Latent Markov Factor Analysis for Exploring Measurement Model Changes in Time-Intensive Longitudinal Studies

Vogelsmeier, Leonie V.D.E., Vermunt, Jeroen K., van Roekel, Eeske, De Roover, Kim

Tilburg University, The Netherlands

Vogelsmeier, Leonie V.D.E
Department of Methodology and Statistics
PO Box 90153 5000 LE Tilburg
l.v.d.e.vogelsmeier@uvt.nl

Vermunt, Jeroen K
Department of Methodology and Statistics
PO Box 90153 5000 LE Tilburg
j.k.vermunt@tilburguniversity.edu

van Roekel, Eeske
Department of Developmental Psychology
PO 90153 5000 LE Tilburg
g.h.vanroekel@tilburguniversity.edu

De Roover, Kim
Department of Methodology and Statistics
PO Box 90153 5000 LE Tilburg
k.deroover@tilburguniversity.edu

The research leading to the results reported in this paper was sponsored by the Netherlands Organization for Scientific Research (NWO) [Research Talent grant 406.17.517; Veni grant 451.16.004]. Correspondence concerning this paper should be addressed to Leonie V. D. E. Vogelsmeier, Department of Methodology and Statistics, PO Box 90153 5000 LE Tilburg. E-mail: l.v.d.e.vogelsmeier@uvt.nl

Abstract

When time-intensive longitudinal data is used to study daily-life dynamics of psychological constructs (e.g., well-being) within persons over time (e.g., by means of experience sampling methodology), the measurement model (MM) – indicating which constructs are measured by which items – can be affected by time- or situation-specific artefacts (e.g., response styles, altered item interpretation). If not captured, these changes might lead to invalid inferences about the constructs. Existing methodology can only test for a priori hypotheses on MM changes, which are often absent or incomplete. Therefore, we present the exploratory method 'latent Markov factor analysis' (LMFA), wherein a latent Markov chain captures MM changes by clustering observations per subject into a few states. Specifically, each state gathers validly comparable observations and state-specific factor analyses reveal what the MMs look like. LMFA performs well in recovering parameters under a wide range of simulated conditions and its empirical value is illustrated with an example.

Keywords: experience sampling, measurement invariance, factor analysis, latent Markov modeling

1 Introduction

Time-intensive longitudinal data for studying daily-life dynamics of psychological constructs (such as well-being, positive affect) within persons allows to delve into time- or situation-specific effects (e.g., stress) on the (e.g., emotional) experiences of a large number of subjects (Larson & Csikszentmihalyi, 2014). The go-to research design to collect such data is Experience Sampling Methodology (ESM; Scollon, Kim-Prieto, & Diener, 2003). Participants repeatedly answer questionnaires at randomized or event-based time-points via smartphone apps, e.g., eight times a day over a few weeks.

While the technology for collecting ESM data is readily available, the methodology to validly analyze this data is lagging behind. This paper provides an upgrade of the methodology by presenting a novel method for tracking and diagnosing changes in measurement models (MMs) over time. The MM is the model underlying a participant's answers and indicates which unobservable or latent variables (i.e., psychological constructs) are measured by which items. Traditionally, it is evaluated by factor analysis (FA; Lawley & Maxwell, 1962), where the factors correspond – ideally – to the hypothesized constructs. Factor loadings express the degree to which each of the items measure a factor and thus how strongly an item relates to an underlying factor. In order to meaningfully compare constructs over time, the MM needs to be invariant across measurement occasions (Adolf, Schuurman, Borkenau, Borsboom, & Dolan, 2014). However, measurement invariance (MI) does not always hold over time because the MM likely changes over the course of an ESM study. First, in ESM, the measurement quality is undermined by time- or situation-specific artefacts such as response styles (RSs; Moors, 2003; Paulhus, 1991). Indeed, participants fill in their questionnaires repeatedly in various, possibly distracting, situations (e.g., during work) or lose motivation to repeatedly answer questions, which may drive the tendency to, for example, use the extreme response categories only (extreme RS; Moors, 2003; Morren, Gelissen, & Vermunt, 2011). Second, substantive

changes may occur over time in what questionnaire items are measuring. For example, depending on the context or mental state, an item may become more important for the measured construct (i.e., loading increases) or (also) an indicator of another construct (i.e., loads strongly on another factor) (reprioritization or reconceptualization; Oort, Visser, & Sprangers, 2005). Moreover, the nature of the measured constructs might change entirely; e.g., when positive and negative affect factors are replaced by high and low arousal factors (Feldman, 1995). In any case, when ignoring changes in the MM, changes in the scores will be interpreted as changes in the psychological constructs, although they are (partly) caused by RSs or changed item interpretation.

To safeguard validity of their time-intensive longitudinal studies, substantive researchers need an efficient approach to evaluate which MMs are underlying the data and for which time-points they apply, so they can gain insight into which artefacts and substantive changes are at play and when. Researchers can take these insights into account when analyzing the data, when setting up future projects or to derive new substantive findings from the MM changes. To meet this need, we present latent Markov factor analysis (LMFA)[1], which combines two building blocks to model MM changes within subjects over time: (1) *Latent Markov modeling* (LMM; Bartolucci, Farcomeni, & Pennoni, 2014; Collins & Lanza, 2010) clusters time-points into states according to the MMs and (2) factor analysis (FA; Lawley & Maxwell, 1962) evaluates which MM applies for each state. Note that LMFA can be applied for single cases, when enough observations are available for that one subject.

Within the states of LMFA, exploratory (EFA) rather than confirmatory (CFA) factor analysis is used. In CFA, users have to specify which items are measuring which factors based on a priori hypotheses. This implies that certain item-factor relations are assumed to be absent

[1] LMFA builds upon Mixture Simultaneous Factor Analysis (MSFA; De Roover, Vermunt, Timmerman, & Ceulemans, 2017), which captures differences in the factor model between groups. Whereas MSFA typically models the data of subjects nested within groups, LMFA specifically deals with observations nested within subjects and it allows subjects to switch between different MMs over time.

and the corresponding factor loadings are set to zero. Thus, for a large part, CFA already imposes a certain MM and thus limits the changes in the MM that can be found. In contrast, EFA estimates all factor loadings and thus, explores all kinds of (unknown) MM changes, including changes in cross-loadings (i.e., items loading on more than one factor) or even in the nature and number of factors (e.g., an additional RS factor). However, if desired, CFA can be used within the states.

An existing method to evaluate whether MI holds over time is longitudinal structural equation modeling (LSEM; Little, Preacher, Selig, & Card, 2007). However, this method merely tests whether MI across time-points holds for all individuals simultaneously, without directly providing insight in for which measurement occasions invariance is violated and what the alternative MMs look like. In contrast to LMFA, LSEM provides no clues for understanding or dealing with the non-invariance. Also, it applies CFA, and thus already assume a certain factor structure, and is thus too restrictive to detect many MM differences. A few methods exist that combine FA with LMM and thus could potentially be useful for identifying violations of MI over time[2] (Asparouhov, Hamaker, & Muthen, 2017; Song, Xia, & Zhu, 2017; Xia, Tang, & Gou, 2016). However, these methods also apply CFA, making them too restrictive to detect all kinds of MM differences. In contrast, factor-analyzed hidden Markov modeling (FAHMM; Rosti & Gales, 2002) is similar to LMFA because it combines EFA with LMM, but was developed merely for accommodating LMM estimation when conditional independence is violated among many variables, using the state-specific FA to reduce the number of parameters of the state-specific covariance matrices rather than being the point of interest (Kang & Thakor,

[2] Note that the overview of existing methods focusses on FA-based methods and thus overlooks switching principal component analysis (SPCA; De Roover, Timmerman, Van Diest, Onghena, & Ceulemans, 2014), which is a deterministic method similar to LMFA that can be used to take the first steps towards detecting MM changes over time, yet only for single-subject data. However, SPCA uses component instead of factor analysis. Although components and factors are similar (Ogasawara, 2000; Velicer & Jackson, 1990; Velicer, Peacock, & Jackson, 1982), components don't correspond to latent variables (Borsboom, Mellenbergh, & van Heerden, 2003) and are thus less ideal for evaluating (changes in) MMs.

2012; Rosti & Gales, 2002). Also, FAHMM cannot analyze multiple subjects simultaneously. Thus, LMFA may be conceived as a multisubject extension of FAHMM, tailored to tackle measurement non-invariance in ESM data.

The remainder of this paper is organized as follows: Section 2 describes the multisubject longitudinal data structure, an empirical example, and the LMFA model specifications and estimation. Section 3 presents a simulation study, evaluating the goodness-of-recovery of states and state-specific MMs under several conditions as well as model selection. Section 4 illustrates LMFA with an application. Section 5 concludes with some points of discussion and directions for future research.

2 Methods

2.1 Data Structure and Motivating Example

Like in ESM, we assume repeated measures data where observations are nested in subjects. For each measurement occasion, data on multiple continuous variables are available. The observed scores are indicated by y_{ijt}, where $i = 1, \dots, I$ refers to subjects, $j = 1, \dots, J$ to items, and $t = 1, \dots, T$ to time-points, where the latter may differ across subjects (i.e., T_i) but we mostly omit the index i for simplicity of notation. The $J \times 1$ vector $\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \dots, y_{iJt})$ contains the multivariate responses for subject i at time-point t and the $T \times J$ dataset $\mathbf{Y}_i = (\mathbf{y}_{i1}', \mathbf{y}_{i2}', \dots, \mathbf{y}_{iT}')'$ contains data for subject i for all time-points T.

To clarify the data structure and illustrate the problem of measurement non-invariance, consider the ESM data of the 'No Fun No Glory' study described in more detail by Van Roekel et al. (2017). In brief, the data contained repeated emotion measures of 69 young adults with persistent anhedonia, which is the diminished pleasure in response to previously enjoyable experiences and one of the core symptoms of depression (American Psychiatric Association, 2013; Treadway & Zald, 2011). Over a course of about three months, every evening, the participants rated on a Visual Analogue Scale, ranging from 0 ("Not at all") to 100 ("Very

much"), how much they had felt each of 18 emotions (listed in Table 3, which is further described in Section 4) in the past 6 hours[3]. The number of repeated measures ranged from 86 to 132 (M = 106.86, SD = 8.21) and resulted in 7373 total observations of which 557 were missing[4]. After the first month, the participants randomly received (1) no intervention (n = 22), (2) a personalized lifestyle advice (PLA) (n = 23), or (3) a PLA and tandem skydive (PLA & SkyD) (n = 24) to potentially reduce anhedonia. After the second month, all participants chose one of the interventions, regardless of their first one (no: n = 3; PLA: n = 17; PLA & SkyD: n = 49). In their original study, Van Roekel et al. (2017) investigated whether the interventions decreased anhedonia, thereby assuming the two underlying factors 'positive affect' ('PA') and negative affect ('NA'). However, if the MM changes over the course of participation (e.g., due to the interventions) conclusions about changes in PA and NA may be invalid. In Section 4, LMFA is used to trace potential MM changes in this data.

2.2 Latent Markov Factor Analysis

In this section, we introduce LMM (2.2.1) before describing LMFA in more detail (2.2.2).

2.2.1 Latent Markov modeling

The LMM (also a hidden Markov or latent transition model; Bartolucci et al., 2014; Collins & Lanza, 2010) captures unobserved heterogeneity or changes over time by means of latent states. In contrast to standard latent class models (Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968), which identify subgroups or so-called latent classes within a population (e.g., high or low risk for depression), a LMM allows respondents to transition between latent states over time and thus to switch between subgroups (e.g., from a high risk to a low risk subgroup). Thus, the states may be conceived as dynamic latent classes. Specifically, the LMM is a probabilistic model where the probability of being in a certain state at time-point t depends

[3] In total, participants rated their emotions three times a day with fixed 6-hour intervals. In the morning and midday, they rated their 'momentary' emotions and in the evening, they rated their emotions 'since the last measure'. To have comparable and evenly spaced measures, we focused on the evening measures.
[4] Missing data will be directly handled in the model estimation by considering only observed values.

only on the state of the previous time-point $t-1$ (*first-order Markov assumption*). Furthermore, the responses at time-point *t* depend only on the state at time-point *t* (*local independence assumption*; Bartolucci, 2006; Vermunt, Langeheine, & Böckenholt, 1999). The joint probability of observations and states for subject *i* is then:

$$p(\mathbf{Y}_i, \mathbf{S}_i) = \overbrace{p(\mathbf{s}_{i1})}^{\text{initial state probabilities}} \prod_{t=2}^{T} \overbrace{p(\mathbf{s}_{it}|\mathbf{s}_{it-1})}^{\text{transition probabilities}} \prod_{t=1}^{T} \overbrace{p\,(\mathbf{y}_{it}|\mathbf{s}_{it})}^{\text{response probabilities}}, \tag{1}$$

where $\mathbf{s}_{it}$ are $K \times 1$ binary variables indicating whether an observation belongs to a state or not and $\mathbf{S}_i = (\mathbf{s}_{i1}, \mathbf{s}_{i2}, \dots, \mathbf{s}_{iT})$ is the subject-specific state membership matrix. In the following, $k = 1, \dots, K$ refers to the states and, if $s_{itk} = 1$, subject *i* is in state *k* at time-point *t*. Equation 1 includes three types of parameters: (1) The *initial state probabilities* indicate the probabilities to start in a certain state, $p(s_{i1k} = 1)$, and thus how the subjects are distributed across the states at $t = 1$. They are often denoted as π_k, with $\sum_{k=1}^{K} \pi_k = 1$, and are gathered in a $K \times 1$ vector $\boldsymbol{\pi}$. (2) The *transition probabilities* indicate the probabilities of being in a certain state at time-point t conditional on the state at $t-1$, $p(s_{itk}|s_{it-1,l})$, where $l = 1, \dots, K$. These may be denoted as a_{lk}, with $\sum_{k=1}^{K} a_{lk} = 1$, and are collected in a $K \times K$ transition probability matrix $\mathbf{A}$. The transition probabilities are often assumed to be *homogeneous* (i.e., invariant) across time (and subjects). The resulting sequence of states is called a *latent Markov chain* (LMC). (3) The *response probabilities* indicate the probability of a certain item response given the state at time-point *t*, $p(\mathbf{y}_{it}|\mathbf{s}_{it})$, which correspond to a the multivariate normal density for continuous responses.

2.2.2 Latent Markov factor analysis

In LMFA, a LMM is used to capture the changes in MMs over time and FA (Lawley & Maxwell, 1962) is applied per state to model the state-specific MMs. The latter is given by:

$$\mathbf{y}_{it} = \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k\, \mathbf{f}_{it} + \mathbf{e}_{it}, \tag{2}$$

where $\mathbf{\Lambda}_k$ is a state-specific $J \times F^k$ loading matrix; $\mathbf{f}_{it}$ is a subject-specific $F^k \times 1$ vector of factor scores at time-point t (where F^k is the state-specific number of factors); $\boldsymbol{\nu}_\mathrm{k}$ is a state-specific $J \times 1$ intercept vector; and $\mathbf{e}_{it}$ is a subject-specific $J \times 1$ vector of residuals at time-point t. The distributional assumptions are: $\mathbf{f}_{it} \sim MVN(0; \mathbf{\Psi}_k)$ and factor scores are thus centered around zero and $\mathbf{e}_{it} \sim MVN(0; \mathbf{D}_k)$, where $\mathbf{D}_k$ contains the unique variances d_{kj} on the diagonal and zeros on the off-diagonal. To partially identify the model, factor variances in $\mathbf{\Psi}_k$ are restricted to one and the remaining rotational freedom is dealt with by means of criteria to optimize the simple structure or between-state agreement of the factor loadings, such as Varimax (Kaiser, 1958), oblimin (Clarkson & Jennrich, 1988) or generalized Procrustes (Kiers, 1997).

From Equation 2, it is clear that the states may differ in terms of their intercepts $\boldsymbol{\nu}_k$, loadings $\mathbf{\Lambda}_k$, unique variances $\mathbf{D}_k$, and/or factor covariances $\mathbf{\Psi}_k$. This implies that LMFA allows to explore all levels of measurement non-invariance at once. This is: (1) *Configural invariance* (invariant number of factors and pattern of zero loadings), (2) *Weak factorial invariance* (invariant non-zero factor loadings), (3) *Strong factorial invariance* (invariant item intercepts), and (4) *Strict factorial invariance* (invariant unique variances). Conveniently, in any case, the strictest level of invariance applies within each state (for more detail see Little et al., 2007; Meredith, 1993; Meredith & Teresi, 2006; Schaie, Maitland, Willis, & Intrieri, 1998). Figure 1 illustrates how LMFA captures the different levels of non-invariance over time based on an example of what might happen in the empirical data by comparing the state-1 MM respectively to the state-2 and state-3 MMs, with dashed lines representing parameter changes.

[Insert Figure 1 about here]

The depicted loadings can be thought of as standardized rotated loadings higher than, e.g., .4 in absolute value (Stevens, 1992). We start by comparing the state-1 MM to the state-2 MM. Here, configural invariance is violated because a third factor (high arousal; HA) appears,

implying that the state-1 items measuring either PA or NA with loadings λ_{141}, λ_{151}, and λ_{162} measure another construct (i.e., HA) in state 2 (now with loadings λ_{242}, λ_{252}, and λ_{262}). This also changes the meaning of the other factors into low arousal PA (LA-PA) and low arousal NA (LA-NA)). Next, we compare the state-1 MM with the state-3 MM. Firstly, weak factorial invariance is violated here because λ_{111} differs from λ_{311} and thus, the items measure PA and NA differently. Secondly, strong factorial invariance is violated because ν_{12} differs from ν_{22}. Note that, when weak invariance appears to hold, properly assessing strong invariance would require re-estimating the model with invariant factor loadings across the states and non-zero state-specific factor means. Finally, strict factorial invariance is violated because e_{11} differs from e_{31}. Usually, strong factorial invariance is said to be sufficient for comparing latent constructs over time, i.e., differences in factor means then correspond to actual changes in the latent variables.

It is important to note that the subjects don't have to go through all the states nor do they have to go through the states in the same order. Relatedly, LMFA does not assume homogeneous transition probabilities across subjects but allows for subject-specific $\mathbf{A}_i$ matrices, implying that some transition probabilities may be zero for a certain subject if that subject does not go through a particular state. This is because subjects likely differ in how stable they respond to questionnaires (e.g., some people might switch more between contexts than others or may be more sensitive to contextual influence or distractions). The transition process $\mathbf{A}_i$ is assumed to be time-homogeneous for each subject, although this is an assumption that might be relaxed in the future.

To conclude, in LMFA, the states indicate for which time-points the data are validly comparable (strict MI applies within each state) and by comparing the state-specific MM parameters one may even evaluate which level of invariance holds for which (pairs of) states and which specific MM parameters are non-invariant.

2.3 Model estimation

To estimate the LMFA model we aim to find the model parameters $\boldsymbol{\theta}$ (i.e., the initial state probabilities $\boldsymbol{\pi}$, the transition probabilities $\mathbf{A}_i$, the intercepts $\boldsymbol{\nu}_k$, and the factor-analyzed covariance matrices $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k + \mathbf{D}_k$) that maximize the loglikelihood function $\log L$. The $\log L$ is derived from Equation 1 by summing over all possible state sequences, taking the logarithm, and considering all the subjects at once:

$$\log L(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{i=1}^{I} \log\left(\sum_{\mathbf{s}_{i1}} \dots \sum_{\mathbf{s}_{iT}} p(\mathbf{s}_{i1}) \prod_{t=2}^{T} p(\mathbf{s}_{it}|\mathbf{s}_{it-1}) \prod_{t=1}^{T} p\,(\mathbf{y}_{it}|\mathbf{s}_{it})\right). \tag{3}$$

Note that the model captures the dependencies only between observations that can be explained by the states but not the autocorrelations of factors within the states. Because the $\log L$ is complicated by the latent states, non-linear optimization algorithms are necessary to find the maximum likelihood (ML) solution (e.g., De Roover et al., 2017; Myung, 2003). LMFA can be estimated by means of LG syntax[5] (Vermunt & Magidson, 2016; Appendix B). Specifically, the ML estimation is performed by an Expectation Maximization (EM; Dempster, Laird, & Rubin, 1977) procedure described in Appendix A. Note that this procedure assumes the number of states K and factors within the states F^k to be known. The most appropriate K and F^k is determined by comparing competing models in terms of their fit-complexity balance. To this end, the Bayesian information criterion (BIC) can be applied, which proved to be effective for both FA (Lopes & West, 2004) and LMM (Bartolucci, Farcomeni, & Pennoni, 2015). Moreover, it may happen that the estimation converges to a local instead of a global maximum. To decrease the probability of finding a local maximum, LG applies a multistart procedure, in which the initial values are automatically chosen based on the loadings and residual variances obtained from PCA (Jolliffe, 1986) on the entire dataset. For each state, randomness is added to get K different sets of initial parameter values (for more details see De Roover et al., 2017).

[5] A user-friendly graphical interface in LG including a tutorial will be developed in the future.

3 Simulation Study

3.1 Problem

To evaluate how well LMFA performs in recovering states and state-specific factor models, we manipulated seven factors that affect state-separation and thus potentially the recovery: (a) number of factors, (b) number of states, (c) between-state difference (consisting of differences in factor loadings and intercepts), (d) unique variance, (e) frequency of transitions, (f) number of subjects, and (g) number of observations per subject and state. For the number of factors (a), we expect the performance to be lower for more factors due to the higher model complexity and the lower level of factor overdetermination (given a fixed number of variables) (MacCallum, Widaman, Preacher, & Hong, 2001; MacCallum, Widaman, Zhang, & Hong, 1999; Preacher & MacCallum, 2002). With respect to the number of states (b), a higher number of states also increases the model complexity and thus, probably, decreases the performance. In case of a Markov model, the increase in model complexity with additional states is suppressed by the level of dependency of the states at consecutive time-points, however. Thus, with respect to (e), we anticipate LMFA to performance worse in case of more frequent state transitions, and thus lower probabilities of staying in a state, because this implies a lower dependence on the state of the previous time-point (Carvalho & Lopes, 2007). With respect to (c), we expect a decrease in performance for more similar factor loadings (De Roover et al., 2017) and/or intercepts across states. Regarding (d), LMFA is expected to perform better with a lower unique variance and thus a higher common variance because this increases the factor overdetermination (Briggs & MacCallum, 2003; Ximenez, 2009; Ximénez, 2006). Factors (f) and (g) pertain to the within-state sample size (i.e., the amount of information) per state in terms of number of subjects and observations per subject and state. We expect a higher performance with increasing information (de Winter, Dodou, & Wieringa, 2009; Steinley & Brusco, 2011). Note that we also tested whether lag-one autocorrelations of factors harm the

performance of LMFA, which was not the case (Appendix C). In addition, for selected conditions, we evaluated the BIC in terms of the frequency of correct model selection.

3.2 Design and Procedure

We crossed seven factors in a complete factorial design with the following conditions[6]:

a. number of factors per state F^k at two levels: 2*, 4*;
b. number of states K at three levels: 2*, 3, 4*;
c. between-state differences at eight levels:
   medium loading difference & no intercept difference,
   low loading difference & no intercept difference,
   medium loading difference & low intercept difference*,
   low loading difference & low intercept difference*,
   no loading difference & low intercept difference,
   medium loading difference & medium intercept difference*,
   low loading difference & medium intercept difference*,
   no loading difference & medium intercept difference;
d. unique variance e at two levels: .2 and .4*;
e. frequency of state transitions at three levels: highly frequent, frequent, infrequent*;
f. number of subjects N at three levels: 2, 5*, 10;
g. number of observations per subject and state T_{ik} at three levels: 50, 100*, 200.

The number of variables J was fixed to 20. The numbers of factors per state F^k was either 2 or 4 (a) and was the same across the states. The two, three or four states (b) differed in factor loadings and intercepts. The degree of the between-state loading difference (c) was either medium, low (i.e., highly similar loadings), or non-existent (i.e., identical loadings across states). Between the state-specific intercepts, there was either no difference, a medium difference or a high difference. The combination of no loading difference and no intercept difference was omitted because this implies no difference in MMs and thus only one state. Note that the degree of the between-state differences was the same for each pair of states.

Regarding the factor loadings $\mathbf{\Lambda}_k$ of the generating model, for all conditions, a binary simple structure matrix was used as a common 'base' (see Table 1). The loading matrices were representative for the ones commonly found in psychological research (cf., the PA and NA structure assumed by the original researchers of the 'no fun no glory study'). In these matrices,

[6] The "*" marks the subset of conditions that is included in the evaluation of model selection.

all variables loaded on one factor only and the variables were equally divided over the factors. In case of two factors, this implied that each factor had ten non-zero loadings, whereas, in case of four factors, each factor consisted of five non-zero loadings. For the 'no loading difference' conditions, the simple structure base matrix was used as $\mathbf{\Lambda}_k$ in all the states, implying no change in loadings across the states. For the low and medium loading difference conditions, the base matrix was altered differently for each state to create the state-specific loading matrices. Thus, no state will have a factor loading structure equal to the base matrix in Table 1. For each state, regardless of the number of factors, we applied the alteration procedure described below.

[Insert Table 1 about here]

Whether $F^k = 2$ or $F^k = 4$, the manipulations were only applied to the first two factors. Thus, for $F^k = 4$, the third and fourth factor are identical across states. For the medium loading difference conditions, the state-specific loading matrices were created by shifting one loading from the first factor to the second one and one loading from the second factor to the first one. This implies that the overdetermination of the factors is unaffected. For the low loading difference condition, the state-specific loading matrices were created by adding cross-loadings of $\sqrt{.5}$ for two variables, i.e., one for factor 1 and one for factor 2, and lowering the primary loading accordingly to $\sqrt{.5}$. This manipulation preserves both the rowwise and columnwise sum of squares (i.e., the variables' common variance and the variance accounted for by each factor). Variables affected by the loading shifts and added cross-loadings differed across states (see Table 1)[7].

To quantify the similarity of the state-specific loadings per condition, a congruence coefficient[8] φ (Tucker, 1951) was computed per factor for each pair of the loading matrices. A φ of one indicates proportionally identical factors (as in the no loading difference

[7] Note that we use EFA and thus, the zero loadings are not fixed but freely estimated in LMFA.

[8] Tucker's (1951) congruence coefficient between column vectors x and y is defined as: $\varphi_{xy} = \frac{x'y}{\sqrt{x'x}\sqrt{y'y}}$

conditions). The grand mean φ_{mean} across all state pairs and factors amounted to .80 for the medium loading difference conditions and .94 for the low loading difference conditions, regardless of the numbers of states and factors. Finally, the matrices were rowwise rescaled such that the sum of squares of each row equaled $1 - e$, where e was either .40 and .20 (g).

Intercept differences were induced as follows. For all variables in all states, the intercept was initially determined to be 5 and kept as such for the no intercept difference conditions. Two of the intercepts (different ones across the states) were increased from 5 to 5.5 for the low intercept difference conditions and from 5 to 7 for the medium intercept difference conditions.

Regarding the frequency of state transitions (e), we manipulated three levels that we considered to be realistic for ESM data. Note that we allowed for between-subject differences in the transition probabilities by randomly sampling each set of subject-specific probabilities from a uniform distribution within a specified range of probabilities. Specifically, the probabilities to stay in a state and to switch to another state were respectively sampled from $U(.73,.77)$ and $U(.01,[\ .27/(K-1)])$ in the highly frequent condition, from $U(.83,.87)$ and $U(.01,[\ .17/(K-1)])$ in the frequent condition and from $U(.93,.97)$ and $U(.01,[\ .07/(K-1)])$ in the infrequent condition. Then, for each resulting matrix, we rescaled the off-diagonal elements of each row to sum to 1 minus the diagonal element of that row, thus maintaining the probabilities to stay in a state and hence also the frequency of switching. As a results, out of the total number of possible transitions (i.e., across subjects (i.e., $\sum_i^I (T_i - 1)$) and across all data sets), a switch to another state occurred for 25% of the possible occasions in the highly frequent condition, for 15% in the frequent condition and for 5% in the infrequent condition.

Depending on the condition, datasets with the above described characteristics were simulated for 2, 5 or 10 subjects (f). Note that limiting our study to such low subject numbers not only confines the computation time but also challenges the method. We expect performance to improve with additional subjects because this accumulates the amount of data within the

states. Furthermore, the number of observations per subject and state, T_{ik}, was either 50, 100 or 200 (g) for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Thus, the total number of observations T_i per subject depended on (b) and (g). Similarly, the within-state sample size per state (over subjects) depended on (f) and (g).

For each subject, a LMC was generated indicating in which state subject i was at each time-point t. The initial state was randomly sampled from a Bernoulli distribution (for $k = 2$) or multinomial distribution (for $k > 2$) with equal initial state probabilities[9]. The remaining LMC was generated by sampling a random sequence of states based on the subject-specific transition probability matrix (i.e., depending on (e)). Note that whenever a state was not represented in a sampled LMC – because the small sample-sizes occasionally led to a data matrix wherein a certain state was not represented – we rejected it and sampled another one, such that parameter estimation was possible for all states.

Given this LMC, a subject-specific data set was generated according to Equation 2 assuming orthogonal factors. Firstly, we sampled a factor score vector $\mathbf{f}_{it} \sim MVN(0; \mathbf{I})$ of length F and a residual vector $\mathbf{e}_{it} \sim MVN(0; \mathbf{D}_k)$ of length J for each of the T_i observations, where the diagonal elements of $\mathbf{D}_k$ are equal to .20 or .40 (g). Subsequently, each vector of observations $\mathbf{y}_{it}$ was created with the loading matrix $\mathbf{\Lambda}_k$ and vector of intercepts $\boldsymbol{\nu}_k$ pertaining to the state that subject i was in at time-point t, according to the subject-specific LMC. Finally, the subject-specific datasets $\mathbf{Y}_i$ were concatenated into one data matrix $\mathbf{Y} = (\mathbf{Y}_1{}', \mathbf{Y}_2{}', \ldots, \mathbf{Y}_I{}')'$ with $\sum_i^I T_i$ rows. Twenty data matrices $\mathbf{Y}$ were generated for each cell of the design. In total, 3 (number of states) × 2 (number of factors) × 8 (between-state difference) × 3 (transition frequency between states) × 3 (number of subjects) × 3 (number of observations per subject and state) × 2 (unique variance) × 20 (replicates) = 51840 simulated data matrices were generated. The data

[9] We had no intention to evaluate the recovery of the initial state probabilities because more than a few subjects are required to validly estimate the distribution of initial starting states (Vermunt & Magidson, 2016). By sampling the initial state from a Bernoulli/multinomial distribution, some randomness in the initial states was obtained.

was generated in the open-source program R (R Core Team, 2002) and communicated to LG (Vermunt & Magidson, 2016) for analysis. LG syntaxes (for details and an example see Appendix B) were used to analyze the data with the correct number of states and factors per state. The average time to estimate a model was 85 seconds on an i5 processor with 8GB RAM. Model selection was evaluated for a subset of the conditions (indicated by "*") and for five replications per condition, i.e., for 80 data sets. The data sets were analyzed with LMFA models with the number of states equal to $K - 1$, K and $K + 1$ states, and the number of factors within the states equal to $F^k - 1$, F^k and $F^k + 1$ and allowed to differ between the states, resulting in 19 models when $K = 2$ and 46 when $K = 4$.

3.3 Results

Sensitivity to local maxima. The estimation procedure described in Appendix A, may result in a locally maximal solution, i.e., the best solution may have a $\log L$ value that is smaller than the one of the global ML solution. The multistart procedure (described in 2.3) increases the chance to find a global ML solution and, in the simulation study—where the global maximum is unknown due to violations of FA assumptions, sampling fluctuations and residuals—we can compare the best solution of the multistart procedure to an approximation (or proxy) of the global ML solution, which we obtain by providing the model estimation with the true parameter values as starting values. A solution is then a local maximum for sure when its $\log L$ value is smaller than the one from the proxy. To exclude mere calculation precision differences, we only considered negative differences with an absolute value higher than .001 as a local maximum. Accordingly, a local maximum was found for 947 out of 51840 simulated data sets (1.83%), which mainly occurred when $K = 4$.

Goodness of state recovery. To investigate the recovery of the state sequence, the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) was computed. The ARI quantifies the overlap between two partitions and is insensitive to permutations of the state labels. It ranges from 0

when the overlap is at chance level and 1 when partitions are identical. In general, the recovery of states was moderate to good (Steinley, 2004) with a mean ARI-value of .78 (SD = 0.28).

Except for the number of states, all manipulated factors had a large influence on the *ARI* (Table 2). In line with our expectations, the recovery improved with a lower number of factors (b), a greater between-state difference (c), a lower frequency of change (d), a higher number of subjects (e), a higher number of observations per subject and state (f) and lower unique variances (g). Figure 2 shows these effects, yet averaged across the number of factors for conciseness. A higher total within-state sample size was especially important for the state recovery in the high unique variance condition when combined with the low and no loading-difference conditions. In contrast, for a low unique variance and a medium loading difference between the states, the state recovery already stabilized at 400 observations. A lower frequency of transitions also further improved the state recovery, but it is most striking that even the most difficult conditions and lowest within-state sample size led to a perfect recovery when there was a medium difference in intercepts between the states.

[Insert Table 2 and Figure 2 about here]

Goodness of loading recovery. To examine the goodness of state-specific loading recovery ($GOSL$), we computed Tucker congruence coefficients φ between the true loading matrices and the estimated loading matrices and averaged across factors and states:

$$GOSL = \frac{\sum_{k=1}^{K}\sum_{f=1}^{F}\varphi(\boldsymbol{\Lambda}_f^k, \widehat{\boldsymbol{\Lambda}}_f^k)}{\sum_{k=1}^{K} F^K}. \tag{4}$$

To deal with the rotational freedom of the factors per state, we rotated the factors prior to calculating the congruence coefficient[10]. Specifically, Procrustes rotation was used to rotate the estimated towards the true loading matrices. To account for the permutational freedom of the

[10] Note that rotation is not yet included in LG, which is why we take the estimated loadings and rotate them in the open-source program R (R Core Team, 2002).

state labels, the state permutation that maximizes the $GOSL$ was retained. The manipulated conditions hardly affected the loading recovery. The overall mean $GOSL$ was .98 (*SD* = 0.05), indicating an excellent recovery. There was a positive correlation between the *ARI* value and the *GOSL* ($r_s = .45, p < 0.001$). Note that the loading recovery was often good despite a bad state recovery because quite similar (to even identical) loading matrices are mixed up.

Goodness of transition matrix recovery. To examine the transition matrix recovery, we calculated the mean absolute difference (MAD) between the true and estimated matrices (applying the best state permutation obtained from the loading recovery):

$$MAD_{trans} = \frac{\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{K}|\mathrm{A}_{ikl} - \widehat{\mathrm{A}}_{ikl}|}{IK^2}. \tag{5}$$

The transition matrix recovery was good with an average MAD_{trans} of .08 (*SD* = 0.10). Overall, the effects of the manipulated conditions were rather small (see Table 2).

Goodness of intercept recovery. To evaluate the recovery of the state-specific intercepts, we calculated the *MAD* between the true intercepts and the estimated ones.

$$MAD_{int} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}|\nu_{kj} - \hat{\nu}_{kj}|}{KJ}. \tag{6}$$

The intercept recovery was moderate with an average MAD_{int} of .12 (*SD* = 0.02). A higher between-state difference of loadings and intercepts (c), more subjects (e), more observations per subject and state (f), and a lower unique variance (g) improved the intercept recovery.

Goodness of unique variance recovery. To examine the recovery of the state-specific unique variances, we calculated the *MAD* between the true and estimated unique variances.

$$MAD_{uniq} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}|d_{kj} - \hat{d}_{kj}|}{KJ}. \tag{7}$$

The unique variance recovery was good with an average MAD_{uniq} of .04 (*SD* = 0.06) and no notable differences across the manipulated conditions. More prominently, the MAD_{uniq} is affected by Heywood cases (Van Driel, 1978), which pertain to “improper” factor solutions

with at least one unique variance that is negative or zero. When a Heywood case occurs, LG fixed the unique variance(s) to a very small number. A Heywood case is considered to be diagnostic of underdetermined factors and/or insufficient sample size (McDonald & Krane, 1979; Rindskopf, 1984; Van Driel, 1978; Velicer & Fava, 1998). Heywood cases occurred for 5877 of the estimated datamatrices (12.19%), where 89% of the Heywood cases indeed occurred in the conditions with the smallest number of observations per subject and state and/or the smallest number of subjects.

Model selection. For 74 out of the 80 data sets (93%), the correct model was selected, when considering the converged models only, and for 78 (98%) the correct model was among the three best models. Five of the six incorrect selections occurred for the data sets with four states and four factors and low loading differences as well as low intercept differences. Specifically, one state too few was selected which is explained by the low state separability in these conditions[11]. We conclude that the BIC performs very well with regard to selecting the most appropriate model complexity for LMFA.

3.4 Conclusions & Recommendations

To sum up, LMFA is promising for detecting MM changes over time and for exploring what the MM differences look like and for which subjects and which time-points the MMs are comparable. However, the performance of the new method in recovering the correct state sequence and the correct state-specific MMs largely depends on model characteristics (i.e., the number of factors, the MM differences between the states, the unique variances, and the frequency of state transitions), and the within-state sample size. Firstly, larger MM differences between states benefit the recovery of the states. Especially intercept differences increased the separability of the states, to the extent that the states were recovered perfectly even under

[11] Note that, in case there is only one MM underlying the data, model selection would indicate that the one-state model fits best, which was confirmed by a small simulation study with the same design as the one of the model selection described in 3.2.. The correct one-state model was correctly chosen for all 10 models (100 %).

difficult conditions. Besides intercept differences, less factors, less frequent transitions between the states and lower unique variances improved the recovery. Finally, all else equal, the within-state sample size greatly enhanced the state recovery. In the following we list recommendations for empirical practice:

- When the above-mentioned model characteristics are unknown (or assumed to be unfavorable), aim for 2000 to 4000 observations in total (subjects × observations) to obtain reliable results for 2 to 4 states.
- When favorable model characteristics are assumed – i.e., when between-state differences are expected to be pronounced (e.g., changes in intercept are expected), transitions to be infrequent (e.g., measurement occasions are closely spaced) and unique variances to be low (e.g., using reliable measurement instruments) – 800 to 1000 observations in total (subjects × observations) are probably enough to obtain reliable results for 2 to 4 states.
- The number of states that can be reliably captured is bound by the total sample size and when the sample size does not allow for the 'true' number of states to be estimated, the obtained results will only convey part of the MM differences present in the data. States that correspond to a few observations only will not be detected.
- Including more subjects in the study might be more feasible than obtaining more measurements from each participants but be aware that in practice, subjects do not necessarily switch between the same set of MMs. LMFA allows for this heterogeneity in MMs, since not all subjects need to go through all states. Nevertheless, it is important to keep this potential heterogeneity in mind since it would imply that additional subjects increase the number of MMs, and thus the number of states. In that case, the number of observations per subject is essential for the sample size per state.
- The BIC is a suitable criterion to decide on the best number of states and factors. However, when differences between the states are subtle, researchers are advised to consider the three

best models and choose one based on interpretability and stability.

4 Application

In order to apply LMFA to the empirical data set introduced in Section 2.1, we first selected the number of states and factors by comparing the BIC among LMFA models with one up to three states and one up to three factors per state. Models with four states or factors in a state failed to converge suggesting sample size limitations or model misspecification. The model [3 3 2] (i.e., three states with 3, 3, and 2 factors) was selected as it had the lowest BIC among the converged models and was the most interpretable. To shed light on the MM differences between the three states, we first looked at the state-specific intercepts (Figure 3). The intercepts are higher for ‘positive affect’ (PA) items than for ‘negative affect' (NA) items in all the states. However, the difference between the PA and NA item scores is most visible in State 3 (hereinafter ‘pleasure state’), intermediate in State 2 (hereinafter ‘neutral state’) and least pronounced in State 1 (hereinafter ‘displeasure state’).

[Insert Figure 3 & Table 3 about here]

Second, we investigated the standardized oblimin rotated loadings (Table 3). As a notable similarity, we see that the positive items are collected into (i.e., load strongly on) a PA factor in all the states, although the strength of the loadings slightly differs between the states. A striking difference is that the pleasure state has a NA factor, whereas the neutral and displeasure states both have a ‘distress’ factor with high loadings of ‘upset’, ‘anxious’, and ‘nervous’ – although they slightly differ in that ‘calm’ has an additional strong negative loading in the displeasure state, whereas ‘gloomy’ and ‘sluggish’ load on the distress factor in the neutral state only. The neutral and displeasure states are further characterized by a third factor. In the neutral state, the third factor pertains to ‘serenity’ with strong loadings of ‘calm’ and ‘relaxed’. In the displeasure state, it is a bipolar ‘drive’ factor indicating that being ‘determined’ (strong negative loading) is inversely related to feeling ‘sluggish’, ‘bored’, and ‘listless’ (strong

positive loadings). This additional drive factor in the displeasure state concurs with theoretical models of anhedonia (Berridge, Robinson, & Aldridge, 2009; Treadway & Zald, 2011), which divide anhedonia in three dimensions: consummatory anhedonia (no longer enjoying pleasurable activities), anticipatory anhedonia (no longer looking forward to pleasurable activities) and motivational anhedonia (no longer experiencing motivation to pursue pleasurable activities). The drive factor confirms that motivation is distinct from general positive affect when individuals are anhedonic. Finally, the state-specific unique variances are listed in Table 3. In general, these are highest in the displeasure state, indicating more emotion-specific variability than in the other states. This concurs with research showing that emotional complexity is associated with higher levels of depression (e.g., Grühn, Lumley, Diehl, & Labouvie-Vief, 2013). In sum, LMFA allowed for us to find substantively meaningful changes in the MM, both in the number and nature of the underlying factors. As an important similarity between the states, it is found that PA is captured in each of the three states[12].

To investigate what potentially triggers the latent states, we explored between-state differences in evening measures on physical discomfort (such as headache) and the occurrence and importance of positive and negative events. We focus on descriptive statistics only since hypothesis testing for MM differences is beyond the scope of the paper. A question was, e.g., “Think about the most unpleasant event you experienced since the last assessment: how unpleasant was this experience?” The scales ranged from 0 (“Not at all”) to 100 (“Very much”). Interestingly, when participants were in the displeasure state, they had experienced more unpleasant (M = 48.64, SD = 24.24) events than when being in the neutral (M = 32.54, SD = 19.85) or the pleasure state (M = 29.52, SD = 18.65). Similarly, being in the pleasure state was related to the occurrence of more pleasant events (M = 64.54, SD = 15.48) in comparison to the

[12] Note that the NA items were mostly right skewed, warranting caution when drawing substantive conclusions because violations of the normality assumption have yet to be investigated (see Section 5). For the purpose of illustrating the aim of LMFA, this is not a problem.

displeasure (*M* = 56.02, *SD* = 20.41) and the neutral state (*M* = 58.95, *SD* = 18.03). These findings are in line with the states' labels.

[Insert Table 4 about here]

Moreover, we inspected how the states related to the interventions (Table 4). Before the intervention (Phase 1), participants were more often in the displeasure or neutral state than in the pleasure state. After the first intervention (Phase 2), the participants in the two intervention groups (i.e., PLA and PLA & SkyD) were more often in the pleasure or neutral state than in the displeasure state. After receiving a second intervention (Phase 3), the distribution across the displeasure and pleasure state stayed about the same or the occurrence of the pleasure state increased. Participants who did not receive an intervention after the first month were distributed equally across the pleasure and displeasure states and were mostly in the neutral state during Phase 2. Notably, in Phase 3, the state membership for these participants – i.e., after receiving their first (self-chosen) intervention – changed in that the pleasure state was now also more frequent than the displeasure state when participants chose PLA & SkyD while it was the opposite for those who chose PLA, perhaps because the more depressed and anhedonic participants were the ones refraining from a SkyD. Looking at the examples of individual transition plots including the individual transition probabilities (Figure 4), it is apparent that participants switched quite often between states, in each phase of the study. This is coherent with previous findings that individuals with anhedonia and depression are often found to experience strong fluctuations in emotional experiences (Heininga, Van Roekel, Ahles, Oldehinkel, & Mezulis, 2017; van Roekel et al., 2015). Some participants switched more often between the states than others, which may pertain to between-subject differences in general stability and experienced events but also to differences in how one reacts to the interventions.

Summarized, the interventions appear to have increased the pleasure state membership and reduced the displeasure state membership, while leaving membership of the neutral state

largely unaffected. It is noteworthy that participants receiving no intervention after the first month also slightly moved toward higher pleasure state membership and a lower displeasure state membership at that point in time. It appears that daily reflections on ones emotions also relieve anhedonia to a certain degree, which was already found in an intervention study using ESM in depressed patients (Kramer et al., 2014). Although these findings are merely exploratory and need to be validated in future research, we demonstrated that LMFA offers valuable insights to applied researchers.

[Insert Figure 4 about here]

5 Discussion

In this paper, we introduced latent Markov factor analysis (LMFA) for modeling measurement model (MM) changes that are expected to be prevalent in time-intensive longitudinal data like experience sampling data. In this way, LMFA safeguards conclusions about changes in the measured constructs. MM changes may pertain to (potentially interesting) substantive changes or may signal the onset of response styles (RSs). Between-state differences in intercepts and loadings might suggest an extreme RS, whereas differences in intercepts only rather suggest an agreeing RS (Cheung & Rensvold, 2000). When one suspects a RS in a specific state, RS detection and correction (e.g., adding an agreeing RS factor; Billiet & McClendon, 2000; Watson, 1992) can be applied to that specific part of the data, rather than to the entire data set. Moreover, the subject-specific transition probabilities of LMFA capture, e.g., to what extent each subject is likely to end up and to stay in an extreme RS state. Even when RSs are hard to distinguish, the fact that LMFA pinpoints MM changes – and thus the reliable and comparable parts of the data – is valuable in itself.

In the future, it would be interesting to go beyond the purely exploratory approach applied in this paper. On the one hand, hypothesis testing to determine which parameters significantly differ between the states might be preferred over visually comparing the state-specific MMs.

To this end, LG already provides the researchers with Wald test statistics when the rotational freedom of the state-specific factors is resolved by a minimal number of restricted loadings (e.g., Geminiani, Ceulemans, & De Roover, 2018). On the other hand, including explanatory variables (i.e., time-constant or time-varying covariates such as personality traits or social contexts) in the model would allow to evaluate whether they significantly predict the state memberships and the transition probabilities.

Moreover, LMFA assumes normally distributed, continuous variables. Categorical Likert-scale ratings are frequently used in psychology, however. Although these data can often be treated as continuous in case of at least five response categories (Dolan, 1994; Olsson, 1979), the ratings are often skewed, thus violating the assumption of normality. Additionally, and even continuous data, such as our application data, might be skewed. Therefore, the robustness of the method to such violations should be examined and, if necessary, possible extensions to deal with non-normality should be considered.

In addition, longitudinal data are often collected in varying time intervals; e.g., when testing the long-term influence of interventions on affect by collecting data in waves. In that case, the transition probabilities can no longer be considered time-homogeneous and continuous time modeling is necessary (Crayen, Eid, Lischetzke, & Vermunt, 2017). Therefore, in future research, we will develop a continuous-time extension of LMFA.

Moreover, a limitation of the method is the assumption that factor scores are centered around zero and have a variance of one. When factor scores evolve over time but the MM stays the same, changes in factor scores would currently be detected as intercept changes and thus, possibly lead to different states according to model selection. In future work, we will investigate necessary LMFA extensions to explicitly model changes in factor means within the states, e.g., depending on time or another covariate.

Next to that, we might consider an extension of LMFA using exploratory dynamic FA

(EDFA; Browne, 2001; Zhang, 2006) within the states, which models the auto- and cross-lagged correlations of the factors at consecutive time-points but comes with important challenges. Firstly, accurately estimating autocorrelations would require more measurement occasions per subject per state (Ram, Brose, & Molenaar, 2012), which may be undesirable. Secondly, in EDFA, factor rotation is more intricate since the auto- and cross-lagged relations between factors need to be rotated towards specified target matrices (Browne, 2001; Zhang, 2006), again necessitating the a priori hypotheses about (changes in) the MM we want to avoid. The LMM in LMFA already partly captures autocorrelations by the states and uncaptured auto- and cross-lagged correlations will not necessarily introduce bias in the estimates of the state-specific MMs (Baltagi, 2011).

Finally, LMFA is a complex model with many assumptions. Therefore, misspecifications can occur and tools to locate local misfit are essential. Local fit measures have been developed in some fields (e.g. bivariate residuals measures for multilevel data; Nagelkerke, Oberski, & Vermunt, 2016) but future work first needs to develop and extensively evaluate them for LMFA.

References

Adolf, J., Schuurman, N. K., Borkenau, P., Borsboom, D., & Dolan, C. V. (2014). Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. *Front Psychol, 5,* 883. doi:10.3389/fpsyg.2014.00883

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA American Psychiatric Publishing

Asparouhov, T., Hamaker, E. L., & Muthen, B. (2017). Dynamic structural equation models. Technical Report.

Baltagi, B. H. (2011). *Econometrics* (5th ed.). Berlin: Springer.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, series B, 68,* 155–178.

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2014). Comments on: Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test, 23,* 473–477. doi:10.1007/s11749-014-0387-1

Bartolucci, F., Farcomeni, A., & Pennoni, F. (2015). *Latent Markov Models for Longitudinal Data*. Hoboken: CRC Press.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics, 41,* 164–171.

Berridge, K. C., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current Opinion in Pharmacology, 9,* 65–73. doi:10.1016/j.coph.2008.12.014

Billiet, J. B., & McClendon, M. J. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608-628. doi:10.1207/s15328007sem0704_5

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (5th ed.). New York: Springer.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219. doi:10.1037/0033-295X.110.2.203

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38*(1), 25-56. doi:10.1207/S15327906MBR3801_2

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*, 111–150. doi:10.1207/S15327906MBR3601_05

Carvalho, C. M., & Lopes, H. F. (2007). Simulation-based sequential analysis of Markov switching stochastic volatility models. *Computational Statistics & Data Analysis, 51*, 4526–4542. doi:10.1016/j.csda.2006.07.019

Cheung, G. W., & Rensvold, R. B. (2000). Assessing Extreme and Acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 187–212. doi:10.1177/0022022100031002003

Clarkson, D. B., & Jennrich, R. I. (1988). Quartic rotation criteria and algorithms. *Psychometrica, 53*, 251–259. doi:10.1007/BF02294136

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.

Crayen, C., Eid, M., Lischetzke, T., & Vermunt, J. K. (2017). A continuous-time mixture latent-state-trait Markov model for experience sampling data. *European Journal of Psychological Assessment, 33*, 296–311. doi:10.1027/1015-5759/a000418

De Roover, K., Timmerman, M. E., Van Diest, I., Onghena, P., & Ceulemans, E. (2014). Switching principal component analysis for modeling means and covariance changes over time. *Psychological Methods, 19*, 113–132. doi:10.1037/a0034525

De Roover, K., Vermunt, J. K., Timmerman, M. E., & Ceulemans, E. (2017). Mixture simultaneous factor analysis for capturing differences in latent variables between higher level units of multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 1–18. doi:10.1080/10705511.2017.1278604

de Winter, J. C., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*, 147–181. doi:10.1080/00273170902794206

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

Dias, J. G., Vermunt, J. K., & Ramos, S. (2008). Heterogeneous hidden Markov models. *COMPSTAT 2008 Proceedings*.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309–326.

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology, 69*, 153–166. doi:10.1037/0022-3514.69.1.153

Geminiani, E., Ceulemans, E., & De Roover, K. (2018). Wald tests for factor loading differences among many groups by means of mixture simultaneous factor analysis: power and type I error. *Manuscript submitted for publication*.

Grühn, D., Lumley, M. A., Diehl, M., & Labouvie-Vief, G. (2013). Time-based indicators of emotional complexity: interrelations and correlates. *Emotion, 13*, 226–237. doi:10.1037/a0030363

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

Heininga, V. E., Van Roekel, E., Ahles, J. J., Oldehinkel, A. J., & Mezulis, A. H. (2017). Positive affective functioning in anhedonic individuals' daily life: Anything but flat and blunted. *Journal of Affective Disorders, 218*, 437–445. doi:10.1016/j.jad.2017.04.029

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Jolliffe, I. T. (1986). *Principal Component Analysis*. NewYork, NY: Springer.

Kaiser, H. F. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187–200. doi:10.1007/BF02289233

Kang, X., & Thakor, N. V. (2012). *Factor analyzed hidden Markov models for estimating prosthetic limb motions using premotor cortical ensembles*. Paper presented at the IEEE RAS/EMBS International Conference on Biomedical Robotics and Biomechatronics, Roma, Italy.

Kiers, H. A. (1997). Techniques for rotating two or more loading matrices to optimal agreement and simple structure: A comparison and some technical details. *Psychometrika, 62*, 545–568. doi:10.1007/BF02294642

Kramer, I., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., . . . Wichers, M. (2014). A therapeutic application of the experience

sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry, 13*, 68–77. doi:10.1002/wps.20090

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In *Flow and the Foundations of Positive Psychology* (pp. 21–34).

Lawley, D. N., & Maxwell, A. E. (1962). *Factor analysis as a statistical method*. London: Butterworth.

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

Lee, S. Y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 44*, 99–113. doi:10.1007/BF02293789

Little, T. D., Preacher, K. J., Selig, J. P., & Card, N. A. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*, 357–365. doi:10.1177/0165025407077757

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica, 14*, 41–67.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*, 611–637.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

McDonald, R. P., & Krane, W. R. (1979). A Monte Carlo study of local identifiability and degrees of freedom in the asymptotic likelihood ratio test. *British Journal of Mathematical and Statistical Psychology, 32*, 121–132. doi:10.1111/j.2044-8317.1979.tb00757.x

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. doi:10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*, 69–77. doi:10.1097/01.mlr.0000245438.73837.89

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity, 37*, 277–302. doi:10.1023/A:1024472110002

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: a restricted latent class factor analysis approach. *Sociological Methodology, 41*, 13–47. doi:10.1111/j.1467-9531.2011.01238.x

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*, 90-100. doi:10.1016/s0022-2496(02)00028-7

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Power and type I error of local fit statistics in multilevel latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 216-229. doi:10.1080/10705511.2016.1250639

Ogasawara, H. (2000). Some relationships between factors and components. *Psychometrika, 65*, 167–185. doi:10.1007/BF02294372

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 481–500. doi:10.1207/s15327906mbr1404_7

Oort, F. J., Visser, M. R., & Sprangers, M. A. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research, 14*, 599–609. doi:10.1007/s11136-004-0831-x

Paulhus, D. L. (1991). Measures of personality and social psychological attitudes. In J. P. Robinson & R. P. Shaver (Eds.), *Measures of Social Psychological Attitudes Series* (Vol. 1, pp. 17–59). San Diego, CA: Academic Press.

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, *32*, 153–161. doi:10.1023/A:1015210025234

R Core Team. (2002). A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Ram, N., Brose, A., & Molenaar, P. C. M. (2012). Dynamic factor analysis: Modeling person-specific process. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods* (Vol. 2, pp. 441–457). New York: Oxford University Press.

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research*, *13*, 109–119. doi:10.1177/0049124184013001004

Rosti, A. I., & Gales, M. J. (2002). Factor analysed hidden Markov models. In *Proceedings of the ICASSP* (Vol. 1, pp. 949–952). Florida, USA.

Schaie, K. W., Maitland, S. B., Willis, S. L., & Intrieri, R. C. (1998). Longitudinal Invariance of Adult Psychometric Ability Factor Structures Across 7 years. *Psychology and Aging*, *13*, 8–20. doi:10.1080/13825580490511134

Scollon, C., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*, 5–34.

Song, X., Xia, Y., & Zhu, H. (2017). Hidden Markov latent variable models with multivariate longitudinal data. *Biometrics*, *73*, 313–323. doi:10.1111/biom.12536

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*, 386–396. doi:10.1037/1082-989X.9.3.386

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods, 16*, 63–79. doi:10.1037/a0022673

Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Treadway, M. T., & Zald, D. H. (2011). Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neurosci Biobehav Rev, 35*, 537–555. doi:10.1016/j.neubiorev.2010.06.006

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies. Personnel Research Section Report No. 984*. Washington, DC: Department of the Army.

Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika, 43*, 225–243. doi:10.1007/BF02293865

van Roekel, E., Bennik, E. C., Bastiaansen, J. A., Verhagen, M., Ormel, J., Engels, R. C., & Oldehinkel, A. J. (2015). Depressive Symptoms and the Experience of Pleasure in Daily Life: An Exploration of Associations in Early and Late Adolescence. *J Abnorm Child Psychol, 44*, 999–1009. doi:10.1007/s10802-015-0090-z

Van Roekel, E., Vrijen, C., Heininga, V. E., Masselink, M., Bos, E. H., & Oldehinkel, A. J. (2017). An exploratory randomized controlled trial of personalized lifestyle advice and tandem skydives as a means to reduce anhedonia. *Behavior Therapy, 48*, 76–96. doi:10.1016/j.beth.2016.09.009

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251. doi:10.1037/1082-989X.3.2.231

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*, 1–28. doi:10.1207/s15327906mbr2501_1

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research, 17*, 371–388. doi:10.1207/s15327906mbr1703_5

Vermunt, J. K., Langeheine, R., & Böckenholt, U. (1999). Discrete-time discrete-state latent markov models with time constant and time-varying covariates. *Journal of Educational and Behavioral Statistics, 24*, 179–207. doi:10.2307/1165200

Vermunt, J. K., & Magidson, J. (2016). *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont Massachussetts.

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale. *Sociological Methods & Research, 21*, 52–88. doi:10.1177/0049124192021001003

Xia, Y.-M., Tang, N.-S., & Gou, J.-W. (2016). Generalized linear latent models for multivariate longitudinal measurements mixed with hidden Markov models. *Journal of Multivariate Analysis, 152*, 259–275. doi:10.1016/j.jmva.2016.09.001

Ximenez, C. (2009). Recovery of weak factor loadings in confirmatory factor analysis under conditions of model misspecification. *Behavior Research Methods, 41*, 1038–1052. doi:10.3758/BRM.41.4.1038

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling, 13*, 587–614. doi:10.1207/s15328007sem1304_5

Zhang, G. (2006). *Bootstrap Procedures for Dynamic Factor Analysis*. Unpublished doctoral dissertation. The Ohio State University, Columbus, OH.

Figures and Tables

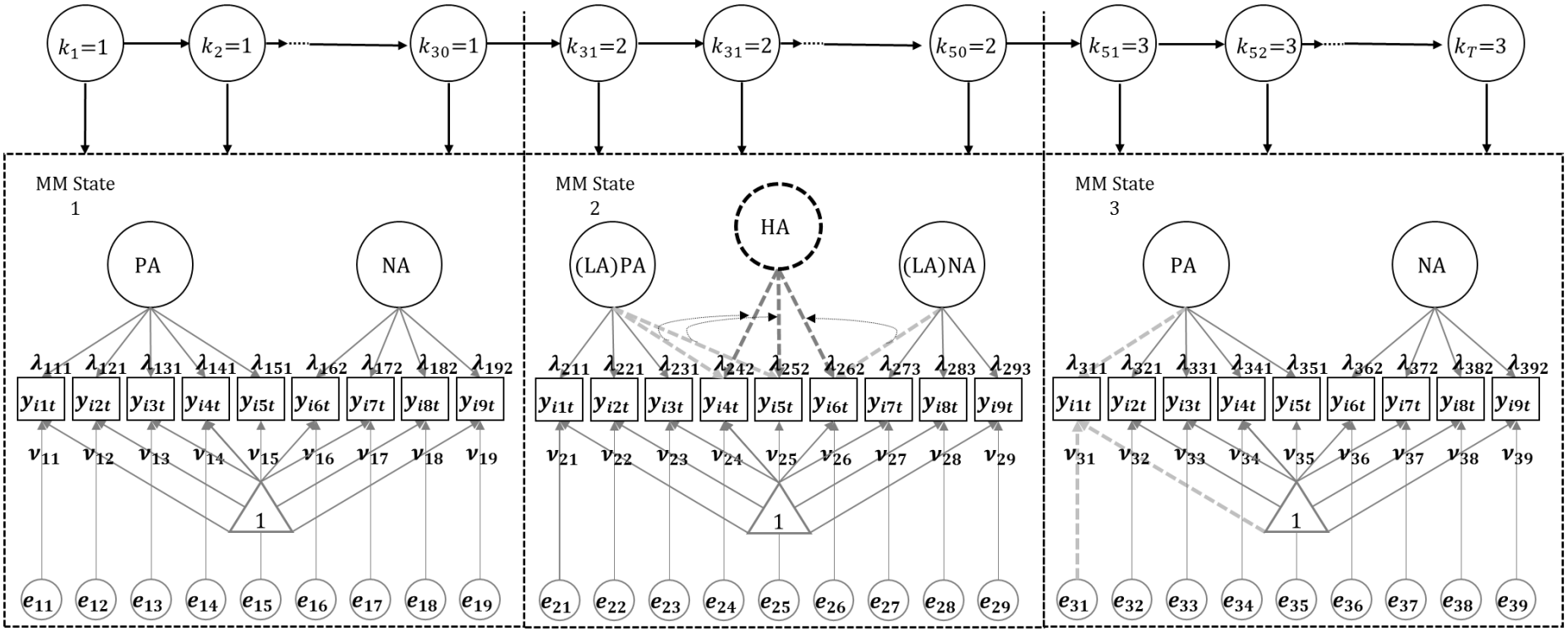

Figure 1. Graphical illustration of a subject-specific LMC from a LMFA model, where the latent states per measurement occasion kt (t = 1, …, T) indicate the measurement model underlying a respondent's observed item scores. Note that to give a clear example, only the standardized factor loadings with an absolute value larger than 0.4 are depicted. Also note that the factor scores (e.g., on PA and NA) are not depicted in the figure since they are not part of the MM, but that they may or may not change within a state over time but average zero.

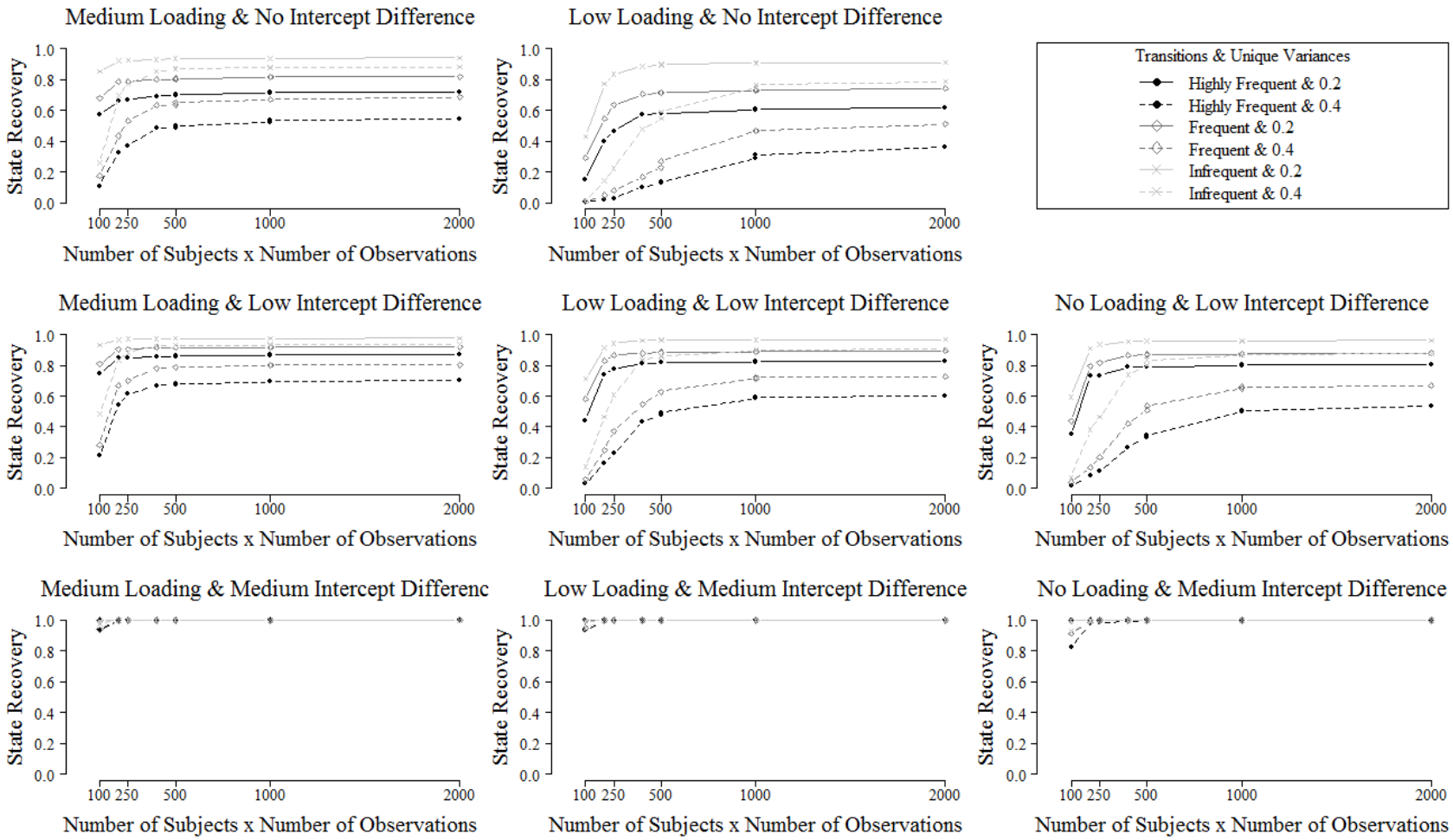

Figure 2. Goodness of state recovery (y-axis) under various manipulations of the between-state difference (loading difference (LD) per column and intercept difference (ID) per row), the unique variances and the frequency of transitions (different lines), and the total number of observations per state (number of subjects × number of observations), while averaging across the number of states and factors. Note that the x-axis value 500 and 1000 occur twice due to two possible combinations (5 × 100 = 10 × 5 and 5 × 200 = 10 × 100, respectively). The dots largely overlap and thus state recovery was not visibly influenced by the combination and thus, it did not matter whether the within-state sample size was increased by additional subjects or additional observations per subject.

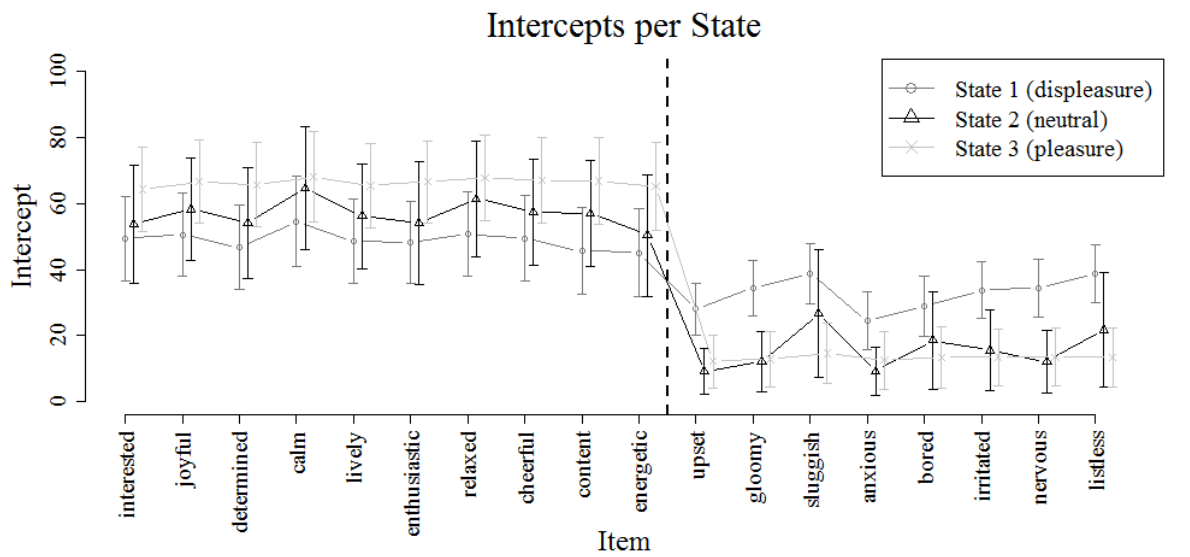

Figure 3. Intercepts and standard deviations of the 18 items per state (positive (left) and lower negative emotions (right)).

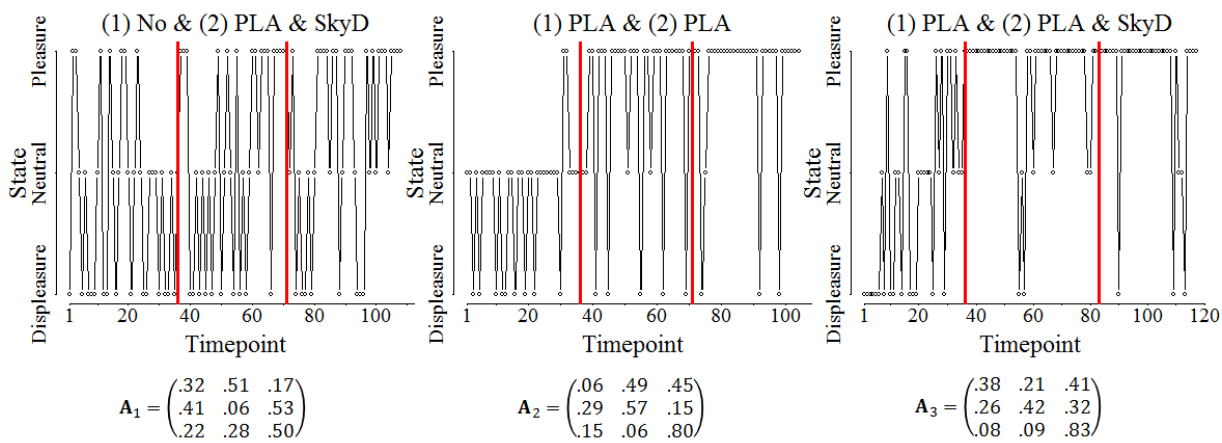

Figure 4. Three examples of individual transition plots (including probabilities) of subjects with three different combinations of interventions.

Table 1. 2-factor base loading matrix and derived loading matrices for state 1 and 2.

	Base Loading Matrix		*State1*		*State 2*	
	Factor 1	*Factor 2*	*Factor 1*	*Factor 2*	*Factor 1*	*Factor 2*
Var.1	1	0	λ_1	λ_2	1	0
Var.2	1	0	1	0	λ_1	λ_2
Var.3	1	0	1	0	1	0
Var.4	1	0	1	0	1	0
Var 5-10	…	…	…	…	…	…
Var.11	0	1	λ_2	λ_1	0	1
Var.12	0	1	0	1	λ_2	λ_1
Var.13	0	1	0	1	0	1
Var.14	0	1	0	1	0	1
Var. 15-20	…	…	…	…	…	…

Note. For the medium loading difference, $\lambda_1 = 0$ and $\lambda_2 = 1$; for the low loading differences, $\lambda_1 = \sqrt{.5}$ and $\lambda_2 = \sqrt{.5}$. Entries of Var. 5-10 and 15-20 equal those of Var.4 and Var.14, respectively. The 4-factor matrices were created by applying the same λ_1 and λ_2 values to other variables because of fewer loadings per factor.

Table 2. Goodness-of-recovery per type of parameter conditional on the manipulated factors

		Goodness of				
		State Recovery (ARI)	Loading Recovery ($GOSL$)	Transition Matrix Recovery (MAD_{trans})	Intercept Recovery (MAD_{int})	Unique Variance Recovery (MAD_{uniq})
Condition	Manipulated Factors					
States	2	0.76	0.97	0.09	0.12	0.04
	3	0.79	0.98	0.08	0.12	0.04
	4	0.8	0.98	0.06	0.12	0.04
Factors	2	0.81	0.98	0.07	0.11	0.03
	4	0.75	0.97	0.09	0.12	0.05
Between-state-difference (Loading Difference-Intercept Difference)	Medium-No	0.69	0.98	0.07	0.09	0.04
	Low-No	0.47	0.94	0.16	0.26	0.08
	Medium-Low	0.81	0.98	0.06	0.08	0.03
	Low-Low	0.68	0.96	0.1	0.16	0.05
	No-Low	0.64	0.96	0.12	0.21	0.06
	Medium-Medium	0.99	1	0.04	0.04	0.02
	Low-Medium	0.99	0.99	0.04	0.05	0.02
	No-Medium	0.99	0.99	0.04	0.04	0.02
Transitions	Highly Frequent	0.71	0.97	0.1	0.13	0.04
	Frequent	0.77	0.98	0.08	0.12	0.04
	Infrequent	0.87	0.98	0.06	0.1	0.04
Subjects	2	0.7	0.95	0.12	0.23	0.07
	5	0.81	0.98	0.07	0.09	0.03
	10	0.84	1	0.05	0.04	0.02
Observations per subject and state	50	0.7	0.95	0.13	0.21	0.07
	100	0.8	0.98	0.07	0.09	0.04
	200	0.84	0.99	0.04	0.05	0.02
Unique Variances	.2	0.88	0.99	0.05	0.07	0.02
	.4	0.69	0.96	0.11	0.16	0.06

Note. For the between-state difference condition, the combination of no loading difference and no intercept difference was not included because this would imply that the MM does not differ across states.

Table 3. Standardized Oblimin rotated factor loadings, unique variances, and proportions of unique variance of the LMFA model with three states and respectively three, three and two factors for the evening emotion questionnaires

	State 1 (Displeasure)					State2 (Neutral)					State 3 (Pleasure)			
	Factors			Unique V.		Factors			Unique V.		Factors		Unique V.	
Items	PA	Distr.	Drive	V	(prop)	PA	Distr.	Serenity	V.	(prop)	PA	NA	V.	(prop)
interested	**0.66**	0.08	-0.13	178.3	(.48)	**0.7**	0.04	0	166.2	(.52)	**0.85**	0.05	52.55	(.32)
joyful	**0.8**	-0.01	-0.07	118.5	(.30)	**0.86**	0	-0.07	63.37	(.26)	**0.92**	0	25.62	(.16)
determined	0.39	0.01	**-0.48**	187.3	(.43)	**0.81**	0.1	0.16	100.7	(.36)	**0.93**	0.04	29.18	(.18)
calm	**0.43**	**-0.43**	0.1	269.4	(.56)	**0.49**	-0.12	**-0.61**	123.3	(.35)	**0.82**	-0.08	47.16	(.25)
lively	**0.73**	0	-0.13	123.4	(.35)	**0.86**	-0.01	-0.07	65.4	(.26)	**0.9**	-0.01	27.62	(.17)
enthusiastic	**0.81**	0.14	-0.11	129.1	(.29)	**0.87**	0.08	-0.01	91.02	(.27)	**0.93**	0.03	23.98	(.16)
relaxed	**0.62**	-0.34	0.22	207.6	(.48)	**0.54**	-0.14	**-0.67**	65.9	(.21)	**0.85**	-0.07	36.89	(.22)
cheerful	**0.84**	0.08	0.04	133.7	(.35)	**0.84**	-0.05	-0.08	69.34	(.27)	**0.87**	-0.02	39.17	(.23)
content	**0.56**	-0.2	-0.09	179.1	(.51)	**0.84**	0	-0.18	65.29	(.26)	**0.93**	0.01	24.2	(.14)
energetic	**0.59**	0.15	-0.31	183.4	(.41)	**0.78**	0.13	0.03	140.8	(.41)	**0.9**	0.01	37.34	(.21)
upset	-0.18	**0.51**	0.12	290.7	(.58)	0.12	**0.87**	0	13.01	(.27)	0.01	**0.89**	13.19	(.21)
gloomy	-0.33	0.27	0.25	301.8	(.59)	-0.03	**0.77**	-0.03	33.19	(.40)	-0.02	**0.91**	11.52	(.16)
sluggish	0.12	0.13	**0.92**	134.8	(.19)	-0.3	0.26	-0.32	268.5	(.71)	-0.02	**0.85**	22.45	(.27)
anxious	0.07	**0.76**	0.17	183.7	(.39)	0.12	**0.86**	0.03	15	(.28)	0.05	**0.91**	14.57	(.19)
bored	-0.09	-0.01	**0.4**	411.1	(.79)	-0.16	0.38	-0.22	163.5	(.74)	-0.01	**0.87**	19.53	(.23)
irritated	-0.14	0.25	0.15	460.3	(.84)	0.05	**0.63**	0.08	92.9	(.61)	-0.01	**0.9**	13.44	(.18)
nervous	0.15	**0.68**	-0.03	326	(.58)	0.1	**0.74**	0.11	42.21	(.48)	-0.01	**0.88**	17.39	(.22)
listless	0.18	0.09	**0.58**	304.2	(.48)	-0.3	0.33	-0.22	216.9	(.72)	0	**0.94**	8.98	(.11)

Note. V. = Variance; prop. = proportion; To enhance interpretation, factor loadings were standardized by dividing the loadings by the state-specific standard deviations of the item. The loadings with an absolute value larger than 0.4 are depicted in boldface. Cor(PA,Distr.) = .-31, Cor(PA,Drive) = .52, and Cor(Distr.,Drive) = -.18 in State 1; Cor(PA,Distr.) = .-15, Cor(PA,Serenity) = -.07, and Cor(Distr.,Serenity) = -.14 in State 2; Cor(PA,NA) = .-42 in State 3.

Table 4. Observed state-memberships per phase and intervention

	State			State			State		
	1 (Displ.)	**2 (Neutral)**	**3 (Pleasure)**	**1 (Displ.)**	**2 (Neutral)**	**3 (Pleasure)**	**1 (Displ.)**	**2 (Neutral)**	**3 (Pleasure)**
	Int. 1 = No Int. (n = 22, T_B = 790 , T_A = 786)			**Int. 1 = PLA** (n = 23, T_B = 820 , T_A = 832)			**Int. 1 = PLA & SkyD** (n = 24, T_B = 850, T_A = 939)		
Before Int. 1	41.7 %	43.0 %	15.3 %	45.0 %	36.3 %	18.7 %	39.3 %	38.1 %	22.6 %
After Int. 1	27.2 %	45.8 %	27.0 %	27.9 %	30.4 %	41.7 %	27.8 %	40.9 %	31.3 %
	Int. after Int. 1 = No Int.			**Int. after Int. 1 = PLA**			**Int. after Int. 1 = PLA & SkyD**		
	Int. 2 = No Int. (n = 0, T = 0)			**Int. 2 = No Int.** (n = 2, T = 55)			**Int. 2 = No Int.** (n = 1, T = 26)		
	/	/	/	30.9 %	16.4 %	52.7 %	26.9 %	26.9 %	46.2 %
After Int. 2	**Int. 2 = PLA** (n = 3, T = 95)			**Int. 2 = PLA** (n = 3, T = 102)			**Int. 2 = PLA** (n = 11, T = 372)		
	51.6 %	37.9 %	10.5 %	8.8 %	22.6 %	68.6 %	26.1 %	41.4 %	32.5 %
	Int. 2 = PLA & SkyD (n = 19, T = 645)			**Int. 2 = PLA & SkyD** (n = 18, T = 631)			**Int. 2 = PLA & SkyD** (n = 12, T = 430)		
	18.5 %	45.1 %	36.4 %	32.5 %	31.7 %	35.8 %	27.2 %	35.4 %	37.4 %

Note. Int. = Intervention, PLA= positive lifestyle advice, SkyD= skydiving. T_B = T ‘Before Int. 1’, T_A = T ‘After Int. 1. The state distributions of the observations in the first two phases (‘Before Int. 1’ and ‘After Int. 1) are presented per intervention type (3 subgroups). The distributions of the observations in the third phase are presented per combination of the first and second intervention (9 subgroups); e.g., the 22 participants who had no first intervention (‘Int. 1 = No Int.’) are divided across ‘Int. 2 = No Int.’ , ‘Int 2. = PLA’, ‘Int. 2 = PLA & SkyD’.

Appendix A

The model is estimated by means of the Expectation Maximization (EM; Dempster et al., 1977) algorithm that uses the so-called *complete-data loglikelihood* ($\log L_c$), i.e., assuming the state assignments of all time-points to be known and thus replacing the difficult maximization by a sequence of easier maximization problems. In the *expectation step (E-step*, see e.g., Bishop, 2006; Dias, Vermunt, & Ramos, 2008), we assume the parameters of interest $\widehat{\boldsymbol{\theta}}$ (i.e., transition probabilities, initial probabilities, and state-specific MMs) to be given (i.e., by a set of initial values or estimates from the previous iteration $\widehat{\boldsymbol{\theta}}^{\mathbf{old}}$, see De Roover et al., 2017; Vermunt & Magidson, 2016) and calculate the posterior probabilities (i.e., conditional on the data) to belong to each of the states and to make transitions between the states, by means of the *forward-backward* algorithm (Baum, Petrie, Soules, & Weiss, 1970). The obtained posterior probabilities are used as expected values of the state assignments to obtain the expected $\log L_c$ ($E(\log L_c)$). Next, in the *maximization step (M-step)*, the parameters $\widehat{\boldsymbol{\theta}}$ are updated such that $E(\log L_c)$ is maximized. The *E-step* (Section A1) and *M-step* (Section A2) are iterated until convergence (Section A3).

A1.E-step

The $\log L_c$ is given by:

$$\begin{aligned}
\log L_c &= \log\left(\prod_{i=1}^{I}\prod_{k=1}^{K}\left[\pi_k^{s_{i1k}}\prod_{t=2}^{T}\prod_{l=1}^{K}a_{ilk}^{s_{it-1,l}\,s_{itk}}\prod_{t=1}^{T}p(\mathbf{y}_{it}|s_{itk})^{s_{itk}}\right]\right) \\
&= \sum_{i=1}^{I}\sum_{k=1}^{K}\left[s_{i1k}\log(\pi_k) + \sum_{t=2}^{T}\sum_{l=1}^{K}s_{it-1,l}\,s_{itk}\log(a_{ilk})\right. \\
&\left. -\frac{1}{2}\sum_{t=1}^{T}s_{itk}\left(J\log(2\pi) + \log(|\boldsymbol{\Sigma}_{\mathrm{k}}|) + (\mathbf{y}_{it}-\boldsymbol{\nu}_k)\boldsymbol{\Sigma}_{\mathrm{k}}^{-1}(\mathbf{y}_{it}-\boldsymbol{\nu}_k)'\right)\right]
\end{aligned} \tag{A1}$$

Since the state memberships are in fact unknown, for each subject and time-point, the expected probability of being in a certain state $\gamma(s_{itk}) = p(s_{itk}|\,\mathbf{Y}_i)$ and the expected probability of the occurrence of two consecutive states $\varepsilon(s_{it-1,l}\; s_{itk}) = p(s_{it-1,l}\; s_{itk}|\mathbf{Y}_i)$ are inserted, in order to obtain $E(\log L_c)$:

$$E(\log L_c) = \sum_{i=1}^{I}\sum_{k=1}^{K}\Bigg[\gamma(s_{i1k})\log(\pi_k) + \sum_{t=2}^{T}\sum_{l=1}^{K}\varepsilon\big(s_{it-1,l}, s_{itk}\big)\log(a_{ilk}) - \frac{1}{2}\sum_{t=1}^{T}\gamma(s_{itk})(J\log(2\pi) + \log(|\boldsymbol{\Sigma}_k|) + (\boldsymbol{y}_{it} - \boldsymbol{\nu}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_{it} - \boldsymbol{\nu}_k)')\Bigg]. \tag{A2}$$

The E-step for a LMC is achieved with the *forward-backward* (or *Baum-Welch*) algorithm (Baum et al., 1970). The algorithm finds the posterior probabilities by making use of the chain rule and the first-order Markov assumption: The joint probability of being in state *k* at time-point *t* and observing the sequence of observations can be expressed as

$$p(s_{itk}, \mathbf{Y}_i) = p(s_{itk}, \mathbf{y}_{i1}, \dots, \mathbf{y}_{it})p(\mathbf{y}_{it+1}, \dots, \mathbf{y}_{iT}|s_{itk}, \mathbf{y}_{i1}, \dots, \mathbf{y}_{it}\,). \tag{A3}$$

The *first-order Markov assumption* implies that we can remove the dependency of the observations at time-point *t* on all previous observations and let them depend only on the state at time-point *t*. Thus, the equation reduces to

$$p(s_{itk}, \mathbf{Y}_i) = p(s_{itk}, \mathbf{y}_{i1}, \dots, \mathbf{y}_{it})p(\mathbf{y}_{it+1}, \dots, \mathbf{y}_{iT}|s_{itk}\,). \tag{A4}$$

The first factor $p(s_{itk}, \mathbf{y}_{i1}, \dots, \mathbf{y}_{it})$ refers to the *forward probabilities* and the second factor $p(\mathbf{y}_{it+1}, \dots, \mathbf{y}_{iT}|s_{itk})$ corresponds to the *backward probabilities*. On the one hand, the forward probabilities, also indicated by $\alpha(s_{itk})$, are the probabilities of observing $\mathbf{y}_{i1:t}$ and to end in state s_{itk} and are calculated by the *forward algorithm*. First, the initial state probabilities are used to calculate the forward probabilities for subject *i* of being in state *k* at time-point 1:

$$\alpha(s_{i1k}) = \pi_k p(\mathbf{y}_{i1}|s_{i1k}). \tag{A5}$$

Then, for the consecutive measurement occasions, we weight the forward probabilities $\alpha(s_{it-1,l})$ of the previous time-point by the corresponding transition probabilities. Next, we sum over all possible ways (i.e., transitions) to come to a specific s_{itk} from any $s_{it-1,l}$ and multiply the values with the corresponding response probabilities:

$$\alpha(s_{itk}) = p(\mathbf{y}_{it}|s_{itk}) \sum_{l=1}^{K} \alpha(s_{it-1,l}) a_{ilk} \,. \tag{A6}$$

On the other hand, the backward probabilities, also indicated by $\beta(s_{itk})$, are the probabilities to be in state s_{itk} and to generate the remaining sequence $\mathbf{y}_{it+1:T}$. Instead of starting at time-point 1, the so called *backward algorithm* starts at time-point *T* and backtracks to time-point *t* + 1. The probability for the backward algorithm to be in final state s_{iTk} and, thus, to produce no outcome ($\emptyset$) anymore is assumed to be 1:

$$\beta(s_{iTk}) = p(\emptyset|s_{iTk}) = 1. \tag{A7}$$

Henceforth, the backward probabilities can be determined in a similar way as the forward probabilities. However, as we go backwards here, we now need to consider the response probabilities of all *K* states at time-point $t + 1$, multiplying them with the backward probabilities $\beta(s_{it+1,l})$ and the respective transition probabilities, prior to summing over *K*:

$$\beta(s_{itk}) = \sum_{l=1}^{K} \beta(s_{it+1,l}) p(\mathbf{y}_{it+1}|s_{it+1,l})\, a_{ilk}. \tag{A8}$$

Next, $p(s_{itk}, \mathbf{Y}_i)$ (Equation A4) is calculated by multiplying the forward and the backward probabilities:

$$p(s_{itk}, \mathbf{Y}_i) = \alpha(s_{itk})\, \beta(s_{itk}) \tag{A9}$$

Subsequently, we can calculate the conditional probability of being in state s_{itk} given the sequence of observations, as we know that:

$$p(s_{itk}, \mathbf{Y}_i) = p(s_{itk} | \mathbf{Y}_i)\, p(\mathbf{Y}_i) \tag{A10}$$

By inserting this into Equation A9, we obtain

$$p(s_{itk} | \mathbf{Y}_i) p(\mathbf{Y}_i) = \alpha(s_{itk})\, \beta(s_{itk}). \tag{A11}$$

It follows that the conditional state probability is equal to

$$p(s_{itk} | \mathbf{Y}_i) = \frac{\alpha(s_{itk})\, \beta(s_{itk})}{p(\mathbf{Y}_i)} = \gamma(s_{itk}). \tag{A12}$$

The denominator can be calculated by marginalizing Equation A9 to $p(\mathbf{Y}_i)$ by summing over the latent states for an arbitrarily chosen *t*:

$$p(\mathbf{Y}_i) = \sum_{k=1}^{K} \alpha(s_{itk})\, \beta(s_{itk}), \tag{A13}$$

which is in its simplest form for *t* = *T*, where $\beta(s_{iTk}) = 1$. Thus, the equation reduces to

$$p(\mathbf{Y}_i) = \sum_{k=1}^{K} \alpha(s_{iTk}). \tag{A14}$$

Finally, we can calculate the joint probability of two successive states by applying Bayes' theorem:

$$p(s_{it-1,l}, s_{itk} | \mathbf{Y}_i) = \frac{p\left(\mathbf{Y}_i \middle| s_{it-1,l}, s_{itk}\right) p\left(s_{it-1,l}, s_{itk}\right)}{p(\mathbf{Y}_i)} \tag{A15}$$

$$= \frac{\alpha\left(s_{it-1,l}\right) p(\mathbf{y}_{it} | s_{itk})\, a_{ilk}\, \beta(s_{itk})}{\sum_{k=1}^{K} \alpha(s_{iTk})} = \varepsilon\left(s_{it-1,l}, s_{itk}\right)$$

A2.M-step

In the M-step, we maximize $E(\log L_c)$ with respect to $\widehat{\boldsymbol{\theta}}$. To maximize, we set the derivatives with respect to the parameters to zero, making use of Lagrange multipliers whenever a constraint, such as $\sum_{k=1}^{K} \pi_k = 1$, needs to be satisfied. The resulting updates are:

$$\pi_k^{new} = \frac{\sum_{i=1}^{I} \gamma(s_{i1k})}{\sum_{k=1}^{K} \sum_{i=1}^{I} \gamma(s_{i1k})} \tag{A16}$$

$$a_{ilk}^{new} = \frac{\sum_{t=2}^{T} \varepsilon\left(s_{it-1,l}, s_{itk}\right)}{\sum_{k=1}^{K} \sum_{t=2}^{T} \varepsilon\left(s_{it-1,l}, s_{itk}\right)} \tag{A17}$$

$$\boldsymbol{\nu}_k^{new} = \frac{\sum_{i=1}^{I} \sum_{t=1}^{T} \gamma(s_{itk}) \boldsymbol{y}_{it}}{\sum_{i=1}^{I} \sum_{t=1}^{T} \gamma(s_{itk})} \tag{A18}$$

The factor models for the state-specific covariance matrices $\boldsymbol{\Sigma}_k^{new} = \boldsymbol{\Lambda}_k^{new} \boldsymbol{\Lambda}_k^{new} + \mathbf{D}_k^{new}$ are estimated by another maximization algorithm within each M-step. Specifically, each observation is weighted by the corresponding $\gamma(s_{itk})$-value, resulting in K weighted data sets $\mathbf{Y}_k$. Fisher scoring (Lee & Jennrich, 1979; Vermunt & Magidson, 2016) is used to perform factor analysis on these weighted data.

A3. Convergence

The convergence can be evaluated either with respect to the $\log L$ or with respect to the parameter estimates. LG applies the latter approach and assesses convergence by computing the following quantity:

$$\delta = \sum_{r=1}^{R} \left| \frac{\hat{\theta}_r^v - \hat{\theta}_r^{v-1}}{\hat{\theta}_r^{v-1}} \right|, \tag{A19}$$

which is the sum of the absolute values of the relative parameter changes where $r = 1, \ldots, R$ refers to the parameters. In this paper, the stopping criterion is $\delta < 1 \times 10^{-8}$. The estimation also stops if the change in the $\log L$ becomes smaller than 1×10^{-10} prior to reaching the stopping criterion above.

Appendix B

```
//LG5.1//
version = 5.1
infile 'Sim1.csv' quote = single


model
options
```

```
   algorithm
      tolerance=1e-008 emtolerance=1e-008 emiterations=3000 nriterations=0;
   startvalues
      seed=0 sets=25 tolerance=1e-005 iterations=100 PCA;
   bayes latent=0;
montecarlo
   seed=0 replicates=500 tolerance=1e-008;
quadrature nodes=10;
   missing  includeall;
//classification added to output
  output
      profile parameters standarderrors estimatedvalues classification probmeans iterationdetails
      WriteParameters = 'results_parameters.csv'
      write = 'results.csv'
      writeloadings='results_loadings.txt';
  outfile
      'classification.csv' classification;

 variables
  caseid id;
  dependent
   V1 continuous, V2 continuous, V3 continuous, V4 continuous, V5 continuous, V6 continuous, V7
continuous, V8 continuous, V9 continuous, V10 continuous, V11 continuous, V12 continuous, V13
continuous, V14 continuous, V15 continuous, V16 continuous, V17 continuous, V18 continuous, V19
continuous, V20 continuous;
 latent
   State nominal dynamic coding=first 2,
   F1 continuous dynamic,
   F2 continuous dynamic;
   independent id nominal;
equations
// factor variances
  (1) F1| State;
  (1) F2| State;
// Markov model
   State[=0] <- 1 ;
   State <- (~tra) 1 | State[-1] id;
//dependent variables determined by state specific Factors
  V1-V20 <- 1 | State + F1 | State + F2 | State;
//Unique variances
  V1-V20 | State;
end model
```

Appendix C

C1. Problem of the Additional Simulation Study

To test whether lag-one autocorrelations of factors not captured by the within-state EFA analyses harm the performance of LMFA, we manipulated factors (a) to (g) as in the main simulation study , leaving out the levels with an already inferior performance, and added an eighth factor (h) specifying the autocorrelation.

C2. Design and Procedure

We crossed eight factors in a complete factorial design[13]:

a. number of factors per state F^k at two levels: 2, 4;
b. number of states K at three levels: 2, 4*;
c. between state difference at two levels:
   medium loading difference & low intercept difference,
   medium loading difference & medium intercept difference*;
d. unique variance e : fixed at .2, .4 ;
e. frequency of transitions between the states at two levels: frequent, infrequent*;
f. number of subjects N at three levels: 2, 10*;
g. number of observations per subject per state T_{ik} at three levels: 50, 100, 200;
h. autocorrelation ϕ at three levels: 0, .3, .7*

The data was generated by means of the orthogonal dynamic factor model which implies that, at time-point *t*, the factors are uncorrelated with one another but a factor's scores at time-point *t* are dependent on its scores at time-point *t*-1. Specifically, they are auto-correlated by the coefficient ϕ as follows:

$$\begin{aligned} \mathbf{y}_{it} &= \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k \mathbf{f}_{it} + \mathbf{e}_{it} \\ \mathbf{f}_{it} &= \phi \mathbf{f}_{it-1} + \boldsymbol{\varepsilon}_{it}, \end{aligned} \tag{C1}$$

where $\boldsymbol{\varepsilon}_{it}$ is a subject-specific $F^k \times 1$ vector of noise at time-point t which is assumed to be multivariate normally distributed with zero mean and the identity matrix as covariance matrix ($\sim MVN(0; I)$). Thus, to generate the subject-specific datasets $\mathbf{Y}_i$, first, the $\mathbf{e}_{it}$ and $\boldsymbol{\varepsilon}_{it}$ vectors were sampled for each observation. Subsequently, we created the autocorrelated factor score vectors $\mathbf{f}_{it}$ by applying a recursive filter (Hamilton, 1994). This filter sets the first noise element as the first

[13] The conditions marked by "*" are the ones that now have less levels than in simulation study 1.

factor score and computes the remaining factor scores as in Equation C1. The resulting factor scores were multiplied by $\sqrt{1-\phi^2}$ to retain an expected variance of 1 (De Roover et al., 2014). Finally, the datasets $\mathbf{Y}_i$ were again merged into one data matrix $\mathbf{Y}$. Note that, for the strength of the autocorrelation (h), we used the values suggested by Cabrieto, Tuerlinckx, Kuppens, Grassmann and Ceulemans (2016). To check how the manipulation played out, we calculated the average autocorrelation across the datasets for each of the three conditions: they amounted to .05, .29 and .68.

For each cell of the factorial design, 20 data matrices **Y** were generated as described above. In total 2 (number of states) × 2 (number of factors) × 2 (between-state difference) × 2 (transition frequency between states) × 2 (number of subjects) × 3 (number of observations per subject and state) × 2 (unique variance) × 3 (autocorrelation) × 20 (replicates) = 5760 simulated data matrices were generated. As in Simulation Study 1, the data was generated in R and analyzed in LG with the same settings and the correct number of states and factors per state.

C3.Results

In general, the recovery was largely unaffected by the autocorrelation conditions (h). Specifically, the recovery of the transition matrices and unique variances was not affected − MAD_{trans} = .05 and MAD_{uniq} = .03 for $\phi = 0, \phi = .3, \phi = .7$ − whereas the recovery of the states, loadings and intercepts was only slightly affected − $ARI = .91$ for $\phi = 0$ and $\phi = .3$ and $.90$ for $\phi = .7$; $GOSL = .99$ for $\phi = 0, \phi = .3$ and $.98$ for $\phi = .7$; $MADint = .07$ for $\phi = 0$, $.08$ for $\phi = .3$ and $.11$ for $\phi = .7$. Note that the mild decrease in intercept recovery with an increased autocorrelation is merely a consequence of the higher variance of the estimated intercepts, since they capture part of the autocorrelation and thus vary more around the population values.