for the $RMdSE$ appears similar for both model types. Values decrease with sample size and are particularly high for $T = 4$ measurement occasions. For $T = 6$ and $T = 8$, values decrease with length of study period. The largest effects for both model types are the main effects of N and T. The pattern for the $CIMd$ is also similar for both model types, with a clear order of the different values for T, a gradual decline with increasing D and a clear decrease for $N > 50$ in the longer study period conditions. Corresponding main effects are large.

Level-1 transition parameters. Again, the overall pattern of the $RMdSE$ is similar for both model types, except for lower values in short study period conditions ($D = 3$) for the CT model. The effect of T is clearly visible, especially for the $T = 4$ condition. The largest effects can be attributed to N and T, and in the case of DT models, also D (.14). For the $CIMd$, the effect of T is clearly visible (larger CI with fewer occasions), with $T = 4$ performing especially bad in short study length conditions ($D = 3$). The values for the CT model are higher in this condition. There is a considerable drop in the $CIMd$ values from $D = 3$ to $D = 7$, but not from $D = 7$ to $D = 10$. Main effects are very large for N, T, and also D.

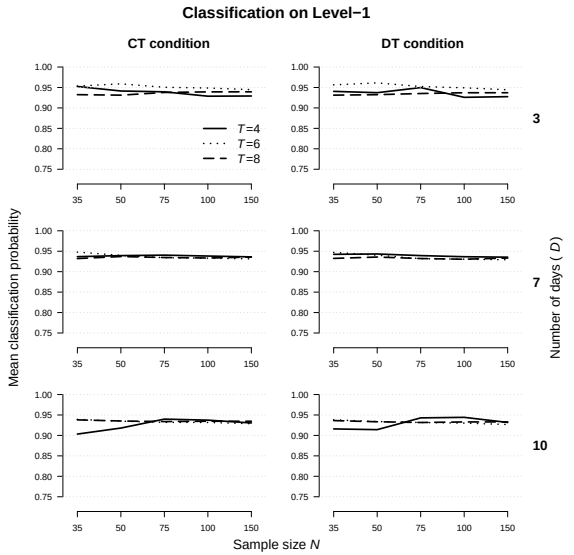


(a) Non-convergence

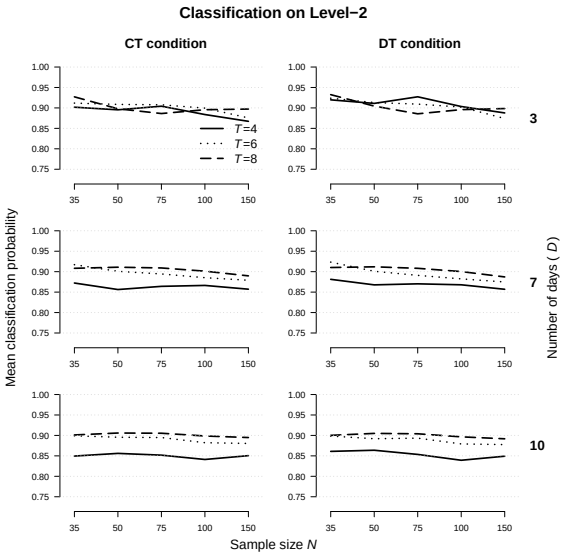


(b) Rank deficiency

Figure 4.2: Proportion of replications with estimation problems



(a) Within-day state class



(b) Between-day day class

Figure 4.3: Mean classification probabilities

Initial state parameters. The parameter estimation bias for these parameters is lower for the CT model. For both model types, values decrease with sample size. The effect of T is not consistent across the length of study period, which is reflected in a large effect for the $T * D$ interaction. The pattern of the $CIMd$ is similar: Lower values for the CT model and large main effects for N and D.

Level-2 transition parameters The bias values are much higher in these parameters than in the Level-1 transition parameters, with a clear effect of N, and D. The effect of T is smaller, and the order of T conditions only consistent for higher-information-conditions (starting with $D = 7$ and $N > 50$). Unusual is a peak in values for the condition with $D = 7$, $T = 6$, and $N = 75$, which appears for both model types. For the $CIMd$, there is a difference between model types for the short study period ($D = 3$) in combination with small samples ($N < 75$). Here, the CT model has larger standard errors. Magnitude decreases for both model types with N and especially D.

Class size parameter The pattern of the $RMdSE$ looks similar for both model types. However, several things are striking about the pattern: First, in the $D = 3$ condition, the value peaks for $T = 4$, $N = 75$ with lower numbers for smaller sample sizes. This is most likely due to the fact that the small number of replications that did converge without rank deficiency for these small sample sizes (see Table 4.6) is a positively selected sample. There were at least twice as many valid replications for $N = 75$, with a larger proportion of biased estimates in replications that did converge. For $D = 7$, the pattern is more in accordance with expectations, with highest values for $T = 4$ and small sample sizes. However, the value increase for $N = 150$. In combination with the $D = 3$ and $D = 7$ conditions, conditions with $T > 4$ show $RMdSE$ values no greater than .5. The most striking result appears for the combination of $D = 10$ and $T = 6$, with values > 1.5 for small sample sizes, falling below .5 only for $N = 150$. Closer inspection of the data revealed that this bias is a negative one, which means that the logit estimate for the class size is extremely negative, implying a very small second class. This very particular result, which does not appear for the $T = 4$ conditions with fewer occasions, remains to be discussed below. In contrast to the $RMdSE$, the pattern for the $CIMd$ is less striking: The model types are very similar. There is a clear effect of the sample size and for the number of occasions, but not for study period. It should be noted that the $D = 10$, $T = 6$ condition only shows slightly elevated values for small

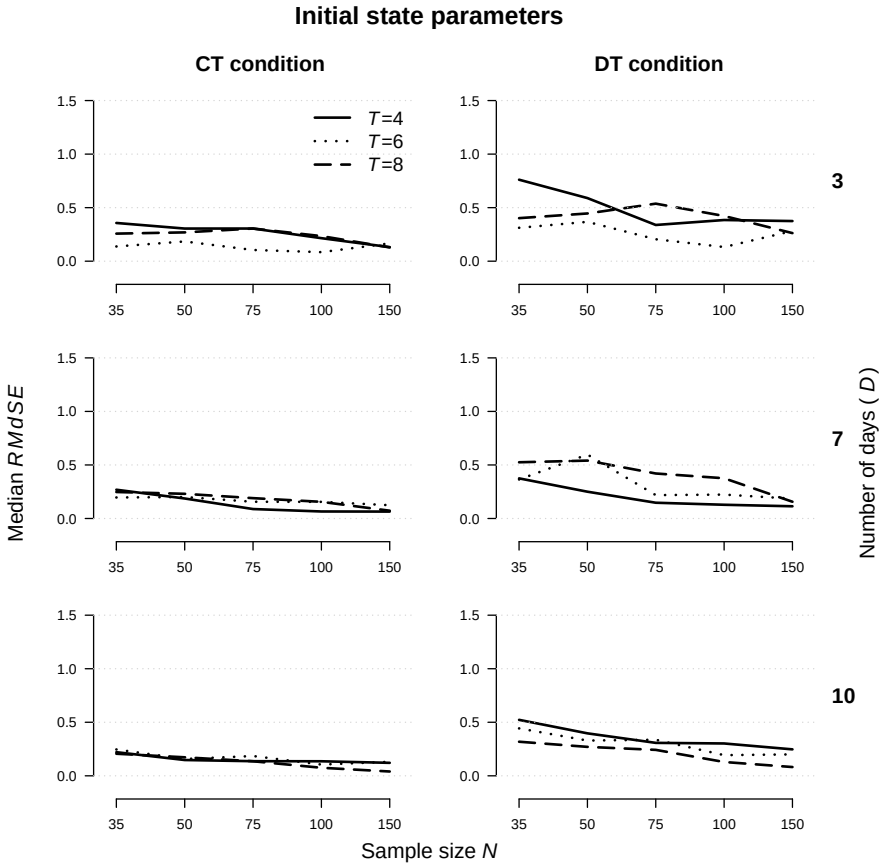


Figure 4.4: Parameter bias for initial state parameters

sample sizes.

4.6.5 Coverage

Tables 4.8 to 4.12 provided in the appendix contain the mean coverage for each condition by parameter type. In general, there are few division in the tables where the target range of $.95 + / - .03$ is consistently met. Apparently, the distributional assumptions are violated. In the following, we will point out some specific sections. For the measurement part parameters, values tend to be too low for $T = 4$ conditions. The Level-1 transition parameters show the largest proportion of an adequate range, with slightly larger values for the CT model. Initial state parameters have high values for $D < 10$ and low values for $D = 10$. For the Level-2 transition

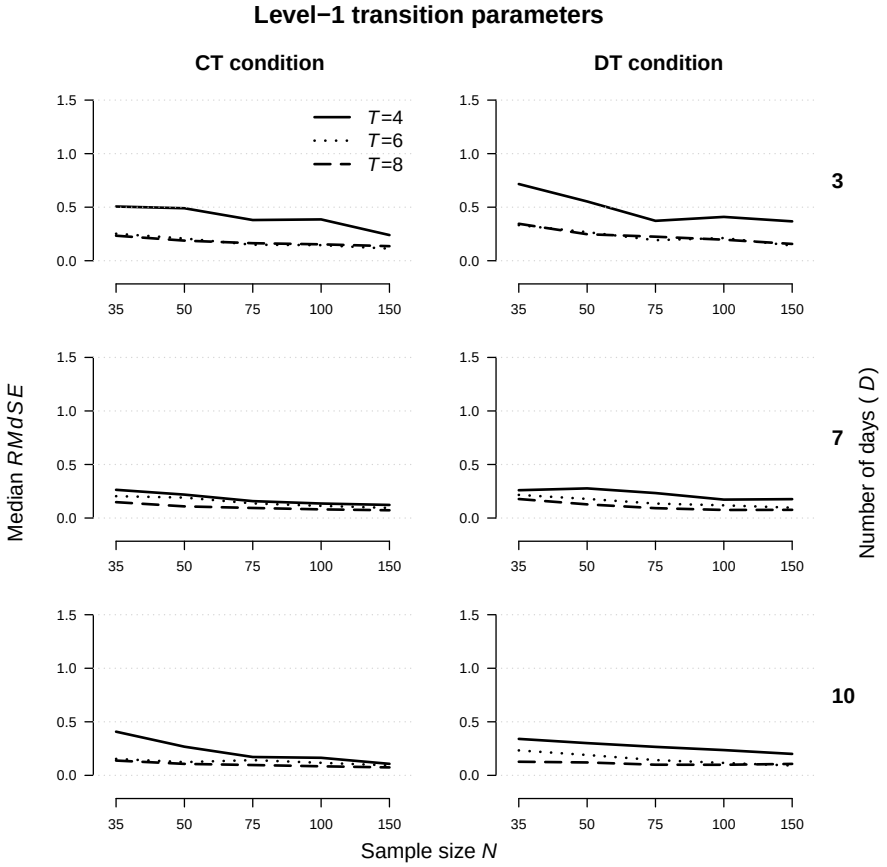


Figure 4.5: Parameter bias for Level-1 transition parameters

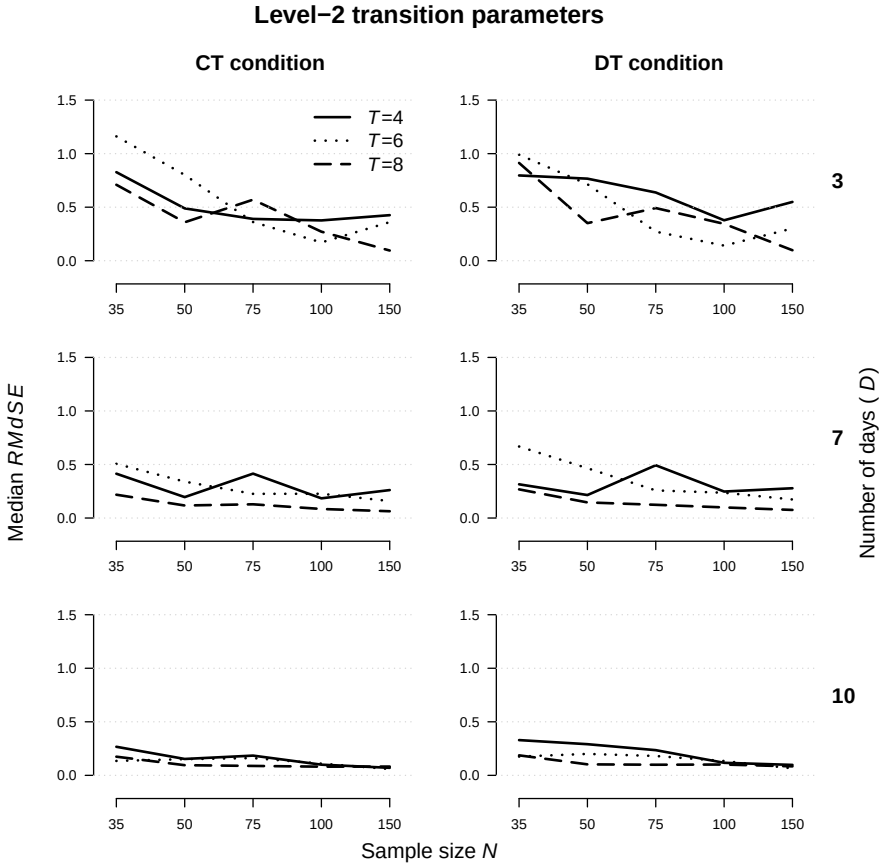


Figure 4.6: Parameter bias for Level-2 transition parameters

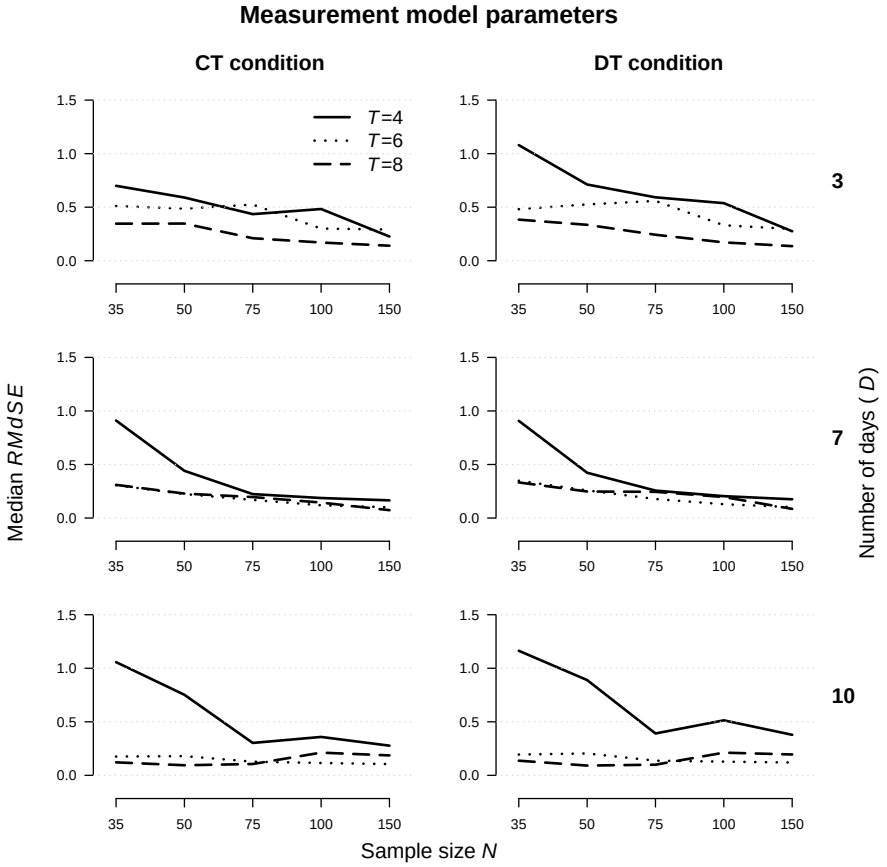


Figure 4.7: Parameter bias for measurement model parameters

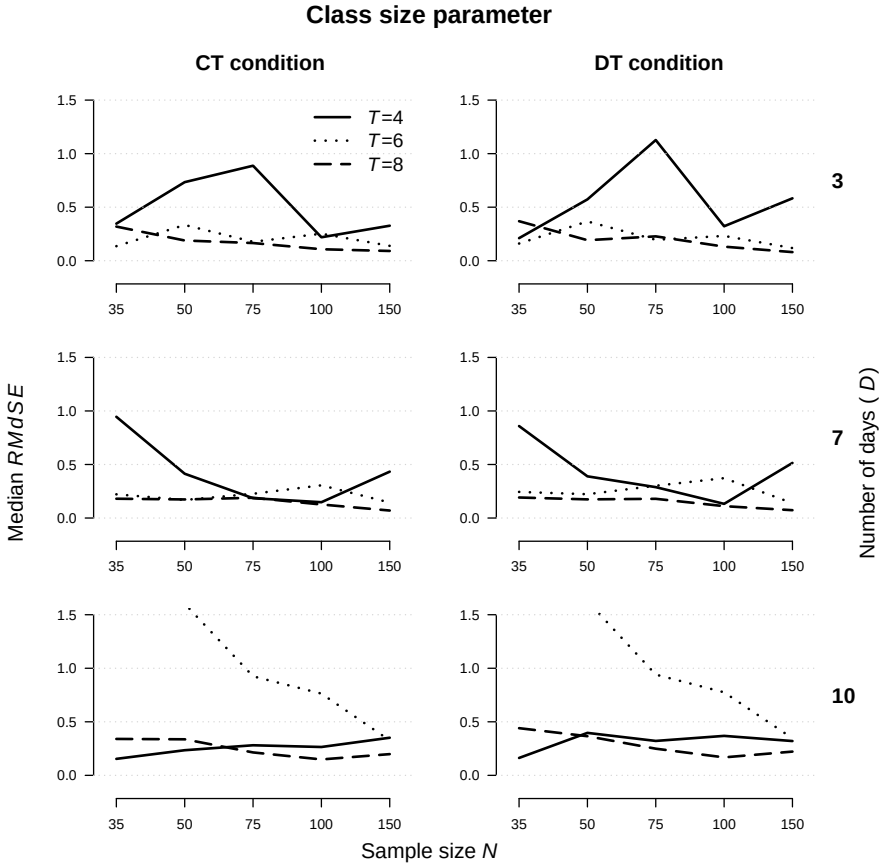


Figure 4.8: Parameter bias for class size parameter

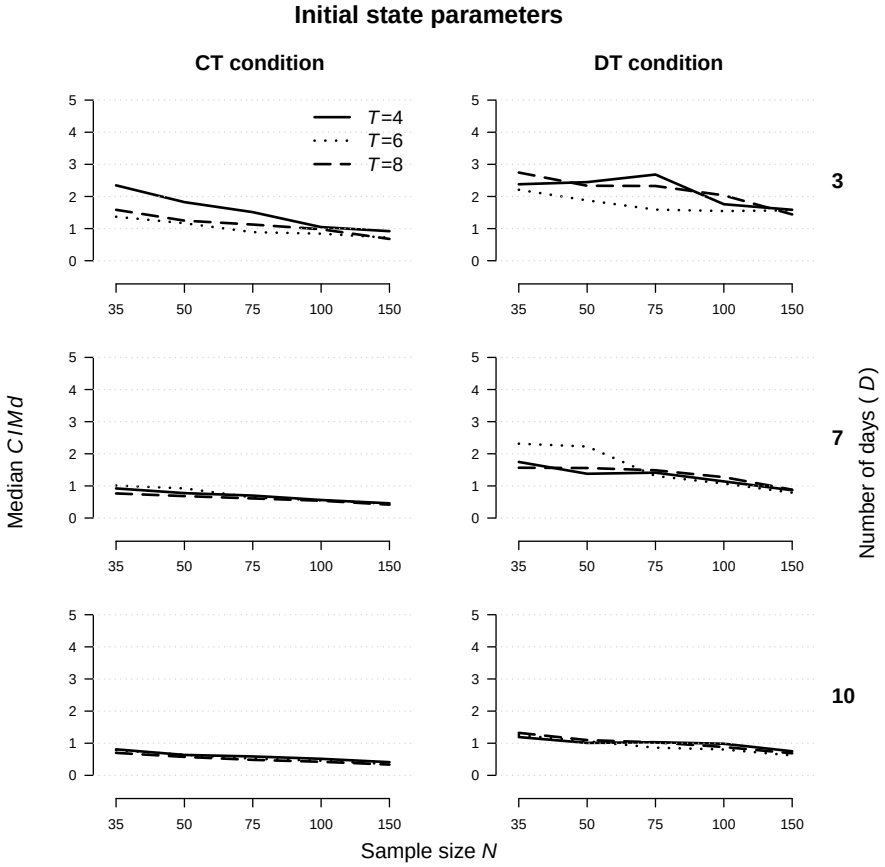


Figure 4.9: Standard error bias for initial state parameters

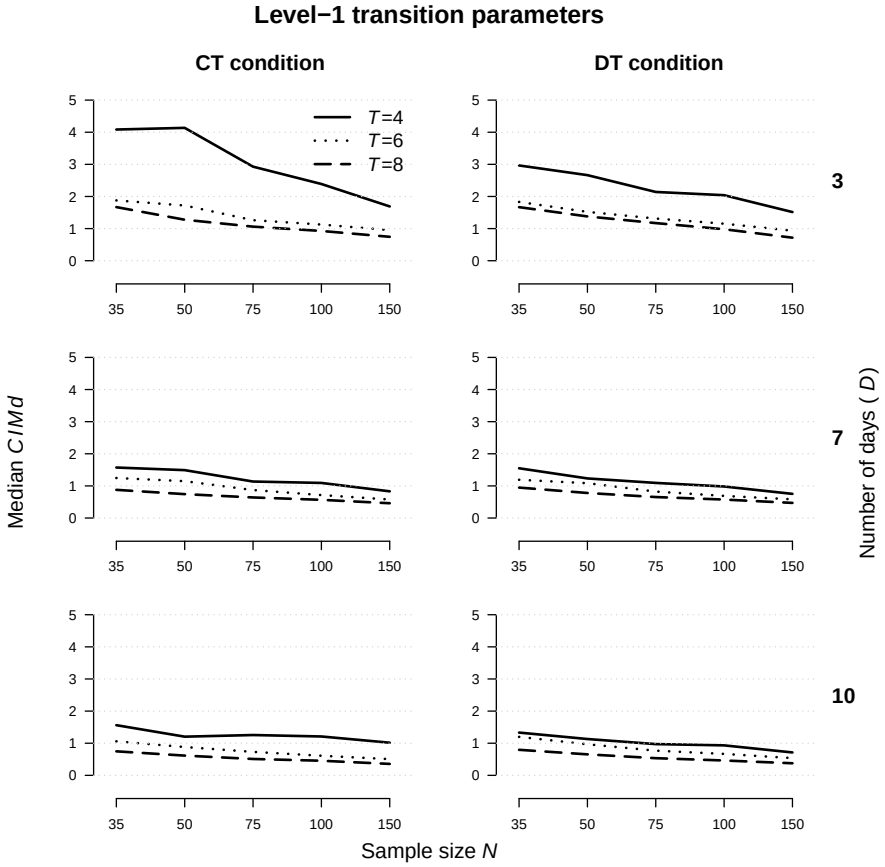


Figure 4.10: Standard error bias for Level-1 transition parameters

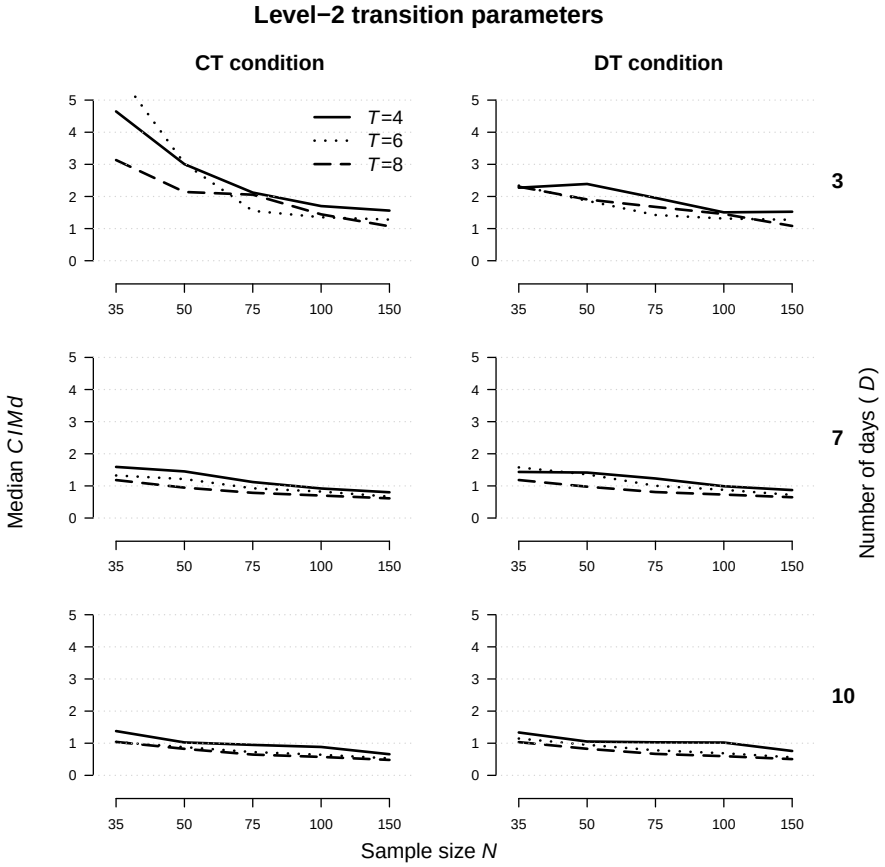


Figure 4.11: Standard error bias for Level-2 transition parameters

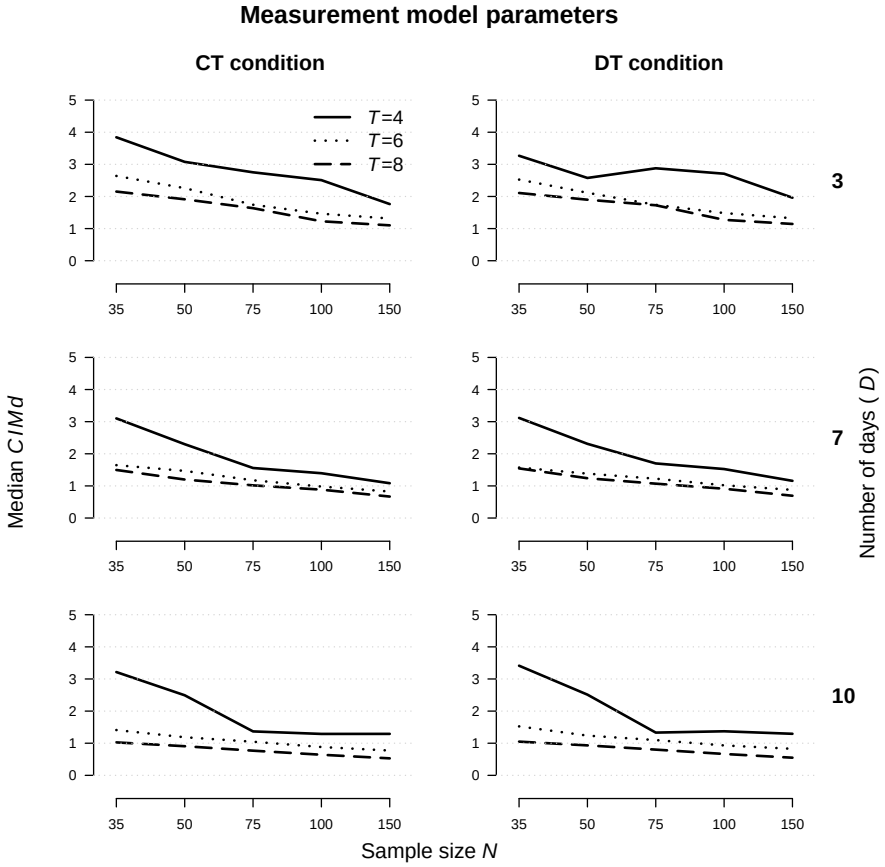


Figure 4.12: Standard error bias for measurement model parameters

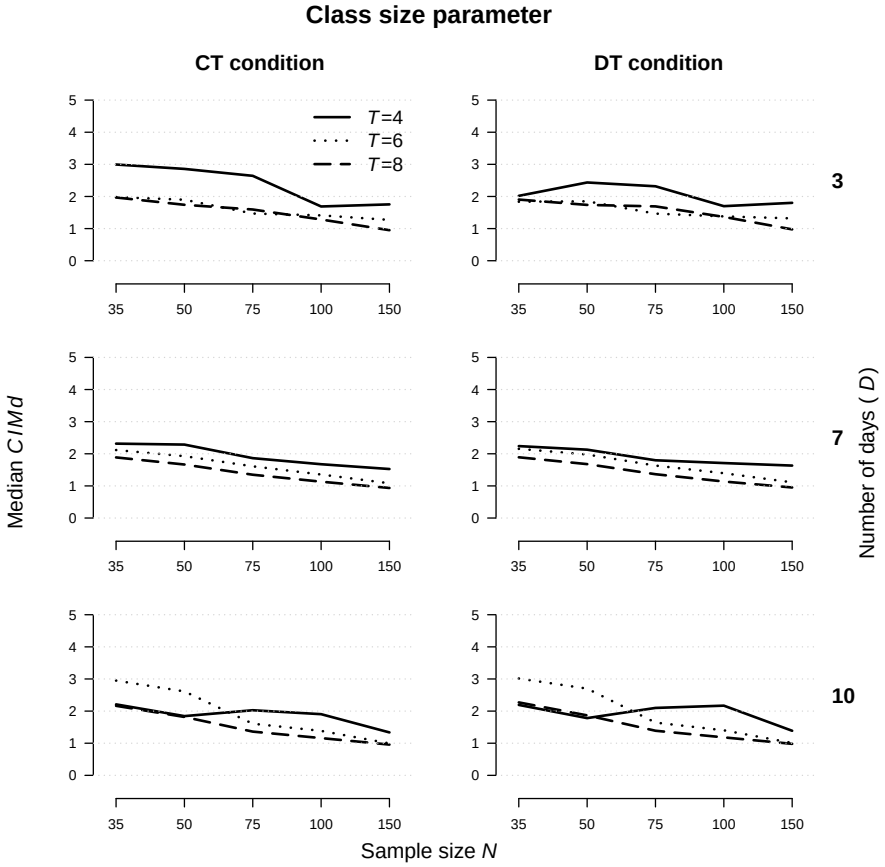


Figure 4.13: Standard error bias for class size parameter

parameters, values for the CT model are mostly too high. In the DT model, low information conditions have low coverage. The class size parameter exhibits mostly unity values, with the dips that were to be expected from the pattern of large $RMdSE$ in combination without enlarged $CIMd$: Very low coverage values (.02-.22) for the $D = 10$, $T = 6$ condition for $N < 150$, and low coverage values for $D = 3$, $T = 4$, $N = 75$ (.57 for DT, .90 for CT).

4.6.6 Summary of Results

The proportion of valid replications was about the same for both model types, with 84% of valid CT model replications compared to 86% for the DT model. The CT model was more strongly affected by non-convergence, while the DT model was more prone to rank-deficiency. The CT model is almost always preferred by IC measures. Classification probabilities are always good and a little higher on Level-1 than on Level-2. Concerning bias measures, there are not many differences in the estimation performance between the model types apart from low-information conditions with few valid replications. Primarily, the CT model performs better in the estimation of initial state parameters, with the overall mean $RMdSE$ for the CT model only about half the size (.153) of the DT one (.286). The $CIMd$ reflects the same, with smaller values for the CT model. There is also a small difference in the estimation accuracy of measurement model parameters and Level-1 transition parameters in favor of the CT model.

Sample size had large effects on estimation problems, as well as on estimation quality for both model and all parameter types, especially on the standard errors. The only exception is the class size parameter (see below). Small sample sizes ($N < 75$) in combination with few occasions proved unstable, especially for the shortest study period. Length of study period affected the estimation of standard errors for all parameter types (save the class size parameter). Length of study period also had the expected large effect on the estimation of Level-2 transition parameters. For the initial state parameters, the effect on the quality of parameter estimation was moderated by number of occasions. A low number of measurement occasions was associated with large standard error bias and low parameter estimation quality, especially in combination with small sample sizes and a short study period. As expected, the number of measurement occasions affects parameter types on the lower level more strongly (Level-1 transition parameters, measurement

model parameters). The class size parameter is also affected by few measurement occasions (for $D < 10$).

4.7 Discussion

There were surprisingly little differences between the model types in parameter recovery. However, the DT model is never superior in estimation and the interpretation of the transition probabilities that are recovered from the DT model is limited to the specific mean time interval in the data set. From a theoretical point of view, the CT model would always be preferred if the underlying process is thought to be continuous in nature and whenever time-varying intervals exist in the data. The convergence problems the CT model exhibited only affect the low empirical information conditions with the combination of few measurement occasions and small sample sizes, especially in combination with a short study period.

4.7.1 Class size parameter

The most striking result concerns the estimation of the class size parameter in the condition with a long study period and six measurement occasions. In contrast to fewer of more occasions, parameter recovery was near zero, reaching acceptable levels only for N=150. In combination with normally sized standard errors and classification probabilities, this looks suspiciously like an artefact. However, data and syntax files were automatically generated for all conditions, so errors should affect all conditions in the same way. Post-hoc analyses for this condition with 300 replications of a large sample ($N = 1000$) showed a similar pattern, with CT and DT models both very close to the population value for $T = 4$ ($RMdSE = 0.04$), a negative bias for $T = 6$ ($RMdSE = 0.20$), which was decreasing for $T = 8$ ($RMdSE = 0.17$). In addition, the class size parameter was the only one of its kind, the other four types had an additional level of aggregation. When looking at parameter recovery on the level of single parameters across conditions, there are some even more extreme examples (see the boxplots for $N = 100$ in the supplemental material).

4.7.2 Limitations of the study

In this context, the limitations of the study become obvious. Specifically, there were only two latent day classes. To be able to reliably identify more classes and class-specific parameters correctly, the sample size would need to be larger. Measurement invariance was quite strict, which could have led to a low separation between the classes and for conditions with many (equal) study days to linear dependencies within the simulated data sets. Just like the initial state probabilities are reset at the beginning of each day, one would probably include a reset of day-class sizes at the beginning of each week when analyzing longer study periods. With a large enough sample, one would eventually have enough power to detect, for example, day-of-the-week effects, which certainly exist (Golder & Macy, 2011).

The study was also limited to a basic version of the model. In theory, the model can be extended in many ways. For example, it is straightforward to carry the idea of day-classes of within-day fluctuation types to the day-level and include a mixture of between-day fluctuation. In case subjects were clustered in higher units, discrete or continuous random effects could be included, truly creating a hierarchical multilevel mixture latent Markov model. There could be more than one Markov process in the model, and there could be time-constant and time-varying covariates.

4.7.3 Conclusion

Böckenholt's (2005) call for an integration of the CT latent Markov model into models with different time scales and heterogenous populations has hardly received attention over the last ten years. In fact, Böckenholt and McShane (2014) just recently re-emphasized the need for the continuous-time approach to be integrated into future latent Markov developments.

We have provided an overview over the features of an continuous-time hierarchical mixture latent Markov model and explored data requirements for a stable estimation. The lower limit conditions did not work well. Given the bias and coverage results, we would discourage using the hierarchical mixture latent Markov models in these conditions. If the AA period is short, estimation stability can be achieved by a high number of (non-missing) within-day occasions. For better results, especially on the between-day level, a longer period, possibly a week, should

be covered. In terms of sample size, 50 are too small for a reliable recovery of measurement parameters, whereas a sample size of 100 produces stable results in most combinations. Monitoring 100 individuals over a week with 6-8 signals per day is a manageable study size, at least for non-clinical samples. We are confident that with the help of sophisticated assessment techniques and the matching methodology, much will be learned about interindividual differences in dynamic processes.

4.8 Acknowledgements

We gratefully acknowledge Martin Schultze's comments on earlier drafts of the results section of this manuscript.

4.9 References

Bartolucci, F., & Lupparelli, M. (2012). *Nested hidden Markov chains for modeling dynamic unobserved heterogeneity in multilevel longitudinal data.* Retrieved from `http://mpra.ub.uni-muenchen.de/40588/`.

Böckenholt, U. (2005). A latent Markov model for the analysis of longitudinal data collected in continuous time: States, durations, and transitions. *Psychological Methods*, 10, 65-83.

Böckenholt, U., & McShane, B. B. (2014). Comments on: Latent Markov models: a review of the general framework for the analysis of longitudinal data with covariates. *Test.* 23:469-472, doi: 10.1007/s11749-014-0388-0.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54, 579-616. doi: 10.1146/annurev.psych.54.101601.145030

Bujarski, S., Roche, D. J. O., Sheets, E. S., Krull, J. L., Guzman, I., & Ray L. A. (2015). Modeling naturalistic craving, withdrawal, and affect during early nicotine abstinence: A pilot ecological momentary assessment study. *Experimental and Clinical Psychopharmacology, 23* (2), 81-89. doi: 10.1037/a0038861

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum. doi: 10.4324/9780203771587

Coleman, J. S. (1981). *Longitudinal data analysis.* New York: Basic Books.

Costa, M., & De Angelis, L. (2010). *Model selection in latent Markov models: a simulation study.* Joint Meeting of the German Classification Society and the Classification and Data Analysis Group of the Italian Statistical Society, Florence, Italy, September 2010.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring dynamics in mood regulation-mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic Medicine, 74*, 366-376. doi: 10.1097/psy.0b013e31825474cb

Dias J. G. (2007). Model selection criteria for model-based clustering of categorical time series data: a Monte-Carlo study. In Decker, R., & Lenz H. J. (Eds.), *Advances in data analysis. Proceedings of the 30th annual conference of the Gesellschaft für Klassifikation.* (pp. 23-30). doi: 10.1007/978-3-540-70981-7_3

Eid, M., Courvoisier, D. S., & Lischetzke, T. (2012). Structural equation modeling of ambulatory assessment data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 384-406). New York, NY US: Guilford Press. doi: 10.5860/choice.50-1159

Finch, W. H., French, B. F. (2014). Multilevel latent class analysis: Parametric and nonparametric models. *Journal of experimental education*, 82, 307-333. doi: 10.1080/00220973.2013.813361

Fox, J., & Weisberg, S. (2011). An R Companion to applied regression, (2nd ed.). Thousand Oaks CA: Sage. `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215-231. doi: 10.1093/biomet/61.2.215

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125. doi: 10.1177/014662169602000201

Hedeker, D., Mermelstein, R. J., Berbaum, M. L., & Campbell, R. T. (2009). Modeling mood variation associated with smoking: an application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data. *Addiction, 104*, 297-307. doi: 10.1111/j.1360-0443.2008.02435.x

Huffziger , S., Ebner-Priemer, U., Zamoscik, V.,Reinhard, R., Kirsch, P., & Kuehner, C. (2013). Effects of mood and rumination on cortisol levels in daily life: An ambulatory assessment study in remitted depressed patients and healthy controls. *Psychoneuroendocrinology, 38*, 2258-2267. doi: 10.1016/j.psyneuen.2013.04.014

Houben, M., Van den Noortgate, W., & Kuppens, P. (2015). The relation between short term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*. doi: 10.1037/a0038822

Kalbfleisch, J. D. & Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association, 80*, 863-871.
doi: 10.1080/01621459.1985.10478195

Kogan, A., Mennin, D., Gruber, J., & Murray, G. (2013). Real-world emotion? An experience-sampling approach to emotion experience and regulation in bipolar I disorder. *Journal of Abnormal Psychology, 122* (4), 971-983. doi: 10.1037/a0034425

Lubke, G. H. & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39. doi: 10.1037/1082-989x.10.1.21

Lukociene, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology, 40*, 247-283. doi: 10.1111/j.1467-9531.2010.01231.x

Lukociene, O. & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In Fink, A., Lausen, B., Seidel, W., & Ultsch, A. (Eds.). *Advances in Data Analysis, Data Handling and Business Intelligence* (pp. 241-249). Springer: Berlin-Heidelberg. doi: 10.1007/978-3-642-01044-6$_2$2

Matthews, G., Jones, D. M., & Chamberlain, A. G. (1990). Refining the measurement of mood: The UWIST mood adjective checklist. *British Journal of Psychology, 81*, 17-42.
doi: 10.1111/j.2044-8295.1990.tb02343.x

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 4, 535-569. doi: 10.1080/10705510701575396

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*, 167-182. doi:10.1007/s11336-007-9001-8

Salovey P., Mayer J. D., Goldman S. L., Turvey, C., & Palfai T. P. (1995). Emotional attention, clarity, and repair: Exploring emotional intelligence using the trait meta-mood scale. In Pennebaker J. W. (Eds.). *Emotion, Disclosure, and Health* (pp. 125-154). Washington, DC: American Psychological Association. doi: 10.1037/10182-006

Shiffman S., Stone A. A. & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1-32. doi: 10.1146/annurev.clinpsy.3.022806.091415

Singer, B., & Spilerman, S. (1976). The representation of social processes by Markov models. *American Journal of Sociology, 82*, 1-54. doi: 10.1086/226269

Singer, J. & Willett, J. (2003). *Applied longitudinal data analysis. Modeling change and event occurrence.* Oxford, NY: Oxford University Press.

Steyer, R., Schwenkmezger, P., Notz, P. & Eid, M. (1994). Testtheoretische Analysen des Mehrdimensionalen Befindlichkeitsfragebogens (MDBF). *Diagnostica, 40*, 320-328.

Stone A. A. & Shiffman S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annual Behavioral Medicine, 16* (3), 199-202.

Van de Pol, F., Langeheine, R. (1990). Mixed Markov latent class models. *Sociological Methodology*, 213-247. doi: 10.2307/271087

Vermunt, J. K. (2010). Longitudinal research using mixture models. In K. van Montfort, J. H. L. Oud, and A. Satorra (eds.), *Longitudinal Research with Latent Variables*, 119-152. Heidelberg, Germany: Springer.

Vermunt, J. K. & Magidson, J. (2013). Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Belmont, Massachussetts: Statistical Innovations Inc.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In: S. Menard (Ed.), Handbook of Longitudinal Research: Design, Measurement, and Analysis (pp. 373-385). Burlington, MA: Elsevier.

Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychological Methods, 17*, 176-192. doi: 10.1037/a0027543

Walls, T. A., Jung, H., & Schwartz, J. (2006) Multilevel models and intensive longitudinal data. In Walls, T. A. & Schafer, J. S. (Eds.). *Models for Intensive Longitudinal Data* (pp. 3-37). New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195173444.003.0001

Yu, H.-T. (2007). *Multilevel latent Markov models for nested longitudinal discrete data.* (Doctoral thesis). Retrieved from `http://search.proquest.com/docview/304854730`

Yu, H.-T. & Park, J. (2014). Simultaneous decision on the number of latent clusters and classes for multilevel latent class models. *Multivariate Behavioral Research, 49*, 232-244.
doi: 10.1080/00273171.2014.900431

4.10 Appendix

4.10.1 R Code for matrix exponential

```
require(Matrix)
# Example values from Eq 4.13
> Q <- matrix(data=c(-.81, .75, .06,
+                    .17, -.32, .15,
+                    .06, .33, -.39),
+             byrow=TRUE, ncol=3, nrow=3)
# Transition prob. for one time unit interval
> deltat=1
> P1 = expm(Q*deltat)
> round(P1,3)
3 x 3 Matrix of class "dgeMatrix"
      [,1]  [,2]  [,3]
[1,] 0.481 0.450 0.069
[2,] 0.103 0.785 0.111
[3,] 0.051 0.252 0.697
# Transition prob. for two time units interval
> deltat=2
> P2 = expm(Q*deltat)
> round(P2,3)
3 x 3 Matrix of class "dgeMatrix"
      [,1]  [,2]  [,3]
[1,] 0.281 0.587 0.131
[2,] 0.137 0.691 0.172
[3,] 0.087 0.396 0.517
```

4.10.2 Coverage Tables

Note for all: Mean coverage by type of parameter. Values below .92 and above .98 appear in boldface. In addition, values below .92 appear in italics.

Table 4.8: Mean coverage for parameters of the measurement part of the model.

			Length of study period (D) and number of measurement occasions (T)								
			$D = 3$			$D = 7$			$D = 10$		
Type	Model	N	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$
Measurement part	CT	35	***.88***	.96	.96	***.84***	.93	.96	***.85***	**1**	**1**
		50	.92	***.88***	.94	***.89***	.98	.97	.94	.98	**.99**
		75	.92	***.82***	**.99**	.95	**1**	.94	***.88***	.98	**.99**
		100	.95	***.80***	**.99**	.93	**1**	.97	***.83***	.96	***.87***
		150	**.99**	***.82***	**.99**	***.91***	**1**	**1**	***.90***	**.99**	***.83***
	DT	35	***.79***	***.91***	.92	***.86***	***.89***	.95	***.86***	**.99**	**1**
		50	***.86***	***.84***	.96	***.88***	.97	.96	***.84***	.95	**1**
		75	***.88***	***.79***	.97	.93	**1**	.93	***.81***	96	**1**
		100	.93	***.78***	**.99**	.92	**1**	.95	***.65***	.93	***.88***
		150	**.99**	***.80***	**.99**	***.91***	**1**	**1**	***.83***	**.99**	***.84***

Table 4.9: Mean coverage for within-day transition parameters.

			Length of study period (D) and number of measurement occasions (T)								
			$D = 3$			$D = 7$			$D = 10$		
Type	Model	N	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$
Level-1 transition	CT	35	**.99**	**1**	.98	.95	.96	.98	.94	.98	.98
		50	.98	**1**	.98	.98	.98	**.99**	.97	.98	.97
		75	.98	**.99**	.97	.96	.98	**.99**	**.99**	.94	.93
		100	**.99**	**.99**	.96	**.99**	.97	**.99**	**1**	.93	.95
		150	**.99**	**1**	.95	.97	.98	.98	**1**	.96	.94
	DT	35	.94	.97	***.90***	.95	.93	.95	.93	.93	.96
		50	.94	**.99**	.94	.93	.95	**.99**	.94	.93	.96
		75	.95	**.99**	***.90***	***.84***	.95	.98	.93	.96	.95
		100	.95	.98	***.86***	.92	***.89***	**.99**	.93	.94	.94
		150	.94	**.99**	***.86***	***.87***	***.84***	.98	.94	.94	***.77***

Table 4.10: Mean coverage for initial state parameters.

			Length of study period (D) and number of measurement occasions (T)								
			$D = 3$			$D = 7$			$D = 10$		
Type	Model	N	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$
Initial state parameters	CT	35	.92	**1**	***.87***	.92	**1**	.93	.97	***.85***	***.91***
		50	.92	**1**	.94	.98	**.99**	***.87***	**.99**	***.91***	***.85***
		75	.97	**1**	***.90***	**.99**	**1**	.92	.98	***.83***	.92
		100	**.99**	**1**	.97	**1**	.98	***.88***	.98	.94	**.99**
		150	**.99**	.97	**1**	**1**	**1**	.96	***.86***	***.80***	**1**
	DT	35	***.80***	**.99**	***.91***	.96	**1**	.92	***.83***	***.81***	.97
		50	***.88***	**.99**	**.99**	.98	**1**	.93	***.90***	***.80***	***.88***
		75	.97	**1**	**.99**	.97	**1**	.97	.94	***.73***	**1**
		100	.98	**1**	**1**	**1**	**1**	.95	.93	.93	**1**
		150	.97	**1**	**1**	**.99**	**1**	.96	***.87***	***.80***	**1**

Table 4.11: Mean coverage for between-day transition parameters.

			Length of study period (D) and number of measurement occasions (T)								
			$D = 3$			$D = 7$			$D = 10$		
Type	Model	N	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$
Level-2 transition	CT	35	**1**	**1**	**1**	**.99**	***.78***	**1**	**1**	**.99**	**1**
		50	**1**	**1**	**1**	**1**	.93	**1**	**1**	**1**	**1**
		75	**.99**	**1**	**1**	***.81***	**.99**	**1**	**1**	**.99**	**1**
		100	.97	**1**	**.99**	**.99**	**.99**	**1**	**1**	**1**	**1**
		150	***.87***	.96	**1**	.95	**1**	**1**	**.99**	**1**	**1**
	DT	35	***.91***	***.71***	***.76***	.95	***.62***	.98	.95	**.99**	**.99**
		50	***.81***	***.80***	**1**	**.99**	***.68***	**1**	.93	**.99**	**1**
		75	***.77***	**1**	.94	***.63***	.96	**1**	**.99**	.95	**1**
		100	***.90***	**1**	.98	.95	.98	**1**	**1**	**1**	**1**
		150	***.65***	.97	**1**	.92	**1**	**1**	**1**	**1**	**1**

Table 4.12: Mean coverage for class size parameter.

			Length of study period (D) and number of measurement occasions (T)								
			$D = 3$			$D = 7$			$D = 10$		
Type	Model	N	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$	$T = 4$	$T = 6$	$T = 8$
Class size	CT	35	**1**	**1**	**1**	***.74***	**1**	**1**	**1**	***.04***	**1**
		50	**1**	**1**	**1**	.97	**1**	**1**	**1**	***.02***	**1**
		75	***.90***	**1**	**1**	**1**	**1**	**1**	**1**	***.12***	**1**
		100	**1**	**1**	**1**	**1**	**1**	**1**	**1**	***.22***	**1**
		150	**.99**	**.99**	**1**	**.99**	**1**	**1**	**1**	**.99**	**1**
	DT	35	**1**	**1**	**.99**	***.79***	**1**	**1**	**1**	***.09***	**1**
		50	.92	**1**	**1**	.98	**1**	**1**	.95	***.02***	**1**
		75	***.57***	**1**	**.99**	**1**	**1**	**1**	**1**	***.15***	**1**
		100	.98	**1**	**1**	**1**	**1**	**1**	**1**	***.20***	**1**
		150	***.88***	**.99**	**1**	.98	**1**	**1**	**1**	.98	**1**

Chapter 5

General discussion

In this final chapter, the three studies of this dissertation will be integrated. The chapter begins with a short summary of the main results of the three studies. Subsequently, implications will be discussed.

5.1 Summary of results

A common goal of the three studies at hand was to illustrate the stepwise modification and extension of complex multivariate models to fit the given data structure and arrive at results for substantial research questions. In study 1, this was demonstrated using a multimethod SEM approach to longitudinal categorical data. It served to test group differences on the latent mean level. Compared to repeated measures ANOVA, measurement invariance was established (not just assumed), and the (ordered) categorical nature of the measures was correctly integrated. In study two, a mixture latent Markov model was used to identify latent subtyps of mood fluctuation. Fifty-six measurement occasions were conveniently summarized by two set of class-specific transition parameters, with a larger class preferring a fairly pleasant mood state, and a smaller class exhibiting stronger transitions into a very pleasant mood state. The classification into fluctuation patterns could be supported by self-report measures. Also, in analyses reported elsewhere (Lischetzke, Eid, Crayen, & Arndt, 2015), time-varying covariates were included in the model. Results indicated a differential sensitivity of the subtypes to positive and negative

events. This adds to the notion that MLM are a powerful tool for accessing the richness of AA data. In the third study, MLM models were extended to incorporate parameters for time-continuous processes. In the simulation study conducted, very low information data conditions led to discouraging results, but a realistic AA sample size of 100 was reliably estimated. The continuous time model was just as stable as the discrete-time one, so there is no reason to not apply this model that is more in accordance with the data. It should be noted, however, that parameter recovery in the MLM models was much worse than what experience with continuous factor models shows (Crayen, 2008; Koch, 2013).

5.2 Implications

The models extended and applied in this work proved useful. That is not to say that they are the only option for this kind of data. The measurement of change in categorical variables has many facets. One fundamental distinction can be made between latent variable models that treat the categorical indicators as quasi-continuous, with the latent occasion-specific variable itself a continous factor, and models with categorical latent variables. In the first case, the categories of the manifest indicators are seen as coarse represantations of an underlying continuous distribution. Change is considered on the level of continuous factors. In latent class models on the other hand, change is represented by transitions between classes. Note that the choice of model and the process under consideration were not crossed in these studies. It was therefore *not* intended to evoke the impression that continuous latent variable models are better suited for the measurement of long-term change, while latent class models are better suited for the assessment of variability. In developmental psychology, for example, maturation is often not linear but characterized by sudden non-reversible transitions into higher ability states.

What was however intended was to illustrate how to modify the complexity of the models to match the complexity of the data and research question. With more powerful software available, the statistical toolbox has become a flexible modeling kit. The ease and freedom of modeling in turn claims responsability and sometimes insecurity. With original models, the literature may not provide all the answers and default options may not match the research question at hand.

5.3 References

Crayen, C. (2008). *A simulation study for investigating the applicability of structural equation models for multitrait-multimethod-multioccasion data.* (Unpublished Diploma thesis), Freie Universität Berlin, Germany.

Koch, T. (2013). *Multilevel structural equation modelling of multitrait-multimethod-multioccasion data.* (Doctoral dissertation), Freie Universität Berlin, Germany.

Lischetzke, T., Eid, M., Crayen, C., & Arndt, C. (2015). *Individual differences in mood regulation processes: Development of indirect measures of affective clarity and mood regulation competencies.* Final report (Abschlussbericht) - DFG Projekt LI 1827/1-2.

List of Tables

List of Figures

Chapter 6

Appendix (in German)

6.1 Zusammenfassung

Diese Dissertation hat zum Ziel, statistische Modelle der Veränderungsmessung für kategoriale Daten im Bereich der Sozial- und Verhaltenswissenschaften zu erweitern und für angewandte Wissenschaftler besser nutzbar zu machen. Veränderung wird auf der Ebene latenter Variablen betrachtet, die für Messfehler korrigiert sind. Das Konzept von Veränderung wird kurz zusammengefasst und es wird auf relevante Aspekte zur Wahl eines geeigneten statistischen Modells eingegangen. Ein wichtiger Aspekt ist der Zeitrahmen, auf den sich der Veränderungsprozess bezieht. Nesselroade (1991) hat die Unterscheidung zwischen langfristiger *Veränderung im engeren Sinne* und kurzfristigen *Fluktuationen* begründet. Beispiele für beide Arten der Veränderung werden hier untersucht. In der ersten Studie wird ein längsschnittliches Strukturgleichungsmodell für mehrere Erfassungsmethoden und Merkmale (Geiser, 2009) auf mehrere Gruppen und kategoriale Indikatoren erweitert. In einer Anwendung auf einen Datensatz mit Einschätzungen von Eltern und Erzieher für 659 Kindergartenkinder wird als Effekt eines Interventionsprogramms der Unterschied in der mittleren Veränderung geschätzt. In der zweiten Studie wird illustriert, wie interindividuelle Unterschiede in der intraindividuellen Stimmungs-Fluktuation durch die Anwendung latenter Markov-Mischverteilungsmodelle (Vermunt, Tran, Magidson, 2008) identifiziert werden können. Dazu dienen Daten einer Studie mit wiederholten Stimmungsmessungen

im Alltag ($N = 164$ Studenten mit bis zu 56 Messzeitpunkten). Das Modell wurde erweitert, um die Schachtelung der Messzeitpunkte in Tagen zu berücksichtigen (Vermunt, 2009). Zwei latente Klassen, die sich in Bezug auf ihr typisches Stimmungsmuster unterscheiden werden identifiziert und in Zusammenhang gesetzt mit Selbstberichten zu Stimmungsregulationskompetenzen. Anders als in der ersten Studie sind hier sowohl die manifesten Indikatoren, als auch die latenten Variablen kategorial. In der dritten Studie wurde das Modell aus der zweiten Studie erweitert, um variierende Zeitintervalle zwischen den einzelnen Messzeitpunkten zu berücksichtigen. Dafür werden zeitkontinuierliche Parameter eingeführt (Böckenholt, 2005). Eine Simulationsstudie zur Qualität der Parameterschätzung in kleinen Stichproben wird berichtet, und es wird ein Vergleich des zeitkontinuierlichen latenten Mischverteilungs-Markov Modells mit dem zeit-diskreten Modell vorgenommen. Vorteile und Beschränkungen der Modelle werden erörtert.

6.2 Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

6.3 Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, Juni 2015 Claudia Crayen