Chronicle of a Measurement Unforetold

Measurement models for intensive longitudinal data

Sebastián Castro Álvarez

Layout: Typeset in LaTeX based on the *MastersDoctoralThesis* template.
Cover and Art Design: Andrea Muñoz Álvarez @vechspot
Printing: Ridderprint, www.ridderprint.nl

university of groningen

Chronicle of a Measurement Unforetold

Measurement models for intensive longitudinal data

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Thursday 22 December 2022 at 9.00 hours

by

Sebastián Castro Álvarez

born on 1 September 1992
in Bogotá, Colombia

Supervisors

Prof. R.R. Meijer
Prof. J.N. Tendeiro

Co-supervisor

Dr. L.F. Bringmann

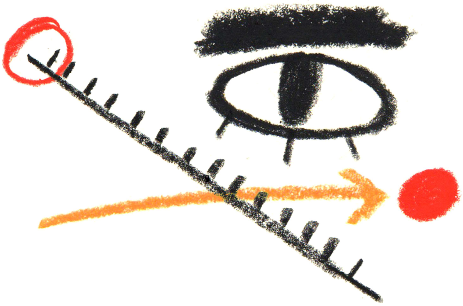
Assessment committee

Prof. C.J. Albers
Prof. A.J. Oldehinkel
Prof. F. Tuerlinckx

Contents

Chapter 1

Introduction



In 2004, the groundbreaking article “A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever” by Molenaar was published. Since then, the interest of many psychological researchers has shifted towards focusing on the individual and the number of studies on psychological dynamics raised dramatically (Bos et al., 2015; Hamaker & Wichers, 2017; Vachon et al., 2019). In short, Molenaar (2004) argued that ergodic processes are rarely observed in psychology. To clarify what an ergodic process is, consider an array of data with three dimensions (individuals, variables, and occasions, also known as Cattell’s data box, 1978). If the process is ergodic, the relationships among the variables between individuals are equivalent to the relationships of those variables within an individual. In other words, if ergodicity holds, one can study a sample of individuals at any occasion or one person during many occasions and draw the same conclusions. Such a process can only be possible under very strict conditions, which are seldom satisfied in real psychological processes (Molenaar, 2004). Therefore, to study how psychological processes unravel at the individual level, one needs to study the psychological dynamics of the individuals.

To being able to study psychological dynamics, researchers collect what is known as time series or intensive longitudinal data. The research methods used to collect intensive longitudinal data are known in the literature with terms such as “ecological momentary assessment”, “experience sampling methods”, and “ambulatory assessment” (Myin-Germeys & Kuppens, 2021; Shiffman et al., 2008; Trull & Ebner-Priemer, 2014). This kind of data implies that an individual (or a sample of individuals) is requested to report on the variables of interest (e.g., negative affect) many times, usually within short periods of time such as one week or one month. Participants might be asked to answer a short questionnaire once every day, or multiple times a day with random prompts, or whenever the behaviour of interest (e.g., smoking) happened.

Furthermore, there are certain advantages of doing intensive longitudinal research when compared to cross-sectional studies, such as: (a) the emphasis on the individuals, (b) the reduction of recall bias, and (c) the higher ecological validity. Firstly, focusing on the individual allows researchers to differentiate between- from within-person variability. This is not the case in cross-sectional research, where the effects of between- and within-person variability are mixed together in the results (Hamaker, 2012). This is especially problematic when the processes of interest are non-ergodic (Hamaker, 2012; Molenaar, 2004) because it may be the case that within-person effects have a completely different direction than the between-person effects. For this reason, being able to differentiate these two sources of variability is helpful to better understand psychological processes. Secondly, the use of self-report questionnaires is widespread in psychological research. However, the response processes of the participants are far from trivial. The final response is influenced, for example, by how the questions are understood, by the questions' format, or by the information that the participant needs to recall to provide an answer (Schwarz, 2012). By using intensive longitudinal methods, the cognitive burden required to answer the questions is reduced because participants are asked to report on their behaviors or feelings at the moment they happened (Schwarz, 2012). As a result, measurements are theoretically more reliable, as the cognitive burden required to answer the questions is lower and measurements should not suffer from a recall bias. Thirdly, intensive longitudinal research offers a higher ecological validity (Reis, 2012) in comparison with research conducted in laboratories because measurements are taken in the natural context of the participants. For all these reasons, intensive longitudinal research allows studying the behaviors and feelings of interest in diverse real-life situations to better grasp the real psychological processes of the persons.

However, in spite of its advantages, intensive longitudinal research comes with its own challenges. In relation to psychological measurement, the question "what are we measuring?" is still the 'elephant' in the room. Firstly, as intensive longitudinal research deals with repeated measurements from the same individuals, one necessarily thinks about the distinction between *traits* and *states*. Broadly, traits are stable dispositions of the individuals and states reflect variability due to the situation. Hence, do intensive longitudinal measurements capture the pure states of the individuals? Or, are the measurements a combination of the traits and the states of the person? Can we distinguish each of these components? Secondly, measurement error is inherent to psychological measurement (Schuurman et al., 2015). Even though measurements in intensive longitudinal research are in theory more reliable as they are collected in the situation and with less recall bias, this is not a guarantee that measurements in intensive longitudinal research are free of measurement error. Thus, key questions are: How can researchers account for measurement error in intensive longitudinal research?, How reliable are intensive longitudinal measurements, and, what are the implications of measurement error in intensive longitudinal research?

1.1 Traits and States

Traits are typically understood as long-lasting and somewhat stable individual characteristics from the persons (Chaplin et al., 1988; Epstein, 1979; Hamaker et al., 2007). In contrast, states are commonly seen as temporary and situational responses that can be observed in a given situation (Chaplin et al., 1988; Hamaker et al., 2007). Yet, evidence of the lack of stability of traits (Burke et al., 1984; Koestner et al., 1992; Roberts & DelVecchio, 2000) led to deep conceptual discussions about the meaning and utility of traits (Allen & Potkay, 1981, 1983; Alston, 1975; Chaplin et al., 1988; Endler & Magnusson, 1976; Hertzog & Nesselroade, 1987; Mischel, 2004; Steyer et al., 1999; Zuckerman, 1983). While some researchers such as Allen and Potkay (1981, 1983) and Mischel (1968, 2004) raised strong criticism about the definition of traits in terms of their arbitrariness and the oversimplification that traits make of the complex reality, others such as Endler and Magnusson (1976) and Steyer et al. (1999) advocated for a more moderate approach that understands psychological concepts as a combination of traits and states.

A first approach used to measure states and traits was to simply rephrase the items to indicate what aspect the researchers were interested in. For example, one can use the term “in general” to refer to trait-like measurements and the term “today” to refer to state-like measurements (e.g., Lubin et al., 2001; Spielberger & Sydeman, 1994; Zuckerman et al., 1983). However, this approach is far more problematic than it may appear. Firstly, adopting this approach implies that traits and states are clearly defined and differentiated by the participants but there is not guarantee that this is the case (Allen & Potkay, 1981; Zuckerman, 1983). Secondly, changing the phrasing of the item may result in the item assessing substantially different experiences (Schwarz, 2012). For example, the interpretation of an item measuring “irritability” might completely change depending on the period time that the question refers to. If the question asks for “today”, participants might answer based on minor annoyances (e.g., the coffee was a bit bitter), but if the question asks for “the last six months”, participants might base their response on major annoyances (e.g., the occasion when the bus was delayed which led to missing a work interview). In this case, the scales are not really capturing the difference between traits and states, but how the person assesses annoying experiences in relation to the time when they happened.

Another approach to the measurement of traits proposed by Epstein (1979, 1981) is the “aggregationist model”. Epstein was concerned about the lack of stability of trait measures, which he attributed to measurement error. To correct for this measurement error, Epstein (1979) suggested to estimate traits by averaging the repeated measurements of an individual. While the aggregationist model is about how to measure traits, it laid the foundation for the conceptual approach towards traits and states that is commonly presumed in intensive longitudinal research (see Hamaker, 2012; Hamaker & Grasman, 2015; Nesselroade, 1991; Nezlek, 2007; Schuurman et al., 2016). Usually, most researchers in intensive longitudinal data assume that an observation from a variable of a subject i at a certain occasion t, y_{it}, can be decomposed into two parts, that is:

$$y_{it} = \mu_i + \delta_{it}, \tag{1.1}$$

where μ_i is the intraindividual mean of the i-th person, which is referred to as the *trait score*, and δ_{it} is the deviation from the intraindividual mean at occasion t, which is referred to as the *state score*. This approach has been extensively used in intensive longitudinal research to study the between- and within-variability (e.g., Hamaker et al., 2007; Moeller et al., 2018; Walls et al., 2006; Zelenski & Larsen, 2000).

Lastly, a framework that was specially proposed for the measurement of traits and states is the so-called latent state-trait theory (LST; Steyer et al., 2015; Steyer et al., 1999). This framework was proposed as a response to the lack of consistency of trait measures and the criticisms raised by Mischel (1968, 2004). The LST theory acknowledges that not only the person but also the situation and the person-situation interaction are relevant sources of variance in the observed scores. In this framework, states, traits, and measurement error are explicit components of the models, which are implemented as structural equation models. In short, in the LST models, the true score is defined as the *latent state variable*, which captures the effects of the person, the situation, and the person-situation interaction. If longitudinal data is available, the *latent state variable* can be decomposed into the *latent trait variable* and the *latent state residual*. The former captures the effects that are stable over time and due to the person, and the latter captures the effects unique to the situation, which are due to the situation and the person-situation interaction. While the LST theory presents a valuable framework for the measurement of traits and states, it has been developed for the analysis of longitudinal data with a few number of waves and it has been rarely used to analyze intensive longitudinal data (see Courvoisier et al., 2010; Eid et al., 2012; Eid et al., 2017; Geiser et al., 2013).

1.2 Measurement Error

Measurement error is a central concept in psychological measurement. Traditionally, in classical test theory (CTT; Crocker & Algina, 1986), the observed score of a test, X, is linearly decomposed into the true score, T, and the error term, E. The true score is defined as the expected score obtained from infinite independent administrations of the test to the same person. The error term is assumed to be unsystematic (i.e., random) and captures the effects of the person, the test, and the situation that are not accounted for by the construct of interest. For example, when assessing the mathematical ability of a person, sources of error may bet the fact that the person was feeling too anxious while answering the test, that many questions were about the topic that the person studied the most, or that the place where the person completed the test was too noisy. Additional assumptions of CTT are that the expected value of the error term is 0, and that the correlation between the random error and the true score is 0.

However, within the context of intensive longitudinal research, key premises of CTT are too restrictive. One of the aims of studying persons intensively is to distinguish between- from within-person variability. Yet, in CTT, the within-person variability is considered to be part of the measurement error (Hamaker, 2012). Also, CTT was developed thinking about measuring traits and the idea that the true score is the expected value of infinite administrations of the test. Associated to these principles, one of the most popular ways to estimate the reliability of a test is by means of the test-retest reliability, but such a definition is not meaningful when doing intensive longitudinal research because we are precisely interested in variability over time. Therefore, there is a need to adjust psychological measurement for intensive longitudinal settings.

For longitudinal settings, the LST theory (Steyer et al., 1999, 2015) offers a probabilistic framework to study psychological measurement which is based on CTT. The most basic LST model capable of distinguishing traits, states, and measurement error is the multistate-singletrait model. In this model, observation Y_{jt} of a variable j (e.g., an indicator or an item) at time t can be decomposed as follows:

$$Y_{jt} = \tau_{jt} + \varepsilon_{jt}, \tag{1.2}$$

where τ_{jt} represents the latent state variable and ε_{jt} represents the random measurement error of variable j at time t. In the LST theory, the measurement error variables at time t are by definition uncorrelated with the latent state variable at time t and the measurement error variables at time u, with $t \neq u$. Then, the latent state variable is further decomposed into:

$$\tau_{jt} = \alpha_{jt} + \lambda_{T_{jt}} \xi + \lambda_{S_{jt}} \zeta_t, \tag{1.3}$$

where α_{jt} is the intercept of the j-th indicator at time t; ξ represent the latent trait variable, which captures the effects of the person; ζ_t represents the latent state residual, which captures the effects of the situation and the person-situation interaction at time t; and $\lambda_{T_{jt}}$ and $\lambda_{S_{jt}}$ represent the factor loadings for the latent trait variable and the latent state residual, respectively. Additionally, the LST theory also defines a set of variance coefficients, which are proportions of the total variance of the observed variable Y_{jt}. The main variance coefficients defined for every LST model are the *reliability*, the *consistency*, and the *occasion-specificity*. These coefficients are particularly useful to study the psychometric properties of the different indicators in the study. Overall, the LST theory seems like a useful framework that can be used to study psychological measurement in intensive longitudinal settings.

Another traditional approach in psychological measurement is the item response theory framework (IRT; Embretson & Reise, 2013; Lord et al., 1968). IRT was developed to overcome several of the limitations of CTT, such as the lack of parameter invariance and the fact that the standard measurement error is assumed to be homogeneous across the sample. In a nutshell, IRT is a nonlinear approach to measurement that allows modeling the interaction between persons and items. In particular, IRT models estimate the probability to endorse an item or a response option given the level on the latent attribute of the person. IRT analyses also define a set of functions such as the item characteristic function, the item information function, and the test information function, which provide an in-depth understanding of the scale and the items. While applications of IRT for longitudinal and intensive longitudinal settings are scarce (e.g., Cai, 2010; Hecht et al., 2019; Kim & Camilli, 2014; Rijn et al., 2010), developing IRT models for intensive longitudinal settings is a promising endeavour that can contribute to improve the measurement of psychological dynamics, because it can inform about the performance and quality of the items and scales used in intensive longitudinal research.

1.3 Outline of the Thesis

Chapter 2 addresses the question about how to distinguish between traits and states in intensive longitudinal data. To do this, we extensively studied three popular longitudinal structural equation models applied to intensive longitudinal settings. The models that we studied were the multistate-singletrait model (MSST; Steyer et al., 2015), the common and unique trait-state model (CUTS; Hamaker et al., 2017), and the trait-state-occasion model (TSO; Eid et al., 2017). While these models are typically used to analyze longitudinal data in wide format, they can be reformulated as multilevel structural equation models to facilitate their application to intensive longitudinal data (Geiser et al., 2013; Hamaker et al., 2017). However, the multilevel version of the models requires additional constraints. For example, in the multilevel version, model parameters cannot vary over time. As a consequence, longitudinal measurement invariance (Meredith, 1993; Meredith & Teresi, 2006) is assumed in the multilevel version. This constraint is not necessary for the standard (single-level) version of the models.

Furthermore, given that the models are or can be encompassed within the LST theory (Steyer et al., 2015; Steyer et al., 1999), in chapter 2, we also studied the psychometric properties of the variables based on the reliability, the consistency, and the occasion-specificity variance coefficients as defined in the LST theory. These coefficients are useful to study how reliable each variable is and to what extent each variable is state- or trait-like. A version of this chapter was published in "Psychological Methods".

Chapter 3 further develops one of the models from Chapter 2. In this chapter, we proposed an extension of the TSO (Eid et al., 2017) to make the model more suitable to analyze intensive longitudinal data. The extended version is named the mixed-effect trait-state-occasion model (ME-TSO). The ME-TSO allows the autoregressive effect to vary across individuals and the inclusion of situational variables. This extension was also based on the LST models for the combination of random and fixed situations approach, which was suggested by Geiser et al. (2015b). In this approach, fixed situations can be defined by characteristics of the situation which are relevant for the research question, such as "being alone" versus "being accompained". Yet, the situations in which a person is alone can vary greatly. We use the term 'random situation' to describe this phenomenon. Hence, the purpose of the LST models for the combination of random and fixed situations is to assess if the overall behaviors or feelings of the participants differ across fixed situations. Additionally, we also defined variance coefficients for the ME-TSO. These variance coefficients allow to study the psychometric properties of the variables per person and per fixed situation. We illustrated how to fit the ME-TSO and how to interpret its results by means of empirical data from the *HowNutsAreTheDutch* (*HoeGekIsNL*) project (van der Krieke et al., 2017; van der Krieke et al., 2016). This chapter was published in "Structural Equation Modeling: A multidisciplinary journal".

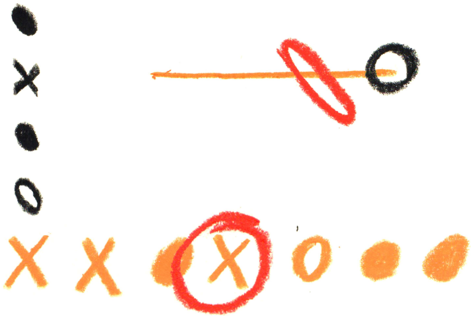
In Chapter 4, we step away from the LST theory (Steyer et al., 2015) and focus on IRT (Embretson & Reise, 2013) instead. A handful of IRT models for intensive longitudinal data have been proposed previously. For example, Rijn et al. (2010) extended the Rasch model and the partial credit model within the state-space modeling framework, Hecht et al. (2019) developed a continuous time Rasch model, and Wang et al. (2013) combined the Rasch model with a growth curve model and a random walk to analyze time series educational data. However, in these previous applications of IRT for intensive longitudinal data, core features of IRT used to study the psychometric properties of the items tend to be ignored. For example, in these previous applications, the authors did not make use of the item characteristic function or the item information function to interpret their results. Because of this, in Chapter 4, we propose an IRT model suitable to analyze time series data while using all the advantages offered within the IRT framework. We refer to this model as the time-varying dynamic partial credit model (TV-DPCM). How to use the TV-DPCM is exemplified by analyzing data openly available in Kossakowski et al. (2017). These data come from one male who filled in a total of 1,473 experience sampling assessments. This Chapter was submitted for publication.

Lastly, in Chapter 5, we considered the topic of goodness-of-fit assessment for the TV-DPCM. Generally, assessing the goodness-of-fit of statistical models is not straightforward, and more research in this area is needed for intensive longitudinal models. Assessing the goodness-of-fit of a model is an important step when analyzing data, to ensure that the model is capable of capturing the main relationships in the data. To achieve this, in Chapter 5, we studied the application of the posterior predictive model checking method (Gelman et al., 1996; Rubin, 1984) to assess the goodness-of-fit of the TV-DPCM. In short, the posterior predictive model checking method is a Bayesian approach that aims to compare features of the observed data with the same features of replicated data. If the differences between the two types of data are too large then there is evidence of model misfit. Based on previous studies of posterior predictive model checking of traditional IRT models, we introduce several test statistics and discrepancy measures that can be used to assess the goodness-of-fit of the TV-DPCM and (potentially) other IRT models for intensive longitudinal data.

The code developed for the different chapters is openly available in various GitHub repositories. For chapter 2, code for the simulation study and for analyzing the empirical data can be found in https://github.com/secastroal/LST_Analyses. For Chapter 3, the code used to analyze the empirical example is available in https://github.com/secastroal/ME-TSO. We do not have authorization to share the empirical data used for Chapters 2 and 3. Finally, code to run the simulations of Chapters 4 and 5, as well as custom functions for the implementation of the posterior predictive model checking method for the TV-DPCM can be found in https://github.com/secastroal/DIRT. Notice that as Chapters 2 through 5 were written as separate articles, some overlap can be found, especially, in the introductory sections of each chapter.

Chapter 2

Using Structural Equation Modeling to Study Traits and States in Intensive Longitudinal Data



This chapter has been published as Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods*, *27*(1), 17–43. https://doi.org/10.1037/met0000393

Abstract

Traditionally, researchers have used time series and multilevel models to analyze intensive longitudinal data. However, these models do not directly address traits and states which conceptualize the stability and variability implicit in longitudinal research, and they do not explicitly take into account measurement error. An alternative to overcome these drawbacks is to consider structural equation models (state-trait SEMs) for longitudinal data that represent traits and states as latent variables. Most of these models are encompassed in the Latent State-Trait (LST) theory. These state-trait SEMs can be problematic when the number of measurement occasions increases. As they require the data to be in wide format, these models quickly become overparameterized and lead to non-convergence issues. For these reasons, multilevel versions of state-trait SEMs have been proposed, which require the data in long format. To study how suitable state-trait SEMs are for intensive longitudinal data, we carried out a simulation study. We compared the traditional single level to the multilevel version of three state-trait SEMs. The selected models were the multistate-singletrait (MSST) model, the common and unique trait-state (CUTS) model, and the trait-state-occasion (TSO) model. Furthermore, we also included an empirical application. Our results indicated that the TSO model performed best in both the simulated and the empirical data. To conclude, we highlight the usefulness of state-trait SEMs to study the psychometric properties of the questionnaires used in intensive longitudinal data. Yet, these models still have multiple limitations, some of which might be overcome by extending them to more general frameworks.

***Keywords**: states and traits, intensive longitudinal data, longitudinal structural equation modeling, measurement error*

The amount of intensive longitudinal studies has increased considerably in the last 20 years (Hamaker & Wichers, 2017). These studies, also known as *ambulatory assessments*, *experience sampling methods* (ESMs), or *ecological momentary assessments* (EMAs), include multiple measurements per person. These measurements are collected in short periods of time and are intended to study the dynamics of psychological processes. Several advantages of intensive longitudinal methods are often highlighted, such as the emphasis on the individual, the higher ecological validity, and the diminution of recall bias; these properties are discussed in depth by Hamaker (2012), Reis (2012), and Schwarz (2012).

When studying intensive longitudinal data, the terms *traits* and *states* often occur, as these terms conceptualize the stability and variability of human behaviors and attitudes across time. Although traits and states have been defined differently throughout the years (Allen & Potkay, 1981; Chaplin et al., 1988; Mischel, 2004; Steyer et al., 1999), traits are typically regarded as relatively stable dispositions and states as the observed variability due to situational conditions (Hamaker et al., 2017; Hamaker et al., 2007). Moreover, some authors have suggested that most psychological variables are not pure traits nor states but they are something "in-between" (Geiser et al., 2017; Hertzog & Nesselroade, 1987).

In order to study states and traits in intensive longitudinal data, researchers have found inspiration in the *aggregationist model* (Epstein, 1979, 1981). In short, the aggregationist model assumes that the mean score over time for each individual (i.e., intra-individual mean) is a good estimate of a person's trait level. Although it is not explicitly mentioned, this conceptual approach has been integrated with the statistical approaches that are generally used to analyze intensive longitudinal data. This includes *time series analysis* when $N = 1$ (Fan & Yao, 2003; Hamaker & Dolan, 2009) and *multilevel modeling* when $N > 1$ (Houben et al., 2020; Nezlek, 2012; Walls et al., 2006). In these approaches, the intra-individual means are considered as trait scores and the deviations from that mean are considered as state scores (e.g., Hamaker & Grasman, 2015; Nesselroade, 1991; Nezlek, 2007; Schuurman et al., 2016). Time series analyses and multilevel models have been extensively used to study traits and states in intensive longitudinal data (e.g., Moeller et al., 2018; Zelenski & Larsen, 2000).

However, the aggregationist approach is not free of criticism. According to Steyer et al. (1992), in the aggregationist model, the situational context is not an integral part of the model. This criticism also applies to some extent to the way traits and states are handled in multilevel models. As the situation is not an explicit part of the model, the state scores are confounded with other sources of variability such as measurement error and other unobserved variables. Moreover, the definitions of traits and states are rather loose, mainly because multilevel models were not especially developed to handle these concepts.

To overcome these drawbacks, longitudinal state-trait structural equation models (i.e., state-trait SEMs) can be used as an alternative approach to analyze intensive longitudinal data. State-trait SEMs are especially suited as (a) their main focus is on differentiating the trait and the state components of the observed variables and (b) they incorporate measurement error. However, applying state-trait SEMs to intensive longitudinal data is not straightforward (Geiser et al., 2013; Geiser et al., 2015b; Nezlek, 2007) because these models have been originally developed for longitudinal data with a limited number of measurement occasions. Furthermore, even though some of these models are supposed to be useful to analyze experience sampling data (Courvoisier et al., 2010; Eid et al., 2012; Eid et al., 2017), there are little to no applications of these models on intensive longitudinal data.

The goal of this study was to explore the suitability of state-trait SEMs to study states and traits in intensive longitudinal data. By doing this, we also provide a comprehensive introduction to these models, which will allow more researchers to apply them to their intensive longitudinal data. To reach this goal, we compared three state-trait SEMs through a simulation study, and we applied them to empirical data. The three models that we considered were the multistate-singletrait (MSST; Geiser et al., 2013; Steyer et al., 2012; Steyer et al., 2015), the common-unique trait-state (CUTS; Hamaker et al., 2017), and the trait-state-occasion model (TSO; Cole et al., 2005; Eid et al., 2017). These models were chosen because all of them require multiple indicators, which is useful when multiple items are used to measure the same construct. Moreover, the selected models are a good representation of the wide variety of state-trait SEMs, as they cover the most general aspects that are commonly accounted for in state-trait SEMs, as we will highlight in this study.

We also discuss important characteristics of state-trait SEMs that have to be addressed when analyzing intensive longitudinal data. For example, we show how state-trait SEMs can be reformulated as multilevel SEMs allowing researchers to deal with a large number of measurement occasions per person. Moreover, we discuss how state-trait SEMs allow studying the psychometric properties of the items included in intensive longitudinal studies. Currently, a multilevel version of the TSO model is lacking, hence we also introduce the multilevel version of this model. Finally, we also show how these models can be easily adjusted to the requirements of the data, such as in the case of multidimensional settings.

In what follows, we explain the theoretical and statistical background of state-trait SEMs. This includes (a) a brief overview of the *latent state-trait theory* (LST; Steyer et al., 2012; Steyer et al., 2015; Steyer et al., 1999), (b) a discussion of the difficulties of using these models with intensive longitudinal data and how they can be overcome by reformulating the models as multilevel SEMs, (c) a detailed description of the models selected for this study, and (d) an explanation of the similarities and differences among the selected models and other models and frameworks. Next, we present the simulation study and its results. After this, the empirical example is introduced by describing and analyzing the data with the three models. This empirical example aims to show practitioners how to analyze and interpret real experience sampling data with state-trait SEMs. Finally, we conclude with a discussion of our findings, and suggestions for future research on this field. Furthermore, in order to help researchers apply these models to their own data, in Appendix A, we provide R code to fit these models (also see https://github.com/secastroal/LST_Analyses).

2.1 State-Trait SEMs

State-trait SEMs are longitudinal structural equation models, which can distinguish the trait and the state components of the observed variables. The observed variables are usually referred to as indicators in the state-trait SEM literature, which are expected to be measured on a continuous scale. They can be, for example, the sumscores of parallel tests or the items' raw scores. Furthermore, an important assumption of state-trait SEMs is that measurement time points are equally spaced over time (Geiser et al., 2013), meaning that the time elapsed between two consecutive observations is the same for all consecutive observations. Finally, an additional assumption that should be considered when analyzing intensive longitudinal data with state-trait SEMs is stationarity. A time series is (weakly) stationary if its means and variances-covariances are equal over time (Song & Zhang, 2014).

Regarding the state-trait distinction, in the state-trait SEMs, not only persons but also situations and person-situation interactions are considered important sources of variability. For example, a person's anxiety is certainly not stable over time, but highly influenced by the situation: How anxious a person feels will be completely different at a romantic dinner than at a job interview. Also, anxiety will not only be affected by the situation but also by the person-situation interaction. For instance, the effect of a stressful situation on a person who has an anxiety disorder will be different from the effect of that same situation on a person who does not have an anxiety disorder.

These sources of variability are captured by the latent variables in state-trait SEMs. Thus, a latent variable is needed to represent the effect of both the situation and the person-situation interaction for each occasion of measurement (latent states), and a general latent variable is needed to represent the stability of the measurements per person across situations (latent trait). In other words, state-trait SEMs aim to describe the structure of the constructs measured in longitudinal studies by differentiating the proportion of the variance that is explained either by trait effects or situational effects (Steyer et al., 1999). Additionally, state-trait SEMs are useful to understand how the trait and the state components of different constructs are related and whether these relationships are different among different subpopulations (Steyer et al., 1999).

An important characteristic of state-trait SEMs is that they are useful to analyze longitudinal data in wide format (Hamaker et al., 2017; Steyer et al., 2015). This means that there are as many columns in the data set as there are measurement occasions for each variable of interest. Therefore, parameters can be allowed to change over time. Technically speaking, state-trait SEMs can be used to test for longitudinal measurement and factorial invariance (Meredith, 1993; Meredith & Teresi, 2006) of the trait and state structures. Longitudinal measurement invariance means that the conditional distribution of the observed variable given a value of the latent variable is equal over measurement occasions. Put differently, assuming longitudinal measurement invariance means that the construct of interest (e.g., positive affect) is measured equally over time (Geiser et al., 2015a). Hence, the latent scores at different occasions can be interpreted in the same way, which allows for meaningful comparisons of latent scores over time. On the other hand, if measurement invariance is not assumed, differences between latent scores over time might be due to changes in the meaning of the scale and not necessarily due to changes in the persons' attributes (i.e., the construct being measured).

Despite their similarities, state-trait SEMs differ in many ways, such as the number of indicators (i.e., variables) needed to identify the latent state variables, or the kind of effects that are accounted for (e.g., autoregressive effects, method effects, trait change). For example, the stable trait autoregressive trait and state model (STARTS; Kenny & Zautra, 1995; Kenny & Zautra, 2001) is one of the few models that only requires one indicator on each occasion to be able to measure a situational latent variable, while most state-trait SEMs require multiple indicators. In addition, state-trait-SEMs might include *method effects* in order to account for the specific time-invariant component of an indicator that is not shared with other indicators. For instance, these effects might be present if a variable was measured through self-report and peer-rated questionnaires. In this case, each questionnaire might have its own effect, which is independent of the situation and unique to each questionnaire. A state-trait SEM that includes this kind of effects is the CUTS model (Hamaker et al., 2017), where method effects are referred to as *unique traits*. Additionally, there are also other approaches to account for method effects, which are broadly discussed by LaGrange and Cole (2008) and Geiser and Lockhart (2012).

Moreover, some state-trait SEMs account for *autoregressive effects* (Cole et al., 2005; Eid et al., 2017; Kenny & Zautra, 1995; Kenny & Zautra, 2001; LaGrange & Cole, 2008). Autoregressive effects are used to study the dependency of the variables on themselves at previous time points. In other words, an autoregressive effect evidences how an observation of a variable at time t can be affected by previous observations of that same variable at time $t-1$, $t-2$, and so on. For example, if a person got a raise in their salary, this would make him/her feel more cheerful during the next few days, in such a way that the person might react more positively to future unpleasant experiences. The STARTS (Kenny & Zautra, 1995; Kenny & Zautra, 2001) and the TSO (Cole et al., 2005; Eid et al., 2017) models are examples of state-trait SEMs that include autoregressive effects to account for the relationship between consecutive occasions. These effects are especially important when studying intensive longitudinal data given the short intervals of time in which observations are collected, as this makes autoregressive effects more likely to occur.

2.1.1 Latent State-Trait Theory

Among the wide variety of state-trait SEMs that have been proposed, it is important to highlight the so-called latent state-trait theory (LST; Steyer et al., 2012; Steyer et al., 2015; Steyer et al., 1999). This theory encompasses several of the state-trait

SEMs while also providing a probabilistic framework to study states and traits. In this probabilistic framework, the LST theory first defines what is the random experiment of interest, which is, in other terms, the set of possible outcomes (Steyer et al., 2015). The random experiment is defined in four steps: (a) A person is sampled at time 0, (b) the person has certain experiences between the moment of sampling and the assessment, (c) the person is in a certain situation at the moment of assessment, (d) the behavior of interest is observed. The last three steps (b through d) are repeated for $t > 1$. Thus, the LST theory acknowledges the dynamic nature of persons because a person at time t is different from the person at time $t+1$ due to the experiences the person has had between consecutive measurement occasions (Steyer et al., 2015).

Secondly, the LST theory defines all the random variables involved in the random experiment. These random variables are the persons at time t (U_t), the situations at time t (S_t), the observed variables j at time t (Y_{jt}), and the conditional expectations of the observed variables given the persons or the situations (e.g., $E[Y_{jt}|U_t]$). These conditional expectations are at the basis of the formal definitions of the latent variables in the model (Steyer et al., 2015). Models encompassed by the LST theory are, for example, the MSST (Steyer et al., 2015) and the TSO (Eid et al., 2017) models.

In addition, the LST theory defines three variance coefficients as proportions of the total variance: The *consistency*, the *occasion-specificity*, and the *reliability* (Steyer et al., 2015). The consistency is the proportion of the variance due to the sources of variability that are stable over time (trait component). The occasion-specificity is the proportion of the variance due to the sources of variability that depend on the situation (state component). The consistency and the occasion-specificity indicate to what extent the observed variables are trait- or state-like, respectively. Lastly, the reliability is the sum of the consistency and the occasion-specificity. In other words, the reliability is the proportion of the variance explained by both, stable and situational sources of variability in the given situation. The proportion of the variance that is not captured by the reliability coefficient is the random measurement error. The consistency, the occasion-specificity, and the reliability are computed for each observed variable Y_{jt}, which means that the coefficients of variable j at time t might be different from the coefficients of the same variable j at time u, for $t \neq u$. Allowing these coefficients to change over time might be interesting if there are clear trends. For example, if a researcher observes that the reliability of an item decreases over time, further inspection of that item would be required.

2.1.2 State-Trait SEM as Multilevel SEM

Although some authors have suggested state-trait SEMs (single-level state-trait SEMs from here on) as a way to analyze intensive longitudinal data (Courvoisier et al., 2010; Eid et al., 2012; Eid et al., 2017), others have recognized that these models become impractical when there are many measurement occasions (Geiser et al., 2013; Geiser et al., 2015b; Nezlek, 2007). In fact, studies that use these models and have intensive longitudinal data at hand consistently use fewer measurement occasions than the ones available due to the requirement of having the data in wide-format. In general, single-level state-trait SEMs are problematic when the number of measurement occasions increases as these models were developed for studies with a limited number of occasions. For instance, if a study measured three variables once a day during ten consecutive days, there will be thirty observed variables in the model. However, when the data are in long format, the number of observed variables in the model stays at three. Therefore, applying single level state-trait SEMs to intensive longitudinal data is difficult because the number of observed variables in the model becomes too large. As a consequence, the syntax to specify the models is extremely long and convergence issues are more likely to happen (Geiser et al., 2013).

To overcome these limitations, state-trait SEMs can be formulated as multilevel SEMs (e.g., Geiser, 2020; Geiser et al., 2013), which requires the data in long format. A multilevel SEM for longitudinal data defines two levels of analysis: The within-level (level 1) and the between-level (level 2). The within-level aims to model the variability within persons and the between-level aims to model the differences between persons. Thus, to specify state-trait SEMs as multilevel SEMs, one has to model the effects of the situations and person-situations interactions as well as the random measurement error at the within-level. Furthermore, one has to model the effects that are independent of the situation, such as the effects of the persons, at the between-level (Geiser et al., 2013). In particular, multilevel state-trait SEMs have as many observed variables as there are indicators in the data, independently of the number of measurement occasions. Considering the previous example, in a study with three indicators and ten measurement occasions, there will be only three observed variables in a multilevel setting, one for each indicator.

Additionally, in the multilevel formulation of state-trait SEMs, all parameters (loadings, intercepts, variances, residual variances, and variance coefficients) are constrained to be equal over time. This restriction implies that some sort of longitudinal measurement and factorial invariance (Meredith, 1993; Meredith & Teresi, 2006) of the trait and state structure is assumed in the multilevel formulation of state-trait SEMs. This can come as a disadvantage because the assumption of measurement invariance cannot be tested. In general, researchers should be cautious when assuming measurement invariance without testing for it. For example, Geiser et al. (2015a) showed in a simulation study that when real data were characterized by a trait change process (e.g., growth curve model), state-trait SEMs assuming measurement invariance might result in well-fitting solutions. In these cases, the trait change process might be masked and the researcher might conclude that there is no trait change at all.

In conclusion, the main advantages of formulating state-trait SEMs as multilevel SEMs are that many measurement occasions are easier to handle, there are fewer parameters to estimate, and the syntax of the model remains unchanged regardless of the number of measurement occasions included in the study. Moreover, the multilevel formulation allows to easily handle missing data (Geiser, 2020; Geiser et al., 2013). However, multilevel state-trait SEMs do not allow testing for longitudinal measurement invariance, which can be tested in single-level state-trait SEMs. Thus, the previous example about observing the reliability of a variable decreasing over time would not be possible in a multilevel setting because the reliability is constrained to be equal over time.

In relation to the models of interest in this study, the multilevel version of the MSST and the CUTS models have been proposed by Geiser et al. (2013) and by Hamaker et al. (2017, Supplementary material), respectively. The multilevel CUTS model is technically a multilevel confirmatory factor analysis (see Roesch et al., 2010). However, there is, to the best of our knowledge, no multilevel version of the TSO model; below, we present such a model.

2.1.3 The Models

In this section, we explain in detail the models that were selected for this study: The MSST, the CUTS, and the TSO models. As mentioned before, these models were selected because all of them require multiple indicators to identify the latent state variables. Therefore, they are useful when multiple items or variables are used to measure one unique construct (e.g., positive affect). Additionally, these models cover the most general characteristics that are studied in state-trait SEMs. In particular, the MSST model is at the foundations of the LST theory. It is the simplest model that allows differentiating states from traits. The CUTS model accounts for method effects straightforwardly and intuitively, while still differentiating between traits and states. Finally, the TSO model takes into account autoregressive effects to study the relation between consecutive measurement occasions, a feature which has been extensively considered when studying intensive longitudinal data. We cover both the single-level and multilevel version of each model. This section is complemented with Appendix A, which provides R code to simulate data based on these models and to fit both versions of each model in Mplus within the R environment.

The multistate-singletrait model (MSST)

The most straightforward model within the LST framework that allows distinguishing between traits and states is the MSST model. Here, we use the version of the MSST model from Geiser and Lockhart (2012). This version does not include the latent trait variable as a second-order factor as it is usually done in the LST theory (Steyer et al., 2015); instead, it is included as a general first-order factor. In general, the MSST model is useful to identify the variance proportions of the indicator variables that are due to either the trait component, the state component, or the random measurement error.

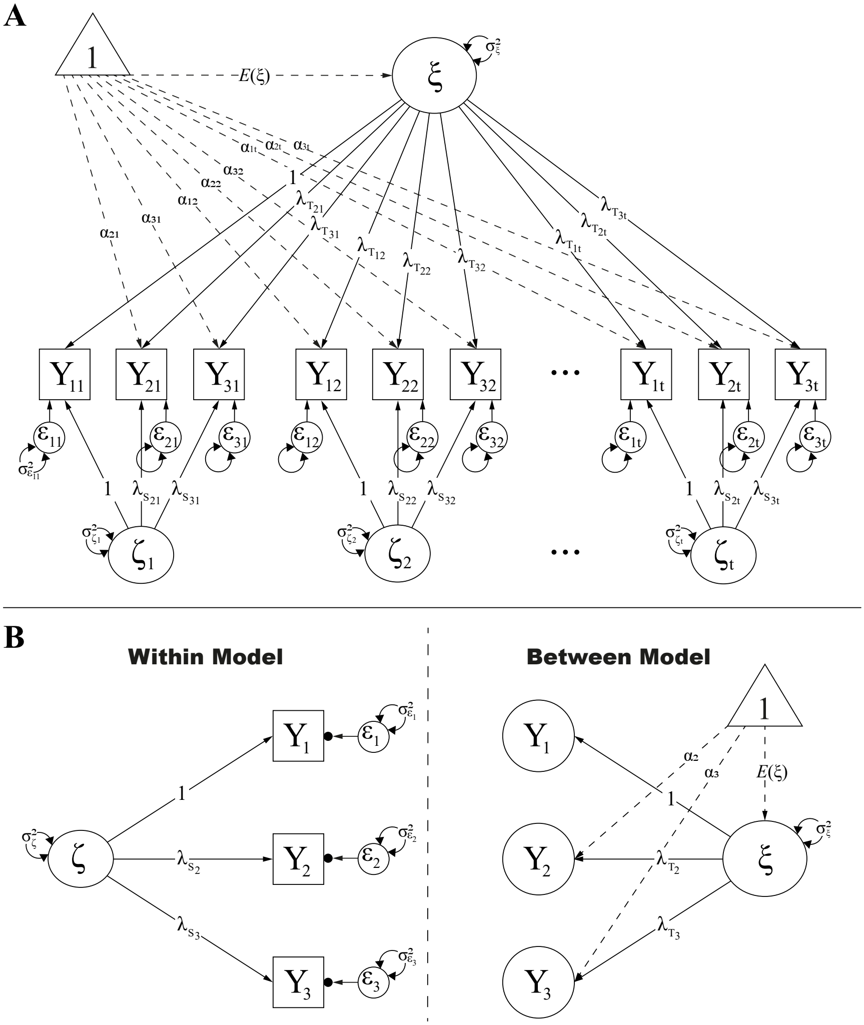
To start, the MSST model requires a set of observed indicator variables Y_{jt} ($j =$ indicator, $t =$ time), which are measured on multiple occasions and aim to measure the same construct (e.g., anxiety, positive affect, etc.) from a sample of size n. Each variable Y_{jt}, which represents an n-variate vector of the sample's responses on the $j-th$ indicator, is decomposed as follows:

$$Y_{jt} = \tau_{jt} + \varepsilon_{jt}. \tag{2.1}$$

where τ_{jt} is an n-variate vector of factor scores of the latent state variable and ε_{jt} is an n-variate vector that captures the deviations of the observed scores from the factor scores of the latent state variable, namely, the random measurement error. The random measurement error variables ε_{jt} are assumed to be normally distributed with mean zero, $\varepsilon_{jt} \sim N(0, \sigma^2_{\varepsilon_{jt}})$. Then, the latent state variable is further decomposed as follows:

$$\tau_{jt} = \alpha_{jt} + \lambda_{T_{jt}}\xi + \lambda_{S_{jt}}\zeta_t, \tag{2.2}$$

Figure 2.1: Path Diagram of the MSST Model



Note. (A) Single level MSST. (B) Multilevel MSST.

where α_{jt} is an n-variate vector with the intercept of indicator j at time t, ξ is an n-variate vector of the factor scores of the latent trait variable, which is assumed to be stable over time, and ζ_t is an n-variate vector that captures the deviations of the latent state variable from the latent trait variable at time t, namely, the latent state residual. The latent state residuals ζ_t reflect both the effects of the situation and the person-situation interaction. Moreover, as well as the random measurement error, the latent state residuals are assumed to be normally distributed with mean zero, $\zeta_t \sim N(0, \sigma^2_{\zeta_t})$. Lastly, $\lambda_{T_{jt}}$ and $\lambda_{S_{jt}}$ are the factor loadings for the latent trait variable and the latent state residual, respectively. Next, by replacing Equation 2.2 into Equation 2.1, the full decomposition of the observed variables given the MSST model can be written as follows:

$$Y_{jt} = \alpha_{jt} + \lambda_{T_{jt}} \xi + \lambda_{S_{jt}} \zeta_t + \varepsilon_{jt}. \tag{2.3}$$

In addition, it is also assumed that the latent trait variable ξ, the latent state residuals ζ_t, and the random measurement error variables ε_{jt} are uncorrelated with each other. The MSST model is identified by fixing one factor loading parameter of the latent trait variable and of each latent state residual to 1. Thus, one factor loading parameter $\lambda_{T_{jt}}$ and the factor loading parameters $\lambda_{S_{jt}}$ associated to one of the indicators Y_j have to be equal to 1. Moreover, one of the intercepts α_{jt} of the same indicator Y_j has to be fixed to 0.

The path diagrams of the single-level and the multilevel MSST models are presented in Figure 2.1. Following the most common way of representing SEMs, observed variables are presented as squares and latent variables are presented as circles. Linear relationships between variables, either observed or latent, are shown through straight arrows and covariances between two variables are shown as curved arrows. The variance of a variable is represented with a looping arrow. Moreover, random measurement error variables and residual variables are represented as latent variables with their own variances. Finally, intercepts and means are represented as the effects of a constant variable 1 (triangle) on the observed or latent variables. Effects and covariances that are set to 0 are not shown in the path diagrams. In addition, in the multilevel SEM representation of the models, a path diagram is used for each level: The within-level model and the between-level model. The path diagrams of each model follow the same conventions mentioned above, the only additional element is a dot at the end of the straight arrows that come from the random measurement error variables. These dots aim to represent the intraindividual means of all individuals on the specified observed variable. These intraindividual means per variable are then represented as latent variables and used to fit the between-level model. Note that in the multilevel SEM diagram, the subscripts t are dropped due to the required assumption that parameters are time-invariant.

The common-unique trait-state model (CUTS)

This model proposed by Hamaker et al. (2017) is actually equivalent to the MSST model with orthogonal methods (Steyer et al., 1992 as cited in Geiser and Lockhart, 2012). Even though it was first proposed within the LST framework, it is not considered a proper LST model because the additional method factors can not be defined within the probabilistic principles of the LST theory (Geiser & Lockhart, 2012, Appendix A). According to the reasoning of Hamaker et al. (2017), the CUTS model is developed based on the *averaged R-technique* and the *pooled P-technique analysis* (Cattell, 1963) and its goal is to distinguish between four sources of variance, namely, the common states, the common traits, the unique states, and the unique traits. The CUTS model is also useful to test whether there is weak factorial invariance between the trait factor structure and the state factor structure (Hamaker et al., 2017). If this assumption holds, it means that the factor structure of the construct of interest obtained in a cross-sectional study adequately describes both the between-factor and the within-factor structures.

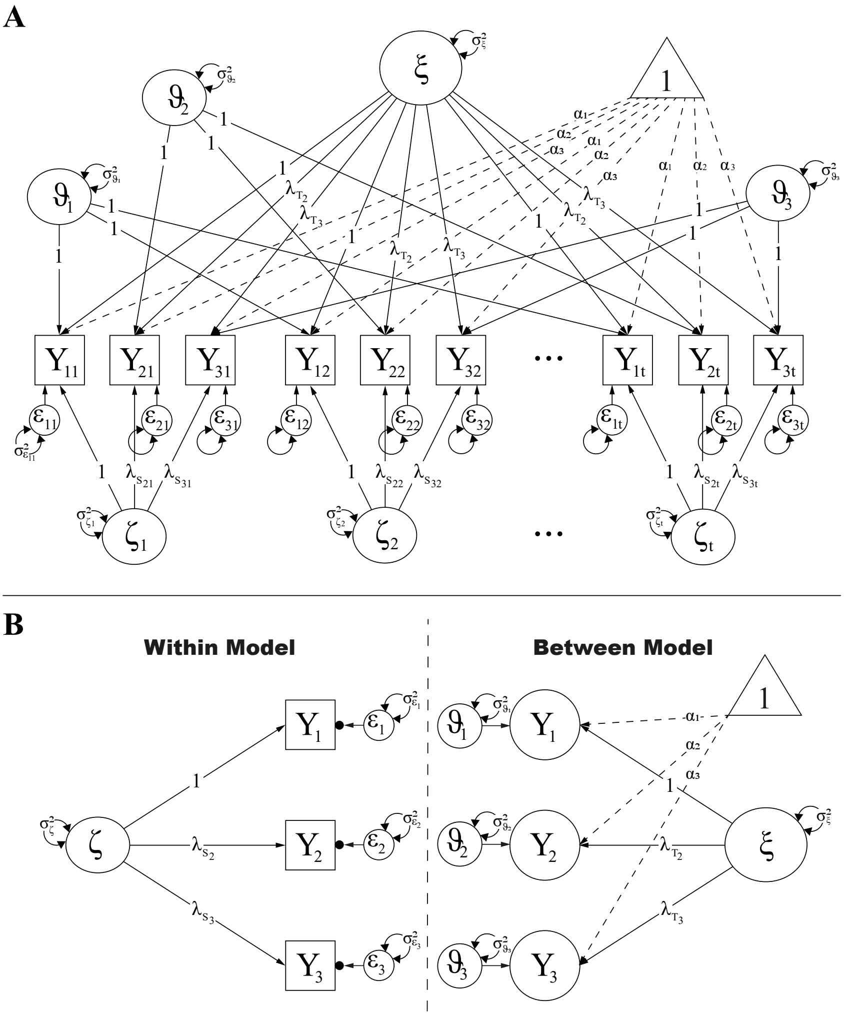
Just like the MSST model, the CUTS model requires a set of observed indicator variables Y_{jt}, which are measured on multiple occasions on a sample of size n and aim to measure the same construct. Each observed indicator variable is an n-variate vector with the observed scores of the variable j at time t. Next, in the CUTS model, the observed variables Y_{jt} are decomposed into *trait-scores* and *state-scores* as shown in Equation 2.4:

$$Y_{jt} = T_j + S_{jt}, \tag{2.4}$$

where T_j is an n-variate vector of the intraindividual means of indicator j, which are time-invariant, and S_{jt} is an n-variate vector of the deviations of the observed scores Y_{jt} from the intraindividual means of indicator j at time t. Then, the trait-scores are further decomposed as follows:

$$T_j = \alpha_j + \lambda_{T_j}\xi + \vartheta_j, \tag{2.5}$$

Figure 2.2: Path Diagram of the CUTS Model



Note. (A) Single level CUTS with time varying intercepts α_{jt} and trait loadings $\lambda_{T_{jt}}$. (B) Multilevel CUTS.

where α_j is an n-variate vector with the grand mean of indicator j, ξ is an n-variate vector of the common trait scores, which represents what is common to all indicators over all time points, λ_{T_j} is the factor loading relating the trait-scores T_j to the common trait scores ξ of indicator j, and ϑ_j is an n-variate vector that contains the part of the trait-score that is not accounted for by the common trait scores and therefore is unique to indicator j (unique trait scores). Both the common trait ξ and the unique trait ϑ_j are assumed to be normally distributed with mean zero and variances σ^2_ξ and $\sigma^2_{\vartheta_j}$, respectively. Moreover, the intercepts α_j and the loadings λ_{T_j} can be allowed to vary over time to test for longitudinal measurement invariance.

Similarly, the state-scores are decomposed as follows:

$$S_{jt} = \lambda_{S_{jt}} \zeta_t + \varepsilon_{jt}, \tag{2.6}$$

where ζ_t is an n-variate vector of the common state scores representing what is common to all indicators at time t, $\lambda_{S_{jt}}$ is the factor loading relating the state-scores S_{jt} to the common state scores ζ_t of indicator j at time t, and ε_{jt} is an n-variate vector that contains the part of the state-score that is not accounted for by the common state scores and therefore it is unique to indicator j at time t (unique state scores). Similarly to the common and unique trait scores, the common state ζ_t and the unique state ε_{jt} are assumed to be normally distributed with mean zero and variances $\sigma^2_{\zeta_t}$ and $\sigma^2_{\varepsilon_{jt}}$, respectively.

The four sources of variance distinguished by the CUTS model are clearly evident when the full model is written in one equation by replacing Equations 2.5 and 2.6 into Equation 2.4 as follows:

$$Y_{jt} = \alpha_j + \lambda_{T_j}\xi + \vartheta_j + \lambda_{S_{jt}}\zeta_t + \varepsilon_{jt}. \tag{2.7}$$

Here, each observation is decomposed into an intercept and four components that are linked to four sources of variability. The first source of variability is the common trait ξ that is invariant over time and variables. The second source is the unique trait ϑ_j that is invariant over time but specific to indicator j and can be interpreted as systematic error. The third source is the common state ζ_t that is common over all the indicators but specific at time t. Finally, the last source of variability is the unique state ε_{jt} that is specific to each indicator and each time point and can be interpreted as random measurement error. This way of disentangling different sources of error variance is similar to the way this is done in generalizability theory (Brennan, 2005).

In the CUTS model, the four sources of variability are independent of each other, which means that common traits, unique traits, common states, and unique states are uncorrelated. The single-level CUTS model can be just-identified with as little as two measurement occasions and three indicator variables when setting one loading of the common trait variables and of each of the common state variables to 1. The path diagrams of the single-level and multilevel versions of the CUTS model are presented in Figure 2.2.

The trait-state-occasion model (TSO)

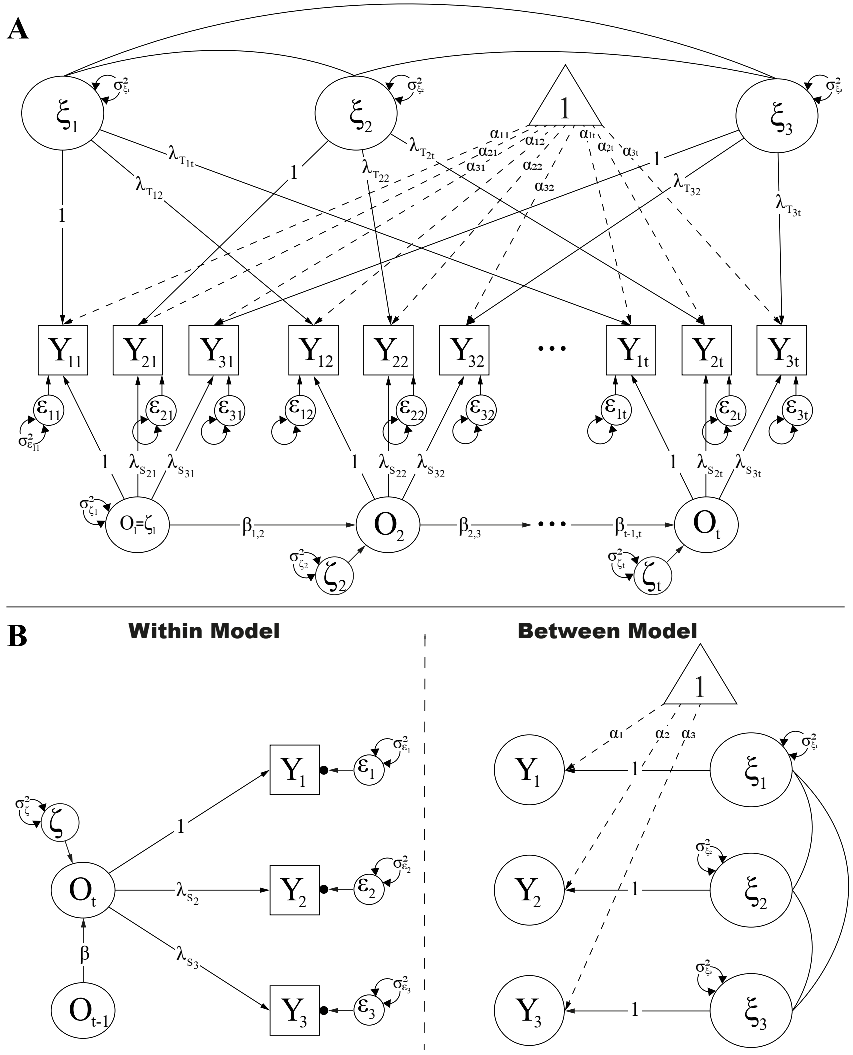
The TSO model was developed in order to account for autoregressive effects within the LST framework. It was first proposed by Cole et al. (2005), inspired by the STARTS model (Kenny & Zautra, 1995; Kenny & Zautra, 2001) and the autoregressive LST model (Steyer & Schmitt, 1994). Also, it was recently reintroduced by Eid et al. (2017) to argue that it is effectively an LST model as all the latent variables are well defined as conditional expectations. The main innovation of the TSO model is that it introduces autoregressive effects among the latent state residuals. This creates two new sets of variables: The occasion-specific variables O_t and the occasion-specific residuals ζ_t. Thus, the TSO model is useful to describe how the trait and the state components of a construct are associated with the indicator variables while controlling for the carry-over effect that might be present between consecutive measurement occasions.

In this study, we use the version of the TSO proposed by Eid et al. (2017, Figure 3B) that does not consider the latent trait variable as a second-order factor, and that uses latent trait-indicator variables instead of a common latent trait variable for all the indicators. The TSO model first decomposes the observed variables Y_{jt} as it is done in the MSST model as in Equation 2.1. The difference is expressed in the decomposition of the latent state variables τ_{jt} as follows:

$$\tau_{jt} = \alpha_{jt} + \lambda_{T_{jt}} \xi_j + \lambda_{S_{jt}} O_t, \tag{2.8}$$

where α_{jt} is an n-variate vector with the intercept of indicator j at time t, ξ_j is an n-variate vector that represents the latent trait-indicator variable of indicator j, O_t is an n-variate vector that represents the latent occasion-specific variable at time t, and

Figure 2.3: Path Diagram of the TSO Model



Note. (A) Single level TSO. (B) Multilevel TSO.

$\lambda_{T_{jt}}$ and $\lambda_{S_{jt}}$ are the factor loadings of the latent trait-indicator variable ξ_j and the latent occasion-specific variable O_t, respectively.

Next, the latent occasion-specific variable at time O_t is regressed on the latent occasion-specific variable at the previous time O_{t-1}. In other words, the TSO model assumes that the effects of the situation and the person-situation interaction are carried-over to future situations, which is represented by the autoregressive effect between consecutive latent occasion-specific variables. This dynamic process has two main implications: (a) It is not possible to regress the latent occasion-specific variable O_1 on any other latent occasion-specific variable, therefore, $O_1 = \zeta_1$ and (b) the latent occasion-specific variable O_t is a linear combination of all the occasion-specific residuals ζ_s for $s = 1, ..., t$. Equations 2.9 and 2.10 show how to define the latent occasion-specific variables O_2 and O_3 in terms of the latent occasion-specific residuals ζ_t:

$$O_2 = \beta_{1,2}O_1 + \zeta_2 = \beta_{1,2}\zeta_1 + \zeta_2, \tag{2.9}$$

$$O_3 = \beta_{2,3}O_2 + \zeta_3 = \beta_{2,3}(\beta_{1,2}\zeta_1 + \zeta_2) + \zeta_3, \tag{2.10}$$

where $\beta_{1,2}$ and $\beta_{2,3}$ are the autoregressive effects between two consecutive latent occasion-specific variables, and ζ_1 to ζ_3 are n-variate vectors that represent the occasion-specific residuals. These occasion-specific residuals capture the effects of the situation and the person-situation interaction, which are not explained by observations in previous situations.

In the TSO model, the latent-trait indicators ξ_j, the latent occasion-specific residuals ζ_t, and the random measurement error variables ε_{jt} are assumed to be normally distributed with mean zero and variances $\sigma^2_{\xi_j}$, $\sigma^2_{\zeta_t}$, and $\sigma^2_{\varepsilon_{jt}}$, respectively. Moreover, latent trait-indicator variables are uncorrelated with the latent occasion-specific residuals and the random measurement error variables but are allowed to correlate among themselves. In addition, the latent occasion-specific residuals and the random measurement error variables are uncorrelated with each other. The path diagrams of the single-level and multilevel TSO model are presented in Figure 2.3.

2.1.4 Similarities and Differences

As mentioned previously, the LST theory defines a set of coefficients that are proportions of the total variance of an indicator. The three main coefficients are reliability, consistency, and occasion-specificity (Steyer et al., 2015). These coefficients are common to every LST model. However, the equations are not necessarily equal among models given how the total variance of an observed variable is defined in each model. For example, while the consistency of an indicator Y_{jt} in the MSST model only accounts for the variance of the latent trait variable $Var(\xi)$, the consistency in the TSO model accounts for both the variance of the latent trait-indicator variable $Var(\xi_j)$ and the variance of the latent occasion-specific variable of the previous measurement occasion $Var(O_{t-1})$. The equations of these variance coefficients are presented in Table 2.1 for each of the three models to show how they differ.

Moreover, additional coefficients can be defined depending on the complexity of the model. In the TSO model, the consistency can be further decomposed into two additional coefficients: The *predictability by trait* and the *unpredictability by trait* (Eid et al., 2017). The predictability by trait is the proportion of the variance that is due to the latent trait-indicator variable ξ. It can be interpreted as the proportion of the variance that is explained by the trait differences between persons on the first occasion of measurement. On the other hand, the unpredictability by trait is the proportion of the variance that is due to the previous occasion-specific residuals ζ_{t-u} for $u = 1,...,t-1$ (recall Equations 2.9 and 2.10). In other words, it is the proportion of the variance that is explained by the carry-over effects observed in the data. The equations of the predictability by trait and the unpredictability by trait are also presented in the end of Table 2.1.

Lastly, the CUTS model is special in this regard because it was not developed within the LST theory. Because of this, this kind of coefficients was not defined for the CUTS model by its authors (Hamaker et al., 2017). However, given the equivalence between the CUTS model and the MSST model with orthogonal methods, we propose to use the coefficients defined by Geiser and Lockhart (2012) with the CUTS model to facilitate the comparison of the models selected in this study (see equations in Table 2.1). Thus, for the CUTS model, the consistency was also decomposed into two additional coefficients: The *common consistency* and the *unique consistency*. The common consistency is the proportion of the variance due to the common trait ξ. Hence, it encompasses the effects that are stable over time and common across all the indicators. On the contrary, the unique consistency is the proportion of the variance due to the unique trait ϑ_j, which is stable over time and unique to one indicator.

In short, there are variance coefficients as proposed in the LST theory that are shared

Table 2.1: Variance Decomposition and Components of the MSST, CUTS, and TSO

Description	**Equation**
	MSST
Total Variance	$Var(Y_{jt}) = \lambda^2_{T_{jt}} Var(\xi) + \lambda^2_{S_{jt}} Var(\zeta_t) + Var(\varepsilon_{jt})$
Reliability	$Rel(Y_{jt}) = 1 - \frac{Var(\varepsilon_{jt})}{Var(Y_{jt})}$
Consistency	$Con(Y_{jt}) = \frac{\lambda^2_{T_{jt}} Var(\xi)}{Var(Y_{jt})}$
Occasion-Specificity	$Spe(Y_{jt}) = \frac{\lambda^2_{S_{jt}} Var(\zeta_t)}{Var(Y_{jt})}$
	CUTS
Total Variance	$Var(Y_{jt}) = \lambda^2_{T_{jt}} Var(\xi) + Var(\vartheta_j) + \lambda^2_{S_{jt}} Var(\zeta_t) + Var(\varepsilon_{jt})$
Reliability	$Rel(Y_{jt}) = 1 - \frac{Var(\varepsilon_{jt})}{Var(Y_{jt})}$
Total Consistency	$TCon(Y_{jt}) = \frac{\lambda^2_{T_{jt}} Var(\xi) + Var(\vartheta_j)}{Var(Y_{jt})}$
Common Consistency	$CCon(Y_{jt}) = \frac{\lambda^2_{T_{jt}} Var(\xi)}{Var(Y_{jt})}$
Unique Consistency	$UCon(Y_{jt}) = \frac{Var(\vartheta_j)}{Var(Y_{jt})}$
Occasion-Specificity	$Spe(Y_{jt}) = \frac{\lambda^2_{S_{jt}} Var(\zeta_t)}{Var(Y_{jt})}$
	TSO
Total Variance	$Var(Y_{jt}) = \lambda^2_{T_{jt}} Var(\xi_j) + \beta^2_{t-1,t} \lambda^2_{S_{jt}} Var(O_{t-1}) + \lambda^2_{S_{jt}} Var(\zeta_t) + Var(\varepsilon_{jt})$
Reliability	$Rel(Y_{jt}) = 1 - \frac{Var(\varepsilon_{jt})}{Var(Y_{jt})}$
Consistency	$Con(Y_{jt}) = \frac{\lambda^2_{T_{jt}} Var(\xi_j) + \beta^2_{t-1,t} \lambda^2_{S_{jt}} Var(O_{t-1})}{Var(Y_{jt})}$
Predictability by Trait	$Pred(Y_{jt}) = \frac{\lambda^2_{T_{jt}} Var(\xi_j)}{Var(Y_{jt})}$
Unpredictability by Trait	$UPred(Y_{jt}) = \frac{\beta^2_{t-1,t} \lambda^2_{S_{jt}} Var(O_{t-1})}{Var(Y_{jt})}$
Occasion-Specificity	$Spe(Y_{jt}) = \frac{\lambda^2_{S_{jt}} Var(\zeta_t)}{Var(Y_{jt})}$

Note. These are the general equations of the variance coefficient components of the single level models. When all parameters are assumed to be invariant over time or when the models are formulated as multilevel SEM, the equations get simplified because the subscripts t are not needed anymore. The equations of the MSST and the CUTS models were retrieved from Geiser and Lockhart (2012), and the equations of the TSO model were retrieved from Eid et al. (2017).

among the three models, which are the reliability, the consistency, and the occasion-specificity. Even though the equations of these coefficients are different from model to model due to how the total variance of an observed variable is computed, their interpretation remains the same across models. Moreover, additional variance coefficients can be defined for a model. In this case, both the CUTS and the TSO models allow decomposing the consistency in two new variance coefficients: The common consistency and the unique consistency for the CUTS, and the predictability by trait and the unpredictability by trait for the TSO. These new coefficients are entirely different between the two models, not only in their equations but also in their interpretations. Functions to compute these coefficients can be found in the R code.

2.1.5 Relation of State-Trait SEMs with Other Models for Intensive Longitudinal Data

An important advantage of formulating state-trait SEMs as multilevel SEMs, especially state-trait SEMs that account for autoregressive effects, is that it allows relating these models to other frameworks that are commonly used to analyze intensive longitudinal data. Specifically, multilevel state-trait SEMs can be compared with multilevel (vector) autoregressive models (multilevel-VAR; e.g., Bringmann et al., 2013; Chow et al., 2007; De Haan-Rietdijk et al., 2016; Ebner-Priemer et al., 2015; Rovine & Walls, 2006), dynamic factor analysis models (DFA; Fuller-Tyszkiewicz et al., 2017; Molenaar, 1985; Song & Zhang, 2014), and more general frameworks such as state-space modeling (Chow et al., 2009; Kalman, 1960; Lodewyckx et al., 2011) and dynamic structural equation modeling (DSEM; Asparouhov et al., 2017, 2018). All these models have in common that they are specifically developed to analyze intensive longitudinal data, and they incorporate autoregressive effects to model persons' dynamics.

Firstly, multilevel-VAR models are an integration of time series and multilevel modeling, which allow modeling auto- and cross-regressive effects when $N > 1$ (Rovine & Walls, 2006). In contrast to autoregressive effects, cross-regressive effects model the dependency of a variable X at time t on a different variable Y at previous time points (e.g., Y at $t-1$ or $t-2$). In this framework, researchers are usually interested in understanding persons' dynamics and their relation with stable variables of the persons (e.g., gender, neuroticism). Moreover, as these models aim to acknowledge the idiosyncrasy of each individual, auto- and cross-regressive coefficients are defined as random effects. Recently, multilevel-VAR models that account for measurement error have been proposed (Schuurman & Hamaker, 2019; Schuurman et al., 2015). These models differentiate the residuals of the model (known as innovations) from the random measurement error by the inclusion of latent variables, which are identified by only one indicator. These one-indicator latent variables are equivalent to the latent state variables used in state-trait SEMs. In particular, the multilevel-VAR model with measurement error (Schuurman & Hamaker, 2019) can be seen as the multilevel version of the STARTS model (Kenny & Zautra, 1995; Kenny & Zautra, 2001) with random auto- and cross-regressive effects.

Secondly, multilevel state-trait SEMs can be framed as multilevel DFA models. Initially, DFA models (Fuller-Tyszkiewicz et al., 2017; Molenaar, 1985) were inspired by the P-technique analysis proposed by Cattell (1963). These are factor analysis models that are applied to the time series of an individual and account for the lagged structure in the data. In recent years, multilevel DFA models have been proposed (Song & Ferrer, 2012; Song & Zhang, 2014), which allow studying multivariate time series when $N > 1$. As well as in multilevel-VAR, multilevel DFA models have an emphasis on the individuals. Therefore, the lagged effects are allowed to randomly vary in the population. The multilevel TSO model, which we presented above, can be seen as a multilevel DFA model where the between factor structure is different from the within factor structure and the autoregressive effect is fixed instead of random.

Thirdly, the state-space modeling approach (Kalman, 1960) is a broader framework, which is especially useful to analyze intensive longitudinal data. In psychology, state-space models have been used to study persons' dynamics (e.g., Chow et al., 2009; Lodewyckx et al., 2011). In general, state-space models are described by two equations: The observation equation and the transition equation (Chow et al., 2010). In the observation equation, the measurement model is specified by incorporating latent variables at each observation, which are denominated as *states*. In the transition equation, the dynamics of the system are modeled by auto- and cross-regressive effects. Note that SEMs are also defined by two equations: The measurement model and the structural model. While the measurement model in SEM is equivalent to the observation equation in state-space modeling, the structural model is not immediately equivalent to the transition equation. As a result, SEM can be seen as a special case of state-space modeling and vice versa (Chow et al., 2010). In relation to multilevel state-trait SEMs, the latent state variables are comparable to the states in state-space modeling.

Finally, multilevel state-trait SEMs can be implemented in the DSEM framework (Asparouhov et al., 2017, 2018; Hamaker et al., 2018; McNeish & Hamaker, 2020). The most general DSEM is the cross-classified model (Asparouhov et al., 2018), which decomposes the observation into three components: The person's component, $Y_{2,i}$; the time component, $Y_{3,i}$; and the deviation of person i at time t, $Y_{1,it}$. A two-level

DSEM model only includes the person's component and the deviation of the person i at time t. In particular, multilevel state-trait SEMs that include lagged relationships (e.g., the multilevel TSO model) can be seen as DSEMs with fixed effects instead of random effects. DSEM has been especially developed to analyze intensive longitudinal data. It integrates time series analyses, multilevel modeling, structural equation modeling (SEM), and time-varying effect modeling (TVEM; Hastie & Tibshirani, 1993; Hoover et al., 1998) into one unified framework. This results in a comprehensive and accessible framework for the analysis of intensive longitudinal data.

To conclude this section, it is also important to highlight the major differences between state-trait SEMs and the models and frameworks that were just mentioned. Firstly, the research questions that can be answered by using state-trait SEMs are considerably different from the research questions that can be answered by using multilevel VAR or state-space modeling. While state-trait SEMs are mostly useful to identify the factor structure of the scales and short questionnaires used in intensive longitudinal data, multilevel VAR and state-space models are more about explaining how a dynamic system is affected by covariates of interest. In relation to DFA models, multilevel state-trait SEMs that account for autoregressive effects can be easily formulated as DFA models, with the difference that DFA models are more flexible by allowing autoregressive effects to be random among persons. Lastly, the DSEM framework allows for bridging the gap between state-trait SEM and multilevel VAR models, and thus for answering a broad range of research questions arising from both approaches.

2.2 Simulation Study

State-trait SEMs have been rarely used to analyze intensive longitudinal data, mainly because most of the models have been proposed as single-level models. As indicated previously, single-level models become harder and practically impossible to use when there are too many measurement occasions. To the best of our knowledge, this chapter presents, for the first time, a simulation study that aims to systematically explore under which conditions state-trait SEMs can be useful to analyze intensive longitudinal data. For this, we selected three state-trait SEMs: The MSST, the CUTS, and the TSO models. While these models are some of the simplest models that allow differentiating states and traits, they also address some of the most important features that are accounted for by more complex state-trait SEMs (i.e., method effects and autoregressive effects). In the simulation, we compared both the single-level and the multilevel formulation of each model. Even though differences in the estimates between the two versions of the same model were not expected because they are mathematically identical, multilevel models can handle large numbers of measurement occasions and missing data (Geiser, 2020; Geiser et al., 2013). It is therefore important to see under which conditions (e.g., how many time points) the single-level version runs into fitting issues, and thus, the multilevel formulation of the models becomes the only possible way to fit the models to the data.

In addition, multiple factors were manipulated, such as the model used to simulate the data, the number of measurement occasions, the proportion of missing values, and the ratio between the trait variance and the state variance. To start with, we simulated separate data sets based on each model to be able to compare the model fits across different data-generating models. We expected that each model would be favored when it matches with the true or data-generating model. Next, we manipulated the number of measurement occasions and the proportion of missing values. Although it is known that the single-level models become impossible to fit as the number of measurement occasions and the proportion of missing values increases, no simulation study has yet aimed to determine the point at which this happens. Finally, the ratio between the trait and the state variances was considered to study how the models would perform if the variables were more trait-like, more state-like, or evenly trait- and state-like. Considering previous results (Cole et al., 2005), we expected no differences in the results due to this factor.

In a nutshell, this simulation study aimed to explore how suitable the selected state-trait SEMs are for analyzing intensive longitudinal data. To explore this, the models were tested under different conditions, which represent realistic settings of intensive longitudinal studies in psychology. The models were evaluated in terms of the number of times the analysis converged to proper solutions and the quality of the estimated parameters in relation to the true population parameters. A secondary goal of the simulation was to identify what is the maximum number of measurement occasions that can be handled by the single-level models. By studying this, we expected to be able to provide practical guidelines on the usage of state-trait SEMs when analyzing intensive longitudinal data.

2.2.1 Method

In the simulation study, we manipulated four factors while keeping the sample size and the number of variables fixed. Thus, the sample size was fixed at 100, as it is a reasonable number of individuals to have in an intensive longitudinal study (e.g., Bos et al., 2015; Bringmann et al., 2013; Kuppens et al., 2010), and the number of indicators was fixed to 4[1]. In relation to the manipulated factors, firstly, we manipulated the model used to generate the data to be able to fairly compare the selected models. In the following sections, the model used to generate data is referred to as the *base model*. Secondly, the number of measurement occasions was set to 10, 15, 20, 30, 60, or 90. The conditions with 20 measurement occasions or less aimed to aid in the identification of the maximum number of measurement occasions that can be handled by single-level state-trait SEMs. Note that single-level state-trait SEMs are commonly used to analyze data that have between two and eight measurement occasions (Geiser & Lockhart, 2012, Appendix B). The conditions with 30 measurement occasions or more aimed to mimic the common number of observations in psychological time series (e.g., Bringmann et al., 2016; Fuller-Tyszkiewicz et al., 2017; Schuurman et al., 2015; Song & Zhang, 2014). Thirdly, three proportions of missing values were selected: 0%, 10%, and 20%[2]. The missing values were assumed to follow an MCAR mechanism (Little & Rubin, 2019; Rubin, 1976). Finally, the approximate ratio between the trait variance and the state variance was set to 1:3, 1:1, and 3:1. This factor was manipulated in a similar way as it was done by Cole et al. (2005) in a simulation study about method effects in the TSO model. In conclusion, the simulation follows a 3 × 6 × 3 × 3 fully crossed design, where 100 replications were generated for each condition.

[1] The number of indicators was kept fixed and small considering that intensive longitudinal questionnaires tend to be short in order to not burden the participant (Eisele et al., 2020). Thus, the number of questions measuring the same construct is likely to be small. Moreover, we ran a short simulation manipulating the number of indicators, which showed that this factor does not seem to affect the performance of the model when it increases. Results from this simulation are available in Appendix B.

[2] These proportions of missing values were kept small due to several practical reasons. Firstly, in previous runs, we observed that increasing the proportion of missing values to around 10% already resulted in very long running times and convergence problems for the single-level models. Secondly, it is known in the literature that a clear advantage of the multilevel formulation is that it is able to handle missing values (Geiser, 2020; Geiser et al., 2013). Finally, these proportions of missing values aimed to resemble the proportion of missingness observed in the filtered empirical data at hand.

In relation to the true parameters, these were defined on the basis of other simulation studies with state-trait SEMs (e.g., Cole et al., 2005; Geiser & Lockhart, 2012; LaGrange & Cole, 2008) and empirical results of the models of interest (e.g., Eid et al., 2017; Geiser & Lockhart, 2012; Hamaker et al., 2017). The proportion of measurement error variance across all models and all indicator variables was set to approximately 0.2, which implies that the true reliabilities of the indicators variables were set to approximately 0.8. All true parameters were assumed to be time-invariant because this is required in the multilevel formulation of the models. In other words, measurement invariance was assumed to hold in all the generated data sets. The true parameters and true variance coefficients are included in Appendix B.

Once a data set was simulated, it was analyzed 11 times (3 models $\times$ 2 versions of each model $\times$ 2 estimation methods $-$ 1 for the multilevel TSO with maximum likelihood estimation). Hence, each replication was analyzed with the single-level and the multilevel version of each model. Moreover, the estimation methods were set to maximum likelihood estimation (MLE) and MCMC Gibbs sampler (Bayes). Note that the multilevel version of the TSO model can only be estimated through Bayesian methods because maximum likelihood estimation is not supported in Mplus when using the LAGGED command or the ampersand (&) command to include lagged variables in the model (Muthén & Muthén, 2017). This is why each generated data set was analyzed 11 instead of 12 times.

The simulation was performed in Mplus 8.2 (Muthén & Muthén, 2017) and R 3.6.1 (R Core Team, 2022). Mplus was used to fit the models to the data and R was used to generate the data and to analyze the results of the simulation. Furthermore, Mplus was called from within the R environment through the package MplusAutomation (Hallquist & Wiley, 2018). All the code for this study is available at https://github.com/secastroal/LST_Analyses. Note that these models can also be fitted in any SEM specialized software, for example with the R package lavaan (Rosseel, 2012). Also, to fit the models within the Bayesian framework, alternative software such as Stan (Carpenter et al., 2017) or JAGS (Depaoli et al., 2016) can be considered. When running the analyses in Mplus, we used a maximum of 50,000 iterations when the estimation method was maximum likelihood, and we used three chains each with a minimum of 5,000 iterations with thinning 10 when the estimation method was Gibbs sampling[3]. These conditions and in general the complexity of the models made the analyses highly computationally demanding, especially when the number of measurement occasions and the proportion of missingness were larger. Therefore, a maximum running time of one hour per model fit was set and 4GB of RAM dedicated per core in order to be able to conduct the simulation study in a reasonable period of time. After the time limit was reached or if all the available memory was used for a specific analysis, it was interrupted and no output was saved.

Outcomes

To compare the performance of the models in the simulation study, we determined (a) the number of successful analyses, that is, the number of times a model was not interrupted, converged and provided a reasonable solution, and (b) the quality of the parameter estimates in relation to the true population parameters. Analyses that did not finish successfully were classified into warnings/errors, non-convergence, and out of resources. The warnings/errors status might be due to negative variances, unreliable standard errors, or non-convergence of the H1 model. The H1 model is the unrestricted model, which is needed to compute the χ^2 and related fit measures in SEM (Muthén & Muthén, 2017). The non-convergence status might be due to the model being unable to compute standard errors or because the algorithm did not converge (exceeding the maximum number of iterations in MLE or not satisfying the convergence criterion in the MCMC algorithm). Finally, the out of resources status implies that the analysis was forced to stop after an hour of running time or the computing process ran out of available memory. To study the quality of the parameter estimates, we computed the average bias ($E[\hat{\theta} - \theta]$), the average relative bias ($E[(\hat{\theta} - \theta)/\theta]$), the average absolute bias ($E[|\hat{\theta} - \theta|]$), and the root mean squared error (RMSE; $\sqrt{E[(\hat{\theta} - \theta)^2]}$). Additionally, in Appendix B, we included results about information criteria indices, which indicate the number of times a model was selected as the best fitting model.

As not all the parameters nor variance coefficients are shared across all the three models, we focused our analyses on the parameters and variances coefficients that are common to the three models. The comparable parameters across all models are

[3] In all Bayesian analyses, we used the default priors provided by Mplus, which are uninformative priors. A sensitivity analysis was done with a random sample of analyses using weak priors. Results from these analyses are included in Appendix B and provided evidence that the weight of the priors of the estimated model is negligible. Furthermore, to be sure that the MCMC algorithm converged, we used the BITERATIONS option of MPlus. This option makes the algorithm run until all the potential scale reduction factors (also known as Rhat) are below 1.05 or 1.1 (when there is a large number of parameters) given a minimum and a maximum number of iterations (Muthén & Muthén, 2017).

the three factor loadings at the within-level λ_{S_j}, the variance of the latent state residual $var(\zeta)$, the four variances of the measurement error component $var(\varepsilon_j)$, the reliabilities $Rel(Y_j)$, the consistencies $Con(Y_j)$, and the occasion-specificities $Spe(Y_j)$.

Table 2.2: Percentages of Successes and Failures per Type of Analysis on 16200 Analyses Each

Model	Est.Method	Successful	Warnings/Errors	Non-convergence	Out of Resources
MSST	MLE	56.06	9.72	1.02	33.2
ML-MSST	MLE	93.91	0.16	5.86	0.06
MSST	Bayes	81.85	0	8.25	9.91
ML-MSST	Bayes	92.81	0	5.35	1.84
CUTS	MLE	21.4	45.31	0	33.29
ML-CUTS	MLE	61.9	38.04	0.01	0.06
CUTS	Bayes	88.49	0	0.04	11.47
ML-CUTS	Bayes	99.44	0	0.04	0.52
TSO	MLE	39.12	33.15	0	27.73
TSO	Bayes	88.59	0	0	11.41
ML-TSO	Bayes	99.48	0	0	0.52

2.2.2 Results

Successful analyses.

The percentage of analyses that finished successfully or failed per type of analysis across all conditions is shown in Table 2.2. This table shows the effect of the formulation of the model and the estimation method. As expected, analyses estimated via MLE were more prone to be interrupted or to present convergence issues or errors when the data were structured in wide format. For example, the single-level CUTS model estimated with MLE only retrieved an interpretable solution in 21.4% of the analyses performed. On the other hand, single-level models estimated via Bayesian methods also failed in multiple replications but mainly due to the analyses running out of time. In contrast, when the data were structured in long format and the multilevel versions of the models were used, the analyses almost never failed independently of the estimation method. However, this is not the case for the multilevel CUTS model estimated by means of MLE, which only converged to an interpretable solution in 61.9% of the analyses.

Figures 2.4 through 2.6 show how many analyses finished successfully in the conditions with trait-state variance ratio of 1:1 given each proportion of missingness. This allows observing in detail how the number of measurement occasions and the proportion of missing values affect the convergence of the analyses. The same plots and particular results for conditions with trait-state variance ratios of 1:3 and 3:1 are presented in Appendix B. Starting with the single-level models, analyses failed to provide a solution when estimated by means of MLE and with 60 or more measurement occasions. This threshold becomes lower when including missing values. With 10% missing values, 30 or more measurement occasions become hard to handle, and with 20% missing values, 20 measurement occasions are already problematic. Moreover, when fitting the single-level CUTS model with MLE to non CUTS data, the analyses completely failed to reach a solution independently of the number of measurement occasions. When the number of measurement occasions was lower than 60, these failures were mainly because of the model estimating negative variances. This could happen due to the fact that there were not real method effects in the data. Similarly, when fitting the single-level TSO model with MLE to MSST data, the single-level TSO model completely failed to provide a reasonable solution. In general, increasing the number of measurement occasions in combination with the proportion of missing values makes the single-level models more likely to fail when estimated by means of MLE.

In contrast, when using Bayesian estimation, single-level models successfully finished independently of the number of measurement occasions or the proportion of missing values. Note that single-level models estimated by means of Bayesian methods only failed to reach a solution with 90 measurement occasions when there were missing values. These failures were mainly due to the imposed time restriction. Hence, without the one-hour time limit, the single-level models estimated via Bayesian methods were likely to converge even with 90 measurement occasions.

In relation to the multilevel formulation of the models, the multilevel models tended to perform relatively well in most conditions independently of the estimation method. As shown in Figures 2.4 through 2.6, the multilevel models finished successfully in almost all of the replications across all conditions. However, this is not true for the multilevel CUTS model when estimated by means of MLE, which failed to provide an interpretable solution when analyzing data generated from another model. These failures were more likely because of negative variances. We conjecture again that the problem is that there are no real method effects in the data generated based on the MSST and the TSO models. As a consequence, it was difficult for the CUTS

Figure 2.4: Number of Successful Analyses per Condition with 0% Missingness and 1:1 Trait-State Variance Ratio

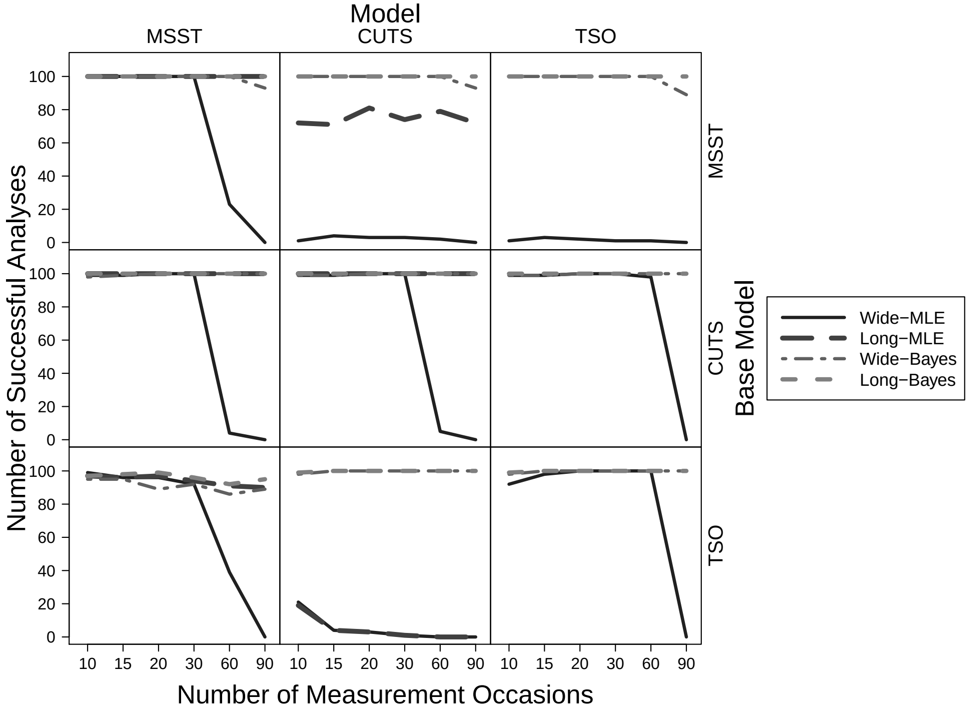
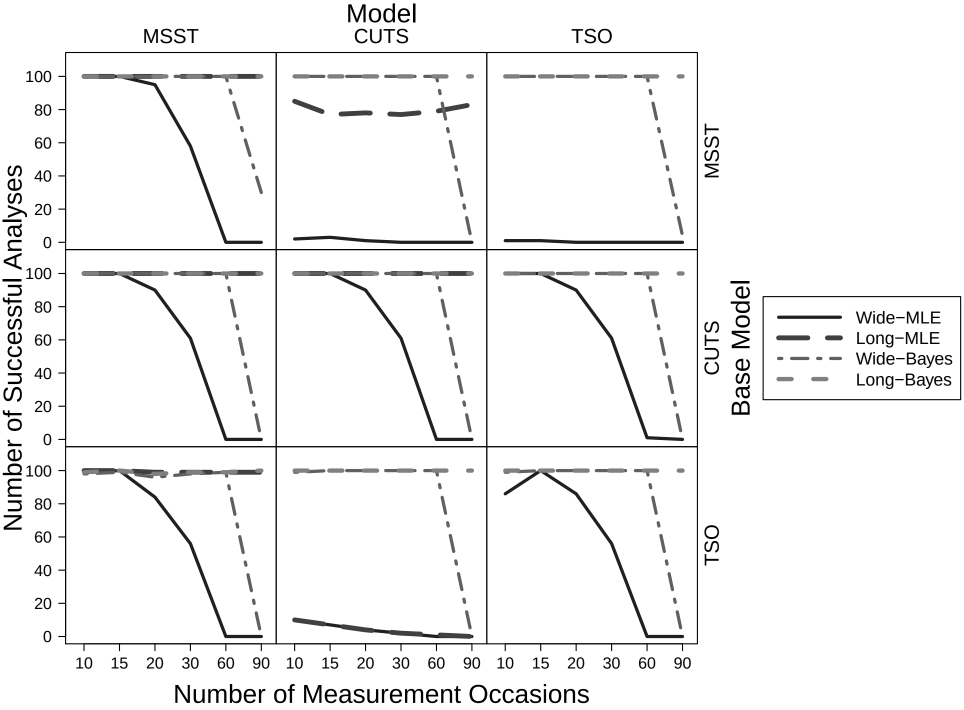
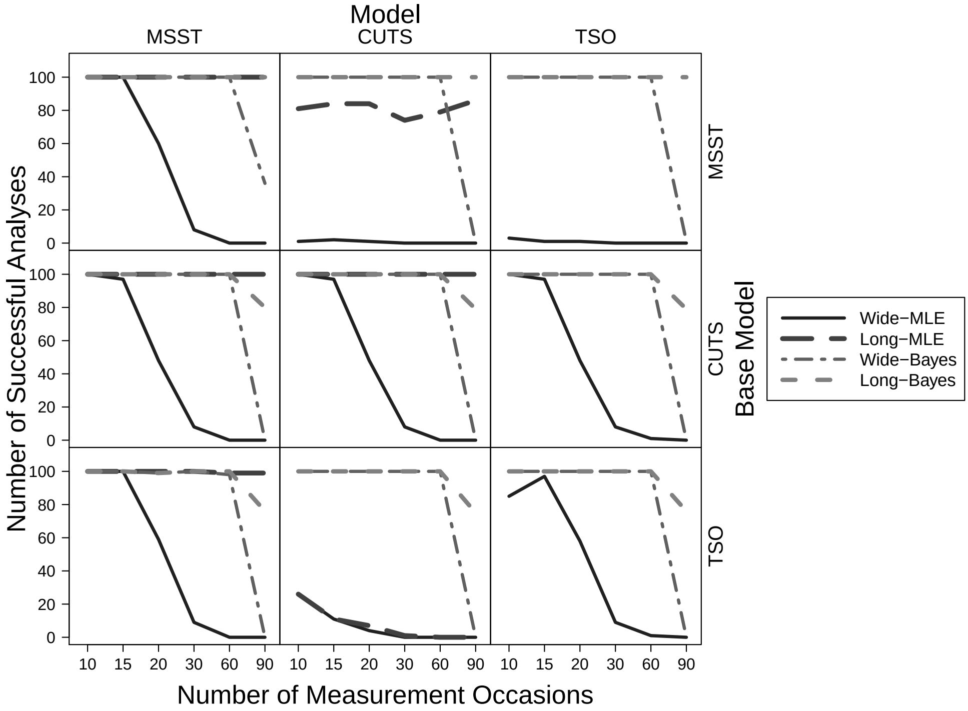


Figure 2.5: Number of Successful Analyses per Condition with 10% Missingness and 1:1 Trait-State Variance Ratio



model to estimate null unique trait variances, resulting in at least one negative estimated variance. Note that the multilevel CUTS model failed completely when there were autoregressive effects in the data. As this was unexpected, we further inspected these analyses by performing some additional simulations while varying some of the true population parameters. In these additional simulations, we found that the covariances of the latent trait-indicator variables ξ_j of the TSO model had an effect on the performance of the multilevel CUTS model estimated via MLE. In other words, had we selected a different variance-covariance matrix to generate the TSO data, we would have had results where the multilevel CUTS model with MLE would converge in most of the replications. Results of these additional simulations are presented in Appendix B.

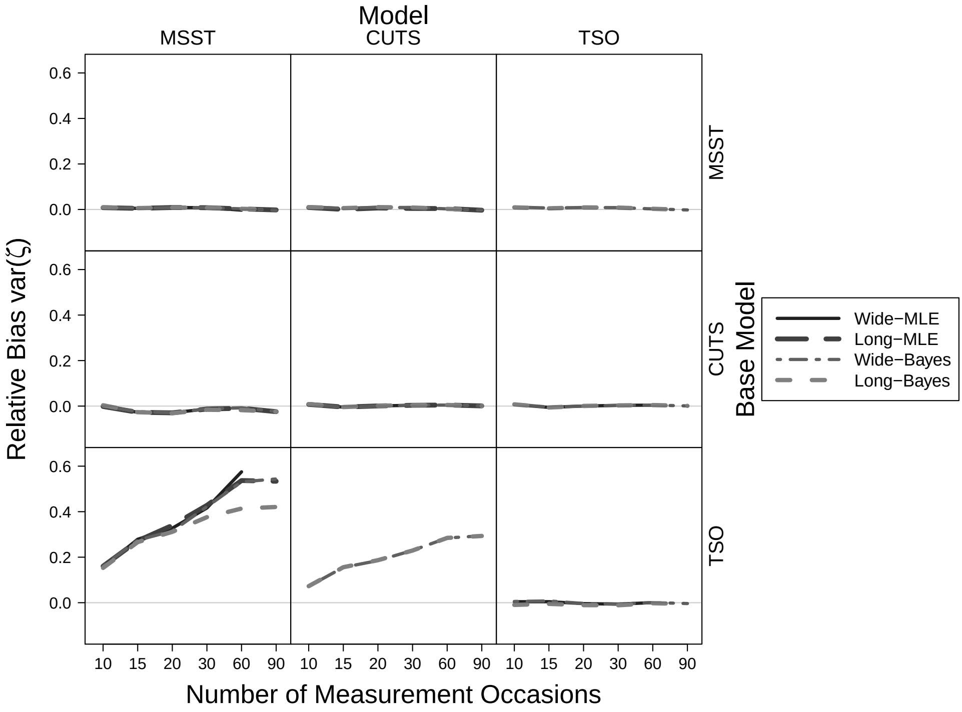
Figure 2.6: Number of Successful Analyses per Condition with 20% Missingness and 1:1 Trait-State Variance Ratio



Estimated Parameters

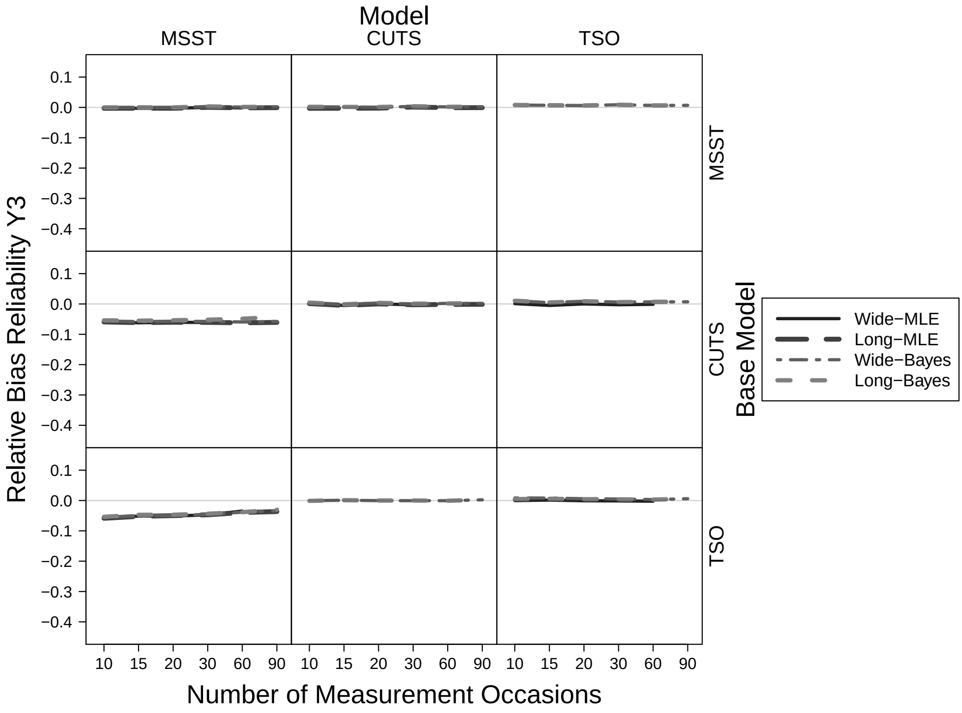
Bias, relative bias, absolute bias, and RMSE were used to compare the estimated parameters with the population parameters. As said previously, we used these statistics to compare the parameters that were common across all three models. Given that these different measures did not lead to different conclusions, the following section is based on the results from the average relative bias. In general, we observed that the three models fit particularly well their own data, showing little deviation from the population parameters. Moreover, the quality of the estimated parameters was not overly affected by the proportion of missing values, the estimation method, or the version of a model, which means that the average bias of a parameter did not vary noticeably when these factors changed.

Figure 2.7: Average Relative Bias of the Variance of the Latent State Residual $var(\zeta)$ per Condition with 1:1 Trait-State Variance Ratio and 0% Missingness



As a whole, the estimated within factor loadings λ_{S_j} showed biases close to 0 in most situations. When looking at the estimated variance of the state residual ($var(\zeta)$) and the estimated variances of the measurement error components $var(\varepsilon_j)$, these estimates tended to show some bias, especially when the generation and estimation models differed. For example, the MSST was shown to perform quite poorly in those cases. Also, the CUTS model was considerably inaccurate at the estimation of the variance of the state residual ($var(\zeta)$) when analyzing data generated based on the TSO model. In contrast, the TSO was consistently accurate independently of the generating model. These observations are highlighted in Figure 2.7, which shows the average relative bias of the variance of the latent state residual ($var(\zeta)$) per condition when the trait-state variance ratio was 1:1 and there were no missing values.

Figure 2.8: Average Relative Bias of the Reliability of the Third Indicator Y_3 per Condition with Trait-State Ratio 1:1 and 0% Missingness



In relation to the variance component coefficients, such as reliability, consistency, and occasion-specificity, the models did a relatively good job at recovering these parameters. For example, in Figures 2.8 and 2.9, we display the average relative bias of the reliability and the consistency for the third variable Y_3 across the conditions without missing values and a trait-state variance ratio of 1:1. For the estimates of the reliability, the MSST model tended to slightly underestimate these parameters when the data were generated based on a different model. This effect was larger when the variables were more trait-like. Similarly, the estimates of the consistency tended to show some bias when using the MSST model to analyze data based on the other models. Yet, this effect was lower in the conditions with a trait-state variance ratio of 3:1. On the other side, the CUTS model was fairly accurate at estimating the reliability regardless of the base model. However, the estimates of the consistency got worse when the model used to generate the data was the TSO and when the variables were more state-like. Finally, the accuracy of the estimates of the TSO model was high independently of the base model, the trait-state variance ratio, and the other manipulated factors.

Additionally, we also looked at the variance component coefficients that were not common across all the three models, such as the unique consistency for the CUTS model and the unpredictability by trait for the TSO model. We expected these coefficients to be estimated close to 0 if the data were generated from a different model. In other words, if the base model was the MSST, and the data were analyzed with the CUTS model, we expected the estimate of the unique consistency to be basically 0. To some extent, our expectations were met. On the one hand, the CUTS model did estimate approximately null unique consistencies when the base model was the MSST, but it did not when the base model was the TSO. Moreover, the unique consistencies, when the base model was the TSO, tended to be larger the more trait-like the variables were.

On the other hand, the estimates of the unpredictability by trait in the TSO model when the data were not generated by itself were basically 0. This is shown in Figure 2.10, which plots the mean estimate across replications of the unpredictability by trait for all indicators Y_j when the model used to generate the data was the CUTS model, the trait-state variance ratio was 1:1, and the proportion of missing values was 0%. Moreover, the same observations were made about the estimates of the autoregressive effect[4]. This is evidence of the robustness of the TSO model. In other words, if there are no autoregressive effects in the data, the TSO is able to identify the lack of these

Figure 2.9: Average Relative Bias of the Consistency of the Third Indicator Y_3 per Condition with Trait-State Ratio 1:1 and 0% Missingness

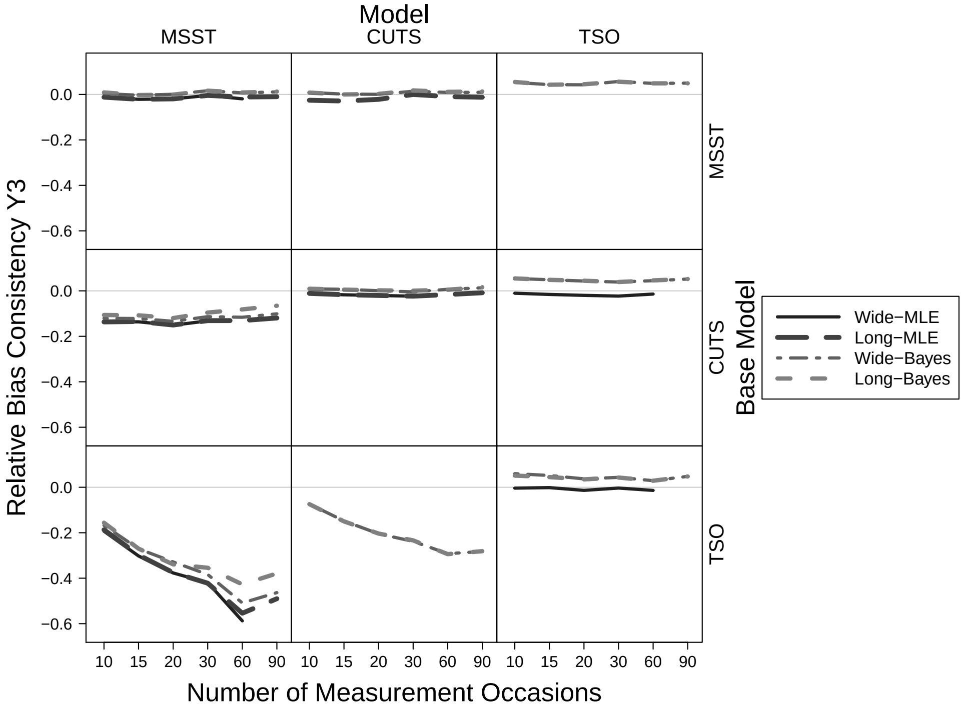
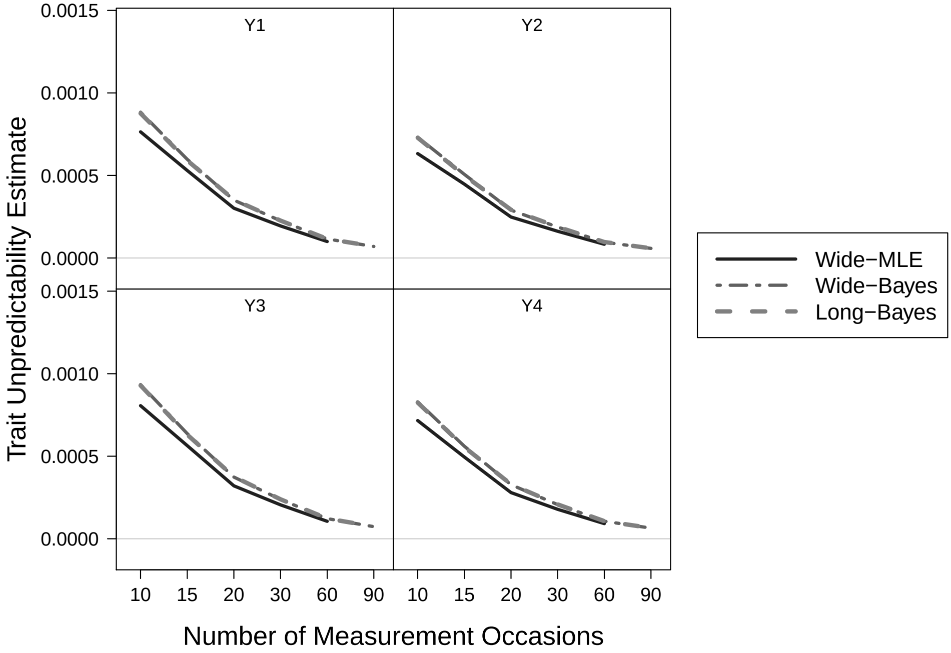


Figure 2.10: Estimated Unpredictability by Trait of all Indicators when the Base Model was the CUTS Model, the Trait-State Variance Ratio was 1:1, and the Proportion of Missing Values was 0%



effects. Thus, the TSO does not estimate unrealistic autoregressive effects even if method effects are present in the data.

2.2.3 Summary

Through the simulation study, we explored the suitability of the selected models at analyzing hypothetical intensive longitudinal data. This also allowed comparing the performance of the three models and to get an idea of which model can be more useful and more reliable when analyzing intensive longitudinal data. As expected, the results showed that the single-level models are harder to fit via MLE when the number of measurement occasions and the proportion of missing values increase. On the contrary, the single-level models were able to converge when estimated via an MCMC algorithm even with 90 measurement occasions and if they failed, it was mainly due to the imposed time limit. Therefore, under the absence of missing values, single-level models can be useful at analyzing data with up to 30 measurement occasions if estimated via MLE and with up to 90 measurement occasions if estimated via an MCMC algorithm. However, when estimated through Bayesian methods, comparing models and thus testing for measurement invariance can be difficult due to the lack of fit statistics, which makes the model impractical when there are too many measurement occasions. In short, using the single-level models can be an alternative if the time series of interest is not too long (around 30 time points), as this would allow researchers to test for measurement invariance.

By any means, the multilevel formulation of the models is undeniably more practical and straightforward to use in comparison with their respective single-level formulation, although multilevel state-trait SEMs could also run into fitting problems when the fitting model did not match the data-generating model. In general, the models did a good job at recovering the population parameters, especially the multilevel TSO model, which consistently performed well across all the conditions of the simulation study. For this reason, we consider the multilevel TSO the most viable model to analyze intensive longitudinal data among the models that were studied.

[4] A plot for the bias of the autoregressive effect is included in Appendix B.

2.3 Empirical Example

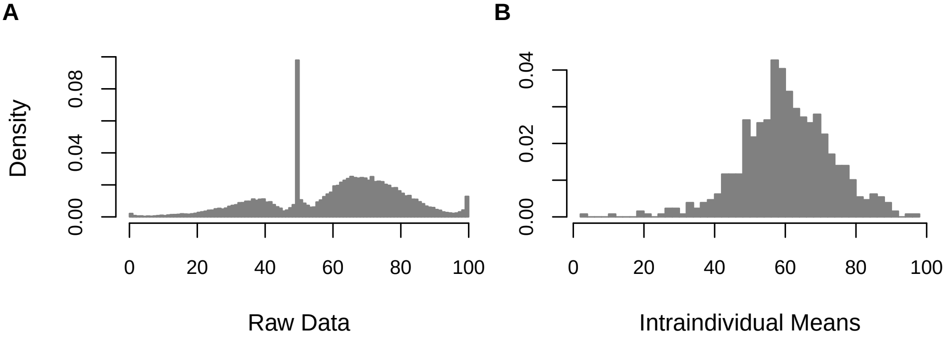
To present a real data application, we analyzed data from the national crowdsourcing study HowNutsAreTheDutch (Dutch: HoeGekIsNL; van der Krieke et al., 2017; van der Krieke et al., 2016), which started in May 2014. The HowNutsAreTheDutch (HND) project aims to study the mental health of the Dutch population by considering the mental health as "a dimensional and dynamic phenomenon" (van der Krieke et al., 2016, p. 124). To capture the dynamics of the mental health, the HND project includes in its design a daily diary study, from which we obtained the data. Specifically, we analyzed the items used to measure positive affect, which are *I feel relaxed*, *content*, *calm*, *energetic*, *enthusiastic*, and *cheerful*. The first three items measure positive affect deactivation, and the last three measure positive affect activation according to the circumplex model of affect (Feldman Barrett and Russell, 1998, and Yik et al., 1999, as cited in van der Krieke et al., 2016). In a nutshell, the circumplex model assumes that affect can be explained by two dimensions: Valence (positive or negative) and activation. In the following analysis, we used the three state-trait SEMs discussed in this chapter to determine if the positive affect items are described best by a one-factor or a two-factor measurement model. Moreover, we aimed to explore the psychometric properties of the items and to determine to what extent the items are trait- or state-like.

2.3.1 Method

In the diary study, participants were followed three times a day, for 30 days. Hence, each participant could have a maximum of 90 measurements. To assess the participants, a link to the questionnaire was sent through a text message every six hours given each participant's daily routine. After receiving the text message, participants had to answer the questionnaire immediately or within the next hour. The questionnaire contained 43 items that measure, for example, subjective well-being, mood, self-esteem, etc. Most of the items were rated on a visual analogue scale (VAS) from 0 to 100. Note that we only considered the six items that measure positive affect.

In this empirical example, we considered the data collected between May 2014 and December 2018. Initially, the data included 115,386 records of 1,396 participants with a mean of 43.13 observations per participant. Most of the participants had less than 20 or more than 60 observations. As some participants did the daily diary study multiple times, we decided to analyze only the first daily diary study of each participant. Furthermore, in some cases, participants have 93 observations because the study period was wrongly programmed to last 31 days during the first month of the study, from which we selected the first 90 observations. Also, given that many participants tended to drop the daily diary study during the early stages of it, we only included participants with at least 65% of the observations (59 or more) in the analyses. This criterion is also used in the HND project to give personalized feedback (van der Krieke et al., 2016, p. 128). After this selection, data were reduced to 57,945 records of 644 individuals (mean age 39.88; 83.86% women) with a mean of 74.9 observations per participant.

Figure 2.11: Histograms of the variable *relaxed.* (A) Histogram of the raw scores. (B) Histogram of the intraindividual means.



2.3.2 Results

The responses on the six positive affect items were spread over the whole range of the scale. The median of the six items varied between 50.0 and 63.8, and the standard deviation varied between 19.05 and 21.43. Moreover, their correlations ranged from 0.36 to 0.79. The distribution of the six items was very similar. Panel A of Figure 2.11 presents the distribution of the item *relaxed.* As shown in the histogram, the item's scores are concentrated around three main points: 40, 50, and 70. This odd distribution can be explained by the large variability of the data and the method of data collection. On the one hand, due to the nature of the variables, it was expected for people to be concentrated just a little above and below the middle point. On the other hand, the visual analogue scales were presented with the middle point pre-selected. If a person was not feeling in any particular way, they were able to just check their response in the middle point. In contrast, the distribution of the intraindividual means follows a more bell-shaped distribution as shown in panel B of Figure 2.11.

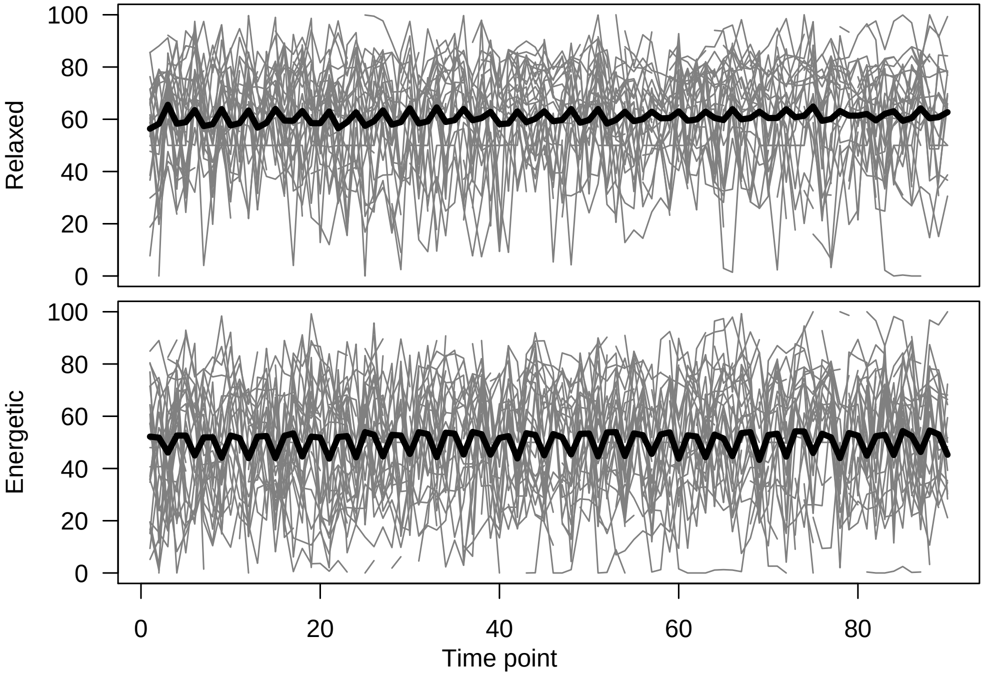
To give an illustration of the raw data, Figure 2.12 presents the trace plots of a random subsample of 30 persons on the variables *relaxed* and *energetic*. As one of the assumptions imposed by the state-trait SEMs discussed in this study is stationarity[5], we tested for it by means of the Kwiatkowski–Phillips–Schmidt–Shin test (Kwiatkowski et al., 1992), where the null hypothesis is that the observable time series is trend-stationary. By excluding individuals with at least one non-stationary time series, the sample size was reduced to 376 persons. The main analyses of this study were performed on the sample with and without the individuals with non-stationary time series. As the results did not show major differences, here, we report the results of the larger sample. The results excluding the individuals with non-stationary time series are available in Appendix B.

The HND data were analyzed with the multilevel versions of the three models: MSST, CUTS, and TSO. All the analyses were performed within the Bayesian framework. This means the models were estimated with the MCMC Gibbs sampler including four chains with thinning 10 and 5,000 iterations per chain. The first half of each chain was considered part of the burn-in phase, which left a total of 10,000 valid samples of the posterior distribution of the parameters. These analyses aimed to identify the factor structure and the variance component distribution of the six positive affect items.

Initially, we tried to fit the models to all items at once, assuming a unique factor of positive affect. However, the models did not fit well the data or did not converge. This was a clear indication that one factor was not enough to capture the structure of the data. Therefore, we decided to analyze the data in two sets of items as is suggested by the circumplex model: The positive affect deactivation (PAD) and the positive affect activation (PAA) sets. The results of these analyses were more promising but the mixing of the chains of the intercepts was poor as shown in Figure 2.13. To

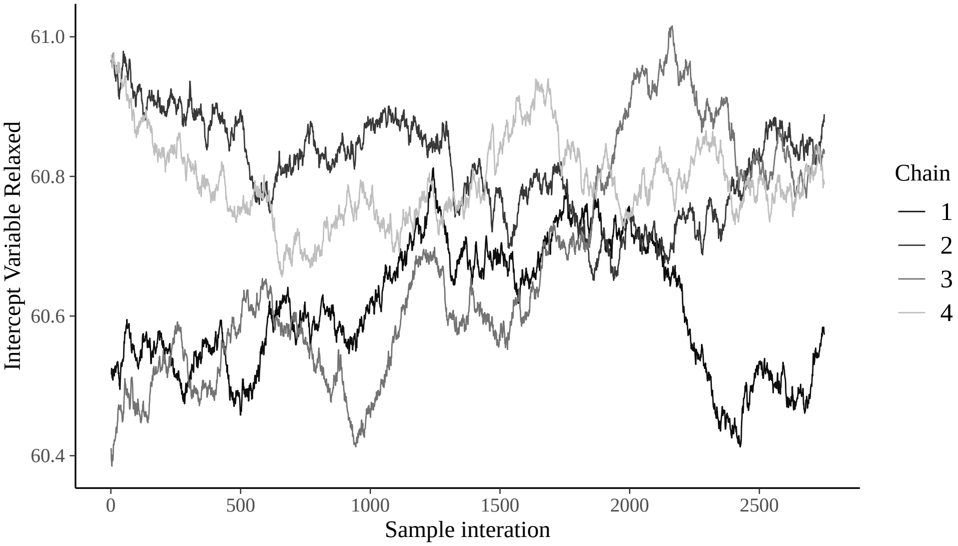
[5] In the literature, different types of stationarity are mentioned, for example, weak stationarity, strict stationarity, trend stationarity, or if the data contained a random walk. Note that it is far from trivial to test for these different kinds of stationarity (see Bringmann et al., 2017). In this specific case, we opted to test for trend stationarity.

Figure 2.12: Time Series of a Sample of 30 Individuals



Note. The variable *relaxed* is displayed on the top and the variable *energetic* is displayed on the bottom. The ticked time series shows the overall means at each time point.

Figure 2.13: Traceplot of the Intercept of the Multilevel TSO



Note. Traceplot of the intercept of variable *relaxed* when fitting the multilevel TSO model to the data.

solve this, we centered the data and constrained the intercepts to 0 to get a better solution. Note that the intercepts do not influence the computation of the variance coefficients. Next, we report the results of the models fitted to each of the centered sets of items. The corresponding posterior predictive p-values (ppp[6]) and Deviance Information Criteria (DIC[7]) are presented in Table 2.3 for each set. Note that the ppp is not available for the TSO due to software limitations. Considering the DIC values, the TSO model seems to fit the data best for both sets of items. Hence, we follow with the interpretation of the results of the TSO analyses.

Table 2.3: ppp and DIC of the Three Models for the Two Sets of Items

		MSST	CUTS	TSO
PAD	ppp	0.583	0.644	-
	DIC	1185142.150	1171133.218	1129157.419
PAA	ppp	0.583	0.657	-
	DIC	1184024.092	1170988.591	1113933.571

To interpret the results, we focused on the variance coefficients of the model: Reliability, consistency, predictability by trait, unpredictability by trait, and occasion-specificity. These estimates are shown in Table 2.4 for the two sets of items. The estimated reliability coefficients show that a large proportion of the variance of the items was explained by reliable sources of variability. In particular, the reliabilities of the items that measure positive affect activation are on average higher than the reliability of the items that measure positive affect deactivation, which means that positive emotions with high arousal are measured more precisely than positive emotions with low arousal.

[6] The posterior predictive p-value (ppp; Asparouhov & Muthén, 2010) is the Bayesian counterpart of the traditional p-value. It tests the model against an unspecified alternative. As implemented in Mplus, low ppp values indicate model misfit. For more details on how ppp values are computed in Mplus see Asparouhov and Muthén (2010, pp. 28-30)

[7] The DIC as computed in Mplus should be interpreted with caution because it can be unstable, especially when the model includes latent variables (Asparouhov et al., 2018, pp. 366-368). Also, note that the DIC values of different models are not always comparable.

Table 2.4: Variance Coefficients of the Three Models for the Two Sets of Items

	Variance Coefficient	Items		
		Relaxed	Content	Calm
PAD	Reliability	0.80	0.68	0.67
	Consistency	0.40	0.41	0.39
	Predictability by Trait	0.34	0.37	0.35
	Unpredictability by Trait	0.06	0.04	0.04
	Occasion-Specificity	0.40	0.27	0.28
		Energetic	Enthusiastic	Cheerful
PAA	Reliability	0.70	0.84	0.79
	Consistency	0.35	0.39	0.39
	Predictability by Trait	0.31	0.34	0.35
	Unpredictability by Trait	0.04	0.05	0.04
	Occasion-Specificity	0.35	0.45	0.40

Regarding the consistency and occasion-specificity, the items *content* and *calm* of positive affect deactivation have a consistency coefficient considerably larger than the occasion-specificity. This means that these items are in general more trait-like than state-like. In other words, emotions such as feeling content or calm are not as influenced by the situation as emotions such as feeling relaxed. Yet, the variation due to the situation is not negligible. On the other hand, the item *relaxed* as well as all the items of positive affect activation appear equally trait-like as they are state-like. The variability due to the person and the variability due to the situation are evenly distributed and equally important. These results can be compared with the intraclass-correlation (ICC; Houben et al., 2020) which is usually computed in an empty multilevel model. The corresponding ICCs of the six items varied between 0.32 and 0.37, which indicates that all the items were more state-like. These contradictory conclusions might be due to the fact that the state-trait SEMs do account for the random measurement error, while traditional multilevel models do not.

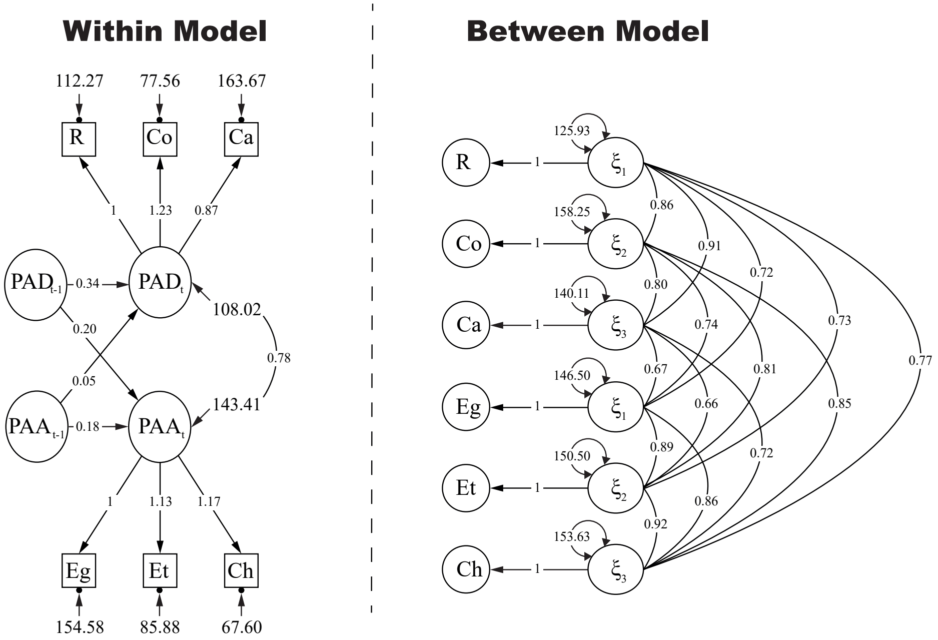
Finally, the unpredictability by trait of all the items was estimated close to 0.05, which means that only 5% of the variability is explained by the carry-over effects. The unpredictability by trait is directly associated with the estimated autoregressive effects, which were 0.367 for the positive affect deactivation items and 0.317 for the positive affect activation items. Despite the fact that the autoregressive effects were relatively large and practically significant, this did not necessarily imply that the proportion of the variance explained by the dynamic process, namely the unpredictability by trait, would be practically significant as well.

Additionally, we fitted a simple extension of the TSO model in order to fit the two sets of items at once and allow including cross-regressive effects on the latent occasion-specific variables. The main goal of fitting this model was to study the possible cross-regressive effects of positive affect deactivation on positive affect activation and vice versa. This model was also fitted to the centered data to avoid divergent MCMC chains. The path diagram of this model and its estimates are presented in Figure 2.14. This analysis suggests that positive affect deactivation had an important effect over time on positive affect activation, but the converse was not true. In other words, when people feel positive emotions with low arousal (e.g., calm and relaxed), then people's positive emotions with high arousal (e.g., energetic and excited) are likely to increase slightly in future situations, but not the other way around.

2.3.3 Summary

To conclude, by analyzing the HND data with the state-trait SEMs, we found evidence supporting the thesis that the factor structure of positive affect consists of two dimensions: Positive affect activation and positive affect deactivation. This evidence supports the circumplex model of affect. Moreover, from the TSO model, we observed that all the items are reliable at measuring positive affect. The analysis also showed that most of the measured emotions are approximately as trait-like as they are state-like. This means that the trait component of these emotions is not negligible even though emotions are prone to fluctuate over time. Finally, the analysis showed that there are important auto- and cross-regressive effects that help explaining the dynamic process that governs the relationship among emotions of positive affect. These results are in concordance with the literature about emotional reactivity (see Kuppens et al., 2010; Suls et al., 1998).

Figure 2.14: Multidimensional TSO with crosslagged effects.



Within Model

Between Model

2.4 Discussion

In this chapter, we explored how to study states and traits in intensive longitudinal data by means of state-trait SEMs. We selected three state-trait SEMs for our study: The MSST, the CUTS, and the TSO models, which represent the variety of state-trait SEMs. Our analyses showed that these models might be especially useful when formulated as multilevel SEM. This facilitates their application to intensive longitudinal data regardless of the number of measurement occasions that are included (see also Geiser, 2020; Geiser et al., 2013). Furthermore, with the empirical example, we also showed how state-trait SEMs allow studying the psychometric properties of the items used in experience sampling studies, and drawing conclusions about the nature of the variables in terms of their state and trait components. As shown in the empirical example, some modeling decisions might be required to facilitate the convergence of the models. For example, centering and constraining intercepts to zero is a safe way to simplify the models. This is possible because intercepts are not needed in the computation of the variance coefficients, which are of major interest in a state-trait SEM analysis.

In relation to our results, out of the three models, the multilevel formulation of the TSO model was the one that performed the best in both the simulation study and the empirical example. Our analyses showed that the multilevel TSO model estimated the parameters with the lower bias across the different conditions and performed equally well even when the real data did *not* have autoregressive effects. For these reasons, and considering that autoregressive effects are likely to occur in intensive longitudinal data, we can conclude that the multilevel TSO model is the most suitable state-trait SEM for analyzing intensive longitudinal data. Note that this is not the case for the single-level TSO model, which not only failed by increasing the number of measurement occasions and the proportion of missing values, but also when the data were generated based on the MSST. In contrast, the performance of both formulations of the MSST and the CUTS models was not as good. For instance, the MSST model introduced bias when there were method or autoregressive effects in the real data. These results are consistent with the results of Geiser and Lockhart (2012), who observed that the MSST model performance is not optimal when there are method effects in the data. Similarly, the CUTS model introduced bias when there were autoregressive effects in the real data.

One of the most important results from the simulation was that both single-level and multilevel TSO models were robust to data that include method effects, meaning that unrealistic autoregressive effects were not estimated even when method effects were present in the data. This result can be explained by the fact that the versions of the TSO used in this study included latent trait indicator variables instead of a common latent trait variable. Including latent trait indicator variables is in fact a way to account for method effects in the state-trait SEM literature (Geiser & Lockhart, 2012). Therefore, this might be the reason why the TSO model as presented in this study was able to correctly deal with data that include real method effects. Basically, the real method effects were captured by the variance of the latent trait indicator variables of the TSO model.

2.4.1 Studying the Psychometric Properties of Intensive Longitudinal Data

One of the main advantages of analyzing intensive longitudinal data with state-trait SEMs is that these models allow studying the psychometric properties of the items used in daily diary studies. As a consequence, it is possible to account for the random measurement error in the data and to study the precision of the items at measuring what they intend to measure. In general, when measuring psychological variables, measurement error is likely to happen (Schuurman & Hamaker, 2019; Schuurman et al., 2015). Therefore, intensive longitudinal data should ideally be analyzed through models that account for measurement error, such as state-trait SEMs.

Furthermore, state-trait SEMs allow quantifying the precision of the measurements in terms of the variance coefficients proposed in the LST theory. As shown in the empirical example, one can estimate the reliability of each item and identify the proportion of random measurement error that is present in the data. In addition, the LST theory estimates the consistency and the occasion-specificity coefficients. These coefficients allow for additional interpretations of the data and they can be used to make decisions about the items for future applications. In the empirical example, it was shown that the positive affect emotions are as trait-like as they are state-like. For example, the items *content* and *calm* had a consistency considerably larger than their occasion-specificity, which meant that they where more trait-like than state-like. Note that other measures used in traditional multilevel modeling such as ICC (Houben et al., 2020) also indicate to what extent a variable is trait- or state-like. However, these kinds of measures might be biased by the amount of random measurement error present in the data (Muthén, 1991; Roesch et al., 2010), which results in overestimation of the variability within individuals. Therefore, models that account for the random measurement error can more accurately determine the trait-state nature of the variables.

Even though one of the advantages of state-trait SEMs is the possibility to study the psychometric properties of the questionnaires used in intensive longitudinal data, this topic has also been addressed from other perspectives. Within multilevel approaches, it is common to study these psychometric properties during the pre-analysis of the data (see Geschwind et al., 2011; Nezlek, 2007). Moreover, alternative reliability coefficients have been recently proposed. Specifically, in the multilevel-VAR model with measurement error (Schuurman & Hamaker, 2019), the authors proposed two reliability coefficients: The between-person reliability and the within-person reliability. Similarly, in the context of confirmatory factor analysis for intensive longitudinal data, Hu et al. (2016) proposed the intraindividual reliability and the interindividual reliability. While these coefficients seem similar to the consistency and the occasion-specificity coefficients of the LST theory, they are not necessarily equal. Both Schuurman and Hamaker (2019) and Hu et al. (2016) acknowledge that the reliability of a questionnaire can vary from person to person, allowing the within-person reliability or the interindividual reliability to vary over persons. This is not the case for the variance coefficients proposed within the LST theory.

2.4.2 Limitations and Future Research

There are some important limitations when applying state-trait SEMs to intensive longitudinal data due to the assumptions that are made. In particular, state-trait SEMs assume that measurement occasions are equally spaced over time, that parameters are equally valid for all the individuals in the sample, that parameters do not vary over time when the models are formulated as multilevel SEM, and that time series are assumed to be stationary. Next, we discuss each of these limitations and possible solutions if available.

Firstly, assuming that measurement occasions are equally spaced over time can be an unrealistic assumption for intensive longitudinal data. This is because, in most studies, individuals are required to fill the questionnaires at random intervals during the day. Therefore, assuming that the measurements are equally spaced over time does not hold and the parameter estimates can be misleading (Crayen et al., 2017; de Haan-Rietdijk et al., 2017b). This is especially problematic when auto- and cross-regressive effects are included because a model like the TSO can only estimate one autoregressive parameter, which is tied to a specific time interval. Thus, if the measurement occasions are unequally spaced over time, such an autoregressive effect would mean that the effect of the variable Y at time $t-1$ on itself at time t is the same regardless of the elapsed time between occasions. One approach to deal with data collected at unequal intervals, suggested within the DSEM framework (Asparouhov et al., 2017, 2018), is to include missing values in such a way that the time interval between occasions becomes approximately equal over time. Alternatively, one can use continuous-time models to deal with unequal time intervals[8], which is strongly suggested as ignoring the violation of this assumption results in biased estimates of the auto- and cross-regressive effects (de Haan-Rietdijk et al., 2017b). Some of the most common continuous-time models used to analyze intensive longitudinal data are encompassed within the hidden Markov model framework. For example, a continuous-time mixture LST Markov model has been proposed by Crayen et al. (2017) for analyzing categorical items.

Secondly, assuming that parameters are equally valid for all the individuals in the sample violates one of the principles of intensive longitudinal methods, which is the emphasis on the individual (Hamaker, 2012). Given this principle, it is logical to expect that relationships between variables or even the factor structures are not exactly the same across all the persons. This is not taken into account when the parameters are estimated for the sample as a whole, as in state-trait SEMs. In general, to account for these individual differences, parameters can be modeled as random effects. This is the standard procedure in the multilevel vector autoregressive (ML-VAR) literature, where auto- and cross-regressive effects are treated as random slopes (Bringmann et al., 2013; De Haan-Rietdijk et al., 2016; Rovine & Walls, 2006; Schuurman et al., 2016). Recent developments also allow for random residual variances (Jongerling et al., 2015). For state-trait SEMs, an important avenue for future research is therefore to allow for random loadings, autoregressive effects, or residual variances, which can be done by extending the models within the DSEM framework (Asparouhov et al., 2017, 2018). Alternatively, some authors have suggested a more *bottom up* approach

[8] Selig et al. (2012) have proposed an interesting approach to deal with unequal time intervals in panel data, which includes the time as a moderator of the lag effect. However, this method still needs to be extended for intensive longitudinal data.

(Adolf et al., 2017; Molenaar, 2004; Ram et al., 2017), where each single-subject time series is first analyzed before pooling out results for the sample.

Thirdly, assuming that parameters are time-invariant and that the time series are stationary implies the type of psychological dynamics that can be studied is very restricted. In other words, these assumptions imply that the dynamic process is static. Therefore, time series collected during a psychological intervention, where the intention is to change the dynamic process, can not be analyzed under these assumptions. Moreover, assuming that the dynamic processes of interest are static may not be accurate in most cases (Bringmann et al., 2017; De Haan-Rietdijk et al., 2016). These limitations not only apply to state-trait SEMs but also to most well-known techniques such as ML-VAR. Alternative methods that relax these assumptions are in development, for example, Bringmann et al. (2017) recently proposed a semi-parametric time-varying autoregressive model to address this problem in the ML-VAR literature. Similarly, De Haan-Rietdijk et al. (2016) proposed a model that estimates a different autoregressive effect given a threshold, which is useful to analyze persons whose dynamics vary given the strength of the psychological variable. The DSEM framework (Asparouhov et al., 2018) also allows relaxing these assumptions with the cross-classified DSEM models, which include time as an additional level of analysis. Lastly, if analyzing time series from a single individual, one can consider the extended Kalman filter with iteration and smoothing estimator (Molenaar et al., 2009). Hopefully, future research will allow extending these methods to multilevel state-trait SEMs.

Finally, notice that in this study, we only covered three fundamental state-trait SEMs. This largely restricts the kind of research questions that can be addressed by these models. However, more complex state-trait SEMs have been proposed over the years. For example, Courvoisier et al. (2007) presented a mixture LST model that can incorporate time-varying covariates. A model like this allows studying the psychological dynamics and how they are affected by other variables, which might answer similar research questions like the ones commonly seen in the multilevel literature. Also, Geiser et al. (2015b) proposed an LST model that accounts for the effects of random and fixed situations. This model allows examining to what extent traits can be specific to a situation. In addition, this model has also been formulated as a multilevel SEM model. To summarize, more complex LST models allow addressing more complex research questions but researchers might need to reformulate them as multilevel SEM in order to be able to study intensive longitudinal data with them.

2.4.3 Conclusion

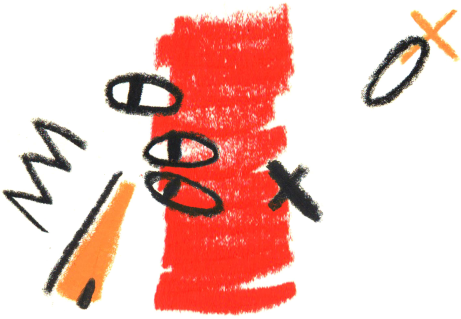
Through this study, we explored the suitability of state-trait SEMs to analyze intensive longitudinal data. Our results showed that out of the three models included in this study, the multilevel TSO model performed the best in both the simulation study and the empirical example. This makes the multilevel TSO model an interesting alternative, which might be useful when analyzing intensive longitudinal data if multiple items were used to measure the same construct. Particularly, we encourage the use of the multilevel versions of state-trait SEMs estimated by means of Bayesian estimation, as the analyses under these settings performed the best in terms of model convergence and accurately recovered the parameters throughout the different conditions. Yet, it might be a disadvantage that there are not enough absolute or relative fit measures that allow comparing the goodness of fit of different models.

In general, for researchers that are interested in applying state-trait SEMs to daily diary data, it is important to highlight that these models are applicable under certain specific conditions such as that (a) there should be a latent construct that is measured with multiple items on the intensive longitudinal study, (b) the variables should preferably be measured on a continuous scale, (c) the time interval between occasions should be equal over time, and (d) the time series is assumed to be stationary. Through these models, researchers can answer research questions about the factor structure of their questionnaires, the proportion if the state and trait components of their variables, and study the psychometric properties of the items used. Beware that fitting state-trait SEMs can be subject to the same modeling decisions that are made when using multilevel modeling such as data centering and item selection. In practice, if the data does not include too many measurement occasions (no more than 30), researchers can consider using the single-level models and allow for time-varying parameters to test whether measurement invariance holds.

In this study, we aimed to present state-trait SEMs as alternative tools that can be considered when analyzing intensive longitudinal data. Moreover, we connected the state-trait SEM literature to a broader literature, showing how these models are related to different frameworks that were especially developed to analyze intensive longitudinal data. We consider this is an important opportunity to integrate these different frameworks, which will allow extending state-trait SEMs to more accurately describe psychological dynamics.

Chapter 3

Mixed-Effects Trait-State-Occasion Model: Studying the Psychometric Properties and the Person-Situation Interactions of Psychological Dynamics



This chapter has been published as Castro-Alvarez, S., Tendeiro, J. N., de Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022). Mixed-Effects Trait-State-Occasion Model: Studying the Psychometric Properties and the Person–Situation Interactions of Psychological Dynamics. *Structural Equation Modeling*, *29*(3), 438–451. https://doi.org/10.1080/10705511.2021.1961587

Abstract

The trait-state-occasion model (TSO) is a popular model within the latent state-trait theory (LST). The TSO allows distinguishing the trait and the state components of the psychological constructs measured in longitudinal data, while also taking into account the carry-over effects between consecutive measurements. In the present study, we extend a multilevel version of the TSO model to allow for the combination of fixed and random situations, namely the mixed-effects TSO (ME-TSO). Hence, the ME-TSO model is a measurement model suitable to analyze intensive longitudinal data that allows studying the psychometric properties of the indicators per individual, the heterogeneity of psychological dynamics, and the person-situation interaction effects. We showcase how to use the model by analyzing the items of positive affect activation of the crowdsourcing study HowNutsAreTheDutch (HoeGekisNL).

Keywords: *trait-state-occasion model, dynamic structural equation modeling, person-situation interaction*

Intensive longitudinal data are rich and complex data, which allow zooming in on the day-to-day life of individuals. This has brought great opportunities but also great challenges to researchers studying psychological dynamics (Hamaker et al., 2015; Hamaker & Wichers, 2017). With intensive longitudinal data, researchers can study in detail the stability and variability of the persons' attributes in short periods of time and how these attributes are related to each other over time. Furthermore, researchers can include time-invariant and time-varying covariates to fully explore how trait-like variables, situational variables, and situational circumstances have an effect on the dynamic process of interest. Yet, intensive longitudinal methods, such as ambulatory assessment (Trull and Ebner-Priemer, 2014, 2020) can also be a burden for the participants due to the frequency of the measurements and because answering the questionnaires can interfere with the activities and interactions in their daily life (Shiffman et al., 2008).

To keep the burden as low as possible, questionnaires tend to be short and in many situations one single question is used as a direct measure of the psychological attribute of interest (e.g., the network dynamics of symptoms, Bringmann et al., 2013). Nevertheless, in some cases researchers do use multiple items to measure one unique construct such as when measuring positive or negative affect (e.g., Geschwind et al., 2011; Snippe et al., 2015). In these cases, a common practice is to compute the sum scores and to study the dynamics of these scores. However, using sum scores can mask the underlying structure of the construct and lead to biased and unreliable results (Fried et al., 2016; McNeish & Wolf, 2020).

To tackle these issues, researchers can use measurement models for intensive longitudinal data. The most common measurement models used to analyze intensive longitudinal data are multilevel structural equation models (multilevel SEM; Geiser, 2020; Geiser et al., 2013; Roesch et al., 2010) and multilevel dynamic factor analysis (DFA; Fuller-Tyszkiewicz et al., 2017; Song & Zhang, 2014). On the one hand, multilevel SEMs are confirmatory factor models that separate the within-person (state) and the between-person (trait) components of the observed variables while accounting for measurement error. Some of these models such as the multistate-singletrait model are encompassed in the so-called latent state-trait theory (LST; Geiser, 2020; Steyer et al., 2015). In particular, the LST theory explicitly defines variance coefficients to study the psychometric properties of the observed indicators. These variance coefficients estimate the reliability of each indicator and to what extent the indicator is trait- or state-like (Steyer et al., 2015). A shortcoming of multilevel SEMs, however, is that these models estimate the parameters of interest for the whole sample ignoring the individual heterogeneity. In contrast, multilevel DFA models fully explore both inter- and intra-individual differences by allowing parameters to vary across individuals, while also including random autoregressive effects to study individual dynamics (Song & Zhang, 2014). Autoregressive effects are of key interest in intensive longitudinal data because they capture the influences of the variables on themselves over time. Note that multilevel DFA models also account for measurement error, hence, reliability estimates per individual have been proposed for these models (Fuller-Tyszkiewicz et al., 2017).

However, most applications of both multilevel SEM and DFA models have been mainly focused on exploring the dynamic latent structure without taking into account time-invariant or time-varying covariates. As a consequence, little can be said about the person-situation interactions with these kind of models. In general, time-invariant covariates, also referred to as between-covariates, include trait-like variables such as neuroticism or optimism. In contrast, time-varying covariates, also known as within-covariates, include situational circumstances like being in a party or at work. Especially time-varying covariates can be of great relevance to understand psychological dynamics because they provide insight into the development of the dynamic process (McNeish & Hamaker, 2020; Ram & Gerstorf, 2009) as well as into the person-situation interactions (Geiser et al., 2015b). In particular, Geiser et al. (2015b) proposed an LST approach for the combination of random and fixed situations to study person-situation interactions, which includes characteristics of the situations as time-varying covariates.

In this chapter, following Geiser et al. (2015b), we introduce a comprehensive measurement model for intensive longitudinal data to study the person-situation interaction. This extension, which we call the mixed-effects trait-state-occasion model (ME-TSO), is fully encompassed within the dynamic structural equation modeling framework (DSEM; Asparouhov et al., 2017, 2018). More specifically, the ME-TSO is an extension of the trait-state-occasion model (TSO; Cole et al., 2005; Eid et al., 2017), which is an LST model (Steyer et al., 2015). In a nutshell, the ME-TSO model allows (a) distinguishing the trait and the state components of the variables, (b) studying individual dynamics by including random autoregressive effects, (c) analyzing the person-situation interactions by adding time-varying and time-invariant covariates, and (d) evaluating the psychometric properties of the items used in intensive longitudinal data.

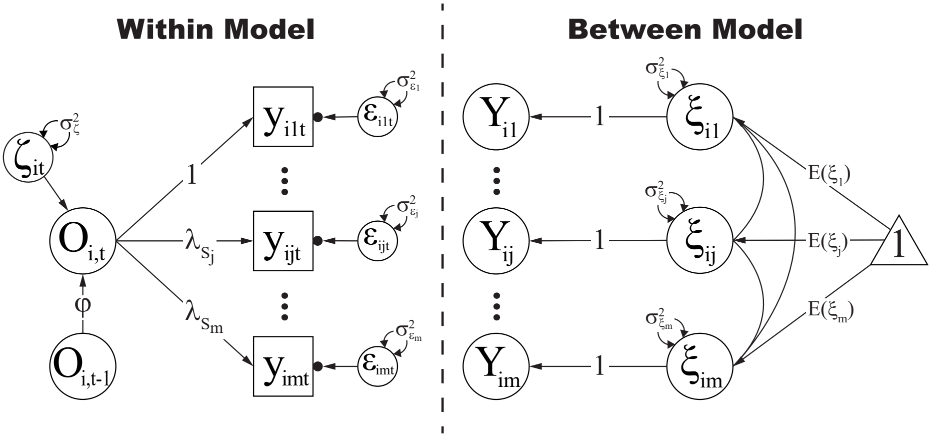
The outline of this chapter is as follows. We first provide a detailed review of the TSO model. Next, we introduce the ME-TSO model, which accounts for the person-situation interaction and the heterogeneity of the individuals. We also discuss the implications of this extension for the definition and computation of the variance coefficient components traditionally defined in the LST theory. Furthermore, we provide an empirical application of the ME-TSO model, in which we analyze the items of positive affect activation from the HowNutsAretheDutch crowdsourcing study (van der Krieke et al., 2017; van der Krieke et al., 2016). Lastly, we discuss the advantages and limitations of this new approach. Code to implement this model is available in the git repository https://github.com/secastroal/ME-TSO.

3.1 The Trait-State-Occasion Model

The TSO model (Cole et al., 2005; Eid et al., 2017) is a model encompassed within the LST theory (Steyer et al., 1992; Steyer et al., 2015). Initially, the TSO model was introduced and applied as a longitudinal SEM (e.g., Cole et al., 2009; Conway et al., 2018; Eid et al., 2017; Musci et al., 2016). This means that it has been presented as a single-level model, which requires the data to be in wide format (i.e., one row per subject and one column for each repeated measure). However, we have previously presented a multilevel version of the TSO model (Castro-Alvarez et al., 2022d), the path diagram of which is shown in Figure 3.1. In contrast to the single-level TSO model, the multilevel TSO model requires the data in long format and is therefore more suitable to analyze intensive longitudinal data. Moreover, while the single-level TSO model allows testing for longitudinal measurement invariance (Meredith, 1993; Meredith & Teresi, 2006), the multilevel TSO model assumes that longitudinal measurement invariance holds. In other words, the multilevel TSO model assumes that the measurement of the construct of interest (e.g., positive affect) over time does not change. Lastly, the multilevel TSO model is easier to apply and estimate than the single-level TSO model because it has a smaller number of free parameters and it allows to easily handle many measurement occasions and missing data (Castro-Alvarez et al., 2022d; Geiser, 2020; Geiser & Lockhart, 2012).

In a nutshell, the TSO model (both single-level and multilevel) acknowledges that psychological variables are not purely traits nor purely states but a combination of both (Cole et al., 2005; Eid et al., 2017). In addition to this, the TSO model accounts for the autoregressive structure likely to be found in longitudinal data. Hence, the

Figure 3.1: Multilevel TSO Path Diagram



Note. Path diagram of the multilevel TSO model. y_{ijt}: Observed scores of person i of the j-th indicator at time t. ε_{ijt}: Measurement error of person i of the j-th indicator at time t. $\sigma^2_{\varepsilon_j}$: Measurement error variance of the j-th indicator. $O_{i,t}$: Latent occasion-specific score of person i at time t. λ_{S_j}: Factor loading from $O_{i,t}$ on the j-th indicator variable. φ: Autoregressive effect between consecutive latent occasion-specific variables $O_{i,t-1}$ and $O_{i,t}$. $\zeta_{i,t}$: Latent occasion-specific residual of person i at time t. σ^2_ζ: Variance of the latent occasion-specific residual. ξ_{ij}: Latent trait score of person i of the j-th indicator. $E(\xi_j)$: Overall mean of the latent trait variable of the j-th indicator. $\sigma^2_{\xi_j}$: Variance of the latent trait variable of the j-th indicator. Furthermore, the dots at the end of the arrows in the within-level model represent the random intercepts of each indicator variable, which are then represented as the latent variables Y_{ij} in the between-level model. Random intercepts are commonly depicted as latent variables in SEM models due to the latent centering of the dependent variables.

TSO model can not only distinguish between the trait and the state components of the psychological variables of interest, but also accounts for the temporal effects from a previous occasion on the next occasion. In other words, autoregressive effects represent the tendency of persons to feel or behave as they were feeling or behaving just the moment before. For example, in case of a positive autoregressive effect, a person that starts their day with a relatively good mood will probably keep being in a good mood throughout the day. One key difference of the TSO model with other LST models is that the trait variables ξ_j represent the trait of the persons on the first measurement occasion and are not necessarily stable over time. This property is extensively explained in Eid et al. (2017).

3.2 Mixed-Effects Trait-State-Occasion Model

While the multilevel TSO model mentioned in the previous section is more suitable to analyze intensive longitudinal data than its single-level counterpart, it is still limited. Firstly, a clear limitation of the TSO model is that it assumes that the different parameters of the model apply to the whole sample. This ignores one of the main principles of intensive longitudinal methods, which is the emphasis on the individual and individual heterogeneity. Considering the autoregressive effect, it is reasonable to expect that this effect will be different among multiple individuals (Kuppens et al., 2010; Nesselroade, 1991). For example, if we are measuring positive affect, persons that are very optimistic might have a higher autoregressive effect than persons that are more pessimistic. This may imply that a high positive affect at the beginning of the day will have a larger impact throughout the day even if something negative happens. On the other hand, persons that are more pessimistic might be more responsive to the different situations throughout the day. As a result, these persons' positive affect might vary more and they would be better described by a lower autoregressive effect on positive affect.

Secondly, another limitation of the TSO model is that the effects of the situation and the person-situation interaction are confounded in the latent occasion-specific residual term ζ_{it}. This means that the model cannot indicate if the variability in the construct of interest is due to specific situations. For example, consider a person who is struggling at work during a daily diary study on positive affect. Probably, the measurements of this person when they were at work will show lower levels of positive affect in contrast to the measurements taken when they were not. As with most measurement models for intensive longitudinal data, these differences are currently not captured or modeled with the multilevel TSO model.

To overcome these limitations, we extended the multilevel TSO model within the DSEM framework (Asparouhov et al., 2017, 2018) in combination with the LST random and fixed situation approach (LST-RF) proposed by Geiser et al. (2015b). First of all, DSEM is a framework that has been especially developed to analyze intensive longitudinal data. To account for the dynamics in the data, DSEM allows to easily include observed and latent lagged variables of any order (Hamaker et al., 2018). Furthermore, it also allows the within-level parameters such as regression coefficients, factor loadings, and residual variances to vary randomly across individuals. Note that all random parameters are then modeled as latent variables in the between-person level model. DSEM is fully implemented within the Bayesian framework in Mplus (Version 8.0 or newer, Muthén & Muthén, 2017) but fitting DSEM models in any other Bayesian software (e.g., JAGS, Stan) is also possible.

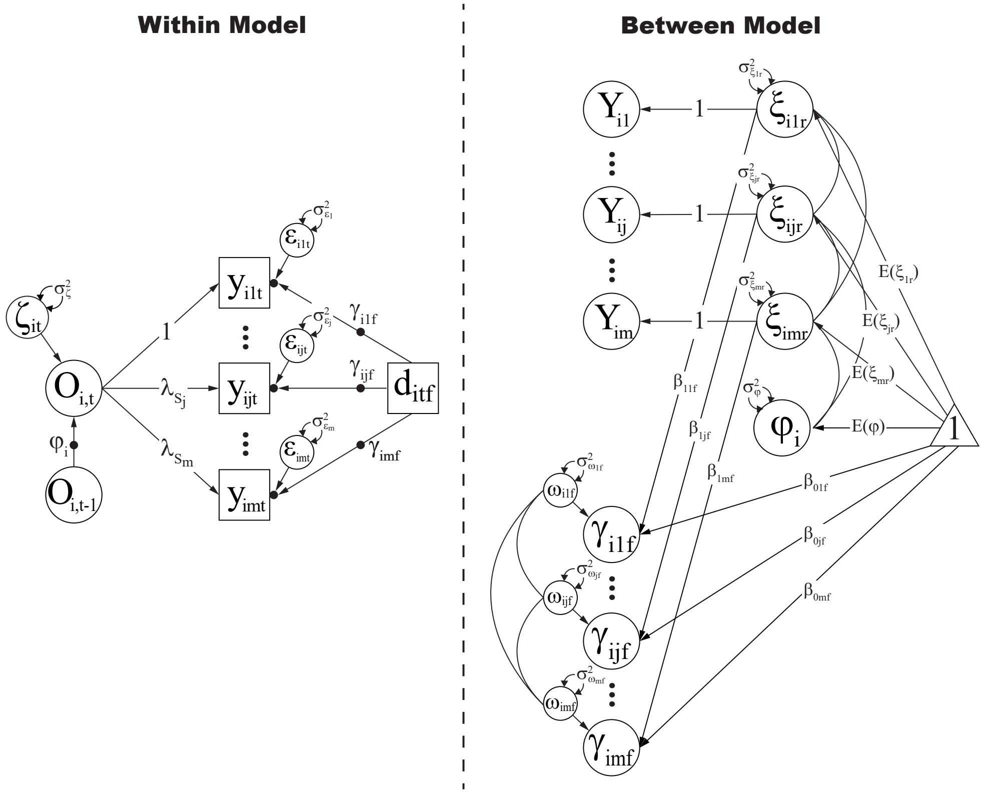
A first step into extending the TSO model is allowing the within-person parameters to vary randomly over the sample. The most flexible and unconstrained extension would be to allow the autoregressive effect φ, the factor loadings of the latent occasion-specific residual λ_S, and the variance of the latent occasion specific residual σ^2_ζ to vary across individuals. This means adding an i subscript to all of these parameters. However, such a complex DSEM model would require a large sample and long time series to deliver reliable estimates (Schultzberg & Muthén, 2018), which are conditions that are rarely satisfied in intensive longitudinal psychological research (Vachon et al., 2019). To strike a balance between flexibility and practical feasibility, and based on the common statistical techniques used to analyze intensive longitudinal data (Bringmann et al., 2013; Kuppens et al., 2010), we prioritize allowing the regression slopes and the autoregressive coefficients to vary randomly across individuals.

Additionally, to distinguish between the effect of the situation and the effect of the person-situation interaction, we consider the LST-RF approach (Geiser et al., 2015b). This approach includes situational variables as dummy variables to identify *situation-specific traits*, *situation effects*, and *person-situation interaction effects*. The LST-RF approach assumes that the situations where the measurements take place can either be random or fixed depending on the study design and on the researcher's knowledge about the situations. In general, random situations are situations of which the specifics are unknown. This can happen by design. For example, in an ambulatory assessment study the measurements are collected at random situations throughout the duration of the study, while no information about the situation was collected. In contrast, fixed situations are situations of which information is available to the researcher. For example, in experimental designs, the situation in which the experiment takes places can be manipulated to create different conditions. Moreover, fixed situations can occur "naturally" if individuals had to report at the moment of measurement whether they were at home, at work, or at some other place (Geiser et al., 2015b).

In particular, in intensive longitudinal studies, random and fixed situations can be combined into the design if details about the situation are collected. In this case, the repeated measurements are observed throughout the study across diverse random situations, which might share some characteristics (e.g., place or time of the day). These random situations that have something in common define a fixed situation that might be of interest for the research question. This means that the random situations are nested in a few fixed situations. For example, during an intensive longitudinal study, participants report their emotions across several days. Of course, the situations in which the measurements take place are very diverse, depending on where the participants were, with whom, if they were hungry, if they just did some exercise, etc. These are random situations. However, if they also report whether they were alone at the moment of measurement, then, we can group the random situations into two fixed situations: Being alone and not being alone. Collecting information about the situations where the measurements happen allows studying the impact of the situation on the behavior or attitudes of interest and the person-situation interaction (Geiser et al., 2015b).

Here, we present how to extend the multilevel TSO within the DSEM framework and the LST-RF approach (see Figure 3.2), namely the mixed-effects TSO model (ME-TSO). The ME-TSO requires a set of indicators (i.e., item scores or sumscores) that measure the same psychological construct over time in an intensive longitudinal study. Consider, for example, an ambulatory assessment study where the participants report m positive emotions such as being cheerful or being enthusiastic multiple times a day for a couple of weeks (with $m \geq 2$). Let y_{ijt} be the observed score of person i on variable j at time t. For example, $y_{2,1,5}$ will be the observed score of the second person on the first emotion at the fifth time point of the ambulatory assessment. To facilitate the presentation of the model, consider that the observed scores y_{ijt} are the responses to a set of items that measure positive affect, with $j = 1, 2, \ldots, m$. Now,

Figure 3.2: ME-TSO Path Diagram



Within Model

Between Model

Note. Path diagram of the ME-TSO model. In the within-level model, the dots at the end of an arrow represent the random intercepts and the dots on an arrow represent the random slopes and the random autoregressive effect. All these random effects are represented as latent variables in the between-level model.

let Y_{it} be an m-variate vector that encompasses the observed scores of the positive emotions of person i at time t, as follows:

$$Y_{it} = \begin{bmatrix} y_{i,1,t} \\ y_{i,2,t} \\ \vdots \\ y_{i,m,t} \end{bmatrix}. \tag{3.1}$$

Additionally, in the ME-TSO model, the situation under which the observations Y_{it} were collected matters, as it allows studying the person-situation interaction. Then, given $l+1$ mutually exclusive fixed situations s_0, s_1, …, s_l, one of the fixed situations is defined as the reference situation and the other l fixed situations are added as dummy variables to the model. Furthermore, as the model is encompassed within the LST theory (Steyer et al., 2015), it assumes that the observed scores of the positive emotions are measured with error. Hence, the measurement model of the ME-TSO model is defined as follows:

$$Y_{it} = \tau_{it} + \varepsilon_{it}, \tag{3.2}$$

where τ_{it} is a m-variate vector with the true scores of the positive emotions of person i at time t. These true scores are referred to, in the LST theory, as the latent states (Steyer et al., 2015). In our example, they represent the error-free positive state emotions of a person at the situation of reference where the measurement took place. In contrast, ε_{it} is an m-variate vector with the deviations of the observed scores from the latent states (true scores). These deviations are known as random measurement error. They capture the unsystematic variability of the observed scores that is not due to the person, the situation, or the person-situation interaction (Steyer et al., 2015). The random measurement errors ε_{it} are assumed to be uncorrelated and normally distributed with means of zero and $m \times m$ diagonal covariance matrix Σ_ε.

Moreover, the true scores τ_{it} are further decomposed as follows:

$$\tau_{it} = \alpha + \Lambda_T \xi_{ir} + \Gamma_i d_{it} + \lambda_S O_{it}, \tag{3.3}$$

where α is an m-variate vector with the intercepts of the observed emotions, which can be interpreted as the grand means. Next, ξ_{ir} is an m-variate vector with the factor scores of the latent indicator- and situation-specific traits of the positive emotions of person i. These trait scores are assumed to represent the trait level of the positive emotions of person i at the first measurement occasion when the person was alone. The trait scores of the first measurement occasion influence all the future latent states τ_{it} with $t > 1$ (Eid et al., 2017). Next, Λ_T is just the $m \times m$ identity matrix. Note that the latent indicator-specific traits ξ_{ir} are latent variables with just one indicator. Therefore, the loadings need to be fixed to 1, otherwise the model is unidentified. Moreover, the latent trait scores are assumed to be normally distributed with a mean of zero and an $m \times m$ covariance matrix Σ_ξ. Alternatively, one can set the intercepts α to zero and estimate the means of the latent traits ξ_{ir} (as shown in Figure 3.2). Then, d_{it} is an l-variate vector of 0s and 1s that indicates the fixed situation of person i at time t, and Γ_i is an $m \times l$ matrix with the effects of the fixed situations on each of the indicators of person i. For example, if the fixed situations are "being alone" and "not being alone", "being alone" can be the reference situation. In this case, d_{it} is just a scalar (0 if the person was alone or 1 if the person was not) and Γ_i is an m-variate vector with the effects of "not being alone" on each of the positive emotions of person i. These coefficients indicate by how much the latent trait scores of the positive emotions increase or decrease when the person is not alone. Hence, they represent the effect of the situation. Note that Γ_i is a matrix of random slopes, which can be further modeled in the between-person equations. Lastly, O_{it} is a scalar that represents the score of the latent occasion-specific variable of person i at time t, which, in our example, is a combination of the state of positive affect at time t and the carry-over effect of states of positive affect from previous time points. The latent occasion-specific variable O_{it} at time t is related to the states of positive emotions τ_{it} at time t via the m-variate vector with the factor loadings λ_S.

Finally, the ME-TSO model acknowledges the dynamic nature of persons by assuming that the latent occasion-specific variables follow an autoregressive structure of order 1. This means that the latent occasion-specific variable at time t is regressed on the latent occasion-specific variable at time $t-1$, that is,

$$O_{it} = \varphi_i O_{i,t-1} + \zeta_{it}, \tag{3.4}$$

where $O_{i,t-1}$ is a scalar that represents the score of the latent occasion-specific variable of person i at time $t-1$; ζ_{it} is a scalar that captures the residual of the autoregressive process of person i at time t, which is referred to, in the LST theory, as the latent occasion-specific residual. This residual represents the pure state of positive affect that is only due to the situation without the influence of the trait, the states of previous measurements, or the interaction between the person and the fixed situations. The latent occasion-specific residual ζ_{it} is assumed to be normally distributed with mean zero and variance σ^2_ζ. Finally, φ_i is the autoregressive effect of person i, which represents the individual carry-over effect between consecutive states of positive affect. In other words, φ_i is a random slope that is normally distributed with mean $E(\varphi)$ and variance σ^2_φ (see Figure 3.2).

Additional advantages of extending the TSO model within the DSEM framework are that observed variables are latent person-mean centered for the analysis and that the model can handle observations that are unequally spaced over time. Firstly, DSEM uses latent centering (Asparouhov et al., 2018; McNeish & Hamaker, 2020), which means that the observed variables are centered based on their latent intraindividual means instead of the observed intraindividual means. This is better, because using latent centering implies that all the fluctuations and random error are captured in the within-level model. As a result, the within-person effects are more meaningful and interpretable in comparison with analyses when no centering or when grand mean centering is used (McNeish & Hamaker, 2020). Furthermore, latent centering avoids Nickell's bias for the autoregressive effects and Lüdtke's bias for the effects of other time-varying covariates (Asparouhov et al., 2018; McNeish & Hamaker, 2020), which can appear when observed person-mean centering is used.

Secondly, a common challenge of intensive longitudinal data is that measurements are not equally spaced over time. When this happens, the auto- and cross-regressive effects, which are parameters of key interest in dynamic models, do not have a clear meaning given that the size of these effects depends on the size of the time interval between the measurements. To handle this issue in DSEM, one can include additional missing values to approximate the measurements to be relatively equally spaced over time (Hamaker et al., 2018; McNeish & Hamaker, 2020). This technique offers results that are similar to the results obtained via continuous time models (de Haan-Rietdijk et al., 2017b) and it retrieves good estimates with a percentage of missing values as large as 80% (Asparouhov et al., 2018).

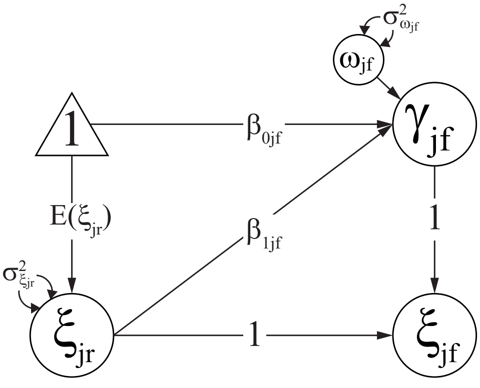
Lastly, to study the person-situation interaction, the LST-RF approach (Geiser et al., 2015b) proposes to regress the random slopes of the effect of the fixed situations on the trait variables at the between-level model. For the ME-TSO model, this means to regress the random slopes in Γ_i on the respective latent indicator- and situation-specific trait variables ξ_{ir}. This regression is expressed as follows given a slope γ_{ijf}

of the matrix Γ_i:

$$\gamma_{ijf} = \beta_{0jf} + \beta_{1jf}\xi_{ijr} + \omega_{ijf}, \tag{3.5}$$

where γ_{ijf} is the effect of the fixed situation f on the indicator j of person i, and ξ_{ijr} is the factor score of the latent indicator- and situation-specific trait variable of indicator j of person i. This regression is described by the coefficients β_{0jf} and β_{1jf}, which are the intercept and the slope, respectively. The slope β_{1jf} can be interpreted as the person-situation interaction effect. Finally, ω_{ijf} is the residual of the regression, which is assumed to be normally distributed with mean zero and variance $\sigma^2_{\omega jf}$. This is the part of the effect of the situation of person i that cannot be explained by the trait scores of the reference situation. Furthermore, one can add additional time-invariant covariates in Equation 3.5 to further explain the variability of the effect of the situation on the daily observations. Note that the random slope γ_{ijf} represents the difference between the trait of the person at the f fixed situation against the trait of the person at the reference situation (i.e., $\gamma_{ijf} = \xi_{ijf} - \xi_{ijr}$). This underlying structure of the indicator- and situation-specific trait variables is shown in Figure 3.3.

Figure 3.3: Structural Diagram of the Indicator- and Situation-Specific Traits.



Note. Path diagram representing the underlying structure of the relationship between the indicator- and situation-specific traits of the reference and the fixed situations.

The key element of Equation 3.5 is the slope β_{1jf} that represents the interaction between the persons' j-th trait with the f-th fixed situation. For example, let "feeling happy" be one of the positive emotions measured in the ambulatory assessment in which "being alone" was the reference situation and "not being alone" was the fixed situation. The slope β_{1jf} represents the interaction between the trait level of happiness during the reference situation with the effect of “not being alone” on the state happiness. In this case, $\beta_{1jf} > 0$ means that persons with higher scores of trait happiness in the reference situation are more likely to have a stronger situation effect in situations when they are not alone. Hence, being not alone implies a higher increase in state happiness, if the trait happiness of the person is also high. On the contrary, $\beta_{1jf} < 0$ means that persons with higher scores of trait happiness in the reference situation are more likely to have a weaker situation effect in situations when they are not alone.

3.3 Variance Coefficients

A fundamental contribution of LST models is the variance coefficients (Steyer et al., 2015). These coefficients allow studying the psychometric properties of the instruments used in longitudinal studies. In a nutshell, they are defined as proportions of the total variance of each indicator that are explained by certain components of the model. Diverse variance coefficients are defined based on the complexity of the model. However, the most essential variance coefficients, which are defined for every LST model, are the consistency, the occasion-specificity, and the reliability. The consistency is the proportion of the variance of an indicator that is explained by the time-invariant sources of variability. In other words, it indicates to what extent the indicators are trait-like. In contrast, the occasion-specificity is the proportion of the variance of an indicator that is explained by the time-varying sources of variability. The occasion-specificity coefficient, thus, indicates to what extent the indicators are state-like. Lastly, the reliability encompasses both, the time-invariant and the time-varying sources of variability. To put it differently, the reliability is the proportion of the variance that is explained by the true score.

For the ME-TSO model proposed in this study, two sets of coefficients can be defined depending on whether they describe the variability across fixed situations or random situations (Geiser et al., 2015b). First of all, the coefficients across fixed situations are the consistency of traits, the situation-specificity of traits, the person-situation interaction coefficient, and the unique situation effect. These coefficients are derived from the assumed underlying structure of the indicator- and situation-specific traits shown in Figure 3.3. The consistency of trait is defined as $[Corr(\xi_{jr}, \xi_{jf})]^2$, which is the squared correlation of the indicator- and situation-specific trait variable of the reference situation with the indicator- and situation-specific trait variable of the fixed situation f. This coefficient indicates the proportion of variance that is shared between the two indicator- and situation-specific traits. Notice that the correlation between the two indicator- and situation-specific traits ($Corr(\xi_{jr}, \xi_{jf})$) is not directly estimated in the model but it has to be computed based on other parameters of the model as shown in Equation 3.6 (for the mathematical derivation of these equations see Appendix C). Next, the situation-specificity of traits is defined as $1 - [Corr(\xi_{jr}, \xi_{jf})]^2$ and represents the proportion of the variance that is unique between the two indicator- and situation-specific traits.

$$\begin{aligned} Var(\xi_{jf}) &= Var(\xi_{jr}) + 2\beta_{1jf}Var(\xi_{jr}) + \beta_{1jf}^2 Var(\xi_{jr}) + Var(\omega_{jf}) \\ Cov(\xi_{jr}, \xi_{jf}) &= Var(\xi_{jr}) + \beta_{1jf}Var(\xi_{jr}) \\ Corr(\xi_{jr}, \xi_{jf}) &= \frac{Cov(\xi_{jr}, \xi_{jf})}{\sqrt{Var(\xi_{jr})Var(\xi_{jf})}} \end{aligned} \tag{3.6}$$

Furthermore, the person-situation interaction coefficient and the unique situation effect are defined as proportions of the variance of the random effect of the situation, γ_{jf}. The total variance of γ_{jf} is defined as follows:

$$Var(\gamma_{jf}) = \beta_{1jf}^2 Var(\xi_{jr}) + Var(\omega_{jf}), \tag{3.7}$$

which is derived from Equation 3.5. Thus, the person-situation interaction coefficient is the proportion of the variance of γ_{jf} that is explained by the indicator- and situation-specific trait variable of the reference situation (i.e., $\beta_{1jf}^2 Var(\xi_{jr})/Var(\gamma_{jf})$). Therefore, it is the proportion of the variance of the situation effect that is due to the person-situation interactions. In contrast, the unique situation effect is the proportion of the variance that is not explained by the person-situation interactions (i.e., $Var(\omega_{jf})/Var(\gamma_{jf})$). This coefficient should decrease towards 0 when adding more time-invariant covariates to the model as they further explain the person-situation interaction.

Additionally, variance coefficients within fixed situations (across random situations) can also be defined. This means that we can compute the traditional variance coefficients of the TSO model (Eid et al., 2017) for each of the fixed situations. This includes the reliability (Rel), the consistency (Con), the occasion-specificity (Ospe), the predictability by trait (Pred), and the unpredictability by trait (Upred). These coefficients are usually defined for each indicator j at time t, allowing the variance coefficients to change over time[1]. However, in the present study, we adapted all these variance coefficients in such a way that they do not change over time and we defined them for each person i, for each indicator j, and for each fixed situation f. This adjustment aims to provide variance coefficients that are more meaningful for the ME-TSO model, by taking into account the emphasis on the individual and the effect of different fixed situations. Therefore, given the time series Y_{ijf} of the variable j, of person i, in the fixed situation f (with $f = r, 1, \dots, l$), the total variance of Y_{ijf} is defined as follows[2]:

$$Var(Y_{ijf}) = Var(\xi_{jf}) + \lambda_{S_j}^2 \frac{\varphi_i^2}{1-\varphi_i^2} Var(\zeta) + \lambda_{S_j}^2 Var(\zeta) + Var(\varepsilon_j). \tag{3.8}$$

The rationale and derivation of this total variance is included in Appendix C.

Once the total variance is defined, defining the variance coefficients becomes trivial as they are just proportions of the total variance. The equations for the five variance coefficients of the ME-TSO model are shown in Equations 3.9-3.13. As mentioned before, the consistency, the occasion-specificity, and the reliability are defined for every LST model. The only difference in this case is that they are defined for each indicator, each person, and each fixed situation. On the other hand, the predictability by trait and the unpredictability by trait are variance coefficients exclusively defined for TSO models. These two coefficients added together are the consistency. In the first place, the predictability by trait is the proportion of the total variance that is explained by the latent indicator- and situation-specific trait variable. It represents the proportion of the variance that is stable over time and predicted by the indicator- and situation-specific trait of the first measurement occasion. In contrast, the unpredictability by trait is the proportion of the total variance that is explained by the

[1] This is the case for the TSO model even when all the parameters are assumed to be time invariant. This happens due to the autoregressive structure of the model, which makes all the variance coefficients to inevitably change over time. In particular, the reliability, the consistency, and the unpredictability by trait increase over time and tend to an upper asymptote. On the other hand, the occasion-specificity and the predictability by trait decrease over time and tend to a lower asymptote. This is further explained in Appendix C.

[2] The total variance and the variance coefficients within fixed situations (Equations 3.8-3.13) are computed for each individual. This means that the individual estimates of φ_i need to be extracted in order to compute these coefficients.

previous states. To put it differently, the unpredictability by trait represents the proportion of the variance that is due to the autoregressive (carry-over) process, hence it represents some sort of stability over time that is not explained by the indicator- and situation-specific trait of the first measurement occasion. In relation to similar reliability coefficients as the ones proposed by Fuller-Tyszkiewicz et al. (2017) and Schuurman and Hamaker (2019), the advantage of the variance coefficients proposed for the ME-TSO model is that the autoregressive structure is taken into account for its computation.

$$Con(Y_{ijf}) = \frac{Var(\xi_{jf}) + \lambda_{S_j}^2 \frac{\varphi_i^2}{1-\varphi_i^2} Var(\zeta)}{Var(Y_{ijf})} \tag{3.9}$$

$$Pred(Y_{ijf}) = \frac{Var(\xi_{jf})}{Var(Y_{ijf})} \tag{3.10}$$

$$Upred(Y_{ijf}) = \frac{\lambda_{S_j}^2 \frac{\varphi_i^2}{1-\varphi_i^2} Var(\zeta)}{Var(Y_{ijf})} \tag{3.11}$$

$$Ospe(Y_{ijf}) = \frac{\lambda_{S_j}^2 Var(\zeta)}{Var(Y_{ijf})} \tag{3.12}$$

$$Rel(Y_{ijf}) = \frac{Var(\xi_{jf}) + \lambda_{S_j}^2 \frac{\varphi_i^2}{1-\varphi_i^2} Var(\zeta) + \lambda_{S_j}^2 Var(\zeta)}{Var(Y_{ijf})} \tag{3.13}$$

3.4 Applying the ME-TSO Model

In this section, we present the application and interpretation of the ME-TSO model via an empirical example. For this, we analyzed daily diary data collected in the HowNutsAreTheDutch study (Dutch: HoeGekIsNL; van der Krieke et al., 2017; van der Krieke et al., 2016). Data were collected between May 2014 and December 2018. The detailed description of the HowNutsAreTheDutch data and the design of the project are available in van der Krieke et al. (2016). For the present application, we analyzed the items used to measure positive affect, which were measured based on the circumplex model of affect (Feldman Barrett and Russell, 1998, and Yik et al., 1999, as cited in van der Krieke et al., 2016). This means that positive affect emotions are divided into two categories: Positive affect activation and positive affect deactivation. More precisely, in this empirical example, we analyzed the items of positive affect activation. Furthermore, to study the person-situation interaction, we considered two situational variables. Firstly, we considered whether something negative happened between the previous and the current measurement. Secondly, we considered whether the persons were alone or in a social situation at the moment of measurement. These analyses are similar to the analyses presented by Geiser et al. (2015b), which also studied, with different data, how daily fluctuations of positive emotions are related with the fixed situations of being alone versus being in a social situation and whether there were person-situation interaction effects.

In these analyses, we aimed to study the psychometric properties of the items of positive affect activation across different fixed situations. Furthermore, we aimed to study the person-situation interaction effects between the positive emotions and the fixed situations of interest while controlling for the lagged structure in the data. Specifically, we wanted to know whether a negative event has a negative effect on the positive affect activation emotions. Similarly, we explored whether being in a social situation leads to an increase of the items of positive affect. Furthermore, we wanted to know whether there are interaction effects such that the trait component of the positive affect emotions in the reference situation predicts the size of the effect of the fixed situations of interest. Lastly, we added optimism as a time-invariant covariate of the situation effects. Hence, optimism was added to further explain the person-situation interactions. Note that we have previously analyzed these data with the TSO model, without considering random autoregressive effects or person-situation interaction effects (Castro-Alvarez et al., 2022d).

3.4.1 Data

The data for these analyses includes the time series of positive emotions of 644[3] Dutch individuals (83.9% women, mean age 39.9). Participants reported their emotions three times a day for 30 days, resulting in time series with a maximum length of 90 observations per individual. The number of observations per individual ranged from 59 to 90 with a mean of 74.9 observations per participant. The mean missingness rate for the selected sample was 16.7%.

[3] The 644 individuals were selected out of 1396 individuals available in the data. Only individuals that responded to at least 65% (59 out of 90) of the beeps in the daily diary study were considered for the analyses. This criterion is also used within the HowNutsAreTheDutch project to provide personalized feedback to the participants van der Krieke et al. (2016). Furthermore, given that participants were able to do the daily diary multiple times, only the first daily diary of each participant was taken into account.

The items related to positive affect activation were measured on a visual analogue scale (VAS) from 0 to 100. Positive affect activation was measured with the following three items: *Energetic*, *enthusiastic*, and *cheerful*. As mentioned before, some characteristics of the situations where the measurements took place were also reported. These variables were included in the analysis as dummy coded variables. In particular, we are interested in the situations where nothing negative happened (0) versus something negative happened (1), which we refer to as variable *event*; and in the situations when the participants were alone (0) versus when they were not alone (1), which we refer to as variable *alone*. These kind of situations are commonly studied in relation to daily fluctuations of affect (see Elmer et al., 2020; van Roekel et al., 2015; Wichers et al., 2010; Wichers et al., 2009). Lastly, we also included the variable *optimism*, which was measured with the Life Orientation Test - Revised (van der Krieke et al., 2016) during the cross-sectional stage of the HowNutsAreTheDutch project. We included *optimism* mainly to showcase how to add additional time-invariant covariates to the model. In short, to show the possibilities of the ME-TSO model, we included a latent construct measured by three indicators (positive affect activation), a time-varying situational variable (event or alone), and a time-invariant covariate (optimism).

3.4.2 Analyses

We considered four models to analyze the data. Model 1 ($\mathscr{M}_1$) is the ME-TSO model with *event* as the situational dummy variable, to study the effect of a negative event on the daily fluctuations of positive affect and the possible interaction between the persons and the situations where something negative happened. Model 1b ($\mathscr{M}_{1b}$) is the same as $\mathscr{M}_1$ but with the addition of *optimism* as a time-invariant covariate, to study whether *optimism* also plays a role in explaining the person-situation interaction. Model 2 ($\mathscr{M}_2$) is the ME-TSO model with *alone* as the situational dummy variable. Note that in $\mathscr{M}_2$, “being alone” is the reference situation, hence the model studied the effect of being in a social situation on the daily fluctuations of positive affect. Lastly, model 2b ($\mathscr{M}_{2b}$) is also equal to $\mathscr{M}_2$ but with *optimism* as a time-invariant covariate. In particular, based on Equations 3.2-3.4, $\mathscr{M}_{1b}$ can be defined by the following equations:

$$\begin{bmatrix} EG_{it} \\ EN_{it} \\ CH_{it} \end{bmatrix} = \begin{bmatrix} \xi_{EG,i1r} \\ \xi_{EN,i2r} \\ \xi_{CH,i3r} \end{bmatrix} + \begin{bmatrix} \gamma_{EG,i11} \\ \gamma_{EN,i21} \\ \gamma_{CH,i31} \end{bmatrix} EV_{it} + \begin{bmatrix} 1 \\ \lambda_{EN,S_2} \\ \lambda_{CH,S_3} \end{bmatrix} O_{PA,it} + \begin{bmatrix} \varepsilon_{EG,i1t} \\ \varepsilon_{EN,i2t} \\ \varepsilon_{CH,i3t} \end{bmatrix}, \tag{3.14}$$

$$O_{PA,it} = \varphi_i O_{PA,i,t-1} + \zeta_{PA,it}, \tag{3.15}$$

where EG_{it}, EN_{it}, and CH_{it} are the observations of person i at time t of the items *energetic*, *enthusiastic*, and *cheerful*, respectively; $\xi_{EG,i1r}$, $\xi_{EN,i2r}$, and $\xi_{CH,i3r}$ are the latent indicator- and situation-specific trait scores of person i for the reference situation; and $\varepsilon_{EG,i1t}$, $\varepsilon_{EN,i1t}$, and $\varepsilon_{CH,i1t}$ are the measurement errors of person i at time t. Next, EV_{it} is the score of the dummy variable that indicates whether something negative happened for person i at time t; and $\gamma_{EG,i11}$, $\gamma_{EN,i21}$, and $\gamma_{CH,i31}$ are the effects of a negative event on the positive emotions of person i. Furthermore, $O_{PA,it}$ and $\zeta_{PA,it}$ are the latent occasion-specific score and residual of *positive affect activation*; and λ_{EN,S_2} and λ_{CH,S_3} are the factor loadings of the latent occasion-specific variables of positive affect. Recall that the first loading is fixed to 1 for identification purposes.

The effects of the fixed situation ($\gamma_{EG,i11}$, $\gamma_{EN,i21}$, and $\gamma_{CH,i31}$) are further decomposed based on Equation 3.5 as follows:

$$\begin{bmatrix} \gamma_{EG,i11} \\ \gamma_{EN,i21} \\ \gamma_{CH,i31} \end{bmatrix} = \begin{bmatrix} \beta_{011} \\ \beta_{021} \\ \beta_{031} \end{bmatrix} + \begin{bmatrix} \beta_{111} & 0 & 0 \\ 0 & \beta_{121} & 0 \\ 0 & 0 & \beta_{131} \end{bmatrix} \begin{bmatrix} \xi_{EG,1r} \\ \xi_{EN,2r} \\ \xi_{CH,3r} \end{bmatrix} + \begin{bmatrix} \beta_{OPT,1} \\ \beta_{OPT,2} \\ \beta_{OPT,3} \end{bmatrix} OPT_i + \begin{bmatrix} \omega_{EG,i11} \\ \omega_{EN,i21} \\ \omega_{CH,i31} \end{bmatrix}, \tag{3.16}$$

where β_{0j1} and β_{1j1} are the intercept and the slope of the j-th indicator for the situation where something negative happened. The slope represents the person-situation interaction that is the effect of the trait of the situation of reference on the situational effect of a negative event on a certain emotion (e.g., energetic). Additionally, OPT_i is the score of person i on the cross-sectional variable *optimism*, and $\beta_{OPT,j}$ is its respective effect on the j-th indicator. Lastly, $\omega_{EG,i11}$, $\omega_{EN,i21}$, and $\omega_{CH,i31}$ are the residuals reflecting the effect of a negative event on a certain emotion that remains unexplained. Estimates for each of the parameters in Equations 3.14-3.16 are included in Appendix C.

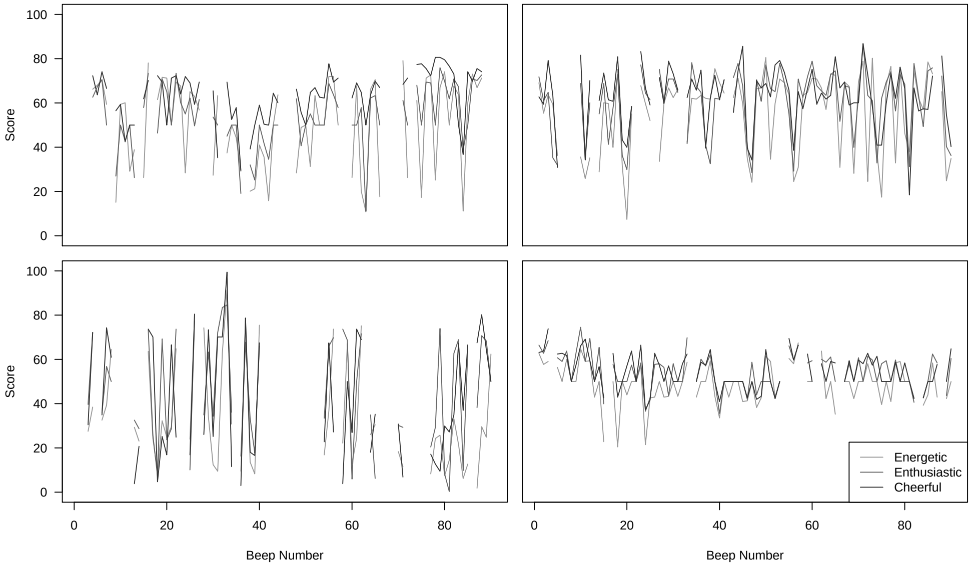
To evaluate the relative fit of the models and to select the best fitting model, we applied the deviance information criterion (DIC). The DIC as well as other relative fit measures indicates that a model fits the data best when the DIC value is the lowest among the competing models. Note that the DICs reported for different DSEM models are not always comparable (Asparouhov et al., 2018). This is a problem especially when comparing DSEM models with several latent variables. To be comparable, the list of latent variables that are treated as parameters needs to be same. This requirement is satisfied in our analyses, given that all four tested model have the same number of latent variables.

3.4.3 Preliminary Steps

As mentioned before, the ME-TSO model requires that multiple items measure the same construct. This is the case in the HowNutsAreTheDutch data with the items of positive affect activation (Castro-Alvarez et al., 2022d). Before fitting the models, we described and visualized the raw data. For example, Figure 3.4 shows the time series of the items of positive affect activation of four individuals. These time series also show that the three items follow similar trends, which is expected because they are supposed to measure the same construct. Furthermore, Figure 3.5 shows the overall differences in the items across the situations when nothing negative happened versus something negative happened. This clearly shows that there is probably a situational effect when something negative happened. By analyzing the data with the ME-TSO model, we can study how this situational effect actually varies across persons and how it might be related to trait-like persons' characteristics.

Moreover, one has to verify that the assumptions of the model are met. In particular, the ME-TSO model assumes that the autoregressive process is stationary and that the observations are equally spaced over time. Regarding stationarity, we used the Kwiatkowski-Phillips-Schmidt-Shin test to study whether the observed time series were trend stationary (Kwiatkowski et al., 1992). This test suggested that 193 individuals had at least one nonstationary time series. However, these kinds of tests tend to be prone to commit Type I errors with short time series ($N \leq 100$; Jönsson, 2011). For this reason, and because the results excluding the individuals with nonstationary time series did not differ substantially from the results with the whole sample, we include the results with the stationary sample in Appendix C. Another assumption is that observations are equally spaced over time. This is not the case in the HowNutsAreTheDutch data due to missing data and the overnight periods. To handle this,

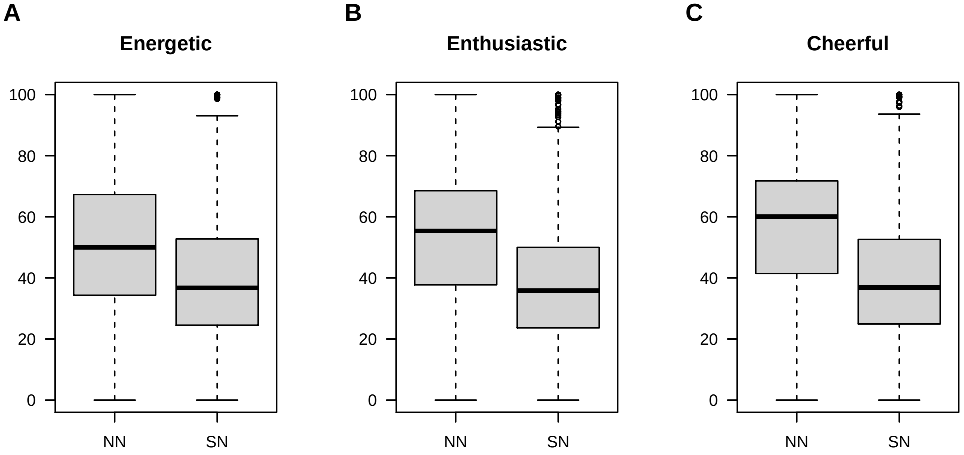
Figure 3.4: Positive Affect Activation Time Series of Four Individuals



Note. Time series of the positive affect activation items of four random individuals. The breaks in the time series mean that the individual did not respond at those beeps.

we approximated the data to be relatively equally spaced over time by including additional missing values. We did this automatically in Mplus with the TINTERVAL command[4](Muthén & Muthén, 2017).

Figure 3.5: Boxplot of Positive Affect Activation Items by Event

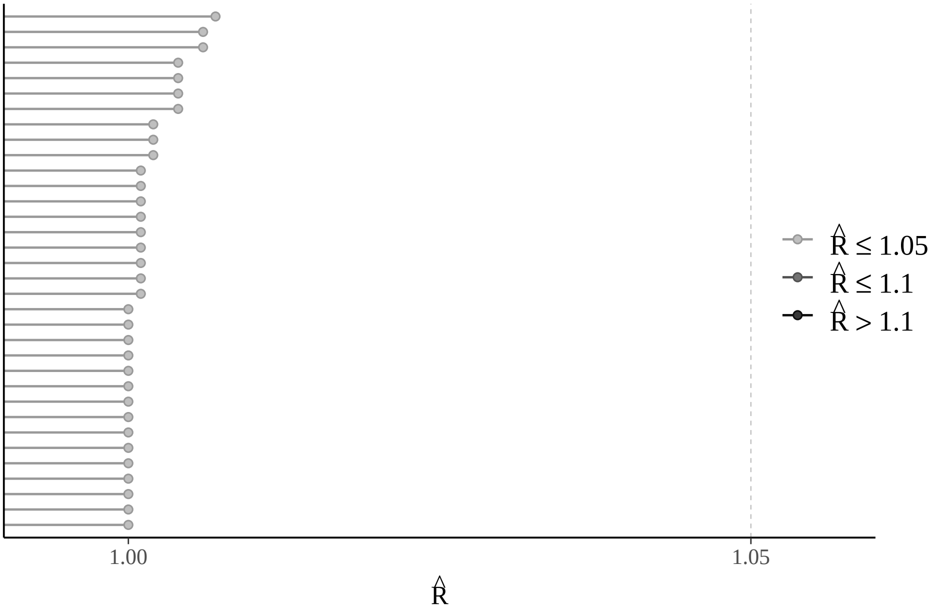


Note. Overall differences in the items of positive affect activation across the variable *event*. NN = Nothing negative happened. SN = Something negative happened.

Finally, as the ME-TSO model is implemented within the Bayesian framework, it is extremely important to verify that the posterior sampling algorithm converged as expected. The convergence of Bayesian models is usually checked via the Gelman-Rubin Statistic ($\hat{R}$; Gelman and Rubin, 1992) and diagnostic plots such as traceplots and autocorrelation plots. Figure 3.6 shows the estimated $\hat{R}$ statistics of $\mathscr{M}_1$, which suggests that the sampling procedure converged. The other tested models also seemed to have converged according to this criterion (See Appendix C).

[4] All the code to run the models is available in the git repository https://github.com/secastroal/ME-TSO.

Figure 3.6: Gelman-Rubin Statistics of $\mathscr{M}_1$



Note. $\mathscr{M}_1$: Model 1. $\hat{R}$: Gelman-Rubin Statistic.

3.4.4 Results

In this empirical example, we disentangled the trait and the state components of the emotions of positive affect activation and we studied how the psychological dynamics of positive mood are influenced by the situation and the person-situation interaction. The results of the tested model are presented in Tables 3.1-3.2. These tables include the estimates and the credibility intervals of the key parameters of the ME-TSO model. Also, the number of free parameters, the DIC, and estimated number of parameters (pD) are reported at the bottom of Tables 3.1-3.2. From this, we can observe that $\mathscr{M}_1$ and $\mathscr{M}_{1b}$ are better at explaining the daily variability of the positive emotions than the models $\mathscr{M}_2$ and $\mathscr{M}_{2b}$. This means that the occurrence of a negative event is more likely to influence the daily fluctuations of positive mood than being alone.

In relation to *optimism*, the analyses showed that adding this variable does not substantially improve the fit of $\mathscr{M}_1$ nor $\mathscr{M}_2$. Additional evidence against $\mathscr{M}_{1b}$ and $\mathscr{M}_{2b}$ is that the amount of unexplained variance of the effects of the situation at the between-level (ω_{jf}) did not decrease (see Tables 3.1-3.2). Therefore, a person's *optimism* typically does not interact with the situation effect (i.e., something negative happening or not being alone) on the daily emotions of the persons.

Additionally, one can look at some key parameters of the model such as the random autoregressive effect φ_i and the interaction effects β_{1jf}. Firstly, the estimated mean autoregressive effect ($E(\varphi)$) in $\mathscr{M}_1$ and $\mathscr{M}_2$ evidenced that there is on average a moderate carry-over effect on the states of positive affect activation. Nonetheless, there are important differences in the lagged relationships across individuals given the estimated variance of the random autoregressive effect ($Var(\varphi) = 0.033$). In other words, there are participants that show little to no carry-over effects on positive affect activation as well as participants that show strong carry-over effects on their positive affect dynamics. Secondly, $\mathscr{M}_1$ showed that there are person-situation interactions between the situational variable *event* and the trait components of each positive emotion ($\beta_{111} = -0.19$, $\beta_{121} = -0.20$, $\beta_{131} = -0.22$). This means that the trait level of the positive affect emotions interacts with the effect of a negative event on the daily emotions of the participants. Therefore, the lower the trait positive emotion of a person, the stronger the negative effect of a negative event on the daily emotions. In other words, the daily emotions of an individual that is not too enthusiastic (trait enthusiastic) will decrease more when something negative happen,

Table 3.1: Unstandardized Estimates of the Key Parameters of the Models using the Situation Variable *Event*

Parameter	$\mathscr{M}_1$ Est. [95% C.I.]	$\mathscr{M}_{1b}$ Est. [95% C.I.]
	Between-level	
Eg-Ev Interaction Effect β_{111}	-0.19 [-0.25, -0.13]	-0.19 [-0.26, -0.10]
En-Ev Interaction Effect β_{121}	-0.20 [-0.26, -0.13]	-0.21 [-0.26, -0.14]
Ch-Ev Interaction Effect β_{131}	-0.22 [-0.29, -0.16]	-0.24 [-0.31, -0.17]
Opt-Eg-Ev Interaction Effect $\beta_{OPT,1}$	—	-0.02 [-0.26, 0.21]
Opt-En-Ev Interaction Effect $\beta_{OPT,2}$	—	0.04 [-0.19, 0.29]
Opt-Ch-Ev Interaction Effect $\beta_{OPT,3}$	—	0.04 [-0.20, 0.29]
AR Effect Mean $E(\varphi)$	0.32 [0.30, 0.34]	0.32 [0.30, 0.34]
Eg-Ev Effect Residual Variance $Var(\omega_{EG,11})$	23.79 [15.05, 34.05]	23.92 [14.76, 34.28]
Et-Ev Effect Residual Variance $Var(\omega_{EN,21})$	41.63 [31.06, 54.34]	41.98 [31.74, 54.52]
Ch-Ev Effect Residual Variance $Var(\omega_{CH,31})$	46.16 [34.97, 59.47]	46.19 [35.00, 59.56]
AR Effect Variance $Var(\varphi)$	0.033 [0.028, 0.039]	0.033 [0.028, 0.039]
	Model Fit Information	
Number of Free Parameters	34	43
DIC	1823224.25	1825196.44
pD	141759.28	141883.55

Note. $\mathscr{M}_1$: Model 1, $\mathscr{M}_{1b}$: Model 1b, Est.: Unstandardized estimate, C.I.: Credibility interval, Eg: Energetic, En: Enthusiastic, Ch: Cheerful, Ev: Event, DIC: Deviance information criterion, pD: Estimated number of parameters.

than the daily emotions of an individual that tends to be enthusiastic. On the other hand, $\mathscr{M}_2$ also showed that there are person-situation interactions between the situation variable *alone* and the trait components of each positive emotion ($\beta_{111} = -0.08$, $\beta_{121} = -0.10$, $\beta_{131} = -0.09$). Therefore, the lower the trait positive emotion of a person, the stronger the positive effect of not being alone on the daily emotions. To put it differently, being in a social situation has a more positive impact on individuals that tend to have low levels of positive emotions when they are alone than on individuals that tend to have high levels of positive emotions. However, the size of these interactions was lower in comparison to the size of the interactions in $\mathscr{M}_1$. Moreover, while the interaction effects in $\mathscr{M}_2$ might be statistically significant, they are not necessarily practically significant.

Finally, we report the variance coefficients of $\mathscr{M}_1$. These coefficients are the added value of the ME-TSO model when compared with more traditional and simpler methods for intensive longitudinal data. In brief, these variance coefficients indicate the strength of the person-situation interaction and allow studying the psychometric properties of the items according to the LST theory. Firstly, the coefficients across fixed situations of the ME-TSO quantify the strength of the person-situation interaction, which, to the best of our knowledge, is not possible in other approaches for intensive longitudinal data. Secondly, the variance coefficients within fixed situations are used to study the psychometric properties of the items and to determine to what extent the items are trait- or state-like. Alternatively, this could be done with for example the between- and the within-reliabilities (Schuurman & Hamaker, 2019) and the intraclass-correlation (Houben et al., 2020). However, these indices might come short when compared with the variance coefficients within fixed situations of the ME-TSO model as they do not account for the autoregressive structure of the data. In what follows, we present and interpret the estimated variance coefficients of $\mathscr{M}_1$.

Table 3.3 shows the estimated variance coefficients across fixed situations. Firstly, the consistency of traits of *energetic* (0.8) was the highest across the three items, which means that the inter-individual differences in *energetic* tend to be consistent across the two situations. Secondly, the person-situation interaction coefficient varied between 13% and 19% across the three items (see the fourth column in Table 3.3). This means that an important part of the variability of the situation effects is due to the person-situation interaction effects. To put it differently, the effect of the situation not only depends on the situation happening but also on the trait level of the positive emotions of the individuals.

Table 3.2: Unstandardized Estimates of the Key Parameters of the Models using the Situation Variable *Alone*

Parameter	$\mathscr{M}_2$ Est. [95% C.I.]	$\mathscr{M}_{2b}$ Est. [95% C.I.]
	Between-level	
Eg-Al Interaction Effect β_{111}	-0.08 [-0.12, -0.04]	-0.04 [-0.10, -0.01]
En-Al Interaction Effect β_{121}	-0.10 [-0.13, -0.07]	-0.08 [-0.12, -0.04]
Ch-Al Interaction Effect β_{131}	-0.09, [-0.12, -0.06]	-0.07 [-0.11, -0.03]
Opt-Eg-Al Interaction Effect $\beta_{OPT,1}$	—	-0.16 [-0.31, -0.01]
Opt-En-Al Interaction Effect $\beta_{OPT,2}$	—	-0.13 [-0.27, 0.00]
Opt-Ch-Al Interaction Effect $\beta_{OPT,3}$	—	-0.10 [-0.23, 0.03]
AR Effect Mean $E(\varphi)$	0.33 [0.31, 0.35]	0.33 [0.31, 0.35]
Eg-Al Effect Residual Variance $Var(\omega_{EG,11})$	18.63 [14.59, 23.11]	18.30 [14.25, 23.08]
Et-Al Effect Residual Variance $Var(\omega_{EN,21})$	13.74 [10.62, 17.53]	13.58 [10.31, 17.36]
Ch-Al Effect Residual Variance $Var(\omega_{CH,31})$	10.42 [7.59, 13.74]	10.44 [7.43, 13.87]
AR Effect Variance $Var(\varphi)$	0.032 [0.027, 0.039]	0.033 [0.027, 0.039]
	Model Fit Information	
Number of Free Parameters	34	43
DIC	1932395.64	1934237.58
pD	150001.81	149947.33

Note. $\mathscr{M}_2$: Model 2, $\mathscr{M}_{2b}$: Model 2b, Est.: Unstandardized estimate, C.I.: Credibility interval, Eg: Energetic, En: Enthusiastic, Ch: Cheerful, Al: Alone, DIC: Deviance information criterion, pD: Estimated number of parameters.

Table 3.3: Variance Coefficients Across Fixed Situations

Item	Consistency of Traits	Specificity of Traits	Person-Situation Interaction Coefficient	Unique Situation Effect
Energetic	0.80	0.20	0.19	0.81
Enthusiastic	0.70	0.30	0.13	0.87
Cheerful	0.66	0.34	0.16	0.84

Furthermore, in the ME-TSO model, one can also estimate the variance coefficients within fixed situations[5]. Note that these variance coefficients are estimated for each item and they also vary across individuals. Therefore, in Table 3.4, we present the average and the standard deviation of the estimated variance coefficients for each item and each fixed situation. In relation to the reliability of the items, the item *enthusiastic* was on average the most reliable item on both fixed situations ($M = 0.83, SD = 0.01$). Moreover, the items *energetic* and *cheerful* seem to be slightly less reliable in the situations where something negative happened. In general, when considering the consistency and the occasion-specificity, the three items seem to be on average as trait-like as they are state-like in both fixed situations because the average consistencies and occasion-specificities tend to be practically equal. However, the items *energetic* and *cheerful* seem to be more trait-like in the situations where nothing negative happened. For example, the difference between the mean consistency and mean occasion-specificity of *energetic* when nothing negative happened is 0.06, while when something negative happened it is 0.01. Lastly, the consistency is divided into the predictability by trait and the unpredictability by trait. On the one hand, the mean predictability by trait of the items *enthusiastic* and *cheerful* were very similar in both fixed situations. In contrast, the mean predictability by trait of *energetic* was lower in the situations when something negative happened (0.28) in comparison with the situations when nothing negative happened (0.33). This means that the trait of *energetic* in the first measurement occasion when nothing negative happened has a larger influence on future situations than the trait of *energetic* in the first measurement occasion when something negative happened. On the other hand, the average unpredictability by trait of all items across fixed situations was between

[5] In order to compute these variance coefficients, we extracted the estimates of the autoregresssive effects per person by using the FSCORES command in Mplus.

5% and 6%. This coefficient also showed the most variability across persons (SD between 0.05 and 0.06). This means that for some individuals the amount of total variance in their daily positive emotions that is due to the autoregressive process or carry-over effects can be as large as 15%.

Table 3.4: Variance Coefficients Within Fixed Situations

Variance Coefficient	Items Energetic M (SD)	 Enthusiastic M (SD)	 Cheerful M (SD)
		Nothing Negative Happened	
Reliability	0.70 (0.02)	0.83 (0.01)	0.79 (0.01)
Consistency	0.38 (0.03)	0.42 (0.04)	0.43 (0.03)
Occasion-Specificity	0.32 (0.02)	0.41 (0.03)	0.36 (0.02)
Predictability by Trait	0.33 (0.02)	0.36 (0.02)	0.37 (0.02)
Unpredictability by Trait	0.05 (0.05)	0.06 (0.06)	0.06 (0.05)
		Something Negative Happened	
Reliability	0.67 (0.02)	0.83 (0.01)	0.77 (0.01)
Consistency	0.33 (0.03)	0.4 (0.04)	0.4 (0.03)
Occasion-Specificity	0.34 (0.02)	0.43 (0.03)	0.37 (0.02)
Predictability by Trait	0.28 (0.01)	0.34 (0.02)	0.34 (0.02)
Unpredictability by Trait	0.05 (0.05)	0.06 (0.06)	0.06 (0.05)

Note. M: Mean, SD: Standard deviation.

3.5 Discussion

The ME-TSO model presented in this study aims to be an additional tool to model psychological dynamics. The model integrates the multilevel TSO (Castro-Alvarez et al., 2022d), the LST-RF approach (Geiser et al., 2015b), and the DSEM framework (Asparouhov et al., 2018). In general, the ME-TSO model allows studying the carry-over effects of psychological constructs, the person-situation interaction effects, and the psychometric properties of the items per individual across fixed situations. Moreover, as it is implemented within the DSEM framework, it can be extended by allowing other parameters (i.e., factor loadings and residual variances) to vary randomly across persons or by including additional within- or between-covariates.

We illustrated how to use the model by means of the empirical example. The results showed that (a) the items of positive affect activation are relatively as state-like as they are trait-like, (b) there are carry-over effects present in the states of positive affect activation, which vary across individuals, (c) the effects of situations when something negative happened better explained the variability of the dynamics of positive affect activation than the effects of the situations when the participants were not alone, and (d) the situations when something negative happened seemed to interact with the trait level of the positive emotions of the individuals. Note that we have previously analyzed these data with the multilevel TSO model (Castro-Alvarez et al., 2022d), where the parameters are fixed for the sample and thus heterogeneity between persons could not be taken into account. The analyses in this study, however, show that variability across persons is non-negligible. Moreover, Geiser et al. (2015b) also studied the person-situation interaction effect between the situation *not being alone* and the emotions *happy*, *energetic*, and *cheerful* with different data. Their results are comparable to our results of $\mathscr{M}_2$, for example, in both studies the interaction effect between trait *energetic* and the situation was -0.08. However, while in Geiser et al. (2015b) this effect was not significant, in our example it seems to be statistically different from 0. This could be partially explained by the fact that our sample size was much larger than the one used by Geiser et al. (2015b). Nevertheless, the size of the effect is what really matters, and an interaction effect of -0.08 does not seem practically significant.

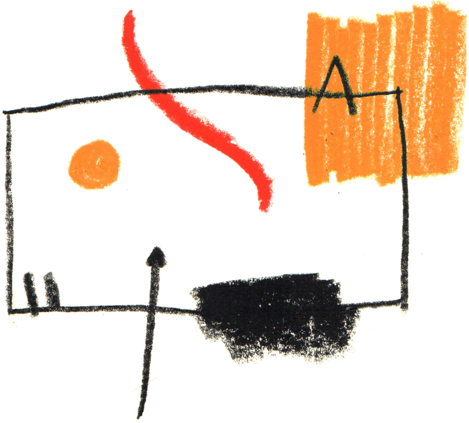
The variance coefficients per individual are key results of the proposed model. These variance coefficients allow studying the psychometric properties of the scales used in intensive longitudinal data by estimating the reliability of each item per individual. The reliabilities per person can be useful to evaluate the factor structure of each person, as suggested by Fuller-Tyszkiewicz et al. (2017). If the reliability of a person is too low, it might be an indication that the assumed factor structure does not fit this person. Hence, a different factor structure might be preferred for these cases. Additionally, the *consistency* and the *occasion-specificity* of the ME-TSO model also allow studying to what extent the variance of an item is due to stable or variable sources of variability per person. Thus, in our empirical example, we could determine whether a positive emotion was more trait-like or state-like for each individual. Similar coefficients at the individual level have been proposed previously (e.g., Fuller-Tyszkiewicz et al., 2017; Hu et al., 2016; Schuurman & Hamaker, 2019). However, the added value of the coefficients proposed within the ME-TSO model is that they also take into account the autoregressive structure of the psychological dynamics. In particular, with the *unpredictability by trait*, researchers can study to what extent the carry-over effects explain the overall variability of an item per person.

As with any model, the ME-TSO model has some limitations. First of all, one of the assumptions is that the autoregressive process is stationary, which can be difficult to test given that this process is unobserved in the ME-TSO model. For example, in the empirical example, we tested whether the observed time series were trend stationary. Yet, the stationarity of the observed time series does not necessarily imply that the latent autoregressive process is also stationary or vice versa. Future research can study how to improve the stationarity tests for the ME-TSO model and similar dynamic factor models (Song & Ferrer, 2012). Secondly and related to the previous point, the ME-TSO model assumes that longitudinal measurement invariance holds for all the parameters. This means that the factor structure as well as the size of the autoregressive effect does not change over time. Yet, this might not be a realistic assumption. For example, it might be the case that persons transition between different measurement models across time (Vogelsmeier et al., 2019b) or that the time dependencies (autoregressive effects) change over time (Bringmann et al., 2017). Thirdly, the model also assumes that configural invariance of the within-level factor model holds. This means that the within-level factor structure is the same for all individuals. Even if the factor loadings and the residual variances are allowed to randomly vary across individuals, there might still be persons for whom the assumed factor structure is not adequate. This drawback can be overcome by, for example, allowing the random measurement variances to correlate, as suggested in the multilevel heterogeneous factor analysis model (Pan et al., 2020). Lastly, in our application of the model we used the default prior distributions available in Mplus. However, it has been shown that the default priors can lead to biased estimates under certain circumstances with latent growth models (Smid et al., 2020). Hence, in the meantime, we recommend practitioners to perform sensitivity analyses when using the ME-TSO model. Alternatively, simulation studies would be required to further investigate the impact of the priors on the estimation of the ME-TSO model.

To conclude, in the present chapter, we presented the ME-TSO model. With this model, researchers can (a) account for the measurement error and study the psychometric properties of the items used in intensive longitudinal data, (b) estimate person-situation interaction effects, and (c) analyze the psychological dynamics of the constructs of interest per individual. We illustrated how to interpret the model with empirical data and we provide the code to fit the model in the git repository https://github.com/secastroal/ME-TSO. With the ME-TSO model, we provided a flexible statistical tool which can be useful to answer some of the research questions that are studied in intensive longitudinal research. We hope that this approach contributes to a better understanding of psychological dynamics. Furthermore, we expect this approach to serve as inspiration for future research to keep developing the statistical methods used to analyze intensive longitudinal data.

Chapter 4

A Time-Varying Dynamic Partial Credit Model to Analyze Polytomous and Multivariate Time Series Data



This chapter has been submitted for publication as Castro-Alvarez, S., Bringmann, L. F., Meijer, R. R., & Tendeiro, J. N. (2022). *A Time-Varying Dynamic Partial Credit Model to Analyze Polytomous and Multivariate Time Series Data*. https://doi.org/10.31234/osf.io/udnbt

Abstract

The accessibility to electronic devices and the novel statistical methodologies available have allowed researchers to better comprehend psychological processes at the individual level. However, there are still great challenges to overcome as, in many cases, collected data are more complex than the available models are able to handle. For example, most methods assume that the variables in the time series are measured on an interval scale, which is not the case when Likert-scale items were used. Ignoring the scale of the variables can be problematic and bias the results. Additionally, most methods also assume that the time series are stationary, which is rarely the case. To tackle these disadvantages, we propose a model that combines the partial credit model (PCM) of the item response theory framework and the time-varying autoregressive model (TV-AR), which is a popular model used to study psychological dynamics. The proposed model is referred to as the time-varying dynamic partial credit model (TV-DPCM), which allows to appropriately analyze multivariate polytomous data and nonstationary time series. We test the performance and accuracy of the TV-DPCM in a simulation study. Lastly, by means of an example, we show how to fit the model to empirical data and interpret the results.

Keywords: *item response theory, time series, psychological dynamics, non-linear trends, splines*

Intensive longitudinal methods such as experience sampling or ecological momentary assessment have allowed researchers to study and unravel the psychological dynamics of individuals (Hamaker et al., 2015; Hamaker & Wichers, 2017). These methods consist of assessing individuals repeatedly during short periods of time. In particular, popular intensive longitudinal designs require participants to fill in short questionnaires of about 10 times a day for 5 to 7 days (Vachon et al., 2019). As a result, psychological time series commonly have between 50 to 100 time points. However, analyzing this kind of data has proven to be a challenging task.

Intensive longitudinal data are complex data with strong dependencies between the measurements due to their closeness in time. Because of this, researchers have applied extensions of the autoregressive model to analyze this kind of data (e.g., Asparouhov et al., 2018; Chatfield, 2003; Hamilton, 1994; Kuppens et al., 2010; Shumway & Stoffer, 2017; Song & Zhang, 2014; Walls & Schafer, 2006). The simplest autoregressive model used to analyze intensive longitudinal data is the autoregressive model of order 1 (AR(1); Chatfield, 2003; Hamilton, 1994), which regresses the dependent variable on a lagged version of itself to represent the relation between two consecutive observations of the dependent variable. This model has been extended, for example, to multilevel and multivariate settings (Bringmann et al., 2013), to account for measurement error (Schuurman & Hamaker, 2019; Schuurman et al., 2015; Song & Zhang, 2014), and to model unequally spaced measurements (i.e., continuous-time modeling, Crayen et al., 2017; Voelkle & Oud, 2013; Voelkle et al., 2012). Furthermore, a comprehensive framework to analyze intensive longitudinal data, known as dynamic structural equation modeling, was recently proposed by Asparouhov et al. (2018).

However, one of the shortcomings of these current methods is that most of these approaches require the data to be continuous, which is not always the case. In particular, to study psychological dynamics, researchers tend to either use visual analogue scales or Likert scales (Vachon et al., 2019). While the former are continuous variables, which are suitable for the mentioned methods, the latter, strictly speaking, are ordinal categorical variables. This is a limitation, especially if there are not many response categories and if the distributions of the item responses are heavily skewed (Vogelsmeier et al., 2020). Furthermore, despite some few exceptions, most of the available statistical methods used to analyze intensive longitudinal data do not account for measurement error, which is likely to be present when measuring psychological constructs (Schuurman et al., 2015). Furthermore, in many intensive longitudinal studies, multiple items are used to measure a unique construct such as positive or negative affect (e.g., Hamaker et al., 2018; van der Krieke et al., 2016) and composite scores are computed before fitting the model. However, ignoring the nature of the variables and the factor structure of the data might lead to biased estimates (Dolan, 1994; McNeish & Wolf, 2020). Hence, measurement models for categorical intensive longitudinal data are needed.

A useful statistical theory that can help to overcome these drawbacks is the item response theory framework (IRT; Embretson & Reise, 2013). In general, IRT models are latent variable measurement models that relate the categorical responses of a set of items to one or multiple latent continuous variables that represent unobservable psychological traits or ability levels (Hambleton & Swaminathan, 1985; Rijn et al., 2010) such as positive affect. Well-known IRT models are, for example, the Rasch model (von Davier, 2016) and the 2-parameter logistic model (van der Linden, 2016) for dichotomous responses, and the partial credit model (Masters, 2016) and the graded response model (Samejima, 1997) for ordered categorical responses. Additionally, IRT as a psychometric theory also allows taking an in-depth look at the quality of the psychological tests and measures. Within IRT, the standard error of measurement differs across scores depending on the characteristics of the items and the latent ability level of the subject (Embretson & Reise, 2013), and measurerement precision can be determined conditional on the latent construct. This means that the quality of the measurements might vary across individuals, given their level on the latent construct.

Although IRT models have been largely developed within educational cross-sectional settings, dynamic IRT models for intensive longitudinal data have also been proposed in recent years (e.g., Hecht et al., 2019; Kropko, 2013; Rijn et al., 2010; Wang et al., 2013). On the one hand, Rijn et al. (2010) proposed a Rasch model and a partial credit model for intensive longitudinal data within the state space modeling framework, which is estimated by means of a Kalman Filter. On the other hand, the approaches by Kropko (2013, item response theory models for time series), Wang et al. (2013, dynamic Rasch model for educational data), and Hecht et al. (2019, continuous time Rasch model) are implemented within the Bayesian framework. The models proposed by Rijn et al. (2010) and Kropko (2013) are of special interest for us as they were developed to analyze psychological time series of one individual. However, these approaches are still limited as they (a) are not suitable for non-stationary time series, (b) have not been systematically tested in simulation studies, (c) lack user-friendly tutorials to be used by practitioners, and (d) do not use the core features of IRT modeling (e.g., item characteristic curves and item information functions) that allow assessing the quality of the scales.

In this chapter, we propose the time-varying dynamic partial credit model (TV-DPCM), which is an IRT model suitable to analyze multivariate time series data of polytomous responses. With this new method, we aim to offer a flexible tool that allows modeling non-linear trends and studying the psychometric properties of the scales used in intensive longitudinal data studies. Also, to facilitate its use by practitioners, we share all the code needed to fit the model in the following git repository: https://github.com/secastroal/DIRT. In particular, the TV-DPCM is useful to analyze intensive longitudinal data of one individual, when a set of Likert scale items that measure the same construct is repeatedly used to measure one subject. The TV-DPCM extends the partial credit model (PCM; Masters, 2016) by assuming that the latent variable follows a time-varying autoregressive model (TV-AR; Bringmann et al., 2017).

This chapter is organized in as follows. Firstly, we introduce the TV-DPCM in detail. This section also covers a brief introduction of the generalized additive model framework. Secondly, we conducted a “proof of concept” simulation to test the performance of the model under diverse conditions, while varying, for example, the number of time points and the size of the true autoregressive effect. Thirdly, we present an empirical application of the model to experience sampling data of self-esteem, which aims to exemplify how to use and interpret the results obtained by means of fitting the TV-DPCM. Lastly, we discuss our findings and how the TV-DPCM can contribute to a better understanding of measurement in intensive longitudinal research. Moreover, we provide some ideas for future methodological research for intensive longitudinal data based on IRT.

4.1 The Time-Varying Dynamic Partial Credit Model

As mentioned before, the TV-DPCM integrates the partial credit model (PCM; Masters, 2016) and the time-varying autoregressive model (TV-AR; Bringmann et al., 2017). Briefly, the PCM is an IRT model for polytomous data, which can be seen as an extension of the Rasch model (Embretson & Reise, 2013; Masters, 2016; Ostini & Nering, 2006). This means that the PCM holds most of the assumptions and properties of the Rasch model such as the assumption of unidimensionality, local independence, and the separability of the person and the item parameters. On the other hand, the TV-AR integrates the standard autoregressive model and the generalized additive model framework (Bringmann et al., 2017; Wood, 2017) to handle non-stationary time series. This means that both the intercept and the autoregressive effect are allowed to smoothly vary over time. By combining these two approaches, we get the TV-DPCM, in which the measurement model is given by the PCM and the dynamic latent process is described by a TV-AR model.

4.1.1 The Basis: The Partial Credit Model

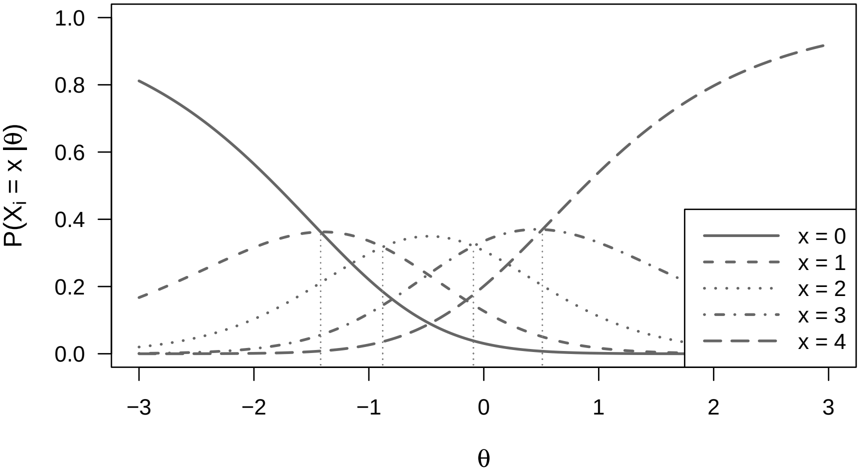
To start, we first introduce the PCM (Masters, 2016), which is an IRT model for polytomous items. The motivation to develop this model was to allow analyzing test items that required multiple sequential steps to find the correct answer, where partial credit is given for completing each of the steps (Embretson & Reise, 2013). Evidently, this model was proposed within an educational assessment context. However, it is also appropriate and it has been widely used to analyze items with ordered response options as found in attitudes and personality tests (Embretson & Reise, 2013).

The PCM is commonly described as a "divide-by-total" (Thissen & Steinberg, 1986) or "direct" (Embretson & Reise, 2013) model because the probability to endorse a certain response option is directly defined as the ratio of the probability of that response option to the sum of the probabilities of all possible response options. Consider that we have a test with I Likert-scale items that is used to measure, for example, positive affect. The items are scored from 0 to m_i, with $i = 1, \dots, I$; which means that item i has $K_i = m_i + 1$ response categories (items might differ in the number of response options). Then, the probability to select response option x of the i-th item given the latent trait of the j-th person, θ_j, can be written as:

$$P(X_i = x|\theta_j) = \frac{exp\left[\sum\limits_{k=0}^{x}(\theta_j - \delta_{ik})\right]}{\sum\limits_{v=0}^{m_i} exp\left[\sum\limits_{k=0}^{v}(\theta_j - \delta_{ik})\right]}, \tag{4.1}$$

where δ_{ik} is the step parameter, also known as threshold parameter, of the k-th category of the i-th item. These threshold parameters δ_{ik} represent the level on the latent continuum at which the probabilities of selecting the response options k and $k-1$ are equal. An example of an item with five response options is presented in Figure 4.1. This shows how the probability of endorsing each response option depends on the level of the latent ability of the subject. Therefore, persons with lower levels of the latent trait are more likely to select the response option 0 of this item (when θ is lower than -1.42). Notice that for notational convenience, when $x = 0$, the summation in the numerator is defined as 0 so that the exponential evaluates to 1. Thus, when there are only two response options (correct or incorrect), the PCM simplifies into the Rasch model.

Figure 4.1: Item Characteristic Curves of an Example Item given the PCM.



Note. Item characteristic curve of an item with five response options and threshold parameters -1.42, -0.88, -0.09, and 0.51. The location of the threshold parameters is shown with the vertical dotted gray lines, which also correspond with the intersection between the curves of adjacent response options.

Moreover, it is important to highlight some of the assumptions and properties of the PCM. First, regarding the assumptions, in a similar way as the most widespread IRT models, the PCM assumes that unidimensionality and local independence hold (Embretson & Reise, 2013). Unidimensionality means that the model assumes that all the items in the test measure a unique latent construct (e.g., positive affect or neuroticism). On the other hand, the assumption of local independence implies that the responses to any pair of items are independent after controlling for the latent variable. Secondly, the PCM also retains two important properties that are shared, in particular, with the Rasch model: The separability of the person and the item parameters and sufficient statistics. The former property means that each type of parameters can be conditioned out from the estimation of the other. The latter property means that the raw scale sum scores are sufficient statistics for the person parameters, so all persons with the same sum score are assumed to display the same value on the latent trait under study.

4.1.2 A Straightforward Extension: Modeling a Dynamic Latent Process

Now, in the context of studying psychological time series, a straightforward extension of the PCM model is to add an autoregressive structure at the latent level. This has been suggested by Rijn et al. (2010) within the state-space modeling framework and by Kropko (2013) within the Bayesian framework. However, to the best of our knowledge, in none of these studies nor in any other studies, the models have been systematically tested in a simulation study. In this chapter, we further extend this model (see the following subsection) and assess its performance in a simulation study. Thus, the model changes as follows:

$$P(X_i = x|\theta_t) = \frac{exp\left[\sum\limits_{k=0}^{x}(\theta_t - \delta_{ik})\right]}{\sum\limits_{v=0}^{m_i} exp\left[\sum\limits_{k=0}^{v}(\theta_t - \delta_{ik})\right]}. \tag{4.2}$$

Notice that the latent variable θ now has a subscript t, which indicates time. In this case, when there are repeated measurements from one individual, the latent variable does not represent the latent trait of a person. Instead, it represents the latent state dispositions of the individual at each measurement occasion. In other words, the latent state disposition represents the attitude or emotion of the person in the situation where the measurement took place. Moreover, we assume that these latent state dispositions follow an autoregressive process of lag-order 1, which is:

$$\theta_t = \alpha + \varphi\theta_{t-1} + \varepsilon_t, \tag{4.3}$$

where α is the intercept of the process and φ is the autoregressive effect of lag-order 1 between consecutive measurement occasions. In particular, the lag-order indicates how many measurements in the past predict the current measurement. With a lag-order 1, only the immediately previous measurement is used to predict the current one. Moreover, the autoregressive effect represents the dependency between consecutive states. This effect is also known as the “innertia” parameter (Kuppens et al., 2010) because the larger this parameter is, the longer it takes the system to return to its equilibrium (i.e, its mean). Lastly, ε_t is the random innovation at time t. The innovations are the part of the current latent state that cannot be explained by the model. Yet, they still influence and are passed along to future states (Schuurman et al., 2015). The innovations are assumed to be normally distributed with mean 0 and variance Ψ.

By extending the PCM in this way, additional assumptions are made about the latent process. Firstly, this extension proposes a discrete-time model for the latent process. This means that the repeated measurements are assumed to be observed in equally spaced time intervals. If this condition is not satisfied, the autoregressive effect might be overestimated and lead to the wrong conclusions (de Haan-Rietdijk et al., 2017b). Secondly, the latent process is assumed to be stationary, which means that its means and its variances-covariances do not change over time (Chatfield, 2003). A necessary but not sufficient condition for stationarity in the autoregressive process in Equation 4.3 is that $|\varphi| < 1$. Lastly, it is assumed that item parameters δ_{ik} are also time invariant. In other words, it is assumed that longitudinal measurement invariance (Meredith, 1993; Meredith & Teresi, 2006) holds.

4.1.3 Dealing with Change: The TV-DPCM

However, assuming stationarity might not be realistic in clinical practice. For example, consider a person that is under psychological treatment and fills in a daily diary questionnaire with Likert-scale items during the whole intervention. If the purpose is to monitor relevant psychological constructs for the intervention such as positive or negative affect, and if the intervention is effective, then we would expect to observe durable changes on the person’s behavior and feelings (e.g., reduction of symptoms or increase in well-being). To allow for such change, we further extended the PCM to allow the latent dynamic process to be non-stationary. We called this extension the TV-DPCM, which aims to model the non-linear change of the latent variable while accounting for the measurement error of the psychological construct.

As with the previous extension, the TV-DPCM is described in two equations: The measurement equation and the structural equation. The measurement equation is the same as Equation 4.2. This equation models the relation between the observed responses and the latent construct based on the PCM. Then, the structural equation, which describes the latent dynamic process, is an extension of Equation 4.3 based on the TV-AR model (Bringmann et al., 2017). In this case, the intercept α is allowed to vary over time[1]. To put it differently, with a time-varying intercept, the TV-DPCM is able to model latent processes that are trend-stationary (i.e., the time series is stationary after detrending). Now, the structural equation is defined like this:

$$\theta_t = \alpha_t + \varphi\theta_{t-1} + \varepsilon_t, \tag{4.4}$$

where α has a subscript t, which indicates that the intercept changes over time. This change is assumed to be described by a smooth function (see the following section).

Moreover, based on the time-varying intercept and the autoregressive effect, it is possible to derive the model-implied mean and variance of the dynamic process in Equation 4.4 (Bringmann et al., 2017; Chatfield, 2003; Giraitis et al., 2014). Firstly, in a TV-AR, the intercept does not have a clear interpretation and what describes the trend of the time series is, in fact, the mean of the dynamic process. Because the intercept varies over time, the mean of the dynamic process also varies over time. Therefore, the mean of the dynamic process at time t can be defined as (see Bringmann et al., 2017):

$$\mu_t \approx \frac{\alpha_t}{1-\varphi}. \tag{4.5}$$

Notice that the approximation in Equation 4.5 applies as long as the change of the intercept is constrained to be gradual[2] (i.e., smooth). The time-varying mean is also known as the attractor (Giraitis et al., 2014). Furthermore, we can also derive the variance of the dynamic process. Because the autoregressive effect is time-invariant,

[1] Ideally, both the intercept and the autoregressive effect should be allowed to vary over time, as proposed in the TV-AR model (Bringmann et al., 2017). By doing this, the model can handle different types of non-stationarity, where the means, the variances, and the autocorrelations change. However, we did not succeed on writing a working TV-DPCM model in Stan that also allowed the autoregressive effect to vary over time. Because of this, we settled with the simpler version in which only the intercept is allowed to vary over time.

[2] The change of the intercept is required to be gradual because an assumption used to derive Equation 4.5 is that μ_t must be approximately equal to μ_{t-1}. This is also why, in Equation 4.5, the approximation sign is used instead of the equal sign.

then, the variance of the dynamic process is also assumed to be time-invariant, and it is shown to be as follows:

$$\sigma^2 \equiv \frac{\Psi}{1-\varphi^2}. \tag{4.6}$$

To summarize, herewith, we propose the TV-DPCM, which is a measurement model useful to analyze psychological time series of one individual. The model keeps most of the assumptions of the PCM and the TV-AR such as (a) unidimensionality, (b) local independence, (c) trend-stationarity of the latent process, and (d) equally spaced observations over time. In contrast, the separability of item and person parameters and the sufficient statistics property, which are properties of the PCM, do not hold for the TV-DPCM.

4.1.4 Estimation: Generalized Additive Models and Bayesian Inference

As with other dynamic IRT models (Hecht et al., 2019; Kropko, 2013; Wang et al., 2013), we implemented the TV-DPCM within the Bayesian framework. This allows estimating all the parameters simultaneously and prior information can be incorporated. Additionally, to estimate the time-varying intercept, we make use of the generalized additive model (Wood, 2017). In what follows, we first do a brief introduction of the generalized additive model (GAM) framework and then we mention the suggested priors required to estimate the model.

Generalized additive models are flexible semiparametric models that define the relation between the dependent variable and the covariates based on "smooth functions" (Wood, 2017). They are specially useful to model nonlinear relationships while keeping a reasonable predictive power. A general representation of a GAM model, given one dependent variable and one covariate is:

$$y_i = f(x_i) + \varepsilon_i, \tag{4.7}$$

where y_i is the dependent variable, x_i is a covariate, $f()$ is a smooth function, and ε_i is the independent and normally distributed random error.

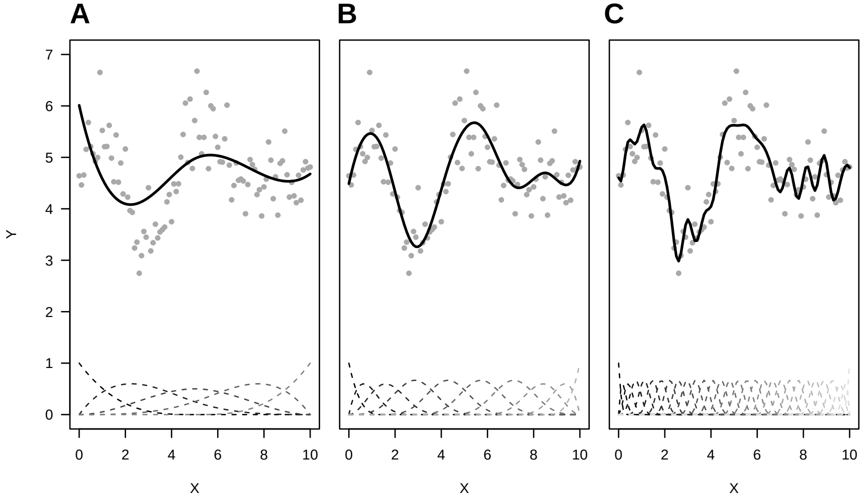
The smooth function is usually the weighted sum of some predefined "basis functions" and is represented as a linear model, as follows:

$$f(x) = \sum_{j=1}^{s} \beta_j b_j(x), \tag{4.8}$$

where $b_j()$, with $j = 1, \ldots, s$, is the j-th basis function, and β_j is the unknown weight for each function. Given this, in the TV-DPCM, the time-varying intercept α_t is modeled as a smooth function of time:

$$\alpha_t = f(t) = \sum_{j=1}^{s} \beta_j b_j(t). \tag{4.9}$$

Figure 4.2: Predicted B-Splines with Different Number of Basis Functions.



Note. The gray dots represent the observed data and the black line represents the predicted non-linear function. At the bottom of each plot, the basis B-splines functions are represented with dashed lines. The number of basis functions are 5 (A), 10 (B), and 30 (C).

However, when using the GAM, one must decide on the type of smoother that is going to be used and how smooth the resulting fit has to be. In our implementation, we opted to use cubic B-splines (Kharratzadeh, 2017; Wood, 2017)[3]. Without going into too much detail, the basis splines or B-splines are a popular smoother in the GAM literature for univariate analysis. B-splines have a polynomial degree p (order of the B-splines is $k = p+1$) and a set of q knots that are typically defined based on the percentiles of the predictor variable. Then, these knots are used to define $q+p-1$ basis functions for the B-splines. Each basis function consist of k pieces of polynomials with degree $k-1$ (except for the ones close to the borders). These pieces of polynomials of each basis function are joined continuously at $k-1$ interior knots and are differentiable $k-2$ times. For the remaining range of the covariate, the basis functions are 0. Most commonly, B-splines of order 4 (i.e., degree 3), which are cubic B-splines, are used. To illustrate this, we simulated data based on cubic B-splines with 10 basis functions as shown in the middle panel of Figure 4.2. The 10 basis functions are depicted at the bottom of the graph. When these functions are weighted by the β_j coefficients and summed together, they result in the nonlinear trend (solid black line) that describes the data.

Figure 4.2 also shows what can happen when too little or too many basis functions are used. Panel A presents the results from a cubic B-splines with 5 basis functions and panel C presents the results from a cubic B-splines with 30 basis functions. While using too little basis functions can result in underfitting, using too many can result in overfitting the data. Because of this, when using GAM, researchers usually use a larger number of basis functions than they would think are needed but impose a penalization on the selected smoother (Bringmann et al., 2017; Wood, 2017). For our implementation, to penalize the cubic B-splines, we used a random-walk prior for the β_j coefficients (Kharratzadeh, 2017). This means:

$$\beta_1 \sim \mathscr{N}(0,1), \qquad \beta_j \sim \mathscr{N}(\beta_{j-1},\tau), \qquad \tau \sim \mathscr{N}(0,1). \tag{4.10}$$

The reasoning behind this prior is that the closer the β_j coefficients are to each other, the smoother the spline function is.

Lastly, to estimate the TV-DPCM within the Bayesian framework, we used relatively informative prior distributions for the different parameters. The following priors were used in both the simulation study and the empirical application. Starting with the

[3] We also wrote an alternative version of the model in JAGS (Depaoli et al., 2016), which can use other kind of smoothers such as thin plate or penalized P-splines based on the *mgcv* (Wood, 2017) package.

threshold parameters δ_{ik}, we used, as it is common in the IRT literature, a standard normal prior (Fox, 2010). For the random innovations ε_t, we first sampled a starting value from the standard normal, which was later scaled in the computation of θ_t given Equation 4.4. Then, the prior for the scaling factor of the innovations (i.e., the standard deviation σ) was a normal distribution with mean 1 and standard deviation 1, which was truncated to be positive. Finally, for the autoregressive effect φ, we used an uniform distribution between -1 and 1 as prior.

4.2 Simulation Study

In this section, we present the design and results of the simulation study that we conducted with the TV-DPCM. The purpose of this simulation was to assess the performance, in terms of convergence and recovery of the population parameters, of the TV-DPCM under common settings seen in the literature.

4.2.1 Data Simulation and Design

Data were simulated based on the TV-DPCM model assuming that the time varying intercepts α_t followed a sinusoidal trend. An example of the simulated latent dynamic process and its trend is presented in Figure 4.3. The same trend was used for all the conditions but it was adjusted to the length of the time series. Moreover, we also kept the variance of the innovations fixed (at 1) and the number of response categories per item (5) equal across all conditions. Regarding the threshold parameters, these were randomly generated in such a way that they were ordered within an item. For example, the threshold parameters for an item with 5 response options in the simulation could be: -1.42, -0.88, -0.09, and 0.51 (recall Figure 4.1 and that the threshold parameters represent where in the latent continuum the item characteristic curves of adjacent response options intersect). Lastly, the latent state disposition of the first measurement occasion was randomly generated for each replication and each condition from a normal distribution with mean $\alpha_1/(1-\varphi)$ and standard deviation $\sqrt{\Psi/(1-\varphi^2)}$. For details on the generation of these parameters, see the code shared on the GitHub repository of this chapter.

Next, for the simulation design, we manipulated four factors. Firstly, the number of time points was varied between 100, 200, 300, and 500. These time points were chosen based on previous simulations with $N = 1$ time series (Bringmann et al., 2017;

Schuurman et al., 2015). In fact, based on preliminary simulations with the TV-DPCM, we do not expect the model to perform well with under 200 time points. Secondly, the number of items was either 3 or 6 items. The reason for this is that scales used in ESM studies tend to be short in order to reduce participants' burden. Next, the size of the autoregressive effect was varied between 0, 0.25, and 0.5, which is similar to the values used in simulations with the VAR model with measurement error (Schuurman & Hamaker, 2019; Schuurman et al., 2015). Lastly, the proportion of missing observations was either 0% or 30%. To recreate the missing data patterns that are commonly seen in ESM data (Silvia et al., 2013), where participants either fill in the complete questionnaire or do not fill it in at all, we randomly sampled 30% of the time points and removed all the observations in those time points. Basically, the simulated missing data mechanism was missing completely at random with the constraint that the observation of the first time point was never removed. The conditions with missing data aimed to test the model under realistic circumstances, as the percentage of missing measurements usually ranges between 20% and 40% (Vachon et al., 2019). To summarize, the simulation had a $4 \times 2 \times 3 \times 2$ fully crossed design, in which we ran 200 replications per condition (i.e., a total of 9,600 analyses).

The models were estimated within a Bayesian framework through the Hamiltonian Monte Carlo algorithm as implemented in Stan (Carpenter et al., 2017). We ran three chains per analysis, each with 2,000 iterations, 500 of which were used for warm-up[4]. To run the analyses, we also adjusted other parameters of the Hamiltonian Monte Carlo algorithm such as the *delta* and the *maximum treedepth* (Stan Development Team, 2022). We increased parameter *delta* from 0.8 (default) to 0.99 and the *maximum treedepth* from 10 (default) to 15, as this was required to facilitate model convergence.

The simulation of the data, the estimation of the model, and the analysis of the results were performed in R (R Core Team, 2022) with the R packages: `rstan` (Stan Development Team, 2020) and `bayesplot` (Gabry & Mahr, 2021). Analyses were run on a high performance computing cluster with Intel Xeon E5 2680v3 CPU (2.5GHz). The maximum RAM usage for an analysis was approximately 500MB.

[4] We conducted preliminary simulations analyses with the model to ascertain that this number of total and warm-up iterations was enough to obtain reliable samples from the posterior distributions.

4.2.2 Output Variables

To assess the performance of the TV-DPCM, we focused on the convergence of the model and the quality of the estimates. In relation to model convergence, we relied on the convergence checks provided in Stan for the Hamiltonian Monte Carlo algorithm. According to these checks, an analysis diverged if the Gelman-Rubin statistic ($\hat{R}$; Gelman & Rubin, 1992) for any of the parameters was larger than 1.05, if there was any divergent transition after warm-up (Stan Development Team, 2022), or if any Bayesian Fraction of Missing Information (BFMI; Betancourt, 2017) was too low. Stan also provides other diagnostic checks about the efficiency of the algorithm that indicate if the maximum tree depth was exceeded or if the effective sample sizes (ESS) were too low. While the latter checks were tracked, no action was taken if, for example, the ESS of an analysis was too low, as these problems do not jeopardize the quality of the estimates and they are usually solved by increasing the number of iterations.

To assess the quality of the estimates, we looked at different accuracy statistics such as bias, absolute bias (abbias), relative bias (rbias), and root mean squared error (RMSE). Suppose that we focus on the set of parameters Θ (e.g., the thresholds, the latent states, or the autoregressive effect) and we run a simulation with M replications per condition. Given a condition c where there are N_c parameters Θ_n with $n = 1,\ldots,N_c$, and their estimates for the m-th replication are $\hat{\Theta}_{nm}$, with $m = 1,\ldots,M$, then, these accuracy statistics are defined as follows:

$$bias = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_c}\sum_{n=1}^{N_c}(\hat{\Theta}_{nm}-\Theta_n)\right], \tag{4.11}$$

$$abbias = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_c}\sum_{n=1}^{N_c}|\hat{\Theta}_{nm}-\Theta_n|\right], \tag{4.12}$$

$$rbias = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N_c}\sum_{n=1}^{N_c}\frac{\hat{\Theta}_{nm}-\Theta_n}{\Theta}\right], \tag{4.13}$$

$$RMSE = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{N_c}\sum_{n=1}^{N_c}(\hat{\Theta}_{nm}-\Theta_n)^2}. \tag{4.14}$$

For parameters such as the item thresholds, the latent state dispositions, and the attractor, we did not compute the relative bias because some of the true values of these parameters were 0 or very close to 0. As a result, the computed relative bias reached infinity or was extremely large, which made the measure unusable. Hence, for these parameters we computed the correlation between the true and the estimated parameters as well as the RMSE. In contrast, with parameters such as the autoregressive effect and the innovation variance, it was possible to compute the relative bias in most of the conditions. Additionally, we also inspected the coverage proportion of the credibility intervals as well as their average width for all the parameters.

4.2.3 Results

In total, 148 analyses of the 9,600 diverged. All the divergent analyses were due to the presence of divergent transitions after warm-up (as indicated by the convergence checks in Stan). Figure 4.4 presents the percentage of convergent replications per condition. This shows that most of the divergences occurred in the conditions with 100 time points and when the true autoregressive effect was the largest. These results indicate that, in general, at least 200 time points seem to be required to fit the TV-DPCM.

Next, to assess the quality of the estimates of the relevant parameters such as the thresholds, the latent state dispositions, and the autoregressive effect, we looked at the different accuracy statistics per each set of parameters. Hereby, we present in detail the results from the threshold parameters and the autoregressive effect. For the other parameters, we summarize the main findings and add supporting Figures in Appendix D. Figure 4.5 shows the average coverage proportions, the average width of the credibility intervals, the average correlation, and the average bias across conditions for the threshold parameters. The intervals around these averages indicate the interquartile range of the measure over the 200 replications per condition. Starting with the width of the credibility intervals across conditions, panel B of Figure 4.5 shows the credibility intervals shrank when there were more time points. This was expected, as usually with IRT models, the estimation of the item parameters improves when the number of subjects (time points in the TV-DPCM) increases and vice versa. Secondly, regarding the coverage proportion, the panel A shows that on average 80% of the credibility intervals included the true parameter. It seemed that the average coverage was slightly lower and spread more when there were more time points. This can be explained by the fact that for some analyses, there were large biases and all the threshold parameters were completely over- or underestimated. This in combination with narrow credibility intervals resulted in lower coverage rates. Thirdly,

Figure 4.3: True Latent Dynamic Process (in gray) and its Trend (in black) of Simulated Data

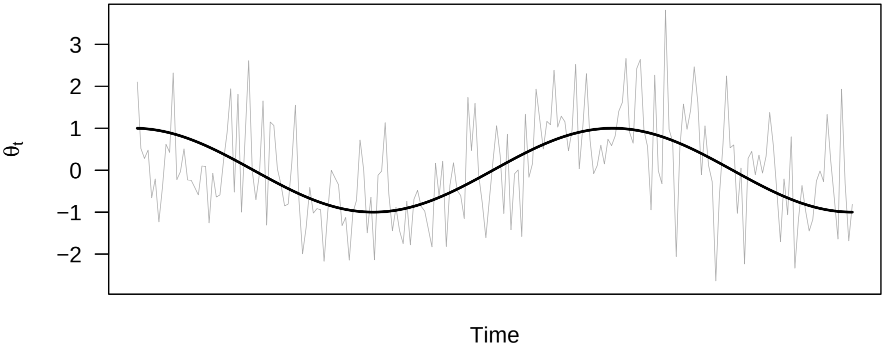
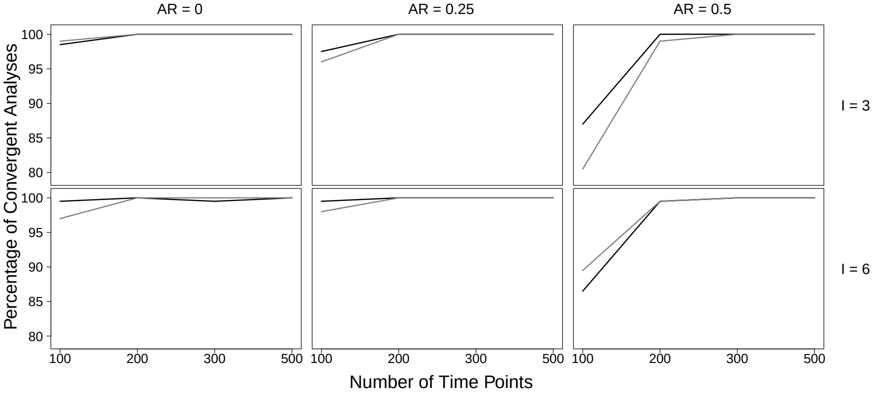


Figure 4.4: Percentage of Analyses that Converged per Condition.

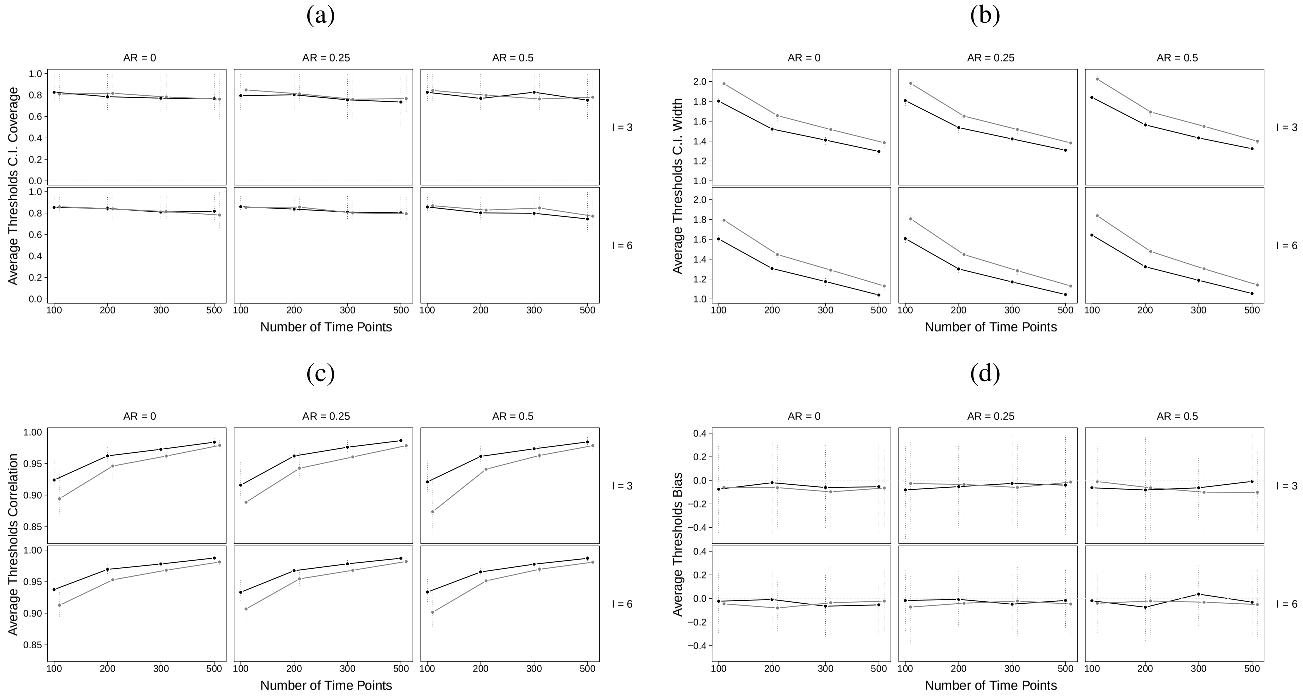


the correlation between the true and the estimated thresholds (panel C) was on average above 0.9 across all conditions and it approached 1 when the number of time points increased. The presence of missing data worsened the correlation between the true and the estimated thresholds in relation to the conditions without missing values. However, these differences became smaller as the number of time points increased. Lastly, the average bias of the threshold parameters was close to 0 across all conditions (panel D). Similar figures for the absolute bias and RMSE of the threshold parameters are included in Appendix D, which show that these measures became smaller as the number of time points increased. Overall, the accuracy of the estimates of the threshold parameters improves, as indicated with the average correlation and average width of the credibility intervals, when the number of time points is larger than 200.

Regarding the autoregressive effect, Figure 4.6 presents the average coverage proportion, credibility interval width, absolute bias, and relative bias of this parameter. On average, the coverage proportion of the autoregressive effect was close to 100% across all conditions (panel A). In contrast, the width of the credibility interval clearly depended on the number of time points and the percentage of missing values as shown in panel B. Furthermore, we present the absolute bias instead of the correlation, as computing the correlation was not adequate or informative. Panel C shows that the average absolute bias ranges between 0.15 and 0.05 across conditions, decreasing when the number of time points increased or when there were no missing values. Regarding the relative bias (panel D), the average is only presented for the conditions in which the true autoregressive effect was different from 0. In general, the relative bias of the autoregressive effect was on average 0 across all conditions. However, when there were 100 time points, 3 items, no missing values, and an autoregressive effect of 0.25, the TV-DPCM tended to overestimate the autoregressive effect about 20% above its true value.

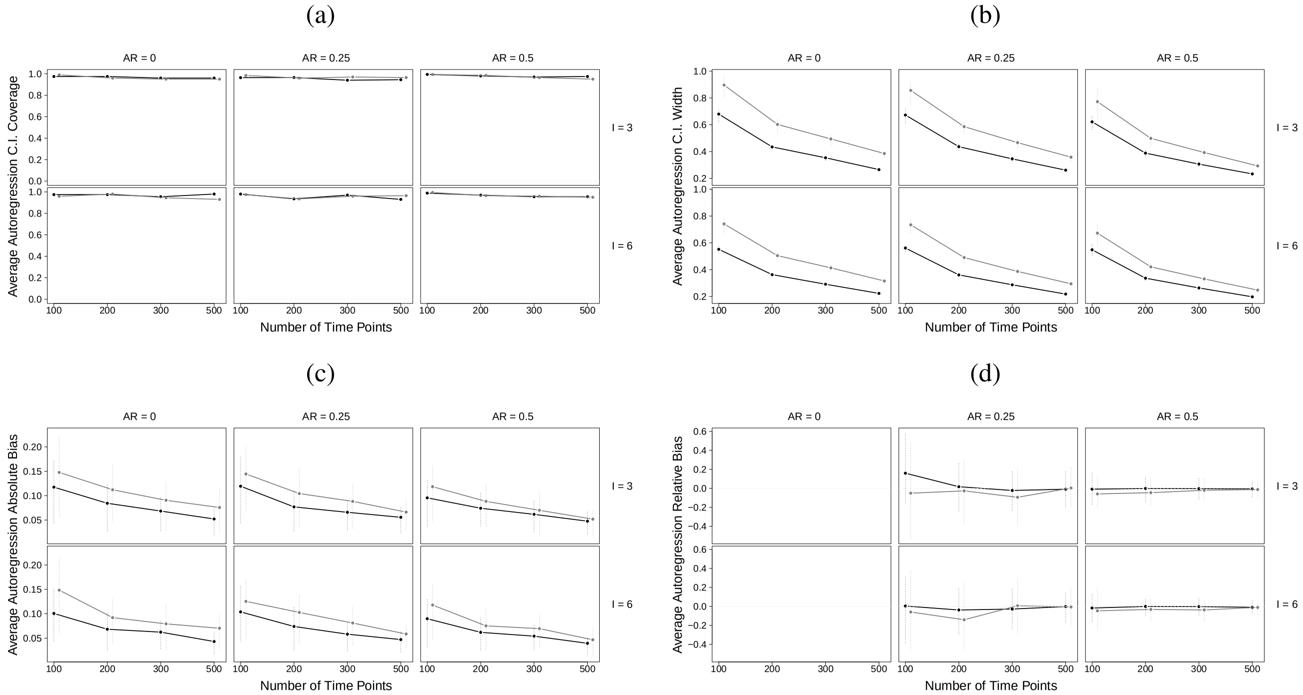
In relation to the other parameters of interest, such as the latent states, the attractor, the variance of the innovations, and the total variance of the dynamic process, we briefly summarize the findings about the recovery of these parameters in Table 4.1 (Figures for these parameters are included in Appendix D). In general, the average coverage rate per condition of these parameters was between 80% and 90%, and the coverage percentage did not seem to be influenced by the manipulated factors. Moreover, the recovery of the attractor showed similar results as the ones seen for the threshold parameters. Similarly, the results of the variance of the innovations and the

Figure 4.5: Parameter Recovery and Accuracy Statistics of the Threshold Parameters.



Note. The black lines represent the conditions where there were no missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average correlation between the true and the estimated thresholds, and (D) average bias per condition.

Figure 4.6: Parameter Recovery and Accuracy Statistics of the Autoregressive Effect.



Note. The black lines represent the conditions where there were no missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average absolute bias, and (D) average relative bias per condition.

variance of the dynamic process in relation to the width of the credibility intervals and the mean absolute bias were similar to the results observed for the autoregressive effect.

Table 4.1: Summary of the Recovery of the Other Parameters of the TV-DPCM

Parameter	**Results Summary**
Latent state dispositions	The width of the credibility intervals shrank when the number of items increased. Also, the average correlation increased when the number of items increased, when the size of the autoregressive effect was larger, and when there were no missing values.
Attractor	Similarly as with the threshold parameters, increasing the number of time points resulted in a slight decrease of the coverage rate, a shrinkage of the credibility intervals, and a raise of the average correlation.
Variance of the innovations and variance of the dynamic process	Just as with the autoregressive effect, the width of the credibility intervals and the average absolute bias decreased when the number of time points increased. However, these parameters tended to be underestimated as their estimates were between 20% (conditions with 100 time points) to 10% (conditions with 500 time points) lower than the true parameter according to the relative bias.

4.2.4 Summary

To conclude, this simulation study showed that the TV-DPCM performs well at recovering its parameters across most of the conditions. In general, the accuracy of the estimates of the TV-DPCM improves when the number of time points increases. The results suggest that at least 200 time points are required for the model to converge and to accurately estimate the parameters. However, given the width of the credibility intervals of some parameters, we tentatively suggest 300 time points as a minimum to estimate the TV-DPCM. Still, in some cases, the model might over- or underestimate some of the parameters of interest. While in such cases the estimates are biased and the coverage of the credibility intervals is poor, the overall pattern is still well recovered as indicated by the high correlations. Regarding the number of items, it seems the TV-DPCM can be estimated with as little as 3 items, with the caveat that the credibility intervals can be very wide, specially in combination with the presence of missing data.

4.3 Empirical example: Using the TV-DPCM to Analyze Self-Esteem

To exemplify how to use and interpret the results of the TV-DPCM, in this section, we analyzed mood data from one subject. These data, collected between August 2012 and April 2013, were retrieved from Kossakowski et al. (2017) and were previously analyzed by, among others, Wichers and Groot (2016). The data come from a 57 years old male (at the time of the study) that had been diagnosed with major depressive disorder. The participant completed up to 10 semi-random assessments per day for 239 days. During this period, the participant also followed a blind gradual reduction of his anti-depressant medication dosage. In what follows, the items of interest and the data collection procedure are described in detail. Then, the data preprocessing procedures are presented. Finally, the TV-DPCM is adjusted to the data in order to study the psychological dynamics of self-esteem and the performances of the items of the ESM questionnaire.

4.3.1 Data Collection and Procedure.

As mentioned before, the participant filled in an ESM questionnaire up to 10 times a day for 239 days. The questionnaire was programmed at random moments within 90-minute intervals that were set between 07:30 AM and 10:30 PM. After the beep signal, the participant had a 10-minute window to complete the questionnaire, which consisted of 50 momentary assessment items that measured different emotions (e.g., feeling enthusiastic or feeling lonely), self-esteem, and descriptions of the situation such as whether the participant was alone or doing something. Furthermore, additional items were used at certain beep signals to measure, for example, sleep quality and depressive symptoms. These items were filled up on a daily or weekly basis. All the momentary assessment items were measured on a 7-point Likert scale from “not feeling the state" to "feeling the state very much". The participant completed a total of 1473 assessments (i.e., on average 6.2 assessments per day). Moreover, the study was divided in 5 phases (Kossakowski et al., 2017): (1) A baseline period of four weeks, (2) a double-blind period without dosage reduction of two weeks, (3) a double blind period with gradual dosage reduction of eight weeks, (4) a post-assessment period of eight weeks, and (5) a follow-up period of twelve weeks.

For this empirical example, we actually fitted the TV-DPCM to different sets of items including some or all the phases. The sets of items were defined based on the results from a principal component analysis on the mood items (Wichers & Groot, 2016), which extracted three components: Positive affect, negative affect, and mental unrest. Furthermore, the set of items that measured self-esteem was also analyzed with the TV-DPCM. Here, we present the results from the best fitting model[5] to illustrate the TV-DPCM, which was when fitting the model to the items of self-esteem including phases 1 and 2 (286 complete beeps).

4.3.2 Data Pre-processing

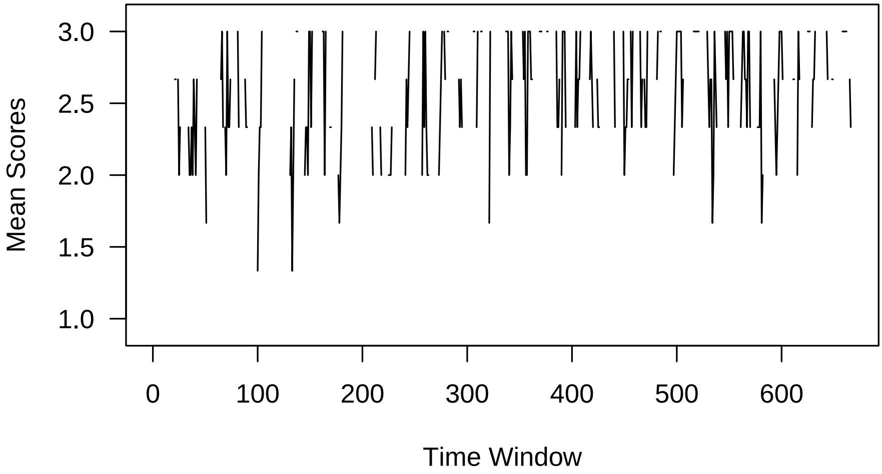
The items of self-esteem were *I like myself* (Self-like), *I am ashamed of myself* (Ashamed), and *I doubt myself* (Self-Doubtful)[6]. The items *Ashamed* and *Self-Doubtful* were reverse-coded to have high scores on the scale represent high levels of self-esteem. Also, given that not all the response categories were selected and that some were selected too few times, several response categories were collapsed. For the item *Self-like*, the response categories lower than 3 and the response categories larger than 5 were collapsed and recoded into response categories 1 and 3, respectively. Also, response category 4 was recoded as 2. For the items *Ashamed* and *Self-Doubtful*, response categories lower than 5 (after reversed coding) were collapsed into response category 1 and response categories 6 and 7 were recoded to 2

[5] When developing the TV-DPCM, we were also working simultaneously on developing posterior predictive model checking methods (PPMC) to assess the goodness of fit of the TV-DPCM. By using preliminary versions of these PPMC methods that we were developing, the analysis of the items of self-esteem including the phases 1 and 2 seemed to lead to the best results. Still, the results with the complete data set are reported in Appendix D.

[6] A fourth item of self-esteem was *I can handle anything*. However, this item was excluded because the scale did not seem to be measuring an unidimensional construct when this item was included according to the preliminary version of the PPMC methods.

and 3, respectively. Therefore, the responses were changed from a 7-point Likert scale to a 3-point Likert scale.

Figure 4.7: Observed Mean Scores of the Self-Esteem Items



Moreover, it is important to note that the TV-DPCM is a discrete time model. This means that the model requires that the time interval between consecutive observations is the same for the whole duration of the data collection. This was clearly not the case with the data at hand due to the random beeps, the missing data, and the overnight time between days. One way to address this issue within the Bayesian framework is to include missing values in order to make the time intervals between observations approximately the same (Asparouhov et al., 2018). This approach has been shown to be useful to deal with unequal time intervals and the results from these kind of analyses are comparable with results from continuous time models (de Haan-Rietdijk et al., 2017b). Given this, we also implemented this approach in the TV-DPCM analysis of the self-esteem items. For this, we divided the days in 90-minute time windows. As a result, there is a total of 16 time windows per day, 6 of which were always missing because they happened during the night. Observations within any of these time windows were considered as a representation of the state of self-esteem of the participant for that time point. When no observations were available, "missing values" were included in the date set. By doing this, we added 380 rows of missing values, for a total number of 666 time windows. The time series of the observed mean scores, after recoding and after including rows of missing values, is presented in Figure 4.7. The mean scores ranged between 1 and 3.

4.3.3 Fitting the TV-DPCM

To fit the TV-DPCM to the data, we used the same setup for the Hamiltonian Monte Carlo algorithm as we did in the simulation study. This means that we ran three chains in parallel, each with 2000 iterations, 500 of which were discarded as warm-up, and we kept the same values for the *adapt_delta* (0.99) and *max_treedepth* (15) parameters. To check convergence of the model, we examined the diagnostics provided in Stan for the HMC algorithm. According to these diagnostics, we found no evidence of divergence. Graphical diagnostics for some selected parameters are presented in Appendix D.

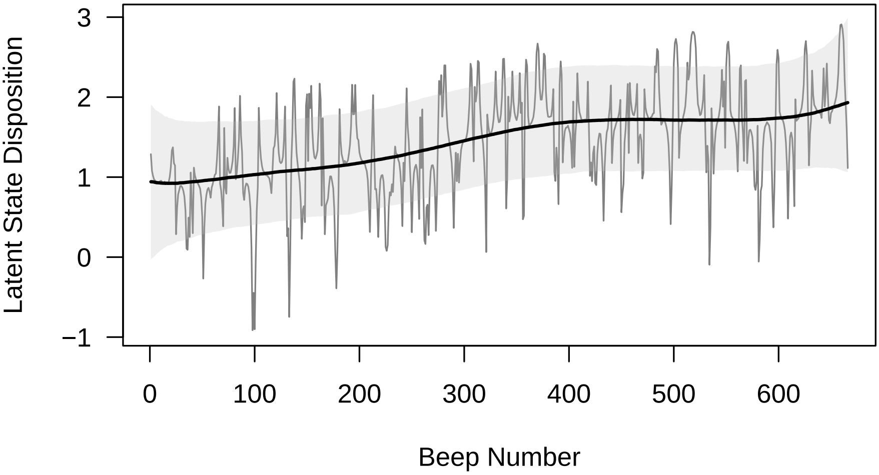
Table 4.2: Estimated Parameters of the TV-DPCM

	Median	SD	C.I.	ESS
$\hat{\delta}_{11}$	-2.55	0.55	$(-3.63, -1.46)$	4474
$\hat{\delta}_{12}$	1.46	0.44	$(0.6, 2.34)$	3110
$\hat{\delta}_{21}$	-1.44	0.63	$(-2.69, -0.21)$	4784
$\hat{\delta}_{22}$	-1.60	0.48	$(-2.55, -0.67)$	3429
$\hat{\delta}_{31}$	0.55	0.44	$(-0.32, 1.41)$	3041
$\hat{\delta}_{32}$	2.54	0.47	$(1.63, 3.48)$	3177
$\hat{\varphi}$	0.47	0.12	$(0.22, 0.69)$	758
$\hat{\Psi}$	1.90	0.56	$(1.03, 3.2)$	1212
$\hat{\sigma}^2$	2.48	0.64	$(1.47, 4.02)$	1401

Note. C.I. = 95% central credible interval.

Table 4.2 shows the estimated values (i.e., the median of the posterior distribution), the standard deviation of the posterior distribution, the 95% credibility interval, and the effective sample size of the threshold parameters, the autoregressive effect, the variance of the innovations, and the total variance of the dynamic process. Note that the threshold parameters are ordered within items 1 and 3 but not for item 2. This means that there is a “reversal” for item 2. Hence, the probability to select response category 2 is always lower than the probability to select either response category 1 or 3 across the latent continuum (see Figure 4.9). Next, the estimated autoregressive effect was 0.47, which implies that there is a medium-strong dependency between consecutive states of self-esteem. Thus, when the person experienced a high level of self-esteem at a certain time point, it was likely that they would keep experiencing high levels of self-esteem for the next measurement. Lastly, the variance of the innovations and the variance of the dynamic process were 1.9 and 2.48, respectively. The first one indicates the variability of the state of self-esteem that cannot be explained by the previous state of self-esteem. The latter represents the total variance of the states of self-esteem across the time series.

Figure 4.8: Estimated Latent Dynamic Process



Note. The estimated latent state dispositions for each beep (observed and missing) are represented with the gray line. The trend of the dynamic process or attractor is represented with the black line alongside with its 95% credibility interval band in light gray.

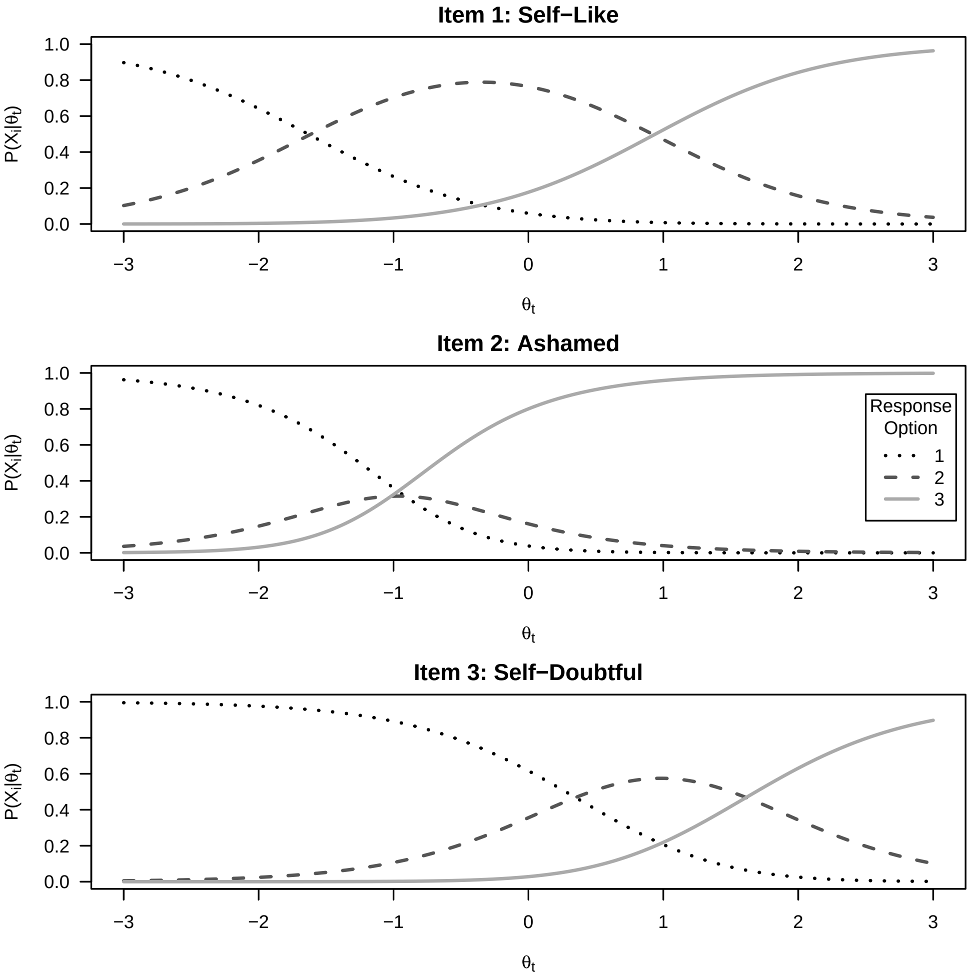
The estimates of the latent state dispositions and the time-varying attractor are presented in Figure 4.8. To facilitate the interpretation, these estimates were previously divided by the standard deviation of the dynamic process (i.e., $\hat{\sigma}$). By doing this, a latent state disposition of 1 means that the latent state of the individual at a certain time point is one standard deviation above the expected mean score on the test. Thus, Figure 4.8 shows that the latent state dispositions varied between -0.91 and 2.91.

The time varying mean or attractor with its credibility interval band shows a slight increasing trend over time[7]. This implies that, on average, at the beginning of the study the mean of the latent states of self-esteem was about one standard deviation above the expected mean score of the questionnaire. Moreover, the mean of the latent states of self-esteem increased in such a way that by the end of the second phase, the mean of the latent states was close to two standard deviations above the expected mean score of the questionnaire.

Importantly, one of the key features of IRT modeling is that IRT models allow studying the properties of the items and the test. In this context, IRT provides the item characteristic functions (ICFs), the item information functions (IIFs), and the test information function (TIF). For the TV-DPCM, we can compute and plot these functions because the model assumes that the item parameters do not change over time (longitudinal measurement invariance holds). Therefore, these functions are defined given the latent state disposition (θ_t) at a certain time point t, namely the states of self-esteem of the individual. Figures 4.9 to 4.11 present the ICFs, the IIFs, and the TIF for the three items of self-esteem. Just as before, to facilitate the interpretation, the estimated latent state dispositions and the estimated thresholds were divided by the standard deviation of the dynamic process $\hat{\sigma}$. Regarding the ICFs, for the items *Self-like* and *Self-Doubtful*, the curves for each response category are nicely ordered and each of the response options gets to have the highest response probability at some point in the latent continuum. On the other hand, for the item *Ashamed* there is a reversal (i.e., the threshold parameters are not ordered for this item). As a consequence, there is no point on the latent continuum where the response category 2 has the highest response probability. Additionally, when inspecting the IIFs, we can see at what levels of the latent continuum the items are more or less informative. Thus, the item *Ashamed* seems to be more useful at measuring lower states of self-esteem and the item *Self-Doubtful* seems to be more useful at measuring higher states of self-esteem. In contrast, the item *Self-Like* is less informative than the other two items. Nonetheless, it seems it is useful to distinguish between very high and very low states of self-esteem but it is not informative in the middle levels of self-esteem. Lastly, the TIF shows that, overall, these three items are the most informative when measuring lower levels of self-esteem (solid line). However, during the study the participant mostly experienced medium and high levels of self-esteem, which means

[7] This trend must be interpreted with caution given the width of the credibility intervals of the attractor parameter, which can also suggest that the real trend is stable.

Figure 4.9: Item Characteristic Functions for the Items of Self-Esteem



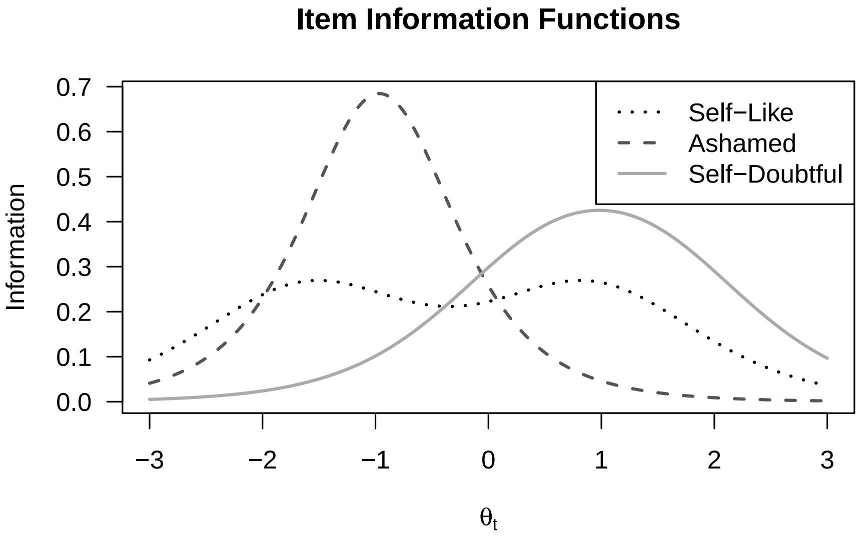
Item 1: Self−Like

Item 2: Ashamed

Item 3: Self−Doubtful

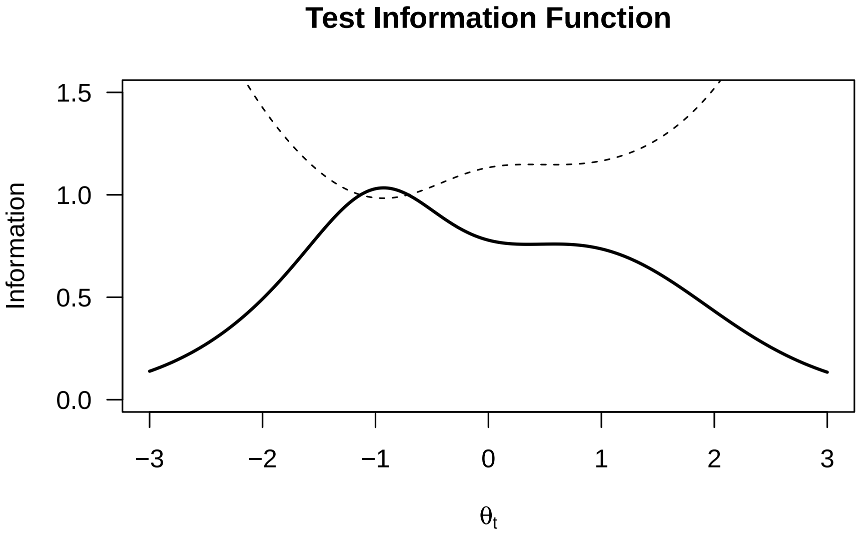
that their self-esteem was measured with high levels of standard measurement error (dashed line). This indicates that more items would be needed to accurately measure the whole spectrum of the participant's self-esteem.

Figure 4.10: Item Information Functions of the Items of Self-Esteem



Finally, we also computed the expected mean scores given the model, which can be interpreted as estimates of the *true scores* (Embretson & Reise, 2013), to compare them with the observed mean scores. This is shown in Figures 4.12 and 4.13. Figure 4.12 shows the nonlinear relation between the estimated latent state dispositions and the observed mean scores. It also displays the expected mean scores given the model (black line) for the observed range of the latent state dispositions. This plot evidences that the observed mean scores are not sufficient statistics for the latent state dispositions. Moreover, Figure 4.13 shows the observed mean scores against the expected mean scores for the last 50 observed beeps. The trajectory of the observed mean scores is closely followed by the trajectory of the expected mean scores. This shows the high predictive value of the TV-DPCM.

Figure 4.11: Test Information Function

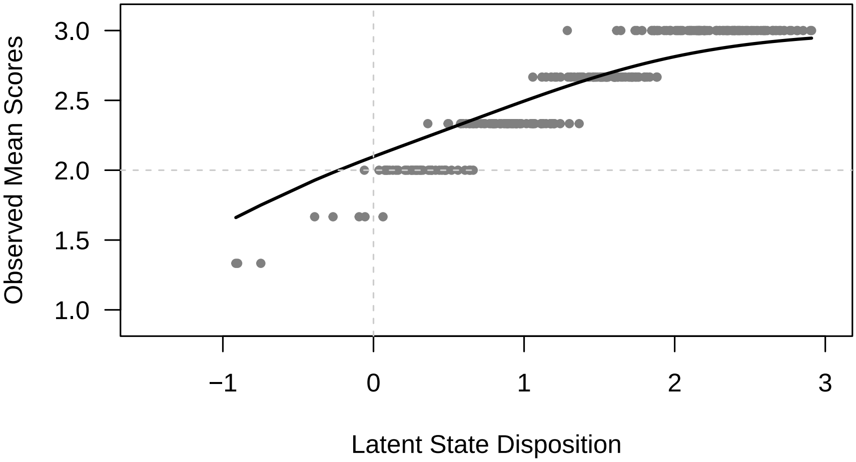


4.4 Discussion

In this chapter, we presented an extension of the PCM to analyze psychological time series, namely, the TV-DPCM. This proposed model integrates the PCM (Masters, 2016) and the TV-AR (Bringmann et al., 2017) to allow studying the quality of the measures of psychological constructs when measured intensively on one subject. We tested the performance of the model in a simulation study while controlling for the number of time points, the number of items, the size of the autoregressive effect, and the presence of missing data. We also illustrated, by means of an empirical example, how to estimate the model and interpret its results. Overall, the TV-DPCM seems to be a promising tool to further understand how psychological measurement works on intensive longitudinal settings.

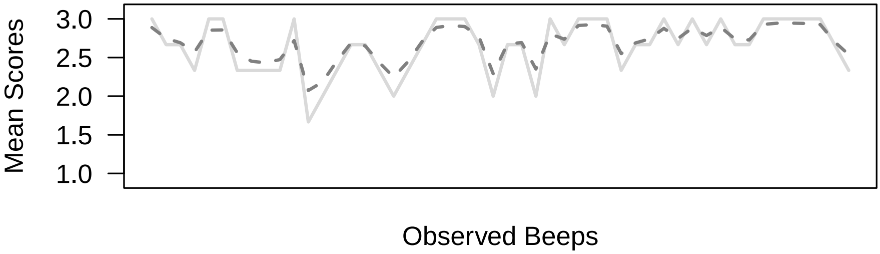
In general, the simulation study indicated that the model requires a large number of time points (more than 200) to converge and to deliver accurate estimates. This is in line with other results from simulation studies with autoregressive models for one individual (Bringmann et al., 2017; Schuurman et al., 2015). In fact, the TV-DPCM might also require more time points due to the increased complexity of the model.

Figure 4.12: Comparison Between the Estimated Latent States Dispositions, the Expected Mean Scores, and the Observed Mean Scores



Note. Comparison between the estimated latent state dispositions and the observed mean scores (gray dots). Also, the expected mean scores given the model in relation to the latent state dispositions are represented by the solid black line.

Figure 4.13: Comparison Between the the Expected Mean Scores and the Observed Mean Scores for the First and the Last 50 Observations.



Note. This figure presents the observed mean scores (solid light gray line) for the last 50 observed beeps of phase 2. The dashed gray line represents the expected mean scores given the model for the same 50 observed beeps.

Given that the credibility intervals of the estimates tend to be wide even with 200 time points, we actually suggest 300 observed time points as a minimum to have less uncertainty about the results and we discourage researchers to use the model with less than 200 time points. Nonetheless, the TV-DPCM seems to be able to accurately recover its parameters across most of the tested conditions.

In relation to the empirical example, we showed how the TV-DPCM allows making a rich interpretation of the scales used in intensive longitudinal settings. By using all the features provided by the IRT framework while accounting for the time dependencies of the data, we were able to take a closer look at the properties of the items and the scale of self-esteem from the empirical data. The information provided by the ICCs, the IIFs, and the TIF allows assessing the quality of the items and the scale, which can give researchers the opportunity to make adjustments for future applications of their experience sampling questionnaires. In this particular example, we noticed that probably more items that measure medium and high levels of self-esteem were needed to reliably measure this individual's self-esteem.

Even though the TV-DPCM can be a useful tool to gather relevant information about the measures in intensive longitudinal data, which can help to improve the scales used, the model still has its limitations. First, as shown in the empirical example, the model requires several steps of data manipulation such as reverse coding and collapsing response options. These steps are required to facilitate the interpretation of the latent variable and to be sure that thresholds are interpretable. Just as in the PCM, all responses options need to be observed to be able to estimate the parameters of the items in the TV-DPCM. If this is not the case, collapsing and recoding some response options becomes necessary, which reduces the variability of the observed data. While this might be a limitation of the model, it also represents a general challenge for researchers that are interested in studying psychological dynamics. This suggest that more research is needed in relation with the wording and the number of response options of the Likert-scale items used in intensive longitudinal settings. In other words, research focused on testing and improving the questionnaires used in intensive longitudinal setting is lacking.

Secondly, while the TV-DPCM is flexible enough to handle (non-linear) trend-stationary time series, which is a specific kind of non-stationarity, the model cannot handle other types of non-stationarity. Initially, a more flexible extension would be to allow the autoregressive effect to also smoothly vary over time (Bringmann et al., 2017). By doing this, the model would be able to handle time series with time varying variances and autocorrelations. Similarly, the TV-DPCM assumes that longitudinal measurement invariance (Meredith, 1993; Meredith & Teresi, 2006) holds. This assumption implies that the items have the same meaning and the same relation with the latent variable for the whole duration of the study. However, if measurement invariance does not hold, the parameters of the items might change, namely item parameter drift (Donoghue & Isham, 1998), and then the latent state dispositions from different measurement occasion would not be comparable. Given this, it would be necessary to extend the model to handle item parameter drift or at least develop statistics to test whether there is item parameter drift on some items. Currently, one way to study measurement non-invariance on intensive longitudinal data from multiple subjects has been proposed by Vogelsmeier et al. (2020) with the latent Markov latent trait analysis. Yet, given the complexity of this model, its use by practitioners might be limited and more research is needed to set guidelines in terms of minimum sample size, number of measurement occasions, or number of response options.

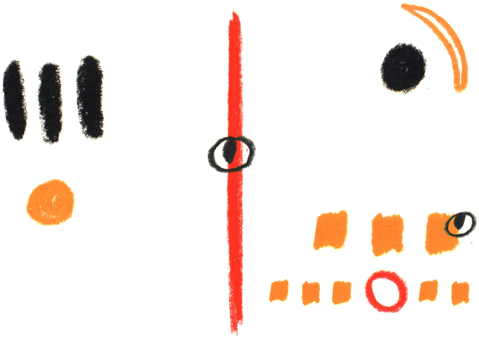
Thirdly, the simulation study showed that the TV-DPCM requires more than 200 time points to perform well. This is considerably above the typical length of the time series observed in intensive longitudinal research of psychological dynamics (Vachon et al., 2019). To overcome this, future research can try to extend the model to multilevel settings. Furthermore, the simulation also showed that the credibility intervals of most parameters tend to be relatively wide. A solution to this would be to increase the number of items in order to have more accurate estimates of the latent dynamic process. Yet, increasing the number of items might be hard to achieve in most intensive longitudinal data settings.

Lastly, in the empirical example, we showed that the TV-DPCM has a reasonable predictive value when comparing the expected mean scores and the observed mean scores. However, this is no guarantee that the model fits the empirical data well. For this, goodness of fit statistics should be developed for the TV-DPCM model and in general for the methods used to analyze intensive longitudinal data. Within the Bayesian framework, a method to assess the goodness of fit of a model is based on posterior predictive model checking methods (Gelman et al., 1996). These methods have also been developed for Bayesian IRT models (Li et al., 2017; Sinharay et al., 2006) but they need to be extended for the TV-DPCM to account for the time dependencies present in intensive longitudinal data. In fact, this is ongoing research, that we expect to be also useful for the TV-DPCM and other IRT model for intensive longitudinal data.

To conclude, bringing IRT with all its features to intensive longitudinal research is a great opportunity for the field. In addition to allowing the study of the psychological dynamics of the individual, it allows assessing the quality of the scales used in intensive longitudinal data, which can provide insight in how to improve these scales. As a result, researchers might be able to make better inferences and comprehend better the psychological dynamics of the individuals.

Chapter 5

Posterior Predictive Model Checking Methods for the Time-Varying Dynamic Partial Credit Model



This chapter is an unpublished manuscript titled Castro-Alvarez, S., Sinharay, S., Bringmann, L. F., Meijer, R. R., & Tendeiro, J. N. (2022). *Posterior Predictive Model Checking Methods for the Time-Varying Dynamic Partial Credit Model*

Abstract

In recent years, new models to analyze intensive longitudinal data have been proposed based on item response theory. One of this new models is the time-varying dynamic partial credit model (TV-DPCM), which is suitable to analyze psychological time series ($N = 1$). The TV-DPCM is a combination of the partial credit model and the time-varying autoregressive model which allows studying the psychometric properties of the items and modeling nonlinear trends at the latent level. Being the TV-DPCM a new model, statistics to assess its model fit are lacking. In this chapter, we propose and develop several test statistics and discrepancy measures based on the posterior predictive model checking method (PPMC) to assess the goodness-of-fit of the TV-DPCM. The proposed tools are based on implementations of the PPMC method for traditional dichotomous and polytomous item response theory models. The test statistics and discrepancy measures aim to identify model misfit when one or more of the assumptions of the TV-DPCM are violated. Simulated and empirical data are used to illustrate the effectiveness of the different measures.

Keywords: *time-varying dynamic partial credit model, partial credit model, autoregressive model, model misfit, posterior predictive model checking, psychological time series*

The use of intensive longitudinal data to study psychological dynamics has steadily grown during the last years (Hamaker & Wichers, 2017; Trull & Ebner-Priemer, 2020). Briefly, intensive longitudinal data consist of measuring individuals several times during short periods of time. For example, typical designs tend to request participants to fill in short questionnaires 10 times a day for one week (Vachon et al., 2019). This kind of studies is also known in the literature as "ecological momentary assessment", "experience sampling", or "daily diaries" (Myin-Germeys & Kuppens, 2021; Shiffman et al., 2008).

The growing interest in has also resulted in the development of new and complex statistical techniques to analyze intensive longitudinal data in psychology. Several new models and frameworks have been proposed to deal with such kind of data. Examples include network models (Bringmann et al., 2016; Bringmann et al., 2013; Moeller et al., 2018), dynamic structural equation modeling (Asparouhov et al., 2017, 2018), dynamic factors models (Fuller-Tyszkiewicz et al., 2017; Molenaar, 1985; Song & Zhang, 2014), unified structural equation modeling (Beltz et al., 2013; Beltz & Gates, 2017), and continuous-time modeling approaches (Voelkle & Oud, 2013; Voelkle et al., 2012).

Furthermore, researchers have also developed models based on the item response theory framework (IRT; Embretson & Reise, 2013) to analyze intensive longitudinal data. For example, the Rasch and the partial credit model were reformulated within the state-space modeling framework by Rijn et al. (2010). The Rasch model was also extended to handle continuous time data by Hecht et al. (2019), and recently, the partial credit model was extended to analyze categorical and multivariate time series data Castro-Alvarez et al. (2022a). These models have been specially developed to handle intensive longitudinal data when the items are dichotomous or polytomous (e.g., Likert-scale items). IRT models allow studying the response processes of the persons given their level on the latent ability. Furthermore, they also allow understanding the properties and the quality of the items and scales by means of the item parameters, the item characteristic function, and the item information function. In this chapter, we are specially interested in the time-varying dynamic partial credit model (TV-DPCM; Castro-Alvarez et al., 2022a) because this model is a promising approach that can handle non-stationary time series. The TV-DPCM is a combination of the partial credit model (PCM; Masters, 2016) and the time-varying autoregressive model (TV-AR; Bringmann et al., 2017). In a nutshell, the TV-DPCM allows modeling nonlinear latent dynamic processes of one individual while accounting for the measurement error of the latent construct of interest (e.g., positive affect). Moreover, as the TV-DPCM is an IRT model, it also allows studying the psychometric properties of the items and the scales used in intensive longitudinal research. However, in empirical settings, practitioners lack tools to assess the goodness-of-fit of the TV-DPCM. In this chapter, we develop goodness-of-fit statistics to assess the fit of the TV-DPCM to data. These goodness-of-fit statistics are derived from the posterior predictive model checking method (PPMC; Gelman et al., 1996; Rubin, 1984).

The PPMC method is a popular model checking tool within the Bayesian framework (Gelman et al., 1996; Rubin, 1984). The method consists of comparing features of the observed data to those from "replicated" data from the fitted model. Typically, these features are either *test statistics* or *discrepancy measures* that can be computed on both the observed and the replicated data. Furthermore, the PPMC method also allows estimating what is known as posterior predictive p-values (also known as Bayesian p-values), which are tail-area probabilities that quantify the similarity of the observed data with the replicated data. Within the IRT framework, the PPMC method has been introduced for dichotomous IRT models by Glas and Meijer (2003) and Sinharay (2005, 2006), and for polytomous IRT models by Li et al. (2017) and Zhu and Stone (2011). Based on these previous studies, we adapt and extend existing test statistics and discrepancy measures to assess the goodness-of-fit of the TV-DPCM (Castro-Alvarez et al., 2022a). The PPMC tools presented in this chapter can also be generalized to other dynamic IRT models.

The remaining of the chapter is organized as follows. Firstly, we briefly introduce the TV-DPCM alongside with its assumptions, practical use, and limitations. Secondly, we present the rationale behind the PPMC method and we list several test statistics and discrepancy measures proposed for the TV-DPCM. Thirdly, we present results from applying the proposed PPMC measures to simulated data. Then, we present results from applying the PPMC measures to empirical data. In this case, we show three examples, two in which the TV-DPCM does not fit the data and one in which it does. Finally, we discuss the implications and practical use of the proposed methods for future research.

5.1 The Time-Varying Dynamic Partial Credit Model

Let us consider a person who, as a part of treatment for depression, is asked by a therapist to report their mood multiple times a day by means of an application on a mobile phone for the whole duration of the treatment. In this application, a set of emotions is presented to the individual on every measurement occasion, and the individual must report the extent to which they are feeling those emotions on a Likert scale (e.g., with 5 response options). For example, to the question "I feel enthusiastic", the individual may choose among the response options "not at all" through "very much".

While variation is expected over the measurement occasions for an individual, we would expect that the measurements follow, for example, an increasing trend if the treatment was effective (with larger scores representing higher well-being). The observations from this person are useful to study psychological dynamics (Hamaker et al., 2015) and constitute a multivariate categorical time series. Temporal dependence, or the degree to which current observations can be predicted by previous observations, is one particularly informative aspect of multivariate categorical time series. A popular set of tools to handling temporal dependence consists of autoregressive (AR) models, which are a family of statistical models in which the structure of the time-dependency in the data is explicitly modeled through regression equations (e.g., Bringmann et al., 2017; Rovine & Walls, 2006). A recent example of an AR model based on IRT is the time-varying dynamic partial credit model that was suggested by Castro-Alvarez et al. (2022a).

The TV-DPCM can be viewed as a combination of the PCM (Masters, 2016) and the TV-AR (Bringmann et al., 2017). Thus, the model captures the assumption that the relationship between the observed responses and the latent states of the individual is described by the PCM, and that this relationship is stable over time (thus, longitudinal measurement invariance is assumed). Furthermore, at the latent-variable level, the TV-DPCM is based on the assumption that the latent states are described by a dynamic nonstationary structure that is represented by the TV-AR model. The purpose of this assumption is to account for the dependencies over time, which might be present between consecutive measurement occasions, while also accounting for durable change over time. Thus, the TV-DPCM is described by two equations: the measurement equation and the structural equation. The measurement equation is technically the same as the general equation of the PCM. Let us assume that the person reported how they felt at each of T time points on I emotions (items), each feeling being expressed in one of $m_i + 1$ response categories. Then, the probability of the person to select the x-th category of item i at time t can be expressed as

$$P(X_i = x|\theta_t) = \frac{exp\left[\sum\limits_{k=0}^{x}(\theta_t - \delta_{ik})\right]}{\sum\limits_{v=0}^{m_i} exp\left[\sum\limits_{k=0}^{v}(\theta_t - \delta_{ik})\right]}, \tag{5.1}$$

where $x = 0,\ldots,m_i$; $i = 1,\ldots,I$; $t = 1,\ldots,T$; θ_t is the latent state disposition of the individual at time t and δ_{ik} is the step parameter, also known as threshold parameter, of the k-th response category of item i. The threshold parameters δ_{ik} represent the level on the latent continuum at which the probabilities of selecting the response options k and $(k-1)$ are equal. In other words, the threshold parameters are the points on the latent continuum at which consecutive category response functions intercept. Furthermore, by definition:

$$\sum_{k=0}^{0}(\theta_t - \delta_{ik}) \equiv 0. \tag{5.2}$$

Note that the expression provided by Equation 5.1, which is the measurement equation of the TV-DPCM, is identical to that of a PCM. Then, the TV-DPCM involves the assumption that the latent states follow a dynamic structure based on the TV-AR model with a time-varying intercept. Thus, the *dynamic equation* under the model is defined as

$$\theta_t = \alpha_t + \varphi\theta_{t-1} + \varepsilon_t, \tag{5.3}$$

where α_t represents the time-varying intercept and φ represents the autoregressive effect between consecutive latent states, which is also known as the *inertia* (Kuppens et al., 2010); ε_t represents the innovation at time t. The innovations are assumed to be normally distributed with mean 0 and variance Ψ. Also, in contrast with Bringmann et al. (2017), in the TV-DPCM, the autoregressive effect is not allowed to change over time. This constrain was kept because the authors of the TV-DPCM did not succeed on writing a working model that also included a time-varying autoregressive effect (Castro-Alvarez et al., 2022a).

To model the time-varying intercepts α_t, the generalized additive modeling framework (GAM, Bringmann et al., 2017; Wood, 2017) is used within the TV-AR model implemented at the latent level. This framework is useful to model smooth nonlinear relationships. In general, generalized additive models (GAMs) imply that the relationship between the dependent variable and the predictors is based on multiple "basis functions", which are functions of the predictor. Then, the basis functions are weighted and summed together to form the nonlinear smooth function that better describes the relationship between the variables. For the TV-DPCM, we assumed that the time-varying intercepts α_t are a function of time based on the GAM without error, that is,

$$\alpha_t = f(t) = \sum_{j=1}^{s} \beta_j b_j(t), \tag{5.4}$$

where $b_j()$, $j = 1, \ldots, s$, is the j-th basis function and β_j is the unknown weight of the j-th basis function.

There exist several methods which can be used to define the basis functions, known as smoothers. Popular smoothers for univariate analyses are, for example, B-splines and thin plate (Wood, 2017). In addition to selecting a smoother, one must also select a penalization factor. Penalization is required to avoid overfitting and to diminish the fit wiggleness. In particular, the TV-DPCM, as implemented by Castro-Alvarez et al. (2022a), uses basic cubic splines to define the basis functions for the intercept and penalizes the fit by using a random walk prior on the β_j coefficients (Kharratzadeh, 2017).

Because of the introduction of a dynamic model component at the latent level, unlike for the PCM, the sumscore is no longer a sufficient statistic for the latent construct in the TV-DPCM. In addition, the TV-DPCM involves the assumptions that the autoregressive process is of order 1, the dynamic process is trend-stationary based on a nonlinear trend, and the item parameters do not change over time (i.e., longitudinal measurement invariance holds).

Castro-Alvarez et al. (2022a) demonstrated, using simulation and real-data results, that the TV-DPCM performs satisfactorily for certain types of multivariate categorical time-series data when there are at least 200 time points. They concluded that the TV-DPCM is a promising tool to further understand how psychological measurement works on intensive longitudinal settings.

While the TV-DPCM showed some promise, before applying the TV-DPCM at a larger scale, one has to demonstrate, using appropriate tools, that this model fits intensive longitudinal data adequately. However, Castro-Alvarez et al. (2022a) noted a lack of research on assessment of fit for models handling intensive longitudinal data, including the TV-DPCM. The main goal of this chapter is to fill that gap. Specifically, this chapter develops several tools, all under the framework of PPMC (Gelman et al., 1996; Sinharay et al., 2006), for assessing the goodness-of-fit of TV-DPCMs.

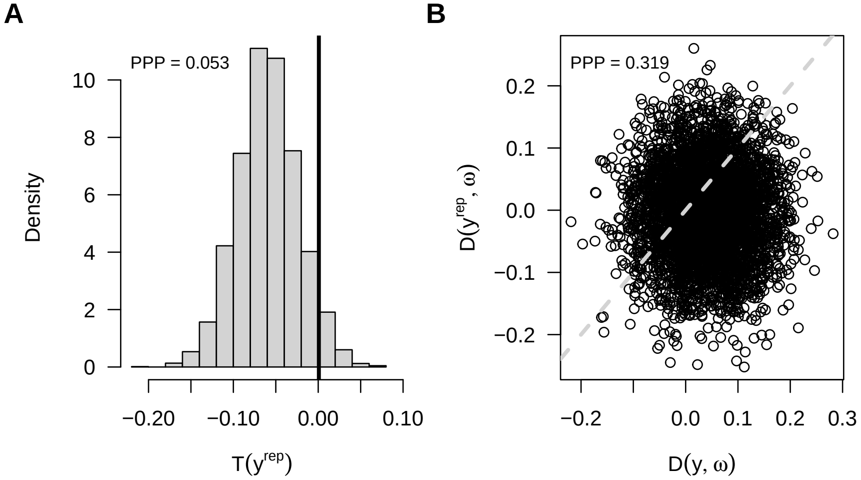
5.2 Posterior Predictive Model Checking

The idea behind the PPMC method is to assess whether various features of data simulated from the posterior predictive distribution are similar to those of the observed data (Gelman et al., 1996; Rubin, 1984). Usually, certain summaries (e.g., descriptive statistics) of the simulated and the observed data are compared. If there are systematic differences between the summaries for the two types of data, then one infers that the model fails to explain an aspect of the data (Sinharay et al., 2006). The PPMC method is in essence a visual diagnostic tool (Gelman et al., 1996). However, one can also use what are known as *posterior predictive p-values* (PPPs), which are tail-area probabilities. PPPs that are too extreme (for example, $< .05$ or $> .95$) are considered to provide evidence of model misfit.

Let $p(y|\omega)$ be the likelihood function of a statistical model with parameters ω applied to data y, and let $p(\omega)$ be the (joint) prior distribution of the parameters in the model. Based on Bayes theorem, the posterior distribution of ω is $p(\omega|y)$, which is proportional to $p(y|\omega)p(\omega)$. Then, let us define y^{rep} as “replicated” data that could have been observed. Conceptually speaking, y^{rep} are the resulting data of replicating the experiment that produced y assuming that the model and the estimated parameters are correctly specified. With this, the posterior predictive distribution of the replicated data can be defined as follows:

$$\begin{aligned} p(y^{rep}|y) &= \int p(y^{rep}, \omega|y)d\omega \\ &= \int p(y^{rep}|\omega, y)p(\omega|y)d\omega \\ &= \int p(y^{rep}|\omega)p(\omega|y)d\omega. \end{aligned} \tag{5.5}$$

Figure 5.1: Posterior Predictive Model Checking Methods



Note. Visual representation of posterior predictive model checking methods by using a test statistic (A) or a discrepancy measure (B). In panel A, the vertical line represents the value of the test statistic of the observed data, the histogram represents the distribution of the test statistic when computed for the replicated data y^{rep}. The proportion of the histogram that is larger than the test statistic of the observed data is an estimate of the PPP. Panel B represents the scatter plot of the pairs $\{D(y,\omega^j);D(y^{rep_j},\omega^j)\}$. The proportion of dots above the dashed gray line is an estimation of the PPP.

The third step in Equation 5.5 follows from the assumption that y and y^{rep} are conditional independent given ω (Gelman et al., 2014).

It is possible to run the PPMC method using *test statistics* or *discrepancy measures* (Gelman et al., 1996). When one uses a test statistic T to perform PPMC, where T is a function of the data only (and not a function of the model parameters), one actually compares the value $T(y)$ to the posterior predictive distribution of $T(y^{rep})$. Here, the posterior predictive distribution of $T(y^{rep})$ refers to the distribution of the statistic T computed from the replicated data, or, y^{rep}'s. The preferable way to compare $T(y)$ to the posterior predictive distribution of $T(y^{rep})$ is to plot the histogram of the posterior predictive distribution of $T(y^{rep})$ against $T(y)$ (see panel A of Figure 5.1). Alternatively, one can also compute the corresponding PPP:

$$PPP = P(T(y^{rep}) \geq T(y)|y) = \iint_{T(y^{rep}) \geq T(y)} p(y^{rep}|\omega)p(\omega|y)dy^{rep}d\omega, \tag{5.6}$$

which is the tail-area probability of $T(y^{rep})$ being larger than $T(y)$.

When one performs PPMC using a discrepancy measure, D, to compare the simulated and the observed data, where the value of D depends on the model parameters ω, one compares the posterior distribution of $D(y,\omega)$ with the posterior predictive distribution of $D(y^{rep},\omega)$. Panel B of Figure 5.1 demonstrates how one can perform the comparison using a graphical plot. The posterior predictive p-value corresponding to this comparison is given by

$$PPP = P(D(y^{rep},\omega) \geq D(y,\omega)|y) = \iint_{D(y^{rep},\omega) \geq D(y,\omega)} p(y^{rep}|\omega)p(\omega|y)dy^{rep}d\omega. \tag{5.7}$$

Because it is difficult to derive many theoretical results using equations 5.5 through 5.7 for all but very simple models, the application of the PPMC method is typically accomplished using simulations, and especially using MCMC algorithms (Gelman et al., 2014; Kruschke, 2014). Given a set of N draws $\omega^1, \omega^2, \ldots, \omega^N$ from the posterior distribution $p(\omega|y)$, implementation of the PPMC involves the following steps for $j = 1, 2, \ldots, N$:

1. Given ω^j, simulate a data set y^{rep_j} from the distribution $p(y|\omega^j)$.

2. Use y^{rep_j} to compute $T(y^{rep_j})$ when using a test statistic or compute $D(y, \omega^j)$ and $D(y^{rep_j}, \omega^j)$ when using a discrepancy measure.

While using a test statistic, the result is a set of N values $T(y^{rep_j})$. One can compare the distribution of these values with $T(y)$ via a histogram. The proportion of $T(y^{rep_j})$ that is larger than $T(y)$ is an estimate of the PPP. When using a discrepancy measure, the result is a set of N pairs $\{D(y, \omega^j); D(y^{rep_j}, \omega^j)\}$, which can be plotted in a scatterplot. Consequently, the estimate of the PPP is the proportion of pairs for which $D(y^{rep_j}, \omega^j)$ is larger than $D(y, \omega^j)$. Note that the PPPs are not a result of hypothesis-testing.

5.2.1 Test Statistics and Discrepancy Measures for the TV-DPCM

The importance of using a variety of test statistics and discrepancy measures to assess the fit of IRT models using the PPMC method has been emphasized by, for example, Sinharay et al. (2006) and Zhu and Stone (2011). Therefore, in this chapter, we use a variety of test statistics and discrepancy measures to assess the fit of the TV-DPCM. Some of the test statistics and discrepancy measures are modifications of those suggested for IRT models by Li et al. (2017), Sinharay et al. (2006), Zhu and Stone (2011), with modifications made to account for the time dependency implicit in time series data and in the TV-DPCM. The test statistics and discrepancy measures are described below. They are categorized into three groups: Test-level measures, item-level measures, and item pairwise measures.

Test-Level Measures

Popular test-level measures for regular IRT models include comparing the distribution of the sumscores (Li et al., 2017; Zhu & Stone, 2011). To account for the time component in the TV-DPCM, we initially used the trajectory of the time series of the sumscores instead. However, assessing the trajectories of the time series was not an effective measure to assess model misfit given preliminary simulations. Because of this, we considered using other test statistics or discrepancy measures that could be computed from the sumscores as test-level measures. Thus, traditional descriptive statistics of time series data (see Chatfield, 2004) such as the autocorrelation (ACF), the autocorrelation of the residuals (RACF), and the mean square successive differences of the sumscores were used as test statistics or discrepancy measures.

Moreover, an adjusted version of the autocorrelation of the residuals based on the residuals of the TV-DPCM is also proposed below.

Autocorrelation The autocorrelation function (ACF) consists of correlating one variable with lagged versions of itself (Chatfield, 2004; Houben et al., 2020). This statistic captures how self-predictive the variable is over time. The order of the autocorrelation is defined by the number of times that the variable was lagged. Usually, the strength of the autocorrelation is expected to fade away as the number of lags increases. To assess the goodness-of-fit of the TV-DPCM, we use the autocorrelations up to order 3 as test statistics.

Autocorrelation of the Residuals When fitting autoregressive models, the residuals are also ordered in time and can be seen as a time series. For this reason, calculating the autocorrelation function of the residuals is a usual procedure to assess the adequacy of autoregressive models (Chatfield, 2004). For example, if the autocorrelation of the residuals at lag 1 is high, it might imply that an autoregressive model of a higher order is needed to analyze the data. Following this idea, we consider the first autocorrelation of the residuals (RACF) as a test statistic for the TV-DPCM. In this case, we fit an autoregressive model to the time series of the sumscores and extract the residuals in order to compute the autocorrelation.

Mean Square Successive Differences This statistic is commonly used in autoregressive models as an indication of instability (Houben et al., 2020; von Neumann et al., 1941) and is a measure of dispersion that takes into account the order of the data. It aims to measure how different the successive observations are. The mean squared successive difference (MSSD) of the sumscores is computed as

$$MSSD(Y) = \frac{\sum_{t=1}^{T-1} (Y_{t+1} - Y_t)^2}{T-1}, \tag{5.8}$$

where Y_t is the sumscore at the t measurement occasion with $t = 1, \ldots, T$.

Autocorrelation of the Latent Residuals We propose another version of the autocorrelation of the residuals, which is calculated with the residuals of the TV-DPCM. For this, we first need to define how to compute the expected scores given the TV-DPCM. The expected score of item i at time t given the latent score of the person at time t is:

$$E_{it} = \sum_{x=0}^{m_i} xP(X_i = x|\theta_t), \tag{5.9}$$

where $P(X_i = x|\theta_t)$ is the probability of the observed response to item i at time t as defined in Equation 5.1. Then, to compute the expected sumscores given the TV-DPCM, one can sum the expected scores across items. The expected sumscores can be subtracted from the observed sumscores to obtain the residuals. Lastly, the autocorrelation is computed based on these residuals. We refer to this statistic as the autocorrelation of the latent residuals (LRACF).

Item-Level Measures

Discrepancy measures at the item level are used to identify problematic items. Typical item-level measures include the item scores distribution (Zhu & Stone, 2011), the item-total correlation (Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011), and item-fit indices such as the Orlando and Thissen item-fit statistic (Sinharay, 2006) and Yen's Q_1 statistic (Li et al., 2017; Zhu & Stone, 2011). In this chapter, we use the item-total correlation, Yen's Q_1, and the autocorrelation of the latent residuals of the item scores as discrepancy measures.

Item-Total Correlation The item-total correlation has been used in the PPMC method to demonstrate misfit of the Rasch model when the discrimination parameters varied across the items (Sinharay et al., 2006) and misfit of the graded response model (GRM) for multidimensional data (Zhu & Stone, 2011). In general, the item-total correlation coefficient is the simple correlation between the scores of an item with the total score of the test. A more rigorous version of this static implies to correlate the scores of an item with the rest scores, which are the total scores minus the scores of the relevant item (Howard & Forehand, 1962). For our analysis, we use the definition of the item-total correlation that uses the rest scores. Depending on the scale of the items, different types of correlations are used (biserial, polyserial, and Pearson). Moreover, to account for the time dependency of the data, we propose two alternative versions of this coefficient. A first modification consists of fitting an autoregressive model to the rest scores and extracting the residuals of this model. Then, we use the correlation coefficient between the item scores and these residuals as a discrepancy measure. The second alternative involves fitting an autoregressive model to both the scores of the item and the rest scores; we use the correlation coefficient between the residuals of the two models as a discrepancy measure.

Yen's Q_1. Yen's Q_1 (Yen, 1984) was proposed as a goodness-of-fit statistic that compares the observed and the predicted score distributions of each item in dichotomous unidimensional IRT models. This goodness-of-fit statistic was later generalized for polytomous items (Zhu & Stone, 2011). Thus, for polytomous items, Yen's Q_1 can be defined as follows:

$$Q_1 = \sum_{g=1}^{G} \sum_{x=0}^{m_i} N_g \frac{(O_{igx} - E_{igx})^2}{E_{igx}}, \tag{5.10}$$

where O_{igx} and E_{igx} are the observed and the expected proportions of responses in category x of item i of group g, and N_g is the size of group g. The G groups (usually 10) are ability subgroups of approximately the same size in which subjects with similar ability are grouped together. In the TV-DPCM, the groups represent a subset of time points in which the participant's experiences had a similar intensity. When used as a discrepancy measure in the PPMC approach, this measure has shown mixed results. For example, Zhu and Stone (2011) did not find this measure effective to identify violations of model assumptions of the GRM. In contrast, Li et al. (2017) found it effective to identify model misspecification for nested models based on the generalized PCM.

We propose an alternative version of the Yen's Q_1 (referred to as "Q_1 alt." later) in which the groups are defined according to the order of the observations; the first set of observations form the first group and so on. By keeping the order of the observations, we expect to account for the time dependency likely to be observed in intensive longitudinal data.

Autocorrelation of the Latent Residuals per Item As with the autocorrelation of the latent residuals of the sumscores, we also use the autocorrelation of the residuals given the TV-DPCM but for each item score. The expected item scores are obtained by means of Equation 5.9. Then, to obtain the residuals, the expected scores are subtracted from the observed item scores. Lastly, the autocorrelation of the residuals is computed per item.

Pairwise-Measures

Models within the IRT framework typically involve the assumption of local independence, which means that the relationships among items are fully captured by the IRT model (Embretson & Reise, 2013). Consequently, when studying the fit of an IRT model using the PPMC method, fit indices that describe the associations between items, such as the odds ratio, Yen's Q_3, and the absolute item covariance residual, have also been used as discrepancy measures (Li et al., 2017; Zhu & Stone, 2011). In this study, we also use these same statistics to assess the goodness-of-fit of the TV-DPCM. Next, we briefly describe each of these measures.

Yen's Q_3 Yen's Q_3 (Yen, 1984) is a popular fit statistic used as a discrepancy measure for polytomous IRT models (Li et al., 2017; Zhu & Stone, 2011). When used as a discrepancy measure, it has been effective in identifying violations of unidimensionality when assessing the GRM (Zhu & Stone, 2011). In simple words, Yen's Q_3 is defined as the correlation between scores of two items after controlling for the latent ability of the persons. In other words, given the residuals scores (d_i) of two items, which are computed based on the expected scores of Equation 5.9, Yen's Q_3 is the correlation of the residuals ($Q_3 = cor(d_i, d_j)$).

Odds Ratio The odds ratio has been used as a discrepancy measure with dichotomous and polytomous IRT models (Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011). When applied to polytomous items, the responses need to be dichotomized to compute this statistic. For dichotomous responses, the odds ratio is defined as

$$OR = \frac{n_{11}n_{00}}{n_{10}n_{01}}, \tag{5.11}$$

where n_{ij} is the number of persons whose responses to the two items in the item pair were i and j, respectively. For the version of the odds ratio that we used to assess the fit of TV-DPCM, n_{ij} denotes the number of time points where the subject's responses were i and j. We also use a modified version of this discrepancy, where the odds ratio between two items is computed for each half of the time series and the difference between the two odds ratio is used as a discrepancy measure. The reason of using this discrepancy, referred to as the *odds ratio difference* or the *OR difference* is to capture changes in the response process that are not accounted for by the TV-DPCM. For example, if the person starts interpreting and responding to the items differently after a certain time point (i.e., longitudinal measurement invariance does not hold), this measure should indicate some sort of misfit given that longitudinal measurement invariance is assumed in the TV-DPCM.

Absolute Item Covariance Residual This discrepancy measure was used by Zhu and Stone (2011) to test violations of unidimensionality and local dependence of the GRM. This measure is defined as the absolute value of the difference between the observed and the expected item covariances after estimating the model. Given two items i and j, the absolute item covariance residual is defined as follows:

$$RESID_{ij} = |cov(X_i, X_j) - cov(E_i, E_j)|. \tag{5.12}$$

Here, E_i is the expected scores on item i across all the time points, which is computed based on Equation 5.9. As with the odds ratio, we compute the absolute item covariance residual for each half of the time series and use the difference between the two estimates as a discrepancy measure in addition to using the residual itself as a discrepancy measure. This alternative is referred to as *RESID difference*.

5.3 Testing the PPMC Method with Simulated Data

We studied the performance of the different test statistics and discrepancy measures under the PPMC framework for assessing the fit of the TV-DPCM to several simulated data sets. In general, we generated data that violated none or some of the assumptions of the TV-DPCM. We then fit the TV-DPCM to the simulated data and computed the values of the test statistics and discrepancy measures. From these values, we computed the PPPs for each measure and evaluated the proportion of PPPs that were too extreme (smaller than 0.05 or larger than 0.95).

5.3.1 Method

We studied the performance of the test statistics and discrepancy measures for 10 different generating models. Two of these did not involve any misfit of the TV-DPCM while eight had different types of misfit. The description of each model is presented in Table 5.1. For all the simulations, we used 200 time points, which is the minimum number of time points advisable to accurately estimate the parameters of the TV-DPCM (Castro-Alvarez et al., 2022a). Moreover, the number of items was

Table 5.1: True Generating Models and Sources of Misfit

Data Generating Model	Source of Misfit
TV-DPCM	None.
DPCM	None.
TV-DPCM(3)	The true dynamic process has 3 lagged effects.
TV-MDPCM	The items are grouped into two correlated latent dimensions.
TV-DGPCM	Two items have discrimination different from 1.
TV-DPCM-IPD	There is homogeneous item parameter drift for all the items on the second half of the measurements.
TV-DPCM-HIPD	Two items show parameter drift for the last third of the observations. One item becomes "harder" and the other becomes "easier".
TV-DPCM-Default	The responses to the last third of the measurements are the same for all the items and time points.
TV-DPCM-Random	The responses to the last third of the measurements are randomly generated for all items.
TV-DPCM-Meaning	The interpretation of one item changes for the last third of the measurements. The simulated item scores of this item are reversed.

varied between 3, 6, and 12. By crossing the generating model and the number of items, we obtained 30 different simulation conditions. We used 10 replications per condition.

The TV-DPCM was fitted to each simulated data set using the Hamiltonian Monte Carlo algorithm (Carpenter et al., 2017). For each data set, we employed three parallel chains and 2,000 iterations per chain (500 iterations of warm-up). Additional parameters of the Hamiltonian Monte Carlo algorithm were adjusted. In particular, the *delta* parameter was increased from 0.8 to 0.99 and the *maximum treedepth* was increased from 10 to 15 (see Stan Development Team, 2020). Then, the PPPs were computed for all the test statistics and discrepancy measures presented in the previous sections.

PPPs that were smaller than 0.05 or larger than 0.95 were considered to be too extreme and evidence of model misfit. In the results, we summarize the evidence of misfit with the proportion of extreme PPPs for a given discrepancy across conditions. For the test-level measures, the proportion of extreme PPPs is the number of PPPs that indicated model misfit divided by the 10 replications. In contrast, for the item-level and the pairwise measures, the proportions were aggregated across items or pairs. In other words, for an item-level measure such as the item-total correlation, the proportion of extreme PPPs for a condition with 6 items was based on 60 PPPs (number of replications $\times$ number of items). Furthermore, for some conditions, the proportions of extreme PPPs of the item-level and pairwise measures were computed for different subsets of items or pairs. For example, when the generating model was the TV-DPCM-HIPD, a proportion of extreme PPPs of the item-level measures was computed for the items that do not show item parameter drift and another proportion was computed for the items that show item parameter drift.

5.3.2 Results

The proportions of extreme PPPs for each statistic and for the conditions with different number of items when the generating model was the TV-DPCM are shown in Table 5.2. The proportions when the generating model is the DPCM are not presented as they are very similar to those shown in Table 5.2. In general for these two conditions, extreme PPPs were rarely observed. In most of the cases the proportions of observed extreme PPPs were 0. When the proportions were larger than 0, they

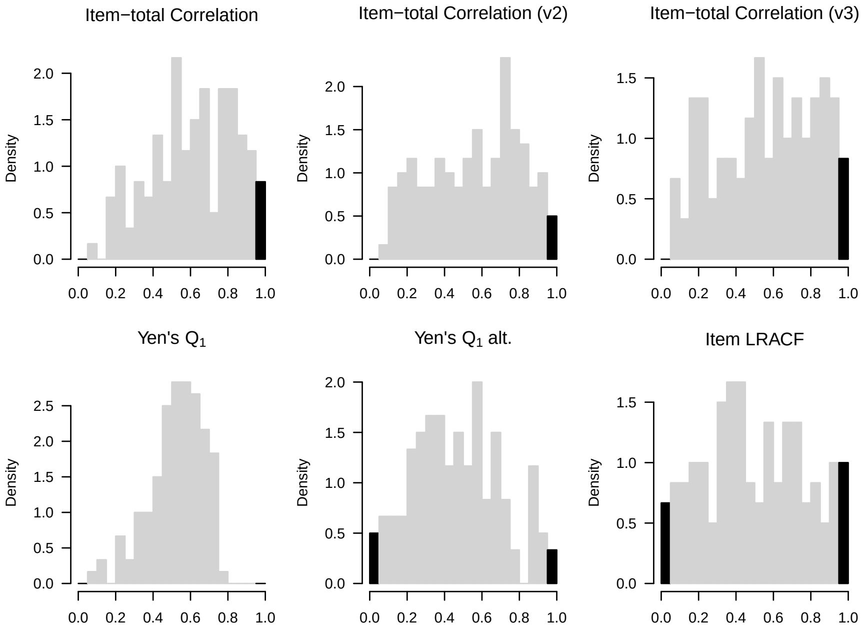
Table 5.2: Proportion of Extreme PPP-Values with the TV-DPCM as the Generating Model

	Number of Items		
	$I = 3$	$I = 6$	$I = 12$
Test-level measures			
ACF lag 1	0.00	0.00	0.00
ACF lag 2	0.00	0.00	0.00
ACF lag 3	0.00	0.00	0.00
RACF	0.00	0.00	0.00
LRACF	0.00	0.00	0.00
MSSD	0.00	0.00	0.00
Item-level measures			
Item-total correlation	0.07	0.03	0.04
Item-total correlation (v2)	0.07	0.00	0.03
Item-total correlation (v3)	0.00	0.02	0.04
Yen's Q_1	0.00	0.00	0.00
Yen's Q_1 alt.	0.00	0.05	0.04
Item LRACF	0.00	0.00	0.08
Pairwise measures			
Yen's Q_3	0.00	0.03	0.03
OR	0.07	0.07	0.07
OR difference	0.07	0.11	0.09
RESID	0.00	0.00	0.00
RESID difference	0.00	0.01	0.02

varied between 0.01 and 0.11 for discrepancy measures such as the OR and the OR difference.

Figures 5.2 and 5.3 present the distributions of the PPPs for the different item-level and pairwise discrepancy measures when the generating model was the TV-DPCM and when there were 12 items. Some of these distributions are far from a uniform distribution between 0 and 1. In particular, the PPPs for Yen's Q_1 are never larger than 0.8 and the PPPs for the RESID are never larger than 0.7. Thus, Table 5.2 and Figures 5.2 and 5.3 imply that the proposed test statistics and discrepancy measures are conservative.

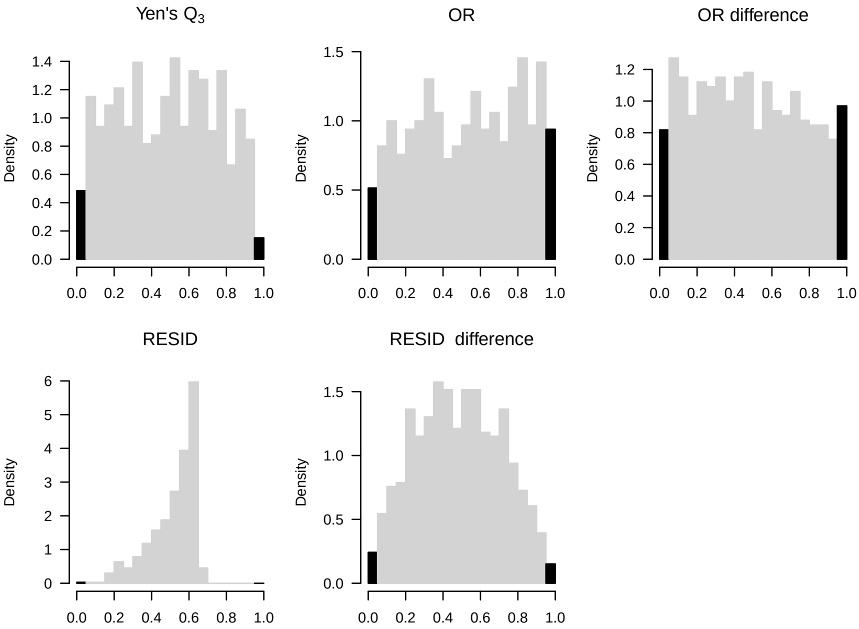
Figure 5.2: Distribution of the PPPs of the Item-Level Measures.



Note. The distributions are based on the simulations when the generating model was the TV-DPCM with 12 items. PPPs that indicate model misfit are drawn in black.

When looking at the conditions in which the generating model included some sort of misfit, none of the test-level measures were effective at detecting misfit. Across all conditions, the proportions of extreme PPPs of the test-level measures were 0 or very close to 0, indicating that the PPPs for test-level measures were rarely extreme.

Figure 5.3: Distribution of the PPPs of the Pairwise Measures.



Note. The distributions are based on the simulations when the generating model was the TV-DPCM with 12 items. PPPs that indicate model misfit are drawn in black.

Furthermore, in the conditions where the generating model was the TV-DPCM(3) or the TV-DPCM-IPD, implying that the latent dynamic process had up to 3 lagged effects and that all the items showed homogeneous item parameter drift, none of the measures were powerful at identifying misfit for these generating models. Across all the different test statistics or discrepancy measures, the proportions of extreme PPPs were 0 or very close to 0. Similarly, the measures were not effective at identifying misfit when the generating model was the TV-DPCM-Random, implying that the responses of the last third of the measurements were randomly selected. For these simulation conditions, the proportion of extreme PPPs among the item-level and the pairwise measures was rarely higher than 40%. While some indication of misfit was observed, the proposed PPMC methods failed to consistently identify misfit under these conditions.

Tables 5.3 to 5.7 present the proportions of extreme PPPs for the measures of the conditions where the generating model were the TV-MDPCM (multidimensional), the TV-DGPCM (discrimination different from 1), the TV-DPCM-Default (default responses), the TV-DPCM-HIPD (item parameter drift for some items), and the TV-DPCM-Meaning (change of meaning for one item), respectively. Under these conditions, some of the discrepancy measures seemed to be effective at identifying the misfit.

Table 5.3 shows that Yen's Q_3 was effective at identifying violations of the unidimensionality assumption. When the items are multidimensional, one can expect that the PPPs for Yen's Q_3 indexes for the pairs of items of the same dimension to be too extreme—the proportions of extreme PPPs for such pairs were around 0.9. When the item pairs include items from the two dimensions, the proportions of extreme PPPs for Q_3 increased from 0.2 for 3 items to 0.74 for 12 items, implying that the measure was more effective when the number of items increases. Other extreme PPPs were observed across the other discrepancy measures but they did not have enough power and did not consistently indicate misfit when the generating model was the TV-MDPCM.

Table 5.4 shows that when the generating model was the TV-DGPCM, that is, when the discrimination parameters of some items were different from 1, only the different versions of the item-total correlations were effective at detecting misfit for the problematic items. From these, the unmodified version of the item-total correlation performed the best, showing proportions of extreme PPPs equal to or larger than

Table 5.3: Proportion of Extreme PPP-Values with the TV-MDPCM as the Generating Model

	Number of Items		
	$I = 3$	$I = 6$	$I = 12$
Item-level measures			
Item-total correlation dim1	0.15	0.03	0.03
Item-total correlation dim2	0.5	0.17	0.05
Item-total correlation (v2) dim1	0.15	0.07	0.03
Item-total correlation (v2) dim2	0.6	0.13	0.08
Item-total correlation (v3) dim1	0.1	0.07	0.05
Item-total correlation (v3) dim2	0.4	0.13	0.1
Yen's Q_1 dim1	0	0	0
Yen's Q_1 dim2	0	0	0
Yen's Q_1 alt. dim1	0.05	0	0.05
Yen's Q_1 alt. dim2	0	0.07	0.08
Item LRACF dim1	0.1	0.33	0.33
Item LRACF dim2	0.4	0.4	0.33
Pairwise measures			
Yen's Q_3 (dim1, dim1)	**0.9**	**0.87**	**0.86**
Yen's Q_3 (dim1, dim2)	0.2	0.57	**0.74**
Yen's Q_3 (dim2, dim2)	-	**0.93**	**0.9**
OR (dim1, dim1)	0.4	0.37	0.5
OR (dim1, dim2)	0.15	0.33	0.28
OR (dim2, dim2)	-	0.47	0.45
OR difference (dim1, dim1)	0.3	0.2	0.29
OR difference (dim1, dim2)	0	0	0.02
OR difference (dim2, dim2)	-	0.23	0.27
RESID (dim1, dim1)	0.6	0.53	0.37
RESID (dim1, dim2)	0.05	0.22	0.32
RESID (dim2, dim2)	-	0.5	0.43
RESID difference (dim1, dim1)	0.2	0.23	0.18
RESID difference (dim1, dim2)	0.05	0.2	0.19
RESID difference (dim2, dim2)	-	0.23	0.21

Note. dim1 denotes the items of dimension 1. *dim2* denotes the items of dimension 2. Proportions larger than 0.7 are highlighted in boldface.

0.85 when there were 6 or more items for the items with discrimination parameters different from 1.

Table 5.5 presents the proportions of extreme PPPs for the conditions where the generating model was the TV-DPCM-HIPD (two items had item parameter drift). For these conditions, the alternative version of Yen's Q_1 and the LRACF of each item were effective at identifying the problematic items. For these two measures, all the proportions were larger than 0.8. Moreover, all the pairwise measures with the exception of the OR difference were successful at identifying misfit, especially for the conditions with more items.

Table 5.6 shows that when the generating model was the TV-DPCM-Default, that is, when the person selected the same response option for all the items during the last third of the study, the alternative version of Yen's Q_1 and the two versions of the odds ratio were powerful in detecting misfit of the TV-DPCM. For these simulation conditions, the alternative Yen's Q_1 always indicated model misfit. On the other hand, the proportions of extreme PPPs for the two versions of the odds ratio were around 0.8 independent of the number of items.

Lastly, Table 5.7 indicates that for the conditions where the generating model was the TV-DPCM-Meaning, all the item-level measures seemed to have more than enough power to identify misfit of the TV-DPCM with proportions of extreme PPPs close to 1. Furthermore, for the pairwise measures such as the odds ratio, the RESID, and the RESID difference, the PPPs are mostly extreme for the pairs that included the item whose meaning changed. In particular, the pairwise measures worked better in the conditions with more items.

Table 5.4: Proportion of Extreme PPP-Values with the TV-DGPCM as the Generating Model

	Number of Items		
	$I = 3$	$I = 6$	$I = 12$
Item-level measures			
Item-total correlation disc1	0.1	0.1	0.1
Item-total correlation disc2	0.25	**0.85**	**0.95**
Item-total correlation (v2) disc1	0	0	0.07
Item-total correlation (v2) disc2	0.25	0.55	**0.7**
Item-total correlation (v3) disc1	0	0	0.08
Item-total correlation (v3) disc2	0.15	0.6	**0.8**
Yen's Q_1 disc1	0	0	0
Yen's Q_1 disc2	0	0.05	0.25
Yen's Q_1 alt. disc1	0	0	0.03
Yen's Q_1 alt. disc2	0	0.1	0.1
Item LRACF disc1	0	0.03	0.02
Item LRACF disc2	0	0	0
Pairwise measures			
Yen's Q_3 (disc1, disc1)	-	0	0.05
Yen's Q_3 (disc1, disc2)	0.2	0.16	0.09
Yen's Q_3 (disc2, disc2)	0	0	0
OR (disc1, disc1)	-	0.03	0.08
OR (disc1, disc2)	0.2	0.3	0.32
OR (disc2, disc2)	0	0	0.2
OR difference (disc1, disc1)	-	0.1	0.1
OR difference (disc1, disc2)	0.05	0.17	0.1
OR difference (disc2, disc2)	0.1	0	0.1
RESID (disc1, disc1)	-	0	0
RESID (disc1, disc2)	0.05	0.12	0.2
RESID (disc2, disc2)	0	0	0
RESID difference (disc1, disc1)	-	0	0.01
RESID difference (disc1, disc2)	0.05	0.1	0.09
RESID difference (disc2, disc2)	0.1	0.1	0.2

Note. disc denotes the items with discrimination parameter equal to 1. *disc2* denotes the items with discrimination parameters different from 1. Proportions larger than 0.7 are highlighted in boldface.

Table 5.5: Proportion of Extreme PPP-Values with the TV-DPCM-HIPD as the Generating Model

	Number of Items		
	$I = 3$	$I = 6$	$I = 12$
Item-level measures			
Item-total correlation drift1	0.5	0.45	0.24
Item-total correlation drift2	0.15	**0.7**	0.5
Item-total correlation (v2) drift1	0.5	0.25	0.07
Item-total correlation (v2) drift2	0.15	0.15	0.4
Item-total correlation (v3) drift1	0.6	0.25	0.17
Item-total correlation (v3) drift2	0.2	0.15	0.25
Yen's Q_1 drift1	0	0	0
Yen's Q_1 drift2	0.1	0.35	0.4
Yen's Q_1 alt. drift1	0	0.03	0.06
Yen's Q_1 alt. drift2	**0.9**	**0.8**	**0.95**
Item LRACF drift1	0	0	0.05
Item LRACF drift2	**0.85**	**0.9**	**1**
Pairwise measures			
Yen's Q_3 (drift1, drift1)	-	0.05	0.06
Yen's Q_3 (drift1, drift2)	0.25	0.05	0.06
Yen's Q_3 (drift2, drift2)	0.4	**0.7**	**1**
OR (drift1, drift1)	-	0.18	0.13
OR (drift1, drift2)	0.15	0.19	0.26
OR (drift2, drift2)	0.4	**0.7**	**1**
OR difference (drift1, drift1)	-	0.22	0.14
OR difference (drift1, drift2)	0.25	0.16	0.12
OR difference (drift2, drift2)	0.4	0.2	0.2
RESID (drift1, drift1)	-	0.03	0
RESID (drift1, drift2)	0.1	0.12	0.22
RESID (drift2, drift2)	0.3	**0.7**	**1**
RESID difference (drift1, drift1)	-	0.12	0.02
RESID difference (drift1, drift2)	0.1	0.12	0.19
RESID difference (drift2, drift2)	0.3	0.5	**0.8**

Note. drift1 denotes the items without parameter drift. *drift2* denotes the items with parameter drift. Proportions larger than 0.7 are highlighted in boldface.

Table 5.6: Proportion of Extreme PPP-Values with the TV-DPCM-Default as the Generating Model

	Number of Items		
	$I = 3$	$I = 6$	$I = 12$
Item-level measures			
Item-total correlation	0.53	0.52	0.53
Item-total correlation (v2)	0.4	0.37	0.37
Item-total correlation (v3)	0.37	0.4	0.46
Yen's Q_1	0.13	0.23	0.39
Yen's Q_1 alt.	**1**	**1**	**1**
Item LRACF	0.5	0.42	0.58
Pairwise measures			
Yen's Q_3	0.43	0.52	0.52
OR	**0.9**	**0.77**	**0.77**
OR difference	**0.77**	**0.85**	**0.8**
RESID	0.23	0.25	0.29
RESID difference	0.37	0.28	0.31

Note. Proportions larger than 0.7 are highlighted in boldface.

Table 5.7: Proportion of Extreme PPP-Values with the TV-DPCM-Meaning as the Generating Model

	Number of Items		
	$I = 3$	$I = 6$	$I = 12$
Item-level measures			
Item-total correlation meaning1	0.25	**0.8**	0.35
Item-total correlation meaning2	**0.9**	**1**	**1**
Item-total correlation (v2) meaning1	0.2	0.58	0.15
Item-total correlation (v2) meaning2	**0.9**	**1**	**1**
Item-total correlation (v3) meaning1	0.3	0.62	0.26
Item-total correlation (v3) meaning2	0.2	**1**	**1**
Yen's Q_1 meaning1	0	0	0
Yen's Q_1 meaning2	0.1	**1**	**1**
Yen's Q_1 alt. meaning1	0.2	0.02	0.03
Yen's Q_1 alt. meaning2	**0.7**	**0.7**	**0.8**
Item LRACF meaning1	0.45	0.06	0.06
Item LRACF meaning2	**1**	**1**	**1**
Item-level measures			
Yen's Q_3 (meaning1, meaning1)	**1**	0.59	0.09
Yen's Q_3 (meaning1, meaning2)	0.55	**0.72**	0.35
OR (meaning1, meaning1)	**0.7**	0.52	0.21
OR (meaning1, meaning2)	0.25	**0.78**	**0.94**
OR difference (meaning1, meaning1)	0.3	0.37	0.17
OR difference (meaning1, meaning2)	0.55	0.2	0.08
RESID (meaning1, meaning1)	**0.9**	0.6	0.04
RESID (meaning1, meaning2)	0.4	**0.88**	**0.9**
RESID difference (meaning1, meaning1)	0.2	0.28	0.07
RESID difference (meaning1, meaning2)	0.4	**0.8**	**0.84**

Note. meaning1 denotes the items for which its meaning did not changed. *meaning2* denotes the item for which its meaning changed. Proportions larger than 0.7 are highlighted in boldface.

5.3.3 Summary

We studied the performance of several test statistics and discrepancy measures in assessing the fit of the TV-DPCM under a PPMC framework using simulated data. In general, the discrepancy measures tended to be conservative and insensitive to certain types of misfit, such as when the generating model was the TV-DPCM(3) or the TV-DPCM-IPD. However, some of the discrepancy measures were effective in detecting other types of misfit. As we tested the measures with simulated data, we have certainty about which type of misfit is being detected, so we know which measures were effective at identifying multidimensionality, item discrimination parameters being different from 1, or changes in the response process of the participant. Table 5.8 presents a summary of the effectiveness of the different test statistics and discrepancy measures under the misfitting generating models. Regarding the different versions of the item-total correlation, the performance of these three test statistics was very similar. Therefore, the traditional item-total correlation seems to be enough for assessing model fit by means of the PPMC method. Extreme values of the item-total correlation might indicate that the item discrimination is different from 1 or that the interpretation of the item changed for the person. The alternative version of Yen's Q_1 was also effective at identifying model misfit for the conditions where the response process of the person changed (i.e., TV-DPCM-HIPD, TV-DPCM-Default, and TV-DPCM-Meaning). For this measure, the simple change to Yen's Q_1 led to an improvement of the traditional Yen's Q_1 when applied to the TV-DPCM. Also, in agreement with Zhu and Stone (2011), Yen's Q_3 statistic was also effective at identifying violations of unidimensionality. Lastly, measures such as the LRACF of the item scores, the OR, the RESID, and the RESID difference were relatively effective at identifying misfit under some of the conditions where the response process of the person changed.

One important finding from the simulations is that there is no single winner among the test statistics and discrepancy measures; that is, there is no single measure that is powerful to detect all types of model misfit. Instead, various measures are powerful for various types of misfit. This result is not a surprise—researchers such as Sinharay et al. (2006) and Zhu and Stone (2011) found similar results in applications of the PPMC method to IRT models, which is exactly why they, as well as other experts, recommended the use of several test statistics and discrepancy measures while performing PPMC for IRT models using a data set.

Table 5.8: Summary of the Effectiveness of the Different Test Statistics and Discrepancy Measures

	TV-MDPCM	TV-DGPCM	TV-DPCM-HIPD	Default	Random	Meaning
Item-level measures						
Item-total correlation		✓	○	○		✓
Item-total correlation (v2)		✓	○	○		✓
Item-total correlation (v3)		✓		○		✓
Yen's Q_1			○*	○*		✓
Yen's Q_1 alt.			✓	✓	○	✓
Item LRACF	○		✓	○	○	✓
Pairwise measures						
Yen's Q_3	✓		✓*	○		○
OR	○	○	✓*	✓		✓*
OR difference				✓	○	
RESID	○		✓*			✓*
RESID difference			✓*			✓*

Note. ✓ indicates that the proportions of extreme PPPs for that condition were generally above 0.7. ○ indicates that the proportions of extreme PPPs for that condition were between 0.3 and 0.7.

*The measure is more effective when the number of items increases.

5.4 Application to Empirical Data

To examine how our suggested measures perform for real data, we assessed the fit of the TV-DPCM to data from a daily diary study conducted by Wichers and Groot (2016). The data come from a 57 years old male (at the time), who completed in total 1.473 experience sampling assessments between August 2012 and April 2013. The participant had to fill in an experience sampling questionnaire up to 10 times a day within 90-minute intervals for 239 days (Kossakowski et al., 2017). On average, the participant completed 6.2 assessments per day. Furthermore, by the time of the study, the participant had a medical history of major depression and was taking medication. The data collection happened in conjunction with a double-blind dose reduction scheme of the participant's antidepressant (Kossakowski et al., 2017; Wichers & Groot, 2016) which included (1) four weeks of baseline, (2) two weeks without dosage reduction, (3) eight weeks with gradual dosage reduction, (4) eight weeks of post assessment, and (5) twelve weeks of follow-up.

The experience sampling questionnaire included 50 items that were asked on every occasion. Overall, the items aimed to measure mood states, self-esteem, physical condition, and the social environment at the moment of the assessment. All the items measuring the different emotions and self-esteem were measured on a 7-point Likert scale raging from "not feeling the state" to "feeling the state very much". We assessed the fit of the TV-DPCM to data for three sets of items included in the study. First, we assessed the fit to data corresponding to the items on negative affect and mental unrest. These items correspond to the emotions "down", "lonely", "anxious", and "guilty" for negative affect; and "irritated", "restless", and "agitated" for mental unrest. Second, we assessed the fit to data corresponding to the items of positive affect, which correspond to "relaxed", "satisfied", "enthusiastic", "cheerful", and "strong". Finally, we assessed the fit to data corresponding to the items on self-esteem, which correspond to "I like myself", "I am ashamed of myself", and "I doubt myself"; Castro-Alvarez et al. (2022a) fitted and interpreted the results of the TV-DPCM to the same items.

Notice that for the analyses, we collapsed several response categories and recoded the responses because some response categories were never chosen by the participant. Thus, the number of response options of the items of negative and positive affect, and the items of mental unrest, were reduced to five response options. The number of response options of the items of self-esteem was reduced to three response options.

Also, the items of self-esteem, “I am ashamed of myself” and “I doubt myself”, were reversed-coded before the analyses so that high scores on the items represent high levels of self-esteem.

Furthermore, to account for the unequal time interval between observations due to missing data and the time overnight, we implemented an approach that is typically used in the dynamic structural equation modeling framework (Asparouhov et al., 2018). This approach involves including missing data between observations to make the time intervals approximately equal across the data. In our case, we divided a day into 16 90-minute time windows. If an observation was made within a specific time window, then the observation was considered representative for the mental state of the participant during that 90-minute time window. Given that the participant reported their emotions up to 10 times a day, data for at least 6 of the 16 time windows were always missing for each day.

Lastly, to fit the TV-DPCM to the empirical data, we used three parallel chains each with 2,000 iterations from which the first 500 were used as warm-up. Then, to assess the fit of the models, we used 4,500 posterior predictive data sets for each model to compute the PPPs for the various test statistics and discrepancy measures.

5.4.1 Assessing Model Fit of the Items of Negative Affect and Mental Unrest

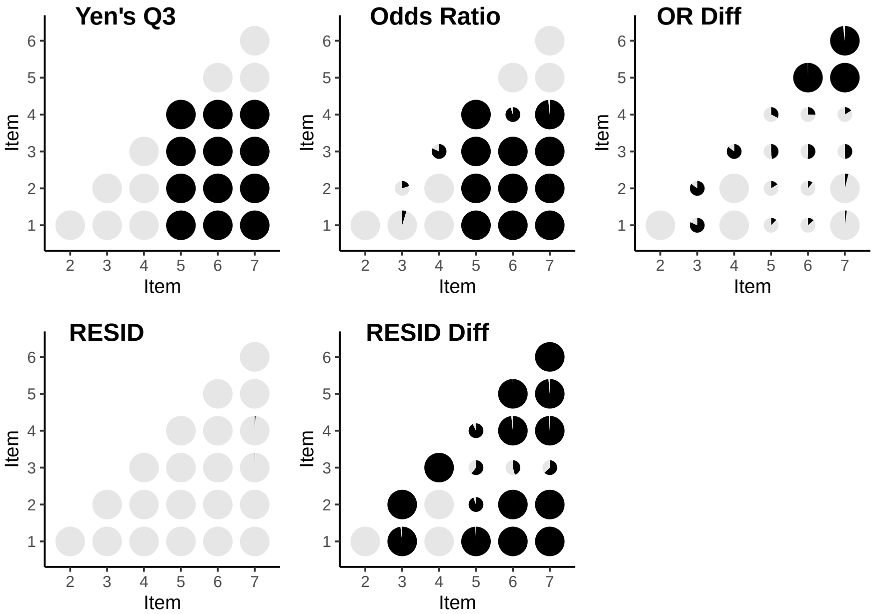
In our analysis of the items of negative affect and mental unrest, the test statistics and discrepancy measures showed strong evidence of model misfit. Table 5.9 presents the PPPs of some item-level measures for each item. Figure 5.4 presents the PPPs of the different pairwise measures. The results for Yen’s Q_3 showed the presence of two distinct dimensions underlying the observations—these dimensions correspond to negative affect and mental unrest, respectively. The results of the OR also point to the presence of these two dimensions. The PPPs of both the item LRACF and the RESID were 0 for all the possible items and pairs, respectively. This result, in combination with the model misfit detected by the item-total correlation, Yen’s Q_1, and the alternative Yen’s Q_1 for some of the items might indicate that the response process of the participant to, for example, items 1, 6, and 7 changed during the study. Therefore, in addition to multidimensionality, item parameters may have drifted or the interpretation of the individual may have changed for some of these items.

Table 5.9: PPPs Item-Level Measures Items of Negative Affect and Mental Unrest

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Item-Total Correlation	**1**	0.321	**0.006**	0.21	**0.975**	**0.006**	**0**
Yen's Q_1	**0.021**	0.133	0.158	**0.002**	**0.039**	0.064	**0.018**
Yen's Q_1 Alternative	**0.001**	**0.04**	0.338	0.067	0.064	**0**	0.064
Item LRACF	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Note. PPPs lower than 0.05 or larger than 0.95 are highlighted in boldface.

Figure 5.4: Pairwise Measures of the Items of Negative Affect and Mental Unrest



5.4.2 Assessing Model Fit of the Items of Positive Affect

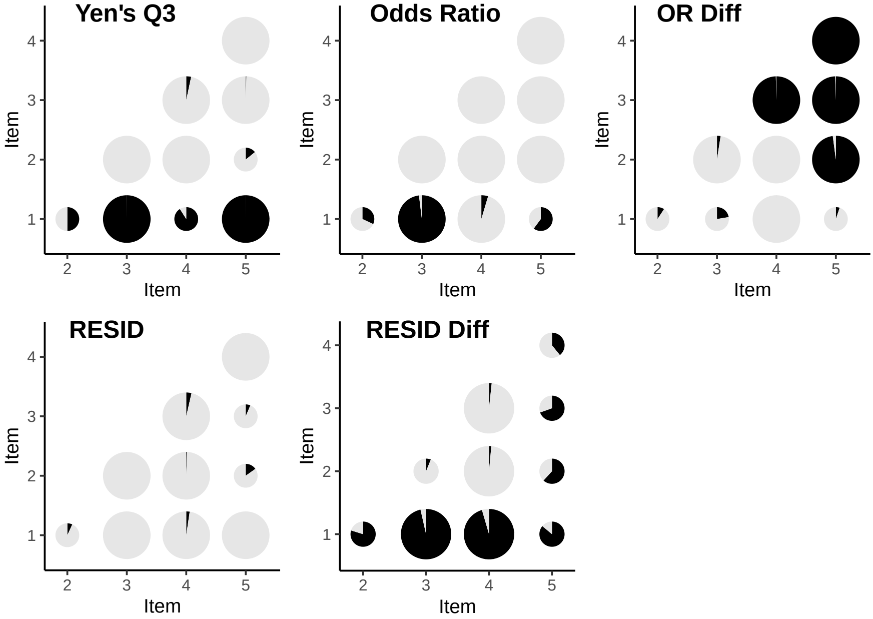
Table 5.10 and Figure 5.5 present the PPPs for the different item-level and pairwise measures, respectively, for the items of positive affect. All the item-level measures showed evidence of model misfit for items 1 ("relax") and 5 ("strong"), which implies that the meaning of these items for the participant may have changed during the study. Also, the fact that the PPPs of the item-total correlation were so extreme for all the items might suggest that a model that does not constraint the discrimination to be equal to 1 might be more appropriate for these data. Moreover, the pairwise measures also showed evidence of model misfit (see Figure 5.5). For example, Yen's Q_3 and the OR showed misfit, which may be interpreted as evidence of multidimensionality. Based on the patterns of the extreme PPPs of these two measures, it seems that items 2-5 were measuring a dimension that was different from the dimension measured by item 1.

Table 5.10: PPPs Item-Level Measures Items of Positive Affect

	Item 1	Item 2	Item 3	Item 4	Item 5
Item-Total Correlation	**1**	**0**	**0**	**0**	**0**
Yen's Q_1	**0.002**	0.07	**0.023**	**0**	**0.027**
Yen's Q_1 Alternative	**0.001**	0.516	0.264	0.162	**0.004**
Item LRACF	**0**	0.676	0.301	0.2	**0.009**

Note. PPPs lower than 0.05 or larger than 0.95 are highlighted in boldface.

Figure 5.5: Pairwise Measures of the Items of Positive Affect



5.4.3 Assessing Model Fit of the Items of Self-Esteem

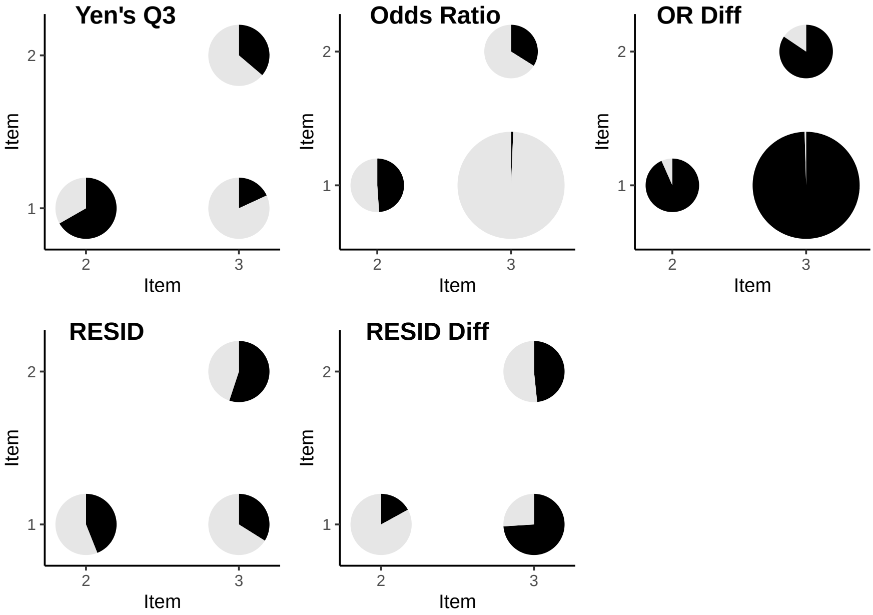
Table 5.11 and Figure 5.6 present the PPPs for the item-level and pairwise measures, respectively, for the items on self-esteem. In this analysis, the evidence of model misfit was scarce. Item 1 showed evidence of misfit according to the alternative Yen's Q_1, and the item pairs including items 1 and 3 showed evidence of misfit given the OR and the OR difference. However, the extent of model misfit does not appear to be severe for the items on self-esteem.

Table 5.11: PPPs Item-Level Measures Items of Self-Esteem

	Item 1	Item 2	Item 3
Item-Total Correlation	0.212	0.676	0.113
Yen's Q_1	0.115	0.164	0.461
Yen's Q_1 Alternative	**0.019**	0.537	0.149
Item LRACF	0.082	0.133	0.171

Note. PPPs lower than 0.05 or larger than 0.95 are highlighted in boldface.

Figure 5.6: Pairwise Measures of the Items of Self-Esteem



5.5 Discussion

Evaluating whether the statistical model is appropriate to explain the data is a key step in statistical modeling. This evaluation can be performed by assessing the goodness-of-fit of the model to the data. However, there is a lack of goodness-of-fit statistics for intensive longitudinal models. Within the context of Bayesian statistical analysis, a popular tool used to assess the goodness-of-fit of a model is the PPMC method. In this chapter, we proposed and studied several test statistics and discrepancy measures under the framework of the PPMC method to assess the goodness-of-fit of the TV-DPCM, which was recently developed by Castro-Alvarez et al. (2022a).

The proposed test statistics and discrepancy measures were inspired by previous applications of the PPMC method to traditional IRT models (e.g., Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011). We modified some of these measures to account for the time dependency likely to be present in intensive longitudinal data. We tested the performance of the different test statistics and discrepancy measures using simulated data. Our results show that there is no single best measure that is effective in detecting several sources of model misfit. Instead, we found that different measures were effective at detecting misfit under different conditions. These results were in agreement with results observed by Sinharay et al. (2006) and Zhu and Stone (2011) and support the recommendation that several test statistics and discrepancy measures should be used when assessing goodness-of-fit under the PPMC framework.

Given the recommendation to apply several measures to assess the goodness-of-fit of a model, researchers might consider making adjustments for multiple comparisons in applications of the PPMC method. However, Gelman et al. (2014) recommended against adjusting for multiple comparison and emphasized that the purpose of applying the PPMC method is not to test for a hypothesis but to understand the limitations of the model in explaining the data at hand.

Overall, we found the item-total correlation, the alternative Yen's Q_1, the LRACF per item, Yen's Q_3, and the OR as key measures for their effectiveness in detecting model misfit of the TV-DPCM. Their utility was also evident with the analyses of the items of positive and negative affect, mental unrest, and self-esteem from the empirical data (Kossakowski et al., 2017; Wichers & Groot, 2016). These test statistics and discrepancy measures can also be useful to assess the goodness-of-fit of other IRT models for intensive longitudinal data, as those proposed by Hecht et al. (2019) and Wang et al. (2013).

In the analyses of the empirical data, we showed two examples in which the TV-DPCM did not seem to fit the data. An important question after finding evidence of model misfit is “What is next?”. One of the challenges of using the PPMC method (or any other goodness-of-fit method) is that when misfit is detected, information is typically lacking on the reason of the misfit. Evidence of model misfit only indicates that certain aspects of the data cannot be replicated by simulated data (Gelman et al., 2014). Then, it is the job of the researcher to hypothesize about the reasons of model misfit and to decide whether they are able to adjust the model accordingly. Also, it is important to highlight that even when extreme PPPs are found, the model can still be useful for some purposes (Gelman et al., 2014). Therefore, one needs to find a balance between the amount of misfit that one is willing to tolerate and the practical utility of the model for the intended purpose. It is important to keep in mind that perfect model fit can be a sign of overfitting, resulting in a less useful model in practical terms (Pitt & Myung, 2002).

To conclude, we highlight the limitations of the current study. Starting with the simulation, we only used 10 replications per condition due to time constraints. Ideally, these test statistics and discrepancy measures should be tested in a simulation study with at least 100 replications per condition to collect clear evidence about their performance. Also, another limitation is due to the model of interest itself. The TV-DPCM requires more than 200 time points for one subject in order to obtain reliable estimates. This condition is rarely satisfied in the study of psychological dynamics, which can prevent the application of the TV-DPCM. In future research, one can consider extending the TV-DPCM to a multilevel framework to be able to analyze psychological dynamics of a sample of individuals. Such a model would be more relevant given that most of the studies in psychological dynamics collect time series from a sample. Also, given that we focused on the TV-DPCM, it is unclear if the test statistics and discrepancy measures proposed in this chapter would be useful to assess the goodness-of-fit of other IRT models for intensive longitudinal data. For example, considering the continuous-time Rasch model proposed by Hecht et al. (2019), most of the proposed measures could be useful if applied to each individual. However, further adjustments would be needed to account for all the individuals simultaneously. Overall, the PPMC method offers a flexible and easy to adapt tool to assess the goodness-of-fit of models implemented within the Bayesian framework.

We encourage researchers interested in psychological dynamics to keep exploring the implementation of the PPMC method to models used to study intensive longitudinal data.

Chapter 6

Discussion

The driving question of this thesis was "what are we measuring in psychological intensive longitudinal research"? This led us into two main topics: Distinguishing between traits and states, and accounting for measurement error. Firstly, distinguishing between traits and states is key in longitudinal research. This allows having a better understanding regarding what the measurements mean and how to better interpret the results based on solid conceptualizations. Secondly, we recognize that psychological measurement is unreliable by nature. Hence, measurements are always executed with some sort of error that needs to be accounted for. To address these topics, we started by studying the latent state trait theory (LST; Steyer et al., 2015; Steyer et al., 1999) and how it can be used for intensive longitudinal settings (see Chapters 2 and 3). Additionally, in Chapters 4 and 5, we also addressed the issue of measurement error in psychological time series by means of an item response theory approach (IRT; Embretson & Reise, 2013; Lord et al., 1968).

6.1 Main Findings

In Chapter 2, we started by studying several longitudinal structural equation models used to study traits and states. We focused on three particular models: The multistate-singletrait model (MSST; Steyer et al., 2015; Steyer et al., 1999), the common and unique trait state model (CUTS; Hamaker et al., 2017), and the trait-state-occasion model (TSO; Eid et al., 2017). From these models, the MSST and the TSO are encompassed within the LST theory (Steyer et al., 2015). We were interested in studying whether the selected models were useful to analyze intensive longitudinal data. To do this, we reformulated the models as multilevel structural equation models and tested the two versions of each model in a simulation study. In general, the multilevel versions of the selected models can be considered additional tools for the analysis of intensive longitudinal data, if one is interested in distinguishing between traits and states. Specially, the multilevel version of the TSO model is advised, as it accounts for the carry-over effects likely to be found in time series data and because it was the model that performed the best in the simulation study. Moreover, these models also offer an approach to study the reliability of the items and the scales used in intensive longitudinal research.

Another interesting result from Chapter 2 was that fitting the single-level version of the models to intensive longitudinal data in wide format was possible with many measurements occasions when using Bayesian estimation. This was unexpected because analyzing the data in wide format implies that the models need to deal with extremely large variance-covariance matrices, which usually results in convergence issues (Geiser et al., 2013). While the single-level version of the models allows for more flexibility because the parameters can be allowed to vary over time, we still recommend against using the single-level models when there are more that 30 measurement occasions. The reason for this is that fitting the single-level models with many measurement occasions is more time-consuming, does not necessarily address the research questions of interest in psychological dynamics, and comes short, as fit statistics for model comparison are lacking.

In Chapter 3, we further extended the TSO model as it was the most promising model for the analysis of intensive longitudinal data. The extension consisted of allowing the autoregressive effect to vary randomly across individuals, and allowing the inclusion of situational variables to account for fixed and random situations, as suggested in the LST theory with fixed and random situations approach (Geiser et al., 2015b). We referred to this new model as the mixed-effects trait-state-occasion model (ME-TSO). Also, by allowing the autoregressive effect to vary randomly across individuals, we also redefined all the variance coefficients within fixed situations such that they are computed per person. This provides a psychometric approach based on the LST theory with an emphasis on the individual level. In a nutshell, the ME-TSO is a model that researchers can use to study (a) the psychometric properties of the items used in intensive longitudinal research per person, (b) the interaction between the persons and the fixed situations of interest, and (c) the dynamic processes of the psychological constructs per person. The usefulness of the ME-TSO was illustrated by analyzing data from the HowNutsAreTheDutch project (van der Krieke et al., 2017; van der Krieke et al., 2016).

Next, in Chapter 4, we introduced the time-varying partial credit model (TV-DPCM), which is an IRT model for the analysis of psychological time series data. This model was specially proposed to analyze intensive longitudinal data from one individual when a set of Likert-scale items is used to measure one psychological construct such as negative affect. We tested the performance of the model in a simulation study while varying the number of time points, the number of items, the size of the autoregressive effect, and the proportion of missing data. Overall, the TV-DPCM performed well under the different conditions. We recommended using the model when there are at least 300 time points. The other manipulated factors did not have major effects on the performance of the TV-DPCM. However, the results showed that the credibility intervals of the different parameters of interest tend to be too wide, especially when there are a few number of items and when the proportion of missing data increases. Nevertheless, by means of the empirical example, where we analyzed the self-esteem items scores of one individual from an experience sampling study (Kossakowski et al., 2017; Wichers & Groot, 2016), the TV-DPCM proved to be a useful approach that provides a rich interpretation of the psychological dynamics and the scales used in intensive longitudinal research. Regarding the psychological dynamics, it allows modeling psychological dynamics described by non linear trends, and it provides information about the situations where the measurements are more or less reliable. In relation to the scales, by making use of core features of IRT, such as the item characteristic function, the item information function, and the test information function, one could identify which items are better and more informative to measure the construct of interest. Also, one can study the overall quality of the scale in order to make adjustments for future applications.

Nowadays, many approaches used to analyzed intensive longitudinal data have been implemented within the Bayesian framework (e.g., Asparouhov et al., 2018; De Haan-Rietdijk et al., 2016; de Haan-Rietdijk et al., 2017a; Hecht et al., 2019; Schuurman & Hamaker, 2019; Schuurman et al., 2015) and the TV-DPCM is not an exception. The advantage of the Bayesian approach is that it facilitates and allows the estimation of very complex models that are challenging to fit with traditional frequentist approaches. However, due to the novelty of these models, assessing their associated goodness-of-fit is not necessarily straightforward and has been overlooked in the literature. In Chapter 5, we were particularly concerned about how to assess the goodness-of-fit of the TV-DPCM. For this reason, we adapted several test statistics and discrepancy measures within the posterior predictive model checking method for the TV-DPCM (Gelman et al., 1996; Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011). These different statistics proved to be effective at identifying misfit under certain simulated conditions. For example, evidence of model misfit via Yen's Q_3 (Yen, 1984) statistic can be a clear indication of multidimensionality of the scale. Although we proposed these measures specifically to assess the goodness-of-fit of the TV-DPCM, they may be also useful to assess the goodness-of-fit of other Bayesian IRT models for intensive longitudinal data, such as the ones proposed by Hecht et al. (2019) and Wang et al. (2013). Furthermore, they can also serve as a basis for the application of the posterior predictive model checking method for other Bayesian approaches for the analysis of intensive longitudinal data, such as the dynamic structural equation modeling framework (Asparouhov et al., 2018).

6.2 Defining Traits and States

Throughout this dissertation, we discussed how the distinction between traits and states is not consistent in the literature (e.g., Allen & Potkay, 1981; Epstein, 1979; Hamaker, 2012; Spielberger & Sydeman, 1994; Steyer et al., 2015). This is problematic for studying psychological dynamics, given that these terms are recurrent in intensive longitudinal research and in many circumstances their definition is taken for granted. We addressed this problem in Chapters 2 and 3, based on the LST theory (Steyer et al., 2015; Steyer et al., 1999). The strengths of the LST theory are that (a) it provides a clear definition of traits and states supported on a probabilistic approach, and (b) it accounts for measurement error. Specifically, the LST theory acknowledges the dynamic nature of individuals, as it considers that the person at time t is not the same person at time $t+1$ due to the new situation and the experiences of the person between measurement occasions (Steyer et al., 2015). Based on this, the LST theory defines a set of random variables that represent the persons and the situations at a given time point, and the latent variables of the model are defined as conditional expectations. In this sense, the latent trait variables capture the effect of the person at each time point and the latent state residuals capture the effects of the situation and the person-situation interaction at each time point. Moreover, the combination of the latent trait variables and the latent state residual form the latent state variables, meaning that in the LST theory, the state of a person at time t is truly considered a combination of the effects of the person, the situation, and the person-situation interaction. Conceptually speaking, I consider that the LST theory approach should be adopted in the field of studying psychological dynamics, as it offers clear definitions of traits and states. We also demonstrated that models based on the LST theory can be developed to analyze intensive longitudinal data, and that they provide additional features that enrich the interpretation of the data.

Naturally, adopting the LST theory approach has consequences on the kind of research questions we ask and how we can approach them. Understanding persons' aspects as a mix of traits and states means that typical 'trait' measures found in cross-sectional research do not really represent a pure 'trait' but a combination of the trait and the state components of the aspect of interest. In other words, as discussed by Hamaker (2012), between- and within-person effects are entangled in cross-sectional research. As a consequence, if we want to distinguish between traits and states, we need to conduct longitudinal or intensive longitudinal research. In this sense, using scales that claim to measure either traits or states based on the wording of the items would not be appropriate. Furthermore, in several longitudinal studies, "trait" measures are included as covariates to explain between-person variability. I think that this is problematic, because adding this kind of "trait" covariates ignores that there are no pure traits or pure states. Therefore, from a theoretical point of view, adding these "trait" covariates is not coherent with the theory. Given this, something that needs to be discussed in future research is what is the meaning of the effect of a "trait" covariate on the psychological dynamic while acknowledging that such effect is a mix of between- and within-person variability.

6.3 Item Response Theory: An Underused Tool

IRT has been extensively developed and used within cross-sectional settings (Embretson & Reise, 2013). Among its advantages over classical test theory (CTT; Crocker & Algina, 1986), one can highlight that IRT models explicitly model the item response behavior of the individuals. As a result, they provide a set of parameters that characterize the properties of the items. Moreover, some properties of IRT models are that (a) unbiased item parameters may be estimated from unrepresentative samples, (b) standard measurement error is a function of the latent ability, meaning that it varies across response patterns, and (c) short tests may be more reliable than longer tests due to the previous property. However, although questionnaires with Likert-scale items are commonly used in intensive longitudinal research (Vachon et al., 2019), applications of the IRT approach to intensive longitudinal settings are rather scarce. Specifically, only a handful of IRT models have been proposed for the analysis of intensive longitudinal data (see Hecht et al., 2019; Kropko, 2013; Ram et al., 2005; Rijn et al., 2010; Vogelsmeier et al., 2020; Wang et al., 2013). In this thesis, in Chapter 4, we also proposed an IRT model, namely the TV-DPCM, to analyze psychological time series data while also emphasizing important features of IRT.

In our implementation of the TV-DPCM, we showed how the TV-DPCM can be used to make a comprehensive analysis and interpretation of the individual, the items, and their relationship in intensive longitudinal studies. This is precisely one of the main contributions of IRT and we consider that the field of psychological dynamics can benefit greatly from this kind of modeling. Nevertheless, there are two clear limitations of the TV-DPCM that can hinder its application. The first limitation is that the model was developed to analyze multivariate time series of one individual. The second limitation is that many time points are required to fit the model. To overcome these limitations, future research can explore extending the TV-DPCM to multilevel settings in order to be able to analyze multivariate time series of a sample of individuals, as it is similarly done by Hecht et al. (2019) and Vogelsmeier et al. (2020). Such an extension would also require the consideration of how to handle the distinction between traits and states within the TV-DPCM model.

6.3.1 A Bright Future

Overall, IRT modeling has many desirable properties for the analysis of intensive longitudinal data. In addition to allowing studying the properties and quality of the items, another important topics in IRT are for example the study of item and person misfit (Glas & Meijer, 2003; Meijer & Sijtsma, 2001; Orlando & Thissen, 2000), differential item functioning (Wang, 2008), item parameter drift (Donoghue & Isham, 1998), and computerized adaptive testing (CAT Meijer & Nering, 1999; Wainer et al., 2000), among others. Each of these topics has great value if adapted to the study of psychological dynamics. One first step to set the basis for these future developments is the development of IRT models suitable to analyze intensive longitudinal data such as the TV-DPCM.

In particular, I think there is a great potential in implementing CAT for the development and analysis of questionnaires in intensive longitudinal research. The aim of CAT is to construct an optimal test to assess each individual (Meijer & Nering, 1999). In short, in CAT, the latent ability of the subject is estimated and monitored while the test is being administered and the subsequent items are selected based on the current estimation of the latent ability. This can allow administering shorter questionnaires which is preferable to reduce the burden on the participants. It can also increase compliance rates given that every administration of the questionnaire will be potentially different from previous administrations, making the task of filling in the experience sampling questionnaire less monotone.

6.4 Stationarity and Measurement Invariance

Most of the statistical methods used to study psychological dynamics imply that psychological dynamic processes are stationary (e.g., Asparouhov et al., 2018; Hamaker, 2005; Schuurman et al., 2015; Song & Zhang, 2014), meaning that the means, the variances, and the autocorrelations of the dynamic processes are stable over time (Hamilton, 1994). While this assumption is convenient to have simpler statistical models, testing that this assumption is satisfied can be hard. Firstly, there are several tests such as the Kwiatkowski-Phillips-Schmidt-Shin test (Kwiatkowski et al., 1992), the augmented Dickey-Fuller test (Mushtaq, 2011), and the Phillips-Perron test (Phillips & Perron, 1988), which test different aspects of stationarity. Secondly, these tests tend to have low detection rates when analyzing short time series (Jönsson, 2011; Leybourne & Newbold, 1999). This is problematic given that the length of time series in the field of psychological dynamics is typically lower than 100 (Vachon et al., 2019). Moreover, the assumption of stationarity in itself can be unrealistic on the context of psychological research (Bringmann et al., 2017; De Haan-Rietdijk et al., 2016). For example, in many situations, researchers are interested on the effectiveness of a psychological treatment or the development of the participants, which are expected to be non-stationary processes by nature.

Furthermore, when using latent variable models as the ones presented in this thesis, one also needs to consider whether longitudinal measurement invariance holds (Meredith, 1993; Meredith & Teresi, 2006). This means that the scale measures the concept of interest equally over time. This assumption is necessary to make sure that the scores on a test on different time points are comparable. Yet, this assumption is rarely tested in intensive longitudinal studies. Besides the work on latent Markov latent trait models by Vogelsmeier et al. (2019a), Vogelsmeier et al. (2020), Vogelsmeier et al. (2019b) and on measurement in longitudinal data by McNeish et al. (2021), research on measurement invariance in psychological dynamics is scarce.

Another limitation in the literature of intensive longitudinal research is that the assumptions of stationarity and measurement invariance are studied as if they were independent. However, I consider that these two assumptions are in fact related, as both of these assumptions are concerned with the stability of the dynamic process. Therefore, when studying measurement models for intensive longitudinal data, one needs to address what are the implications of the assumptions of stationarity and measurement invariance for the model and the data at hand.

6.4.1 Handling Stationarity and Measurement Invariance

Both stationarity and measurement invariance imply that there is some stability present in psychological dynamic processes. These assumptions are often mentioned as limitations when studying psychological dynamics, as they may be too strict or unrealistic. Yet, at the same time, assuming stationarity and measurement invariance is helpful to keep the models simpler. We also highlighted these assumptions as limitations on Chapters 2 and 3, and even one of the aims of Chapter 4 was to handle non-stationary time series. From my research, I consider that these two properties are interconnected when studying measurement models for intensive longitudinal data. Nonetheless, more research is needed to understand the implications of these assumptions on each other. Some important questions that need to be addressed are: Does stationarity imply longitudinal measurement invariance when using measurement models for intensive longitudinal data?, can a dynamic process be stationary while violating the assumption of longitudinal measurement invariance or vice versa?, when assuming stationarity in measurement models for intensive longitudinal data, do we assume that the observed time series is stationary, or that the latent time series is stationary, or both?, if it is observed that the time series is stationary, does it imply that the latent time series is stationary as well?

6.5 Latent Variables: Are They Worth the Struggle?

Latent variables have been central to psychological measurement and lay at the foundations of CTT and IRT. In a nutshell, a latent variable aims to represent an unobserved theoretical construct that is a common cause of a set of observed indicators or items. In other words, there exists a causal relationship between the latent variable and the observed indicators. Moreover, in most of the cases, we assume that the shared variance among the indicators is captured by the latent variable sufficiently well. Overall, latent variable models imply several strong and restrictive assumptions that come into play when using this kind of models. In the context of intensive longitudinal research, one can argue that there are several limitations to the use of latent variables. One limitation is the requirement that multiple indicators need to be used to identify the latent variable. This is a practical limitation because questionnaires in intensive longitudinal research tend to be short in order to reduce the burden on the participants. Thus, having multiple items measuring the same construct may not be feasible in many circumstances. Furthermore, when measurement models are used, the rationale for using them is mostly based on the statistical justification rather than on the conceptual theory (Rhemtulla et al., 2020). This is problematic because perfect fitting models may still be wrong if they are not supported on solid theories. Also, by adding the time component, problems related to longitudinal measurement invariance also arise.

These issues can discourage researchers from using latent variable models in general. In fact, the latent variable approach have been criticized and complex network approaches has been suggested as an alternative conceptualization of psychological phenomena (Borsboom & Cramer, 2013; Cramer et al., 2012; Epskamp et al., 2018). Nevertheless, the distinction between latent variable models and network approaches is not clear cut (Bringmann & Eronen, 2018). Bringmann and Eronen (2018) argue that latent variable models are not inherently problematic. Furthermore, it has been shown that some latent variable models and some network models are mathematically equivalent (Epskamp et al., 2018; van Bork et al., 2021).

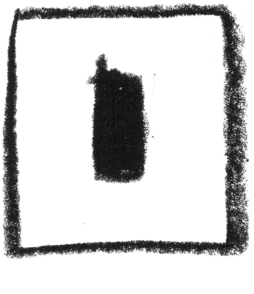
Lastly, in spite of the criticism towards latent variable models, I consider that latent variable models are here to stay, also in intensive longitudinal research. In most of the cases, the limitations that are mentioned in relation to the assumptions of latent variable models can be overcome by making the models more flexible. Also, another advantage is that latent variable models allow accounting for measurement error, which is of utter importance in psychological research. In addition to the models studied in this dissertation, several latent variable models have been proposed to study psychological dynamics (e.g., Asparouhov et al., 2018; Chow et al., 2011; Hecht et al., 2019; Molenaar, 1985; Schuurman & Hamaker, 2019; Schuurman et al., 2015; Song & Ferrer, 2012; Vogelsmeier et al., 2019b). In particular, some interesting latent variable models for the study of psychological dynamics were proposed by Schuurman and Hamaker (2019) and Schuurman et al. (2015). These models are autoregressive models with measurement error in which the latent variables are identified with just one indicator. This development is very useful, given that in intensive longitudinal research, it is harder to have multiple items measuring the same construct.

6.6 Beyond Measurement Models

To conclude, I want to emphasize that understanding what and how we are measuring in intensive longitudinal research goes beyond the use of measurement models.

As shown in this thesis, measurement models aid on understanding the strengths and limitations of the questionnaires used in intensive longitudinal research, but they do not fix poor quality measurements. Therefore, complementary to the use of measurement models, qualitative research is also needed to understand the response processes of the individuals (e.g., why they choose a certain response option) and which type of items are better suited for self-report in intensive longitudinal research (e.g., Likert versus visual analogue scales). Naturally, research about measurement models and the development of questionnaires need to go 'hand-by-hand', where results on one area are used for the improvement of the other. One idea to get more information about persons' response processes would be to ask participants to explain the reasoning of their responses to a couple of key items every now and then during an experience sampling study. This can be helpful to monitor how the participant interprets the item and to assess whether the interpretation is stable or changes over time. Although much more research is needed, the methods presented in this thesis offer a solid basis to understanding how psychological measurement is done in intensive longitudinal research. In this sense, in the future, by assessing the scales that are currently used, researchers might be able to improve how we measure psychological processes over time.

Appendix A



In this appendix of Chapter 2, we intend to show how to easily analyze data with the MSST, CUTS, and TSO models in Mplus from within the R environment. Clearly, both Mplus and R should be installed. We also suggest to install RStudio.

This material includes the following sections:

- Prepare enviroment, which mentions which R packages are needed for the analyses and how to load the functions stored in the git repository associated to this article.
- Multistate-singletrait model (MSST).
- Common-unique trait-state model (CUTS).
- Trait-state-occasion model (TSO).

Moreover, the sections of each model include three subsections: (a) Generating data, (b)fitting the single level model, and (c) fitting the multilevel model. The subsection generating data shows how to simulate data based on each model. This subsection can be skipped, as their main purpose is to generate data to use in the following subsections. Own data should be load into R through the function `read.table` or similar functions from the foreign package. Moreover, the function `reshape` can be used to change the data from wide to long format or vice versa. Finally, the subsections fitting the single level model and fitting the multilevel model show how to fit both versions of each model to the data by means of bayesian estimation in Mplus.

A.1 Prepare environment

To prepare the environment for the next analyses three packages have to be loaded:

- RCurl, which is needed to load the functions from the git repository.
- MASS, which is needed for some of the written functions.
- MplusAutomation, which allows using Mplus within R.

Next, the functions developed within this study have to be loaded into R. These functions allow generating data and writing the Mplus syntaxes based on each model. There are two ways to load these functions. Firstly, you can download the files in your working directory and use the `source` function to load them in R. Alternatively, you can run the following code, which loads all the functions directly from the git repository without saving the files in your hard drive:

```
rfiles <- c("sim.data.cuts.R",
            "sim.data.msst.R",
            "sim.data.tso.R",
            "var.coeff.R",
            "write.cuts.to.Mplus.R",
            "write.mlcuts.to.Mplus.R",
            "write.msst.to.Mplus.R",
            "write.mlmsst.to.Mplus.R",
            "write.tso.to.Mplus.R",
            "write.mltso.to.Mplus.R",
            "write.Mplus.options.R")
for( i in 1:length(rfiles)){
eval(
 parse(
  text =
   getURL(
    paste0(
     "https://raw.githubusercontent.com/secastroal/LST_Analyses/master/R/",
       rfiles[i]), ssl.verifypeer = FALSE)))
}
rm(i, rfiles)
```

A.2 Multistate-singletrait model (MSST)

We will be generating data from the MSST model for 100 individuals, 3 items, and 3 measurement occasions. Two graphical representations of the model are shown in Figures A.1 and A.2 for the single level MSST and the multilevel MSST, respectively.

Figure A.1: Single level MSST

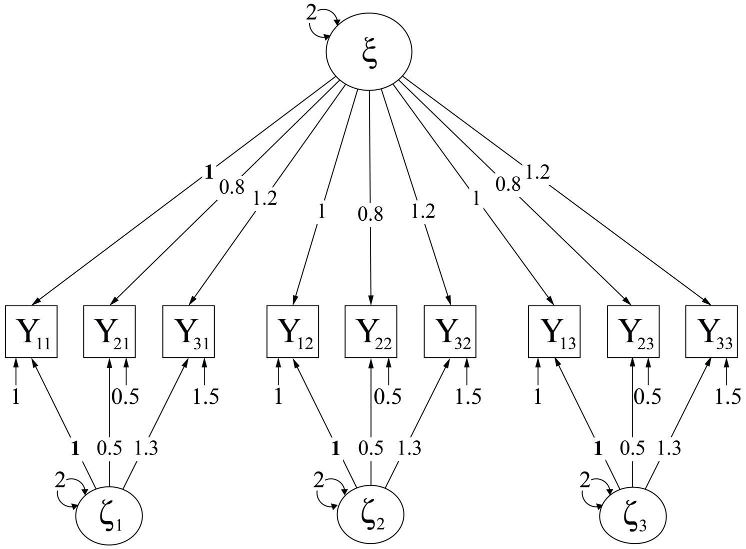
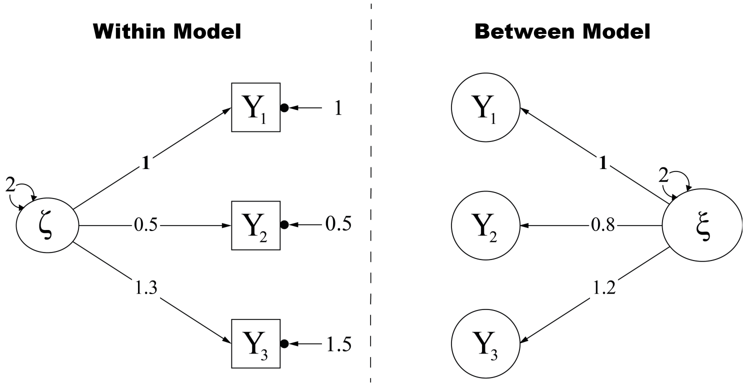


Figure A.2: Multilevel MSST



A.2.1 Generating MSST data

To generate data based on the MSST, use the function `sim.data.msst`. This function requires the number of persons, the number of items, the number of measurement occasions, and two lists that contain the within and the between parameters. Let us define these parameters as follows:

```
N  <- 100 # Number of persons
I  <- 3   # Number of items
nT <- 3   # Number of measurement occasions

# Within Parameters

loadings.state <- c(1, 0.5, 1.3) # Loading parameters for the latent state
# variables.
error.var      <- c(1, 0.5, 1.5) # Variance of measurement errors for each
# item.
state.var      <- 2              # Variance of the latent state residual.

within.parameters <- list(loadings  = loadings.state,
                          state.var = state.var,
                          error.var = error.var)

# Between Paramaters

loadings.trait <- c(1, 0.8, 1.2) # Loading parameters for the latent trait
# variable.
intercepts     <- rep(0, I)      # Intercepts.
trait.var      <- 2              # variance of the latent trait variable.
trait.mean     <- 4              # Mean of the latent trait variable.

between.parameters <- list(loadings   = loadings.trait,
                           intercepts = intercepts,
                           trait.mean = trait.mean,
                           trait.var  = trait.mean)
```

Next, once all the parameters are defined, we can simulate data based on the model:

```
data <- sim.data.msst(N, nT, I,
                      within.parameters  = within.parameters,
                      between.parameters = between.parameters)
```

The last code will store the simulated data in the object `data`. This object returns a list that includes the within and the between parameters, and the simulated data in wide and long format. These are the first five rows of the simulated data in wide and long format:

Table A.1: First 5 Rows of the MSST Simulated Data in Wide Format

	subjn	y11	y21	y31	y12	y22	y32	y13	y23	y33
1	1	1.80	1.29	0.93	2.07	1.40	2.65	1.90	1.64	2.98
4	2	3.02	2.67	3.45	2.86	1.19	-0.62	1.83	3.22	5.44
7	3	5.66	5.34	7.40	5.51	3.88	8.43	6.04	5.59	8.68
10	4	5.67	4.49	6.11	3.82	3.66	7.34	5.27	3.59	9.56
13	5	2.62	2.20	0.64	4.06	3.31	5.58	4.58	4.79	5.81

Table A.2: First 5 Rows of the MSST Simulated Data in Long Format

subjn	time	y1	y2	y3
1	1	1.80	1.29	0.93
1	2	2.07	1.40	2.65
1	3	1.90	1.64	2.98
2	1	3.02	2.67	3.45
2	2	2.86	1.19	-0.62

A.2.2 Fitting the single level MSST model

To fit the different models, we will follow basically the same procedure with small differences. Firstly, we will write the whole syntax of the model in R and export it to Mplus. Secondly, we will run the model in Mplus by using the function `runModels` and we will read the output file in R with the function `readModels`. Finally, we will take the output in R to extract the estimated parameters to be able to compute the variance coefficients.

To create the syntax in R for the single level MSST model the next functions are needed:

- `prepareMplusData`, which exports the data to be compatible with Mplus and creates a basic input file.
- `write.Mplus.options`, which allows specifying additional options for the analysis such as the estimation method and the number of iterations. This function does not include all the options available in Mplus.
- `write.msst.to.Mplus`, which creates a basic syntax to fit the single level MSST model in Mplus.

Prepare the data and the syntax files for Mplus

Here we export the data to Mplus in the file **slmsst.dat** and create the syntax file **slmsst.inp**:

```
# Write data in Mplus format and write input file template to the working
# directory:
prepareMplusData(data$data.wide, paste0(getwd(), "/slmsst.dat"),
                 inpfile = TRUE)

# Write additional options:
options_syntax <- write.Mplus.options(
                    usevariables  = names(data$data.wide)[-1],
                    analysis_type = "GENERAL",
                    estimator     = "BAYES",
                    iterations    = 5000)

# Write Mplus syntax of the single level MSST:
analysis_syntax <- write.msst.to.Mplus(
  data$data.wide[, -1],
  neta                        = nT,
  ntheta                      = 1,
  equiv.assumption            = list(tau = "cong", theta = "cong"),
  scale.invariance            = list(lait0 = TRUE, lait1 = TRUE,
                                     lat0 = TRUE, lat1 = TRUE),
  homocedasticity.assumption = list(error = TRUE, state.red = TRUE),
  second.order.trait          = FALSE)

# Additional options important for Bayesian analyses to save MCMC samples and
# Rhat statistics:
output_syntax <- "\nSAVEDATA: BPARAMETERS=samples_slmsst.dat;\nOUTPUT: TECH8;"

# Overwrite the basic syntax file to include additional options and the
#  syntax of the model:
write(options_syntax,  paste0(getwd(), "/slmsst.inp"), append = TRUE)
write(analysis_syntax, paste0(getwd(), "/slmsst.inp"), append = TRUE)
write(output_syntax,   paste0(getwd(), "/slmsst.inp"), append = TRUE)
```

Run the analysis in Mplus and import the results in R

Next, we can run the analysis in Mplus with the following code:

```
runModels(paste0(getwd(), "/slmsst.inp"))
```

This can take a while, especially if the estimation method is “BAYES”. By default, the function `write.Mplus.options` uses 4 processors, 4 chains, and 10 thinning. Once the analysis is done, the results, which are in the file **slmsst.out**, can be read in R as follows:

```
fit <- readModels(paste0(getwd(), "/slmsst.out"))
```

This object (`fit`) is a list that stores all the information available in the output file from Mplus. To access the estimated parameters you can type:

```
fit$parameters$unstandardized
```

This is a data frame that includes the estimated parameters, the posterior standard deviations, and the credibility intervals. The first five rows will look like this:

Table A.3: Mplus Estimates of the Single Level MSST Model

paramHeader	param	est	posterior_sd	pval	lower_2.5ci	upper_2.5ci	sig
ETA1.BY	Y11	1.00	0.00	0	1.00	1.00	FALSE
ETA1.BY	Y21	0.60	0.05	0	0.50	0.71	TRUE
ETA1.BY	Y31	1.44	0.12	0	1.23	1.71	TRUE
ETA2.BY	Y12	1.00	0.00	0	1.00	1.00	FALSE
ETA2.BY	Y22	0.60	0.05	0	0.50	0.71	TRUE

Compute variance coefficients

The variance coefficients defined in the LST theory are the most important output of these models because they provide the information about the psychometric properties of the items. To estimate the variance coefficients of the MSST model, we will use the function `msst.var.coeff`. For this function, we need the within and the between estimated parameters in lists as it was done in Section A.2.1. Doing this requires some little coding, but once this coding is done, the variance coefficients can be computed like this:

```
msst.var.coeff(within.parameters  = within.estimates,
               between.parameters = between.estimates)
```

Table A.4: Variance Coefficients of the Single Level MSST Model

Coefficent	Item 1	Item 2	Item 3
Consistency	0.49	0.63	0.44
Occasion-Specificity	0.33	0.20	0.42
Reliability	0.82	0.83	0.86

A.2.3 Fitting the multilevel MSST model

Prepare the data and the syntax files for Mplus

To fit the multilevel MSST model, we use the function `write.mlmsst.to.Mplus` instead of `write.msst.to.Mplus`. In the following code, we export the data to Mplus in the file **mlmsst.dat** and we create the syntax file **mlmsst.inp**:

```
# Write data in Mplus format and write input file template to the working
# directory:
prepareMplusData(data$data.long, paste0(getwd(), "/mlmsst.dat"),
                 inpfile = TRUE)

# Write additional options:
options_syntax <- write.Mplus.options(
  usevariables  = names(data$data.long)[-(1:2)],
  cluster       = names(data$data.long)[1],
  analysis_type = "TWOLEVEL",
  estimator     = "BAYES",
  iterations    = 5000)

# Write Mplus syntax of the multilevel MSST:
analysis_syntax <- write.mlmsst.to.Mplus(data$data.long[, -(1:2)])

# The previous syntax has some variances constrained to 0, which is
# undesirable when doing bayesian estimation. Hence, those constraints
# are changed to 0.001.
analysis_syntax <- gsub("@0;", "@0.001;", analysis_syntax)


# Additional options important for Bayesian analyses to save MCMC samples and
# Rhat statistics:
output_syntax <- "\nSAVEDATA: BPARAMETERS=samples_mlmsst.dat;\nOUTPUT: TECH8;"

# Overwrite the basic syntax file to include additional options and the
# syntax of the model:
write(options_syntax,  paste0(getwd(), "/mlmsst.inp"), append = TRUE)
write(analysis_syntax, paste0(getwd(), "/mlmsst.inp"), append = TRUE)
write(output_syntax,   paste0(getwd(), "/mlmsst.inp"), append = TRUE)
```

Run the analysis in Mplus and import the results in R

The analysis is run in Mplus with the following code:

```
runModels(paste0(getwd(), "/mlmsst.inp"))
```

Once the analysis is done, we import the results in R from the file **mlmsst.out** as follows:

```
fit <- readModels(paste0(getwd(), "/mlmsst.out"))
```

The estimates can be found in:

```
fit$parameters$unstandardized
```

The first five rows of this data frame look like this:

Table A.5: Mplus Estimates of the Multilevel MSST Model

paramHeader	param	est	posterior_sd	pval	lower_2.5ci	upper_2.5ci	sig	BetweenWithin
ETA.BY	Y1	1.00	0.00	0	1.00	1.00	FALSE	Within
ETA.BY	Y2	0.60	0.05	0	0.50	0.71	TRUE	Within
ETA.BY	Y3	1.44	0.12	0	1.22	1.71	TRUE	Within
Variances	ETA	1.86	0.29	0	1.37	2.50	TRUE	Within
Residual.Variances	Y1	1.02	0.14	0	0.76	1.31	TRUE	Within

Compute variance coefficients

To estimate the variance coefficients of the multilevel MSST model, we need the same function `msst.var.coeff`. Hence, we will need to extract the estimates from `fit` and store them in two lists. When this is ready, the variance coefficients van be computed like this:

```
msst.var.coeff(within.parameters  = within.estimates,
               between.parameters = between.estimates)
```

Table A.6: Variance Coefficients of the Multilevel MSST Model

Coefficient	Item 1	Item 2	Item 3
Consistency	0.50	0.63	0.45
Occasion-Specificity	0.33	0.20	0.41
Reliability	0.82	0.83	0.86

A.3 Common-unique trait-state (CUTS)

We will generate data from the CUTS model for 100 individuals, 3 items, and 3 measurement occasions. The graphical representations of the single level CUTS and multilevel CUTS are shown in Figures A.3 and A.4, respectively.

Figure A.3: Single level CUTS

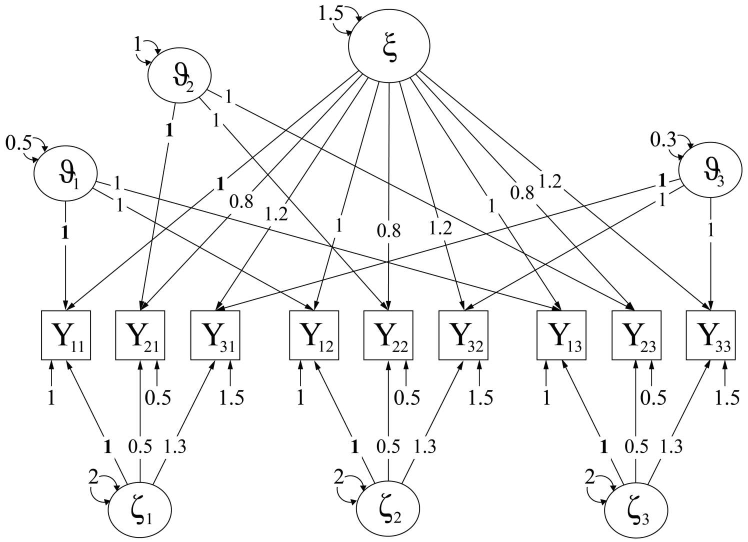
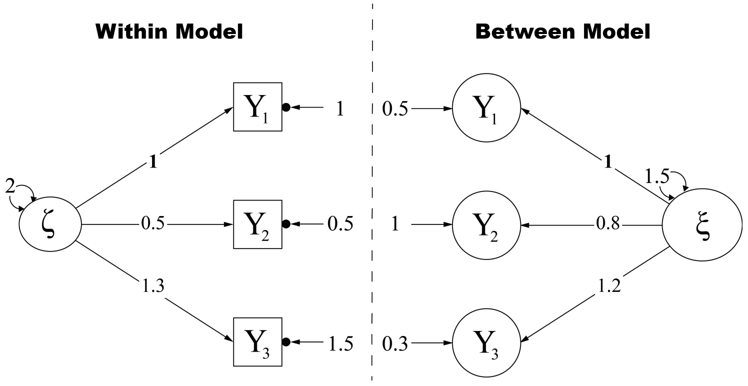


Figure A.4: Multilevel CUTS



A.3.1 Generating CUTS data

To generate data based on the CUTS, we need the function `sim.data.cuts`. Let us define the parameters to simulate the data as follows:

```
N  <- 100 # Number of persons
I  <- 3   # Number of items
nT <- 3   # Number of measurement occasions

# Within Parameters

loadings.state <- c(1, 0.5, 1.3) # Loading parameters for the latent common
# state.
US.var         <- c(1, 0.5, 1.5) # Variance of the latent unique states.
CS.var         <- 2              # Variance of the latent common state.

within.parameters <- list(loadings = loadings.state,
                          CS.var   = CS.var,
                          US.var   = US.var)

# Between Paramaters

loadings.trait <- c(1, 0.8, 1.2) # Loading parametes for the latent common
# trait.
intercepts     <- rep(0,I)       # Intercepts.
UT.var         <- c(0.5, 1, 0.3) # Variance of the latent unique traits.
CT.var         <- 1.5            # Variance of the latent common trait.

between.parameters <- list(loadings   = loadings.trait,
                           intercepts = intercepts,
                           CT.var     = CT.var,
                           UT.var     = UT.var)
```

Next, we can simulate data based on the model like this:

```
data <- sim.data.cuts(N, nT, I,
                      within.parameters  = within.parameters,
                      between.parameters = between.parameters)
```

This function returns the simulated data in both wide and long format. These are the first five rows of these data:

Table A.7: First 5 Rows of the CUTS Simulated Data in Wide Format

	subjn	y11	y21	y31	y12	y22	y32	y13	y23	y33
1	1	-2.02	0.89	-2.27	-4.38	0.09	-3.76	-2.78	1.19	-3.79
4	2	-0.02	0.97	-0.37	-0.52	-0.24	1.00	-1.84	0.66	-2.99
7	3	1.53	1.19	-0.36	1.90	3.51	3.30	1.20	1.37	-0.37
10	4	0.91	0.46	-0.06	1.60	2.94	2.08	0.01	2.80	-0.60
13	5	0.32	-0.69	2.03	-0.77	-0.04	-1.19	0.34	-0.73	-1.40

Table A.8: First 5 Rows of the CUTS Simulated Data in Wide Format

subjn	time	y1	y2	y3
1	1	-2.02	0.89	-2.27
1	2	-4.38	0.09	-3.76
1	3	-2.78	1.19	-3.79
2	1	-0.02	0.97	-0.37
2	2	-0.52	-0.24	1.00

A.3.2 Fitting the single level CUTS model

Prepare the data and the syntax files for Mplus

To fit the single level CUTS model to the data, we will create the syntax for this model with the function `write.cuts.to.Mplus`. Through the next code we export the data to Mplus in the file **slmsst.dat** and create the syntax file **slmsst.inp**:

```
# Write data in Mplus format and write input file template to the working
# directory:
prepareMplusData(data$data.wide, paste0(getwd(), "/slcuts.dat"),
                 inpfile = TRUE)

# Write additional options:
options_syntax <- write.Mplus.options(
                 usevariables  = names(data$data.wide)[-1],
                 analysis_type = "GENERAL",
                 estimator     = "BAYES",
                 iterations    = 5000)
```

```
# Write Mplus syntax of the single level MSST:
analysis_syntax <- write.cuts.to.Mplus(
  data$data.wide[,-1],
  nstate                       = nT,
  method.trait                 = "om",
  scale.invariance             = list(int = TRUE, lambda = TRUE),
  state.trait.invariance       = FALSE,
  fixed.method.loadings        = TRUE,
  homocedasticity.assumption = list(error = TRUE,
                                       cs.red = TRUE,
                                       ut.red = FALSE))


# Additional options important for Bayesian analyses to save MCMC samples and
# Rhat statistics:
output_syntax <- "\nSAVEDATA: BPARAMETERS=samples_slcuts.dat;\nOUTPUT: TECH8;"

# Overwrite the basic syntax file to include additional options and the
# syntax of the model:
write(options_syntax,  paste0(getwd(), "/slcuts.inp"), append = TRUE)
write(analysis_syntax, paste0(getwd(), "/slcuts.inp"), append = TRUE)
write(output_syntax,   paste0(getwd(), "/slcuts.inp"), append = TRUE)
```

Run the analysis in Mplus and import the results in R

Next, run the analysis in Mplus:

```
runModels(paste0(getwd(), "/slcuts.inp"))
```

The results from this analysis are stored in the file **slcuts.out**, which can be read in R as follows:

```
fit <- readModels(paste0(getwd(), "/slcuts.out"))
```

To access the estimated parameters, type:

```
fit$parameters$unstandardized
```

From which the first five rows look like this:

Table A.9: Mplus Estimates of the Single Level CUTS Model

paramHeader	param	est	posterior_sd	pval	lower_2.5ci	upper_2.5ci	sig
CS1.BY	Y11	1.00	0.00	0	1.00	1.00	FALSE
CS1.BY	Y21	0.50	0.05	0	0.40	0.61	TRUE
CS1.BY	Y31	1.29	0.13	0	1.06	1.56	TRUE
CS2.BY	Y12	1.00	0.00	0	1.00	1.00	FALSE
CS2.BY	Y22	0.50	0.05	0	0.40	0.61	TRUE

Compute variance coefficients

For the single level and multilevel CUTS model, we use the function `cuts.var.coeff` to compute the variance coefficients as follows:

```
cuts.var.coeff(within.parameters  = within.estimates,
               between.parameters = between.estimates)
```

Table A.10: Variance Coefficients of the Single Level CUTS Model

Coefficient	Item 1	Item 2	Item 3
Common Consistency	0.16	0.19	0.17
Unique Consistency	0.12	0.37	0.04
Total Consistency	0.28	0.56	0.21
Occasion-Specificity	0.47	0.22	0.56
Reliability	0.75	0.78	0.77

A.3.3 Fitting the multilevel CUTS model

Prepare the data and the syntax files for Mplus

To fit the multilevel CUTS model, use the function `write.mlcuts.to.Mplus`. The next code exports the data to Mplus in the file **mlcuts.dat** and creates the syntax file **mlcuts.inp**:

```
# Write data in Mplus format and write input file template to the working
# directory:
prepareMplusData(data$data.long, paste0(getwd(), "/mlcuts.dat"),
                inpfile = TRUE)
```

```
# Write additional options:
options_syntax <- write.Mplus.options(
  usevariables  = names(data$data.long)[-(1:2)],
  cluster       = names(data$data.long)[1],
  analysis_type = "TWOLEVEL",
  estimator     = "BAYES",
  iterations    = 5000)

# Write Mplus syntax of the multilevel MSST:
analysis_syntax <- write.mlcuts.to.Mplus(data$data.long[, -(1:2)])

# Additional options important for Bayesian analyses to save MCMC samples and
# Rhat statistics:
output_syntax <- "\nSAVEDATA: BPARAMETERS=samples_mlcuts.dat;\nOUTPUT: TECH8;"

# Overwrite the basic syntax file to include additional options and the
# syntax of the model:
write(options_syntax,  paste0(getwd(), "/mlcuts.inp"), append = TRUE)
write(analysis_syntax, paste0(getwd(), "/mlcuts.inp"), append = TRUE)
write(output_syntax,   paste0(getwd(), "/mlcuts.inp"), append = TRUE)
```

Run the analysis in Mplus and import the results in R

Next, run the analysis in Mplus:

```
runModels(paste0(getwd(), "/mlcuts.inp"))
```

Finally, read the output file **mlcuts.out** in R as follows:

```
fit <- readModels(paste0(getwd(), "/mlcuts.out"))
```

Look at the first five estimated as follows:

Table A.11: Mplus Estimates of the Multilevel CUTS Model

paramHeader	param	est	posterior_sd	pval	lower_2.5ci	upper_2.5ci	sig	BetweenWithin
CS.BY	Y1	1.00	0.00	0	1.00	1.00	FALSE	Within
CS.BY	Y2	0.50	0.05	0	0.40	0.61	TRUE	Within
CS.BY	Y3	1.29	0.13	0	1.06	1.57	TRUE	Within
Variances	CS	2.08	0.39	0	1.45	2.98	TRUE	Within
Residual.Variances	Y1	1.06	0.20	0	0.71	1.49	TRUE	Within

Compute variance coefficients

Now, the function `cuts.var.coeff` is used to compute the variance coefficients:

```
cuts.var.coeff(within.parameters  = within.estimates,
               between.parameters = between.estimates)
```

Table A.12: Variance Coefficients of the Multilevel CUTS Model

Coefficient	Item 1	Item 2	Item 3
Common Consistency	0.14	0.18	0.16
Unique Consistency	0.13	0.37	0.04
Total Consistency	0.27	0.55	0.20
Occasion-Specificity	0.48	0.23	0.57
Reliability	0.75	0.78	0.77

A.4 Trait-state-occasion model (TSO)

Finally, we will generate data from the TSO model for 100 individuals, 3 items, and 3 measurement occasions. The graphical representations of the single level and multilevel TSO model are shown in Figures A.5 and A.6.

Figure A.5: Single level TSO

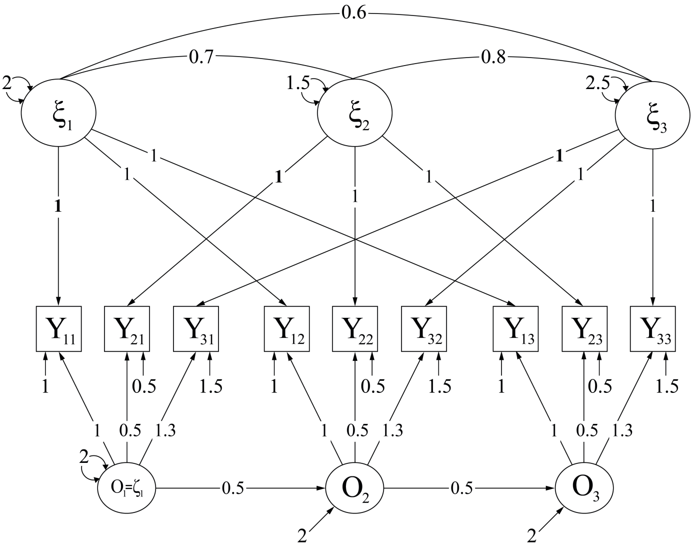
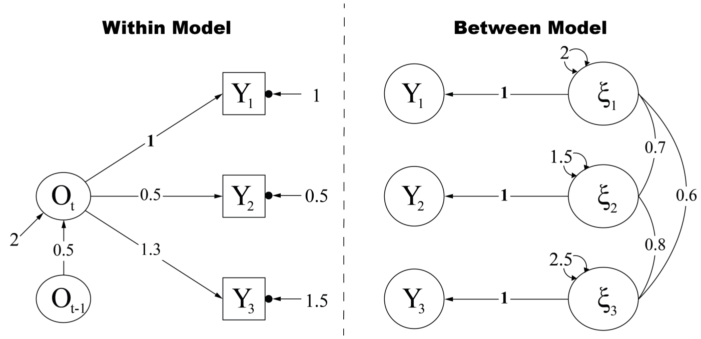


Figure A.6: Multilevel TSO



A

A.4.1 Generating TSO data

Here, we define the parameters to simulate data based on the TSO model:

```
N  <- 100 # Number of persons
I  <- 3   # Number of items
nT <- 3   # Number of measurement occasions

# Wihtin Parameters

loadings.state <- c(1, 0.5, 1.3) # Loading parameters for the latent
                                 # occasion specific variable.
error.var      <- c(1, 0.5, 1.5) # Variance of the latent measurement errors.
state.var      <- 2              # Variance of the latent state residual
# variable.
ar.effect      <- 0.5            # Autoregressive effect.

within.parameters <- list(loadings  = loadings.state,
                          ar.effect = ar.effect,
                          error.var = error.var,
                          state.var = state.var)


# Between Paramaters

trait.ind.var <- c(2, 1.5, 2.5) # Variance of the latent indicator trait
# variables.
intercepts    <- rep(0, I)      # Intercepts.
cor.matrix    <- matrix(c(1.0, 0.7, 0.6,
                          0.7, 1.0, 0.8,
                          0.6, 0.8, 1.0), 3) # Correlation matrix of the latent
                                             # indicator trait variables.



between.parameters <- list(intercepts    = intercepts,
                           trait.ind.var = trait.ind.var,
                           cor.matrix    = cor.matrix)
```

Next, the function `sim.data.tso` is used to generate the data:

```
data <- sim.data.tso(N, nT, I,
                     within.parameters  = within.parameters,
                     between.parameters = between.parameters)
```

This function also returns the data in both wide and long format. The first five rows are like this:

Table A.13: First 5 Rows of the TSO Simulated Data in Wide Format

	subjn	y11	y21	y31	y12	y22	y32	y13	y23	y33
1	1	0.52	-1.61	1.22	-0.88	-1.40	0.06	-0.35	-0.32	-1.06
4	2	-0.43	-1.49	-1.79	-3.07	-2.22	-2.71	-2.50	-1.08	-3.54
7	3	-3.44	-2.06	-2.88	-4.33	-2.40	-5.63	-0.55	-1.93	-0.45
10	4	-3.27	0.10	-2.62	-0.93	-0.80	-0.36	1.56	-0.97	0.57
13	5	-0.76	0.58	0.21	-1.66	-1.36	1.59	-2.47	-0.67	-1.84

Table A.14: First 5 Rows of the TSO Simulated Data in Long Format

subjn	time	y1	y2	y3
1	1	0.52	-1.61	1.22
1	2	-0.88	-1.40	0.06
1	3	-0.35	-0.32	-1.06
2	1	-0.43	-1.49	-1.79
2	2	-3.07	-2.22	-2.71

A.4.2 Fitting the single level TSO model

Prepare the data and the syntax files for Mplus

To prepare the files to fit the single level TSO model, we will use the function `write.tso.to.Mplus`. The following code will export the data to Mplus in the file **sltso.dat** and create the syntax file **sltso.inp**:

```
# Write data in Mplus format and write input file template to the working
# directory:
prepareMplusData(data$data.wide, paste0(getwd(), "/sltso.dat"),
                 inpfile = TRUE)

# Write additional options:
options_syntax <- write.Mplus.options(
                 usevariables  = names(data$data.wide)[-1],
                 analysis_type = "GENERAL",
                    estimator     = "BAYES",
                    iterations    = 5000)

# Write Mplus syntax of the single level MSST:
analysis_syntax <- write.tso.to.Mplus(
  data$data.wide[,-1],
  nocc                         = nT,
  figure                       = "3b",
  equiv.assumption             = list(occ = "cong", theta = "equi"),
  scale.invariance             = list(int = TRUE, lambda = TRUE),
  homocedasticity.assumption = list(error = TRUE, occ.red = TRUE),
  autoregressive.homogeneity = TRUE)

# Additional options important for Bayesian analyses to save MCMC samples and
# Rhat statistics:
output_syntax <- "\nSAVEDATA: BPARAMETERS=samples_sltso.dat;\nOUTPUT: TECH8;"

# Overwrite the basic syntax file to include additional options and the
# syntax of the model:
write(options_syntax,  paste0(getwd(), "/sltso.inp"), append = TRUE)
write(analysis_syntax, paste0(getwd(), "/sltso.inp"), append = TRUE)
write(output_syntax,   paste0(getwd(), "/sltso.inp"), append = TRUE)
```

Run the analysis in Mplus and import the results in R

Run the analysis in Mplus:

```
runModels(paste0(getwd(), "/sltso.inp"))
```

Read the results from the outfile **sltso.out** in R as follows:

```
fit <- readModels(paste0(getwd(), "/sltso.out"))
```

We can access the first five rows of the estimated parameters like this:

Compute variance coefficients

Now, compute the variance coefficients of the TSO model by using the function `tso.var.coeff`. In the case of the TSO, these variance coefficients will vary over time due to the autoregressive effect. The function `tso.var.coeff` computes these variance coefficients for each occasion like this:

Table A.15: Mplus Estimates of the Single Level TSO Model

paramHeader	param	est	posterior_sd	pval	lower_2.5ci	upper_2.5ci	sig
OCC1.BY	Y11	1.00	0.00	0	1.00	1.00	FALSE
OCC1.BY	Y21	0.55	0.07	0	0.43	0.71	TRUE
OCC1.BY	Y31	1.16	0.14	0	0.90	1.48	TRUE
OCC2.BY	Y12	1.00	0.00	0	1.00	1.00	FALSE
OCC2.BY	Y22	0.55	0.07	0	0.43	0.71	TRUE

```
tso.var.coeff(I = I, nT = nT,
              within.parameters = within.estimates,
              between.parameters = between.estimates)
```

Table A.16: Variance Coefficients of the Single Level TSO Model

Coefficient	Time 1	Time 2	Time 3
Predictability by trait Item 1	0.40	0.34	0.32
Predictability by trait Item 2	0.56	0.51	0.50
Predictability by trait Item 3	0.41	0.35	0.34
Unpredictability by trait Item 1	0.00	0.14	0.19
Unpredictability by trait Item 2	0.00	0.09	0.12
Unpredictability by trait Item 3	0.00	0.13	0.17
Consistency Item 1	0.40	0.48	0.51
Consistency Item 2	0.56	0.60	0.62
Consistency Item 3	0.41	0.48	0.51
Occasion-Specificity Item 1	0.40	0.35	0.33
Occasion-Specificity Item 2	0.24	0.22	0.21
Occasion-Specificity Item 3	0.37	0.32	0.31
Reliability Item 1	0.80	0.83	0.84
Reliability Item 2	0.81	0.82	0.83
Reliability Item 3	0.78	0.81	0.82

A.4.3 Fitting the multilevel TSO model

Prepare the data and the syntax files for Mplus

To finish, we will write the syntax of the multilevel TSO model with the function `write.mltso.to.Mplus`. The following code exports the data to Mplus in the file **mltso.dat** and creates the syntax file **mltso.inp**:

```
# Write data in Mplus format and write input file template to the working
# directory:
prepareMplusData(data$data.long, paste0(getwd(), "/mltso.dat"),
                    inpfile = TRUE)

# Write additional options:
options_syntax <- write.Mplus.options(
  usevariables  = names(data$data.long)[-(1:2)],
  cluster       = names(data$data.long)[1],
  analysis_type = "TWOLEVEL",
  estimator     = "BAYES",
  iterations    = 5000)

# Write Mplus syntax of the multilevel MSST:
analysis_syntax <- write.mltso.to.Mplus(data$data.long[, -(1:2)])

# Additional options important for Bayesian analyses to save MCMC samples and
# Rhat statistics:
output_syntax <- "\nSAVEDATA: BPARAMETERS=samples_mltso.dat;\nOUTPUT: TECH8;"

# Overwrite the basic syntax file to include additional options and the
# syntax of the model:
write(options_syntax,  paste0(getwd(), "/mltso.inp"), append = TRUE)
write(analysis_syntax, paste0(getwd(), "/mltso.inp"), append = TRUE)
write(output_syntax,   paste0(getwd(), "/mltso.inp"), append = TRUE)
```

Run the analysis in Mplus and import the results in R

Run the analysis in Mplus:

```
runModels(paste0(getwd(), "/mltso.inp"))
```

Read the output file **mltso.out** in R:

```
fit <- readModels(paste0(getwd(), "/mltso.out"))
```

In R, the first five rows of the estimated parameters look like this:

Table A.17: Mplus Estimates of the Multilevel TSO Model

paramHeader	param	est	posterior_sd	pval	lower_2.5ci	upper_2.5ci	sig	BetweenWithin
ETA.BY	Y1	1.00	0.00	0	1.00	1.00	FALSE	Within
ETA.BY	Y2	0.55	0.07	0	0.43	0.70	TRUE	Within
ETA.BY	Y3	1.12	0.13	0	0.88	1.40	TRUE	Within
ETA.ON	ETA&1	0.64	0.13	0	0.32	0.83	TRUE	Within
Residual.Variances	Y1	1.02	0.20	0	0.64	1.43	TRUE	Within

Compute variance coefficients

The variance coefficients for the multilevel TSO model are also computed with the function `tso.var.coeff`. This function retrieve these coefficients for each measurement occasion like this:

```
tso.var.coeff(I = I,
              nT = nT,
              within.parameters = within.estimates,
              between.parameters = between.estimates)
```

Table A.18: Variance Coefficients of the Multilevel MSST Model

Coefficient	Time 1	Time 2	Time 3
Predictability by trait Item 1	0.34	0.28	0.27
Predictability by trait Item 2	0.54	0.49	0.47
Predictability by trait Item 3	0.38	0.33	0.32
Unpredictability by trait Item 1	0.00	0.15	0.21
Unpredictability by trait Item 2	0.00	0.09	0.13
Unpredictability by trait Item 3	0.00	0.13	0.18
Consistency Item 1	0.34	0.44	0.47
Consistency Item 2	0.54	0.58	0.60
Consistency Item 3	0.38	0.47	0.49
Occasion-Specificity Item 1	0.44	0.37	0.35
Occasion-Specificity Item 2	0.25	0.23	0.22
Occasion-Specificity Item 3	0.37	0.32	0.30
Reliability Item 1	0.78	0.81	0.82
Reliability Item 2	0.79	0.81	0.82
Reliability Item 3	0.75	0.78	0.80

Appendix B

This appendix includes complementary results of the simulation study and the empirical example. In relation to the simulation study, we provide (a) the true population parameters used during the simulation to generate the data, (b) a sensitivity analysis of the Bayesian models, (c) the results of the simulation that was performed to manipulate and explore the effect of the number of indicators, (d) the extended results of the simulation taking into account the information criteria and the particular effect of the trait-state variance ratio in some conditions, (e) the results of the additional simulation that was performed to explore why the multilevel CUTS model failed to fit to TSO data, and (f) a plot of the average bias of the autoregressive effect for the conditions where the TSO model was fitted to TSO data. Regarding the empirical example, we present the results of the analyses when non-stationary time series were excluded from the analyses, which reduced the sample size to 376 individuals. In particular, we added four tables that summarize these results: The first table shows the ppp and the DIC of the models fitted to these data, the second and third tables show the estimates of the multilevel TSO model and their credibility intervals for each set of items, and the fourth table shows the estimates of the variance coefficients.

B.1 True Population Parameters

Table B.1: MSST True Parameters per Trait-State Variance Ratio

	MSST 1:3	MSST 1:1	MSST 3:1
Within loading λ_{S_1}	1.00	1.00	1.00
Within loading λ_{S_2}	0.50	0.50	0.50
Within loading λ_{S_3}	1.30	1.30	1.30
Within loading λ_{S_4}	0.80	0.80	0.80
State variance $var(\zeta)$	1.80	1.20	0.60
Error variance $var(\varepsilon_1)$	0.60	0.60	0.60
Error variance $var(\varepsilon_2)$	0.25	0.25	0.25
Error variance $var(\varepsilon_3)$	0.70	0.70	0.70
Error variance $var(\varepsilon_4)$	0.50	0.50	0.50
Between loading λ_{T_1}	1.00	1.00	1.00
Between loading λ_{T_2}	0.50	0.50	0.50
Between loading λ_{T_3}	1.30	1.30	1.30
Between loading λ_{T_4}	0.80	0.80	0.80
Intercept α_1	0.00	0.00	0.00
Intercept α_2	0.20	0.20	0.20
Intercept α_3	0.40	0.40	0.40
Intercept α_4	0.60	0.60	0.60
Trait variance $var(\xi)$	0.60	1.20	1.80
Trait mean $\hat{\xi}$	4.00	4.00	4.00

Table B.2: CUTS True Parameters per Trait-State Variance Ratio

	CUTS 1:3	CUTS 1:1	CUTS 3:1
Within loading λ_{S_1}	1.00	1.00	1.00
Within loading λ_{S_2}	0.50	0.50	0.50
Within loading λ_{S_3}	1.30	1.30	1.30
Within loading λ_{S_4}	0.80	0.80	0.80
Common State variance $var(\zeta)$	1.80	1.20	0.60
Unique state variance $var(\varepsilon_1)$	0.60	0.60	0.60
Unique state variance $var(\varepsilon_2)$	0.25	0.25	0.25
Unique state variance $var(\varepsilon_3)$	0.80	0.80	0.80
Unique state variance $var(\varepsilon_4)$	0.50	0.50	0.50
Between loading λ_{T_1}	1.00	1.00	1.00
Between loading λ_{T_2}	0.50	0.50	0.50
Between loading λ_{T_3}	1.30	1.30	1.30
Between loading λ_{T_4}	0.80	0.80	0.80
Intercept α_1	2.00	2.00	2.00
Intercept α_2	2.50	2.50	2.50
Intercept α_3	3.00	3.00	3.00
Intercept α_4	3.50	3.50	3.50
Common Trait variance $var(\xi)$	0.40	1.00	1.60
Unique trait variance $var(\vartheta)$	0.20	0.20	0.20
Unique trait variance $var(\vartheta)$	0.10	0.10	0.10
Unique trait variance $var(\vartheta)$	0.25	0.25	0.25
Unique trait variance $var(\vartheta)$	0.15	0.15	0.15

Table B.3: TSO True Parameters per Trait-State Variance Ratio

	TSO 1:3	TSO 1:1	TSO 3:1
Within loading λ_{S_1}	1.00	1.00	1.00
Within loading λ_{S_2}	0.50	0.50	0.50
Within loading λ_{S_3}	1.30	1.30	1.30
Within loading λ_{S_4}	0.80	0.80	0.80
Occasion-specific residual variance $var(\zeta)$	1.80	1.20	0.60
Error variance $var(\varepsilon_1)$	0.60	0.60	0.60
Error variance $var(\varepsilon_2)$	0.25	0.25	0.25
Error variance $var(\varepsilon_3)$	0.70	0.70	0.70
Error variance $var(\varepsilon_4)$	0.50	0.50	0.50
Autoregressive effect β	0.50	0.50	0.50
Intercept α_1	2.00	2.00	2.00
Intercept α_2	2.50	2.50	2.50
Intercept α_3	3.00	3.00	3.00
Intercept α_4	3.50	3.50	3.50
Latent trait indicator variance $var(\xi_1)$	0.30	0.80	1.60
Latent trait indicator variance $var(\xi_2)$	0.10	0.20	0.55
Latent trait indicator variance $var(\xi_3)$	0.40	1.30	2.20
Latent trait indicator variance $var(\xi_4)$	0.20	0.50	1.20
$Cov(\xi_1,\xi_2)$	0.14	0.32	0.75
$Cov(\xi_1,\xi_3)$	0.31	0.92	1.69
$Cov(\xi_2,\xi_3)$	0.16	0.41	0.88
$Cov(\xi_1,\xi_4)$	0.22	0.57	1.25
$Cov(\xi_2,\xi_4)$	0.10	0.22	0.57
$Cov(\xi_3,\xi_4)$	0.20	0.56	1.14

Table B.4: MSST True Variance Coefficient Components per Trait-State Variance Ratio

	MSST 1:3	MSST 1:1	MSST 3:1
Reliability Y_1	0.80	0.80	0.80
Reliability Y_2	0.71	0.71	0.71
Reliability Y_3	0.85	0.85	0.85
Reliability Y_4	0.75	0.75	0.75
Consistency Y_1	0.20	0.40	0.60
Consistency Y_2	0.18	0.35	0.53
Consistency Y_3	0.21	0.43	0.64
Consistency Y_4	0.19	0.38	0.57
Occasion Specificity Y_1	0.60	0.40	0.20
Occasion Specificity Y_2	0.53	0.35	0.18
Occasion Specificity Y_3	0.64	0.43	0.21
Occasion Specificity Y_4	0.57	0.38	0.19

Table B.5: CUTS True Variance Coefficient Components per Trait-State Variance Ratio

	CUTS 1:3	CUTS 1:1	CUTS 3:1
Reliability Y_1	0.80	0.80	0.80
Reliability Y_2	0.72	0.72	0.72
Reliability Y_3	0.83	0.83	0.83
Reliability Y_4	0.76	0.76	0.76
Total Consistency Y_1	0.20	0.40	0.60
Total Consistency Y_2	0.22	0.39	0.56
Total Consistency Y_3	0.19	0.41	0.62
Total Consistency Y_4	0.20	0.38	0.57
Common Consistency Y_1	0.13	0.33	0.53
Common Consistency Y_2	0.11	0.28	0.44
Common Consistency Y_3	0.14	0.35	0.57
Common Consistency Y_4	0.12	0.31	0.50
Unique Consistency Y_1	0.07	0.07	0.07
Unique Consistency Y_2	0.11	0.11	0.11
Unique Consistency Y_3	0.05	0.05	0.05
Unique Consistency Y_4	0.07	0.07	0.07
Occasion Specificity Y_1	0.60	0.40	0.20
Occasion Specificity Y_2	0.50	0.33	0.17
Occasion Specificity Y_3	0.64	0.43	0.21
Occasion Specificity Y_4	0.56	0.37	0.19

Table B.6: TSO True Variance Coefficient Components per Trait-State Variance Ratio

	TSO 1:3	TSO 1:1	TSO 3:1
Reliability Y_1	0.82	0.80	0.80
Reliability Y_2	0.74	0.71	0.75
Reliability Y_3	0.86	0.85	0.84
Reliability Y_4	0.78	0.75	0.77
Consistency Y_1	0.27	0.40	0.60
Consistency Y_2	0.26	0.35	0.60
Consistency Y_3	0.27	0.42	0.60
Consistency Y_4	0.26	0.37	0.60
Predictability by Trait Y_1	0.09	0.27	0.53
Predictability by Trait Y_2	0.11	0.24	0.55
Predictability by Trait Y_3	0.08	0.28	0.52
Predictability by Trait Y_4	0.09	0.25	0.54
Unpredictability by Trait Y_1	0.18	0.13	0.07
Unpredictability by Trait Y_2	0.16	0.12	0.05
Unpredictability by Trait Y_3	0.20	0.14	0.08
Unpredictability by Trait Y_4	0.17	0.13	0.06
Occasion Specificity Y_1	0.55	0.40	0.20
Occasion Specificity Y_2	0.47	0.35	0.15
Occasion Specificity Y_3	0.59	0.43	0.24
Occasion Specificity Y_4	0.52	0.38	0.17

B.2 Sensitivity Analysis

When doing Bayesian estimation, researchers have to be careful in the selection of their prior distributions because they can influence the estimates of the model. In general, the effect of the prior distributions is diminished the larger the data are. For the simulation study, we used the default prior distributions available in Mplus for our Bayesian analyses. These default priors are uninformative priors, for example, loadings are given normal priors $N(0,10^{10})$. To verify that a different selection of priors would not affect our results, we conducted a sensitivity analysis on a random sample of the analyses of the simulation. We selected one random replication from 20 random conditions, and fitted the Bayesian models using the default priors and some weak priors selected by us (e.g., $N(0,5)$). The results from these analyses showed that the estimates are basically the same regardless of the priors. Differences between the estimates at the within-level were not larger than 0.01 and differences between the estimates at the between-level were not larger than 0.15. These larger differences in the between-level were mainly associated to the variances, which are harder to estimate because there was less information at this level due to the sample size. As an example, Table B.7 presents the estimates obtained by using both uninformative and weak priors of one of the models in one of the replications.

Table B.7: Estimates of the Multilevel TSO Model using Uninformative and Weak Priors with the Base Model as the TSO, 30 Measurement Occasions, 0% Missing Values, and a Trait-State Variance Ratio of 1:1.

	Uninformative Priors	Weak Priors
Within loading λ_{S_1}	**1.00**	**1.00**
Within loading λ_{S_2}	0.479	0.479
Within loading λ_{S_3}	1.258	1.259
Within loading λ_{S_4}	0.780	0.780
Occasion-specific residual variance $var(\zeta)$	1.347	1.348
Error variance $var(\varepsilon_1)$	0.599	0.599
Error variance $var(\varepsilon_2)$	0.263	0.262
Error variance $var(\varepsilon_3)$	0.644	0.642
Error variance $var(\varepsilon_4)$	0.510	0.510
Autoregressive effect β	0.493	0.495
Intercept α_1	1.930	1.935
Intercept α_2	2.477	2.479
Intercept α_3	2.903	2.918
Intercept α_4	3.479	3.481
Latent trait indicator variance $var(\xi_1)$	0.829	0.696
Latent trait indicator variance $var(\xi_2)$	0.209	0.180
Latent trait indicator variance $var(\xi_3)$	1.295	1.110
Latent trait indicator variance $var(\xi_4)$	0.505	0.438
$Cov(\xi_1,\xi_2)$	0.339	0.279
$Cov(\xi_1,\xi_3)$	0.927	0.773
$Cov(\xi_2,\xi_3)$	0.424	0.352
$Cov(\xi_1,\xi_4)$	0.570	0.477
$Cov(\xi_2,\xi_4)$	0.235	0.192
$Cov(\xi_3,\xi_4)$	0.528	0.422

B.3 Simulation: Effect of the Number of Indicators

When using the single-level state-trait SEMs, increasing the number of indicators has a similar effect as increasing the number of measurement occasions in relation to the number of observed variables that are included in the models. For example, a study with six indicators and five measurement occasions has the same number of observed variables as a study with three indicators and ten measurement occasions. Because of this, it is reasonable to think that increasing the number of indicators also impacts the performance of the models. However, this factor was kept fixed in the simulation design because intensive longitudinal studies are not likely to include a lot indicators to measure the same construct. Nevertheless, to verify how the number of indicators might affect the performance of the models, we ran a small-scale simulation where we manipulated this factor. In these analyses, the number of indicators was varied between 4, 7, and 10; and the number of measurement occasions was varied between 10, 20, 30, and 60. Moreover, we only analyzed the data with the same model that was used to generate them.

The number of analyses that finished successfully given each condition are shown in Figures B.1 to B.3. These plots clearly show that the number of indicators does not impact the performance of the models but the number of measurement occasions does. This difference can be explained by how each of these factors directly affects the models. On the one hand, increasing the number of measurement occasions makes the models more complex because it introduces more latent variables that have to be modeled for each additional occasion. For example, if we are using the CUTS model in a design with 3 indicators and increase the number of measurement occasions from 10 to 20, the number of latent variables in the model increases from 14 to 24. On the other hand, while increasing the number of indicators increases the number of loadings and within-residual variances, it barely increases the number of latent variables that have to be modeled. For example, if we are applying the CUTS model and increase the number of indicators from 3 to 6 in a design with 10 measurement occasions, we go from a model with 14 to a model with 17 latent variables. In conclusion, with large datasets, single-level state-trait SEMs are more likely to run into convergence issues the more latent variables there are in the model. Therefore, as increasing the number of indicators barely increases the number of latent variables in the model, including this factor in the simulation study was unnecessary.

Figure B.1: Number of Successful Analyses varying the Number of Indicators with the MSST Model

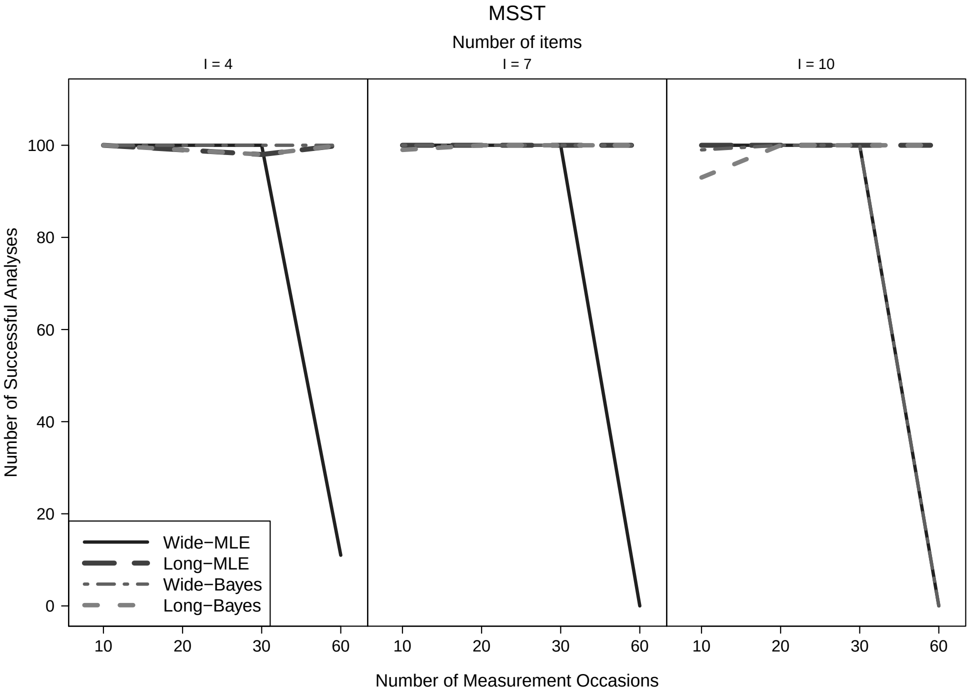


Figure B.2: Number of Successful Analyses varying the Number of Indicators with the CUTS Model

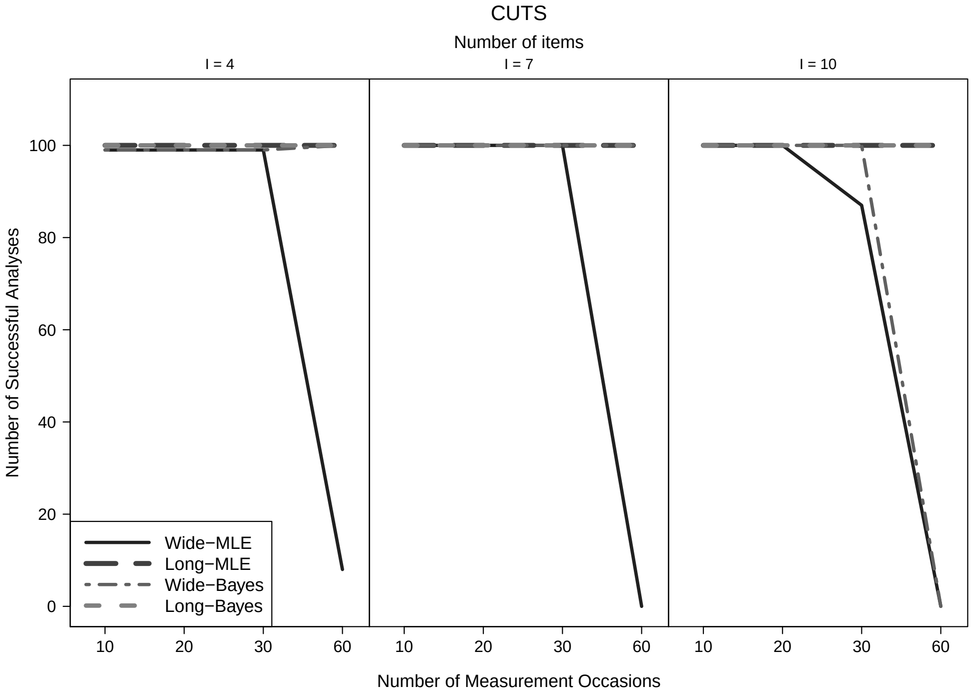
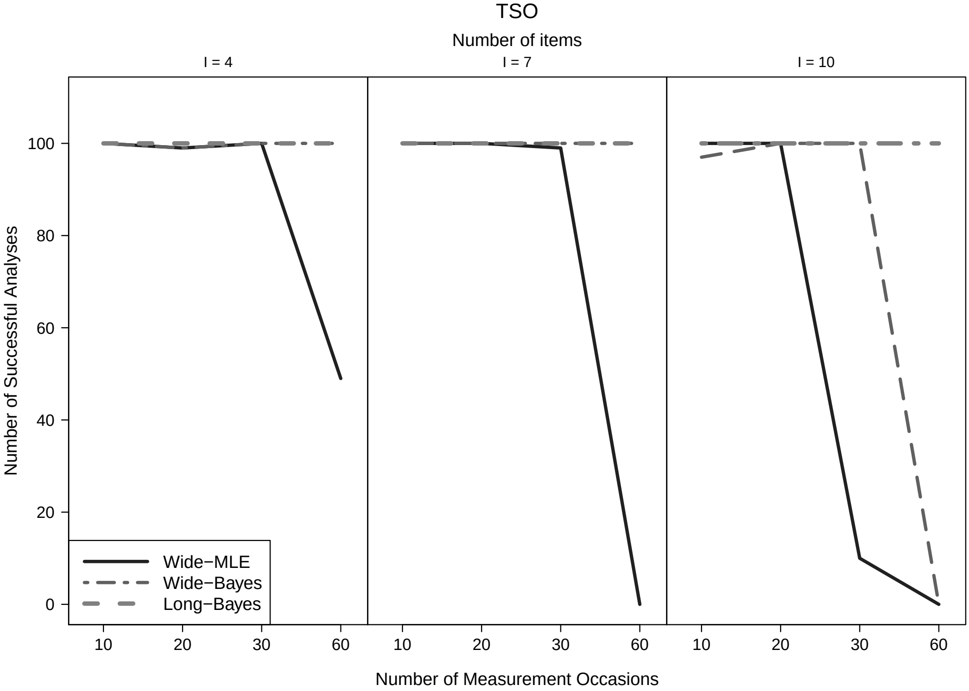


Figure B.3: Number of Successful Analyses varying the Number of Indicators with the TSO Model



B.4 Simulation: Extended Results

In this section, we provide extended results of the simulation study. Firstly, we give a brief summary of the information criteria indices, which are useful to decide which model fitted the data best. Secondly, we include the results of the number of successful analyses and the quality of the estimates related to the conditions with a trait-state variance ratio different from 1:1.

B.4.1 Information Criteria

To decide which model fitted the data best, we used the information criterion indices available in Mplus. This includes the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the adjusted Bayesian Information Criterion (aBIC) when the estimation method was MLE; and the Deviance Information Criterion (DIC) when the estimation method was MCMC Gibbs sampling.

Concerning the information criteria, we were only able to compare models estimated with the same method. When the estimation method was MLE, we selected among five models (MSST, ML-MSST, CUTS, ML-CUTS, and TSO) by means of the AIC, the BIC, and the aBIC. The percentage of the number of times a model was selected as the best model is shown in Table B.8 given the model used to generate the data. In most of the cases, the correct model was selected as the best model regardless of whether it was the single level or the multilevel version. However, when the data were generated based on the TSO model, the model selected as the best model was actually the ML-MSST in about a third of the analyses. This happened because when the number of measurement occasions was large and the base model was the TSO, the only model that converged by means of MLE was the ML-MSST model (see Figures B.4 to B.9). Hence, it was the only model to pick from as the best model.

When the estimation method was Bayesian, we compared the models by means of the DIC. The ML-TSO model was almost always selected as the best model across all the conditions independently of the base model. For example, when the data were generated based on the MSST model, the ML-TSO model was selected as the best model 5383 times out of 5400 regardless of the number of measurement occasions, the proportion of missing values, or the trait-state variance ratio. Note that if the data were in wide format and had 30 measurement occasions or more, Mplus was unable to compute the DIC.[1]This means that when the number of measurement occasions

Table B.8: Percentage of Times a Model was Selected as Best Model According to the Information Criteria Indices Available with Maximum Likelihood Estimation

Base Model	Information Criterion	Fitted Model				
		MSST	ML-MSST	CUTS	ML-CUTS	TSO
MSST	AIC	35.9	63.6	0.1	0.4	0
	BIC	58.4	41.6	0	0	0
	aBIC	58.2	41.6	0.2	0	0
CUTS	AIC	0	0	50.1	42.4	7.4
	BIC	0	0	56.5	38.4	5.1
	aBIC	0	0	42.9	38.4	18.7
TSO	AIC	0.1	42.8	0.7	1.1	55.4
	BIC	6.6	36.2	0.7	1.1	55.4
	aBIC	6.6	36.2	0.7	1.1	55.4

was 30 or more, the best model according to the DIC was selected only from the multilevel models.

B.4.2 Effect of the Trait-State Variance Ratio

We included the trait-state variance ratio in the simulation design expecting little to no differences when manipulating this factor. While this was generally true, there are some particular conditions where the trait-state variance ratio interacts with other manipulated factors and shows some effects in the performance of the models and the quality of the estimates. In this section, we further explain these results.

First of all, in Figures B.4 to B.9, we present the number of analyses that finished successfully in the conditions with a trait-state variance ratio different from 1:1 given

[1] When doing the analyses in wide-format by means of the Gibbs sampling algorithm, the following message was always printed in the Mplus output: "Problem occurred in the computation of the posterior predictive p-value. This may be due to a singular sample variance-covariance matrix, such as with zero sample variances." This happened because of the large number of observed variables that were included in some of our analyses (120, 240, 360), which led to huge sample variance-covariance matrices. These huge matrices might make it impossible for Mplus to compute the deviance and the posterior predictive p-value of the model.

each proportion of missingness. In particular, Figures B.7 to B.9 show how having more state-like variables can affect the performance of the MSST when analyzing CUTS data and the performance of the TSO model when analyzing its own data. Firstly, the single-level and the multilevel MSST models performed very poorly when analyzing CUTS data across all conditions. This means that the MSST model is problematic if the variables of interest are more state-like and if real method effects are present in the data. Secondly, the single-level TSO model with MLE seemed to improve as the number of measurement occasions increased. One possible explanation is that, because the variables were more state-like, more data was needed to correctly capture the structure of the latent trait indicator variables (the trait component of the variables).

Figure B.4: Number of Successful Analyses per Condition with 0% Missingness and 3:1 Trait-State Variance Ratio

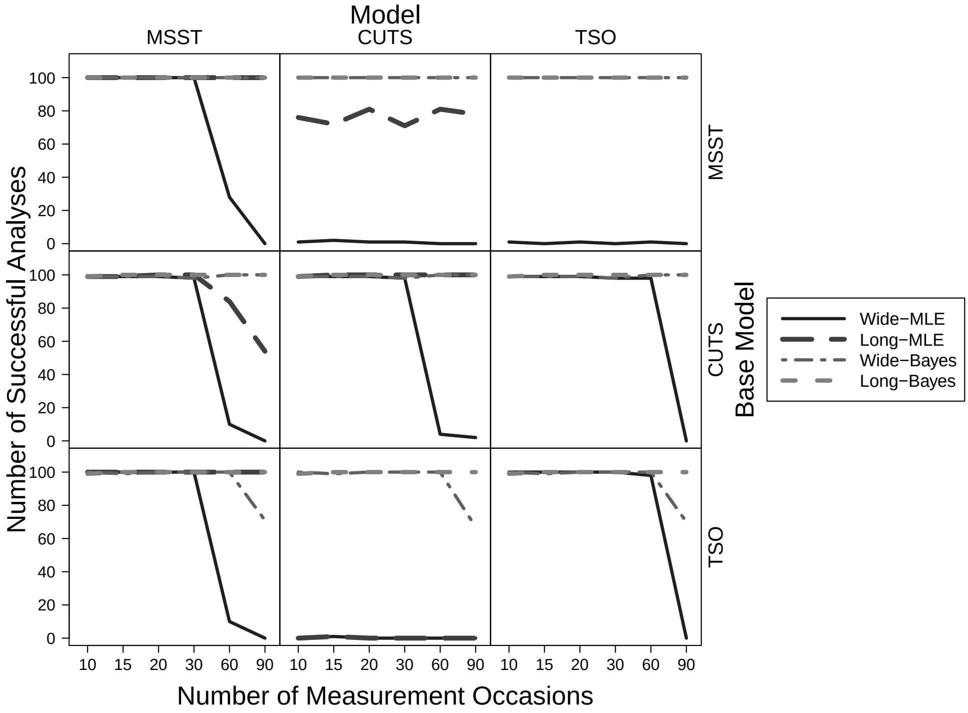


Figure B.5: Number of Successful Analyses per Condition with 10% Missingness and 3:1 Trait-State Variance Ratio

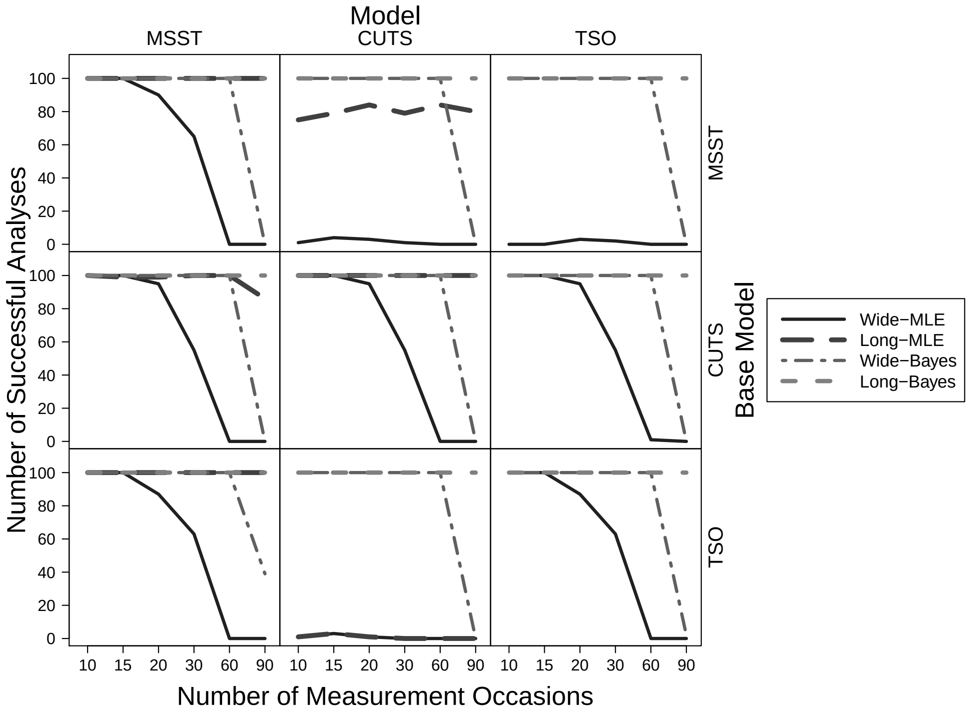


Figure B.6: Number of Successful Analyses per Condition with 20% Missingness and 3:1 Trait-State Variance Ratio

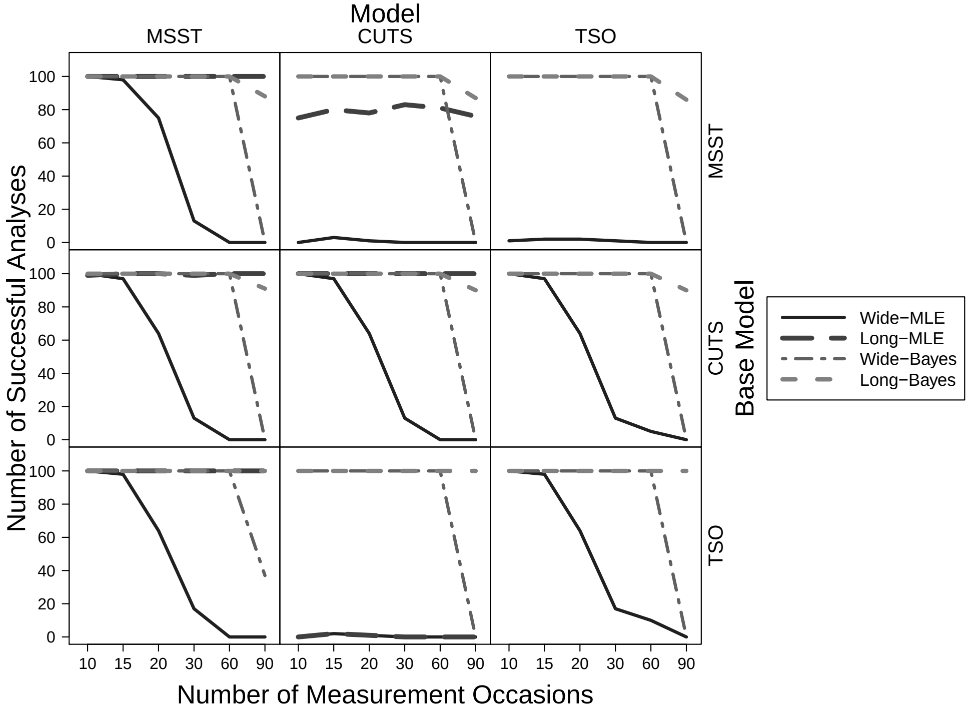


Figure B.7: Number of Successful Analyses per Condition with 0% Missingness and 1:3 Trait-State Variance Ratio

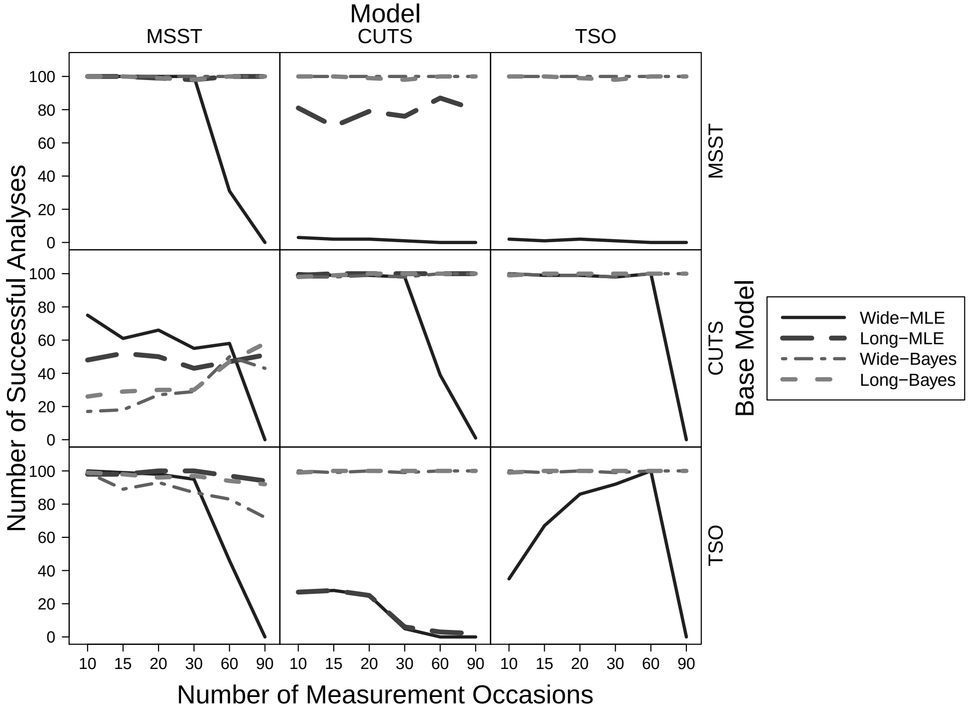


Figure B.8: Number of Successful Analyses per Condition with 10% Missingness and 1:3 Trait-State Variance Ratio

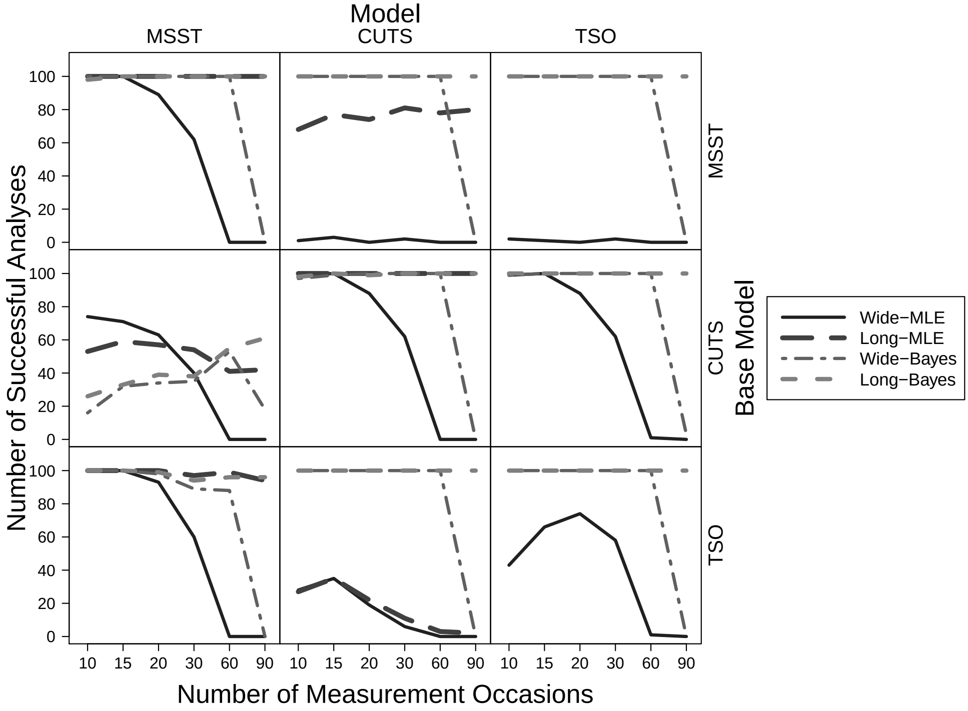
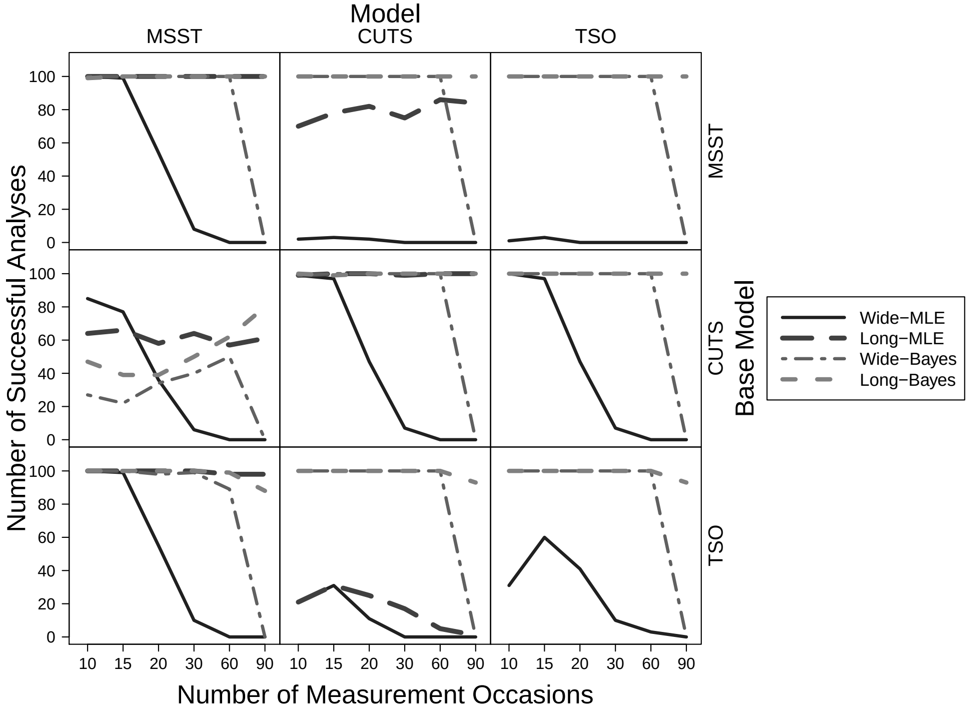


Figure B.9: Number of Successful Analyses per Condition with 20% Missingness and 1:3 Trait-State Variance Ratio



Finally, varying the trait-state variance ratio also had some specific effects in the quality of the estimates of some of the parameters. Firstly, while the estimates of the within factor loadings λ_{S_j} were practically unbiased in most conditions, they tended to be overestimated when analyzing TSO data in conditions that had a trait-state variance ratio of 3:1 with the MSST model. Moreover, as mentioned in the main text, the estimates of the consistencies tended to show more bias when the variables were more state-like. In contrast, the estimates of the occasion-specificities tended to show more bias when the variables were more trait-like.

B.5 Simulation: Fitting the Multilevel CUTS model to TSO data

In the simulation study, it was unexpected that the multilevel version of the CUTS model failed to converge to a reasonable solution in all the replications of certain conditions. Particularly, the CUTS model failed when it was estimated with maximum likelihood estimation (MLE) and the TSO model was used to generate the data. As a consequence, we decided to further investigate these results.

To start with, we repeated some analyses under the problematic conditions. We noted that the CUTS model was estimating at least one negative variance, which resulted in an improper solution. To find what was causing these results, we repeated part of the simulation changing the true parameters of the TSO model. Specifically, we decided to modify the correlation matrix used to simulate the latent trait indicator variables of the TSO. We followed the next process: (a) Generate a new correlation matrix based on a specific seed, (b) simulate 50 datasets with 30 measurement occasions and without missing values, (c) fit the multilevel CUTS model to the data by means of MLE, (d) count how many analyses failed. This procedure was repeated 14 times, using seven different seeds and with high (i.e., 0.7, 0.8, 0.9) and low (i.e., 0.5, 0.6, 0.7) correlations.

The results of the analyses with high correlations are shown in Table B.9. This table presents the seed used to generate the correlation matrix, the correlation matrix, its determinant, and the number of analyses that failed out of the 50 replications. The first row is actually the correlation matrix used in the simulation of this study, which failed in 50 out of 50 analyses. Next, by generating the correlation matrix based on a different seed, the results change dramatically. In some cases, the CUTS model did an excellent job (seeds 13002 and 666), in other cases, the CUTS model was mediocre (seeds 13003, 13004, and 2019), and in others, it failed completely (seeds 13001 and 13014).

We also repeated the previous simulations but with low correlations. The results are shown in Table B.10. In these cases, the CUTS model did, generally speaking, a good job. There were only two situations were there was at least one analysis that failed, and the number of analyses that failed was rather small.

To sum up, it is a fact that the correlation matrix used to generated the TSO data has an effect on the number of analyses that fail when fitting the CUTS model by

Table B.9: Results of Simulations Varying the Correlation Matrix of High Correlations used to Generate TSO Data

	High Correlations		
Seed	Correlation matrix	Determinant	# of Analyses that Failed
13001	$\begin{bmatrix}1&&&\\0.8&1&&\\0.9&0.8&1&\\0.9&0.7&0.7&1\end{bmatrix}$	0.0077	50
13002	$\begin{bmatrix}1&&&\\0.7&1&&\\0.8&0.9&1&\\0.8&0.8&0.8&1\end{bmatrix}$	0.0168	1
13003	$\begin{bmatrix}1&&&\\0.8&1&&\\0.8&0.7&1&\\0.9&0.9&0.8&1\end{bmatrix}$	0.0117	15
13004	$\begin{bmatrix}1&&&\\0.9&1&&\\0.7&0.9&1&\\0.9&0.8&0.7&1\end{bmatrix}$	0.0032	9
13014	$\begin{bmatrix}1&&&\\0.9&1&&\\0.8&0.9&1&\\0.7&0.9&0.8&1\end{bmatrix}$	0.0045	49
666	$\begin{bmatrix}1&&&\\0.8&1&&\\0.7&0.7&1&\\0.7&0.7&0.7&1\end{bmatrix}$	0.066	0
2019	$\begin{bmatrix}1&&&\\0.7&1&&\\0.7&0.8&1&\\0.7&0.7&0.9&1\end{bmatrix}$	0.0288	7

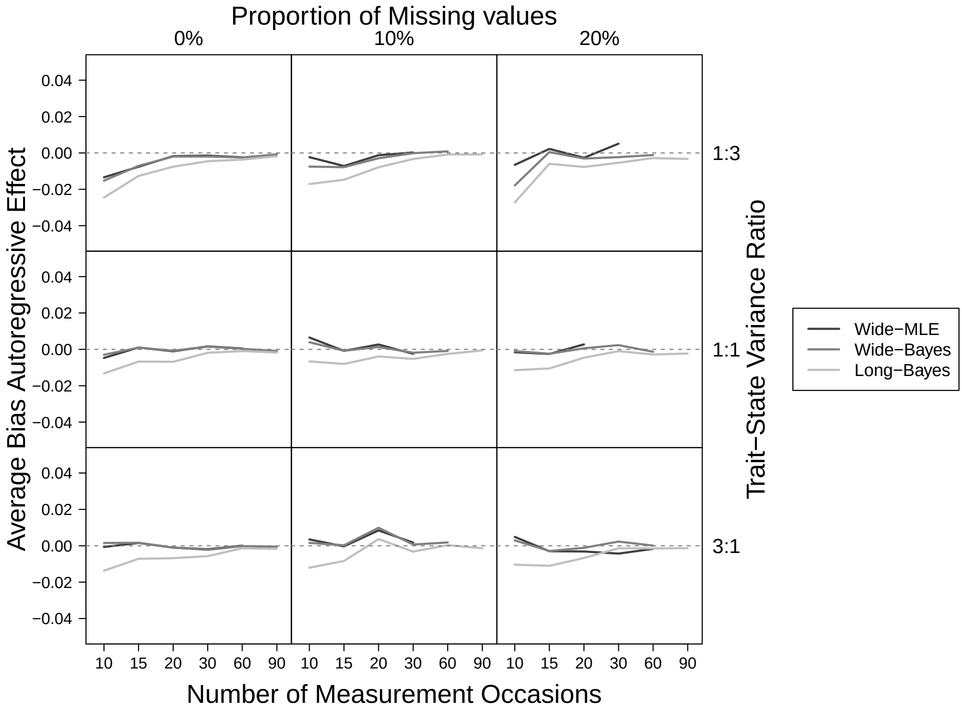
Table B.10: Results of Simulations Varying the Correlation Matrix of Low Correlations used to Generate TSO Data

Low Correlations			
Seed	Correlation matrix	Determinant	# of Analyses that Failed
13001	$\begin{bmatrix} 1 & & & \\ 0.6 & 1 & & \\ 0.7 & 0.6 & 1 & \\ 0.7 & 0.5 & 0.5 & 1 \end{bmatrix}$	0.1469	3
13002	$\begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ 0.6 & 0.7 & 1 & \\ 0.6 & 0.6 & 0.6 & 1 \end{bmatrix}$	0.1616	0
13003	$\begin{bmatrix} 1 & & & \\ 0.6 & 1 & & \\ 0.6 & 0.5 & 1 & \\ 0.7 & 0.7 & 0.6 & 1 \end{bmatrix}$	0.1421	0
13004	$\begin{bmatrix} 1 & & & \\ 0.7 & 1 & & \\ 0.5 & 0.7 & 1 & \\ 0.7 & 0.6 & 0.5 & 1 \end{bmatrix}$	0.1236	0
13014	$\begin{bmatrix} 1 & & & \\ 0.7 & 1 & & \\ 0.6 & 0.7 & 1 & \\ 0.5 & 0.7 & 0.6 & 1 \end{bmatrix}$	0.1205	1
666	$\begin{bmatrix} 1 & & & \\ 0.6 & 1 & & \\ 0.5 & 0.5 & 1 & \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}$	0.28	0
2019	$\begin{bmatrix} 1 & & & \\ 0.5 & 1 & & \\ 0.5 & 0.6 & 1 & \\ 0.5 & 0.5 & 0.7 & 1 \end{bmatrix}$	0.21	0

means of MLE. Importantly, if the correlations in the correlation matrix are high then it is more likely that analyses will fail. However, it is unclear what exactly in the correlation matrix results in more failed analyses.

B.6 Average Bias of the Autoregressive Effect

Figure B.10: Average Bias of the Autoregressive Effect when fitting the single-level and multilevel TSO models to TSO data.



B.7 Empirical Example Excluding Non-stationary Time Series

Stationarity is an important assumption of state-trait SEMs when analyzing intensive longitudinal data. For this reason, the HND data were analyzed a second time but excluding individuals with at least one non-stationary time series. To test for stationarity, we used the Kwiatkowski–Phillips–Schmidt–Shin test (Kwiatkowski et al., 1992). This procedure reduced our sample size from 644 to 376 individuals. Next, we present the fit measures of the MSST, the CUTS, and TSO models in Table B.11. The DICs indicated that the TSO model fit best the data. Moreover, Tables B.12 and B.13 present the estimates of the TSO model for the two samples: The sample with 644 individuals and the sample with 376 individuals that excludes non-stationary time series. Finally, Table B.14 presents the estimates of the variance coefficients for the TSO model on the sample without non-stationary time series.

Although there were some small differences in the estimates, they were not large enough to change our interpretation of the results. For example, the autoregressive effect of the multilevel TSO model fitted to the items of positive affect deactivation went from 0.37 with the whole sample to 0.32 with the stationary sample (see B.12). However, the fact that the differences were not substantial does not mean that stationarity is not a required assumption for state-trait SEMs. Simply, in this particular data the violations of stationarity were not large enough to bias our results. In general, we advice practitioners to always test for stationarity and to fit the models with both the whole sample and the stationary sample. If the results from these analyses lead to different interpretations, the results from the stationary sample should be preferred.

Table B.11: ppp and DIC of the Three Models for the Two Sets of Items

		MSST	CUTS	TSO
PAD	ppp	0.614	0.655	-
	DIC	688083.833	680948.219	656303.508
PAA	ppp	0.603	0.650	-
	DIC	687287.995	680842.489	647618.688

Table B.12: Estimates of the Multilevel TSO for the Items of Positive Affect Deactivation

	N = 644	N = 376
Parameter	$\hat{\theta}$ (Credibility Interval)	$\hat{\theta}$ (Credibility Interval)
Within Loading R	**1**	**1**
Within Loading Co	0.89 (0.88, 0.9)	0.89 (0.87, 0.91)
Within Loading Ca	0.89 (0.87, 0.9)	0.86 (0.84, 0.88)
Autoregressive Effect	0.37 (0.36, 0.38)	0.32 (0.3, 0.33)
Within Variance R	74.15 (72.01, 76.23)	71.74 (68.87, 74.58)
Within Variance Co	135.45 (133.12, 137.79)	136.67 (133.63, 139.93)
Within Variance Ca	130.34 (128.14, 132.6)	132.21 (129.34, 135.27)
Latent Occasion-Specific Residual Variance: PAD	143.56 (140.3, 146.98)	153.02 (148.56, 157.57)
Latent Trait Indicator Variance: R	124.38 (111.26, 139.93)	104.89 (90.66, 122.35)
Latent Trait Indicator Variance: Co	157.47 (141.13, 176.96)	141.34 (121.83, 164.55)
Latent Trait Indicator Variance: Ca	138.43 (123.97, 155.18)	121.22 (104.68, 141.38)
Covariance R-Co	119.75 (106.39, 135.84)	104.95 (89.11, 123.77)
Covariance R-Ca	119.04 (105.89, 134.48)	100.89 (86.2, 118.49)
Covariance Co-Ca	117.98 (104.05, 134.32)	104.87 (88.3, 124.26)

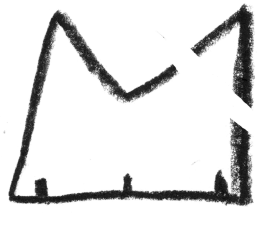
Table B.13: Estimates of the Multilevel TSO for the Items of Positive Affect Activation

	N = 644	N = 376
Parameter	$\hat{\theta}$ (Credibility Interval)	$\hat{\theta}$ (Credibility Interval)
Within Loading Eg	**1**	**1**
Within Loading Et	1.13 (1.11, 1.14)	1.13 (1.11, 1.14)
Within Loading Ch	1.05 (1.04, 1.06)	1.06 (1.05, 1.08)
Autoregressive Effect	0.32 (0.31, 0.33)	0.28 (0.26, 0.29)
Within Variance Eg	140.1 (137.91, 142.36)	140.66 (137.75, 143.57)
Within Variance Et	68.83 (66.92, 70.72)	68.51 (66, 71.04)
Within Variance Ch	92.11 (90.21, 94.06)	92.41 (89.9, 94.99)
Latent Occasion-Specific Residual Variance: PAD	158.55 (155.06, 162.17)	163.11 (158.39, 167.91)
Latent Trait Indicator Variance: Eg	144.93 (129.81, 162.75)	119.77 (103.65, 139.86)
Latent Trait Indicator Variance: Et	148.82 (133.23, 167.41)	128.13 (110.17, 149.22)
Latent Trait Indicator Variance: Ch	152.34 (136.63, 170.84)	125.63 (108.52, 146.89)
Covariance Eg-Et	130.74 (116.22, 147.84)	110.42 (94.28, 129.97)
Covariance Eg-Ch	127.22 (112.71, 144.05)	105.07 (89.49, 124.2)
Covariance Et-Ch	138.49 (123.49, 156.24)	115.51 (98.71, 135.75)

Table B.14: Variance Coefficients of the Three Models for the Two Sets of Items

	Variance Coefficient	Items		
		Relaxed	Content	Calm
PAD	Reliability	0.79	0.67	0.65
	Consistency	0.35	0.38	0.35
	Predictability by Trait	0.30	0.34	0.32
	Unpredictability by Trait	0.05	0.03	0.03
	Occasion Specificity	0.44	0.29	0.30
		Energetic	Enthusiastic	Cheerful
PAA	Reliability	0.68	0.84	0.78
	Consistency	0.30	0.34	0.34
	Predictability by Trait	0.27	0.30	0.30
	Unpredictability by Trait	0.03	0.04	0.04
	Occasion Specificity	0.37	0.49	0.44

Appendix C

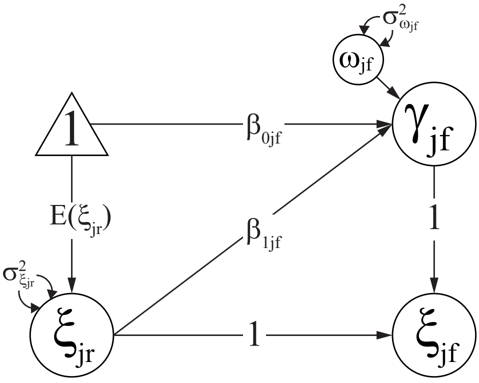


In Chapter 3, we extended the multilevel trait-state-occasion model (TSO; Castro-Alvarez et al., 2022d; Eid et al., 2017) based on the dynamic structural equation modeling framework (DSEM; Asparouhov et al., 2018) and the latent state-trait theory for the combination of random and fixed situations (Geiser et al., 2015b). This appendix includes (a) the mathematical derivation of the variances and covariances of the latent indicator- and situation-specific traits; (b) the rationale for the development of the variance coefficients within fixed situation per person of the ME-TSO model; (c) plots of the convergence of the models $\mathscr{M}_{1b}$, $\mathscr{M}_{2}$, and $\mathscr{M}_{2b}$; and (d) the unstandardized estimates of key parameters of the models when excluding individuals that had at least one nonstationary time series.

C.1 Variances and Covariances of the Latent Indicator- and Situation-Specific Traits

The ME-TSO model provides a set of variance coefficients that allow studying the psychometric properties of the items used in intensive longitudinal data. In particular, the model defines a set of coefficients across fixed situations, namely, the consistency of traits, the situation-specificity of traits, the person situation-interaction coefficient, and the unique situation effect. These coefficients are derived based on the path diagram presented in Figure C.1, which is the assumed structural model that relates the indicator- and situation-specific traits (ξ_{jf}) with the trait of the reference situation (ξ_{jr}).

Figure C.1: Structural Diagram of the Indicator- and Situation-Specific Traits.



Next, we derive the variance of the indicator- and situation-specific trait variable ξ_{jf} as well as its covariance and correlation with the indicator-specific trait variable of the reference situation ξ_{jr}. These results are required to compute the variance coefficients proposed within the ME-TSO model. Note that the model assumes the following:

$$\begin{aligned} \xi_{jf} &= \xi_{jr} + \gamma_{jf} \\ &= \xi_{jr} + \beta_{0jf} + \beta_{1jf}\xi_{jr} + \omega_{jf} \\ E[\omega_{jf}] &= 0 \\ Cov(\xi_{jr}, \omega_{jf}) &= E[\xi_{jr}\omega_{jf}] = 0. \end{aligned} \tag{C.1}$$

Variance of the indicator- and situation-specific trait variable ξ_{jf}:

$$
\begin{aligned}
Var(\xi_{jf}) &= E\Big\{\big[\xi_{jr} + \beta_{0jf} + \beta_{1jf}\xi_{jr} + \omega_{jf} - E(\xi_{jr} + \beta_{0jf} + \beta_{1jf}\xi_{jr} + \omega_{jf})\big]^2\Big\} \\
&= E\Big\{\big[\xi_{jr} + \beta_{0jf} + \beta_{1jf}\xi_{jr} + \omega_{jf} - E(\xi_{jr}) - \beta_{0jf} - \beta_{1jf}E(\xi_{jr}) - \\
&\quad E(\omega_{jf})\big]^2\Big\} \\
&= E\Big\{\big[\xi_{jr} + \beta_{1jf}\xi_{jr} + \omega_{jf} - E(\xi_{jr}) - \beta_{1jf}E(\xi_{jr})\big]^2\Big\} \\
&= E\Big\{\xi_{jr}^2 + \beta_{1jf}^2\xi_{jr}^2 + \omega_{jf}^2 + E(\xi_{jr})^2 + \beta_{1jf}^2E(\xi_{jr})^2 + 2\beta_{1jf}\xi_{jr}^2 + 2\xi_{jr}\omega_{jf} - \\
&\quad 2\xi_{jr}E(\xi_{jr}) - 2\beta_{1jf}\xi_{jr}E(\xi_{jr}) + 2\beta_{1jf}\xi_{jr}\omega_{jf} - 2\beta_{1jf}\xi_{jr}E(\xi_{jr}) - \\
&\quad 2\beta_{1jf}^2\xi_{jr}E(\xi_{jr}) - 2\omega_{jf}E(\xi_{jr}) - 2\beta_{1jf}\omega_{jf}E(\xi_{jr}) + 2\beta_{1jf}E(\xi_{jr})^2\Big\} \\
&= E(\xi_{jr}^2) + \beta_{1jf}^2E(\xi_{jr}^2) + E(\omega_{jf}^2) + E(\xi_{jr})^2 + \beta_{1jf}^2E(\xi_{jr})^2 + 2\beta_{1jf}E(\xi_{jr}^2) + \\
&\quad 2E(\xi_{jr}\omega_{jf}) - 2E(\xi_{jr})^2 - 2\beta_{1jf}E(\xi_{jr})^2 + 2\beta_{1jf}E(\xi_{jr}\omega_{jf}) - \\
&\quad 2\beta_{1jf}E(\xi_{jr})^2 - 2\beta_{1jf}^2E(\xi_{jr})^2 - 2E(\omega_{jf})E(\xi_{jr}) - 2\beta_{1jf}E(\omega_{jf})E(\xi_{jr}) + \\
&\quad 2\beta_{1jf}E(\xi_{jr})^2 \\
&= E(\xi_{jr}^2) + \beta_{1jf}^2E(\xi_{jr}^2) + E(\omega_{jf}^2) + E(\xi_{jr})^2 + \beta_{1jf}^2E(\xi_{jr})^2 + 2\beta_{1jf}E(\xi_{jr}^2) - \\
&\quad 2E(\xi_{jr})^2 - 2\beta_{1jf}E(\xi_{jr})^2 - 2\beta_{1jf}E(\xi_{jr})^2 - 2\beta_{1jf}^2E(\xi_{jr})^2 + 2\beta_{1jf}E(\xi_{jr})^2 \\
&= E(\xi_{jr}^2) - E(\xi_{jr})^2 + 2\beta_{1jf}E(\xi_{jr}^2) - 2\beta_{1jf}E(\xi_{jr})^2 + \beta_{1jf}^2E(\xi_{jr}^2) - \\
&\quad \beta_{1jf}^2E(\xi_{jr})^2 + E(\omega_{jf}^2) \\
&= \big[E(\xi_{jr}^2) - E(\xi_{jr})^2\big] + 2\beta_{1jf}\big[E(\xi_{jr}^2) - E(\xi_{jr})^2\big] + \beta_{1jf}^2\big[E(\xi_{jr}^2) - \\
&\quad E(\xi_{jr})^2\big] + E(\omega_{jf}^2) \\
&= Var(\xi_{jr}) + 2\beta_{1jf}Var(\xi_{jr}) + \beta_{1jf}^2Var(\xi_{jr}) + Var(\omega_{jf}).
\end{aligned}
$$

Covariance of the trait variables of the reference situation and a fixed situation f:

$$\begin{aligned}
Cov(\xi_{jr},\xi_{jf}) &= E\Big\{\left[\xi_{jr}-E(\xi_{jr})\right]\left[\xi_{jr}+\beta_{0jf}+\beta_{1jf}\xi_{jr}+\omega_{jf}-E(\xi_{jr}+\beta_{0jf}+\right.\\
&\qquad \left.\beta_{1jf}\xi_{jr}+\omega_{jf})\right]\Big\}\\
&= E\Big\{\left[\xi_{jr}-E(\xi_{jr})\right]\left[\xi_{jr}+\beta_{0jf}+\beta_{1jf}\xi_{jr}+\omega_{jf}-E(\xi_{jr})-\beta_{0jf}-\right.\\
&\qquad \left.\beta_{1jf}E(\xi_{jr})-E(\omega_{jf})\right]\Big\}\\
&= E\Big\{\left[\xi_{jr}-E(\xi_{jr})\right]\left[\xi_{jr}+\beta_{1jf}\xi_{jr}+\omega_{jf}-E(\xi_{jr})-\beta_{1jf}E(\xi_{jr})\right]\Big\}\\
&= E\Big\{\xi_{jr}^2+\beta_{1jf}\xi_{jr}^2+\xi_{jr}\omega_{jf}-\xi_{jr}E(\xi_{jr})-\beta_{1jf}\xi_{jr}E(\xi_{jr})-\xi_{jr}E(\xi_{jr})-\\
&\qquad \beta_{1jf}\xi_{jr}E(\xi_{jr})-\omega_{jf}E(\xi_{jr})+E(\xi_{jr})^2+\beta_{1jf}E(\xi_{jr})^2\Big\}\\
&= E(\xi_{jr}^2)+\beta_{1jf}E(\xi_{jr}^2)+E(\xi_{jr}\omega_{jf})-E(\xi_{jr})^2-\beta_{1jf}E(\xi_{jr})^2-\\
&\qquad E(\xi_{jr})^2-\beta_{1jf}E(\xi_{jr})^2-E(\omega_{jf})E(\xi_{jr})+E(\xi_{jr})^2+\beta_{1jf}E(\xi_{jr})^2\\
&= E(\xi_{jr}^2)-E(\xi_{jr})^2+\beta_{1jf}E(\xi_{jr}^2)-\beta_{1jf}E(\xi_{jr})^2\\
&= Var(\xi_{jr})+\beta_{1jf}Var(\xi_{jr}).
\end{aligned}$$

Correlation of the trait variables of the reference situation and a fixed situation f:

$$\begin{aligned}
Corr(\xi_{jr},\xi_{jf}) &= \frac{Cov(\xi_{jr},\xi_{jf})}{\sqrt{Var(\xi_{jr})Var(\xi_{jf})}} \qquad\qquad (C.2)\\
&= \frac{Var(\xi_{jr})+\beta_{1jf}Var(\xi_{jr})}{\sqrt{\left[Var(\xi_{jr})\right]\left[Var(\xi_{jr})+2\beta_{1jf}Var(\xi_{jr})+\beta_{1jf}^2Var(\xi_{jr})+Var(\omega_{jf})\right]}}.
\end{aligned}$$

C.2 Total Variance Decomposition in the ME-TSO

In the ME-TSO model, we propose the variance coefficients within fixed situations per individuals. These coefficients are proportions of the total variance decomposition of the variables Y_{ijf}, which represent the time series of person i on the j-th indicator in the fixed situation f. This total variance decomposition was developed based on the total variance decomposition of the TSO model (Castro-Alvarez et al., 2022d; Eid et al., 2017). To start with, consider the observed variable Y_{jt} that represents the scores across the sample of indicator j at time t. In the TSO model assuming longitudinal measurement invariance, Y_{jt} is decomposed as follows:

$$Y_{jt} = \alpha_j + \lambda_{T_j}\xi_j + \lambda_{S_j}O_t + \varepsilon_{jt}, \tag{C.3}$$

where α_j is a constant vector that represents the intercept, ξ_j represents the latent indicator-specific trait variable of indicator j, λ_{T_j} is the loading of the trait variable ξ_j and is equal to 1 for identification purposes, O_t is the latent occasion specific variable, and λ_{S_j} is the j-th loading of the occasion specific variable. Moreover, the autoregressive structure is imposed on the latent occasion specific variables, which means that a occasion specific variable O_t at time t is regressed on the occasion specific variable of the previous occasion O_{t-1}. As a result, the first three occasion specific variables are defined as follows:

$$\begin{aligned}
O_1 &= \zeta_1, && \text{(C.4)}\\
O_2 &= \varphi O_1 + \zeta_2 \\
&= \varphi \zeta_1 + \zeta_2, && \text{(C.5)}\\
O_3 &= \varphi O_2 + \zeta_3 \\
&= \varphi^2 \zeta_1 + \varphi \zeta_2 + \zeta_3. && \text{(C.6)}
\end{aligned}$$

This can be generalized to any occasion specific variable O_t as follows:

$$\begin{aligned}
O_t &= \varphi O_{t-1} + \zeta_t \\
&= \sum_{u=1}^{t} \varphi^{t-u} \zeta_u. && \text{(C.7)}
\end{aligned}$$

Next, to define the variance coefficients, we first need to define the decomposition of the total variance. Note that because of the longitudinal measurement invariance,

$Var(\zeta_1) = Var(\zeta_2) = \cdots = Var(\zeta_t) = Var(\zeta_{t+1}) = \ldots$. Hence, the variance of the latent occasion specific residual ζ_t is simply denoted as $Var(\zeta)$. Thus, the variances of the first three occasion specific variables are defined as follows:

$$Var(O_1) = Var(\zeta) \tag{C.8}$$

$$Var(O_2) = \varphi^2 Var(\zeta) + Var(\zeta) \tag{C.9}$$

$$Var(O_3) = \varphi^4 Var(\zeta) + \varphi^2 Var(\zeta) + Var(\zeta). \tag{C.10}$$

Again, this can be generalized to any occasion specific variable O_t:

$$Var(O_t) = Var(\zeta) \sum_{u=0}^{t-1} \varphi^{2u}. \tag{C.11}$$

The previous result, along with the assumption that the latent indicator-specific traits ξ_j, the occasion-specific variables O_t, and the random measurement errors ε_j are uncorrelated, allows defining the total variance decomposition of an observed variable Y_{jt} (see Equation C.3) as follows:

$$\begin{aligned} Var(Y_{jt}) &= Var(\xi_j) + \lambda_{S_j}^2 Var(\zeta) \sum_{u=0}^{t-1} \varphi^{2u} + Var(\varepsilon_j) \\ &= Var(\xi_j) + \lambda_{S_j}^2 Var(\zeta) \sum_{u=1}^{t-1} \varphi^{2u} + \lambda_{S_j}^2 Var(\zeta) + Var(\varepsilon_j). \end{aligned} \tag{C.12}$$

Given this definition, we can see that the total variance of a variable Y_{jt} is larger than the total variance of the variable $Y_{j,t-1}$ (i.e., $\ldots < Var(Y_{j,t-1}) < Var(Y_{j,t}) < Var(Y_{j,t+1}) < \ldots$). Moreover, if $|\varphi| < 1$, the total variance of Y_{jt} will approach an horizontal asymptote as t increases. As a result, the variance coefficients of the TSO model change over time even when longitudinal measurement invariance is assumed.

Now, let us define the total variance decomposition of the ME-TSO model. To do this, there are two essential changes in relation to the TSO model. Firstly, the latent indicator-specific trait variables ξ_j are now latent indicator- and situation-specific trait variables ξ_{jf}. Secondly, the autoregressive effect φ is allowed to vary across individuals, hence we add an i subscript to it. With these changes, we can define the "total variance" of an observation Y_{ijtf} of person i on indicator j at time t and fixed situation f as follows:

$$Var(Y_{ijtf}) = Var(\xi_{jf}) + \lambda_{S_j}^2 Var(\zeta) \sum_{u=1}^{t-1} \varphi_i^{2u} + \lambda_{S_j}^2 Var(\zeta) + Var(\varepsilon_j). \tag{C.13}$$

However, defining the total variance of an observation is not really meaningful because there is no variability when there is only one datum. Therefore, in the ME-TSO model, we propose to drop the time component by approaching t to infinity and we define the total variance for the time series Y_{ijf} of person i on the j-th indicator at the fixed situation f instead. These changes make sense within the ME-TSO model because there are many measurement occasions and because the model acknowledges the individuals' heterogeneity. Hence, the total variance of Y_{ijf} is defined as follows:

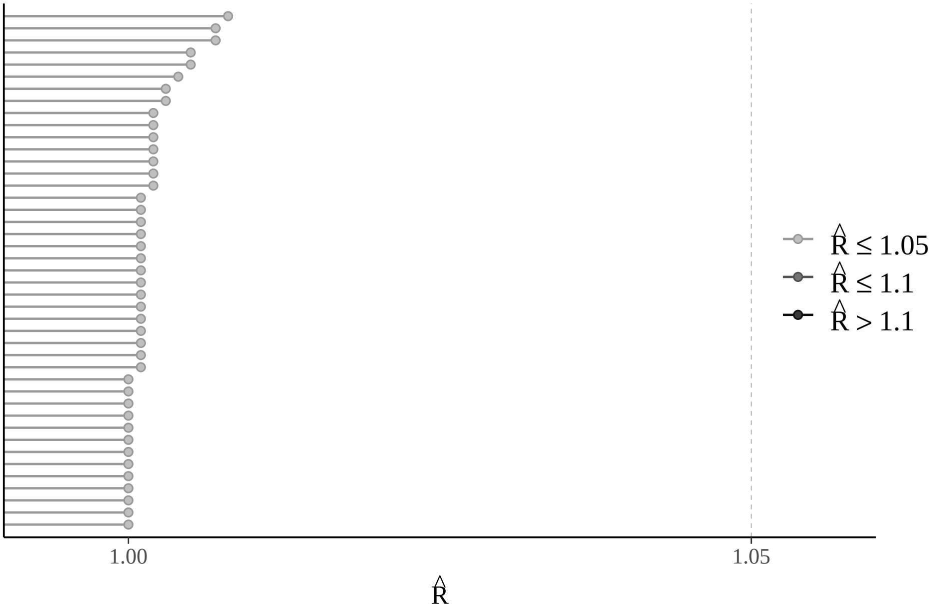
$$Var(Y_{ijf}) = Var(\xi_{jf}) + \lambda_{S_j}^2 Var(\zeta) \sum_{u=1}^{\infty} (\varphi_i^2)^u + \lambda_{S_j}^2 Var(\zeta) + Var(\varepsilon_j). \tag{C.14}$$

This can be simplified, by noticing that $\sum_{u=1}^{\infty}(\varphi_i^2)^u$ is a convergent series if $|\varphi| < 1$. This should be the case, given that the ME-TSO model assumes that the time series are stationary. Thus, when solving for the infinite convergent series, we get that the total variance of Y_{ijf} is:

$$Var(Y_{ijf}) = Var(\xi_{jf}) + \lambda_{S_j}^2 \frac{\varphi_i^2}{1-\varphi_i^2} Var(\zeta) + \lambda_{S_j}^2 Var(\zeta) + Var(\varepsilon_j). \tag{C.15}$$

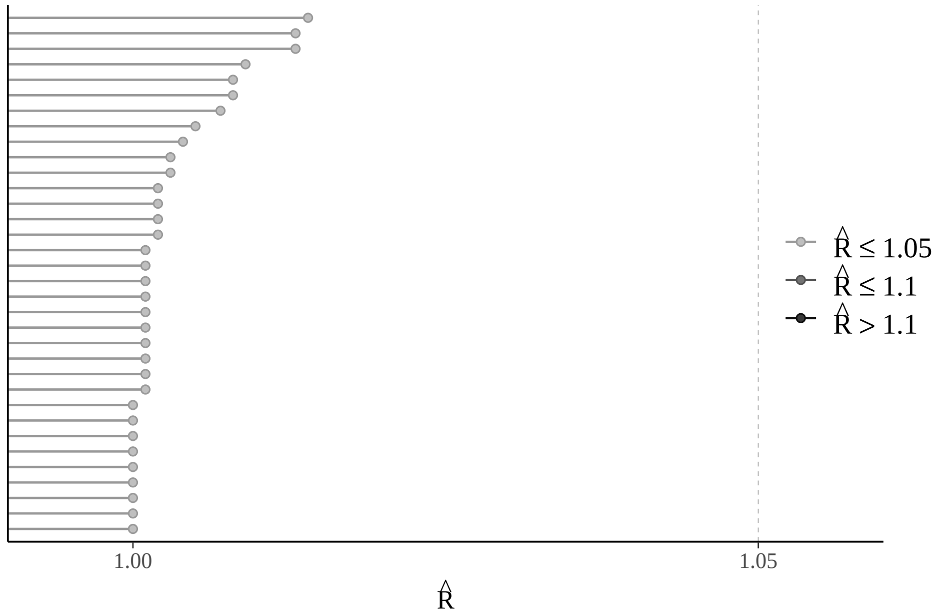
C.3 Model Convergence: Plots of the Gelman-Rubin Statistic

Figure C.2: Gelman-Rubin Statistics of $\mathscr{M}_{1b}$



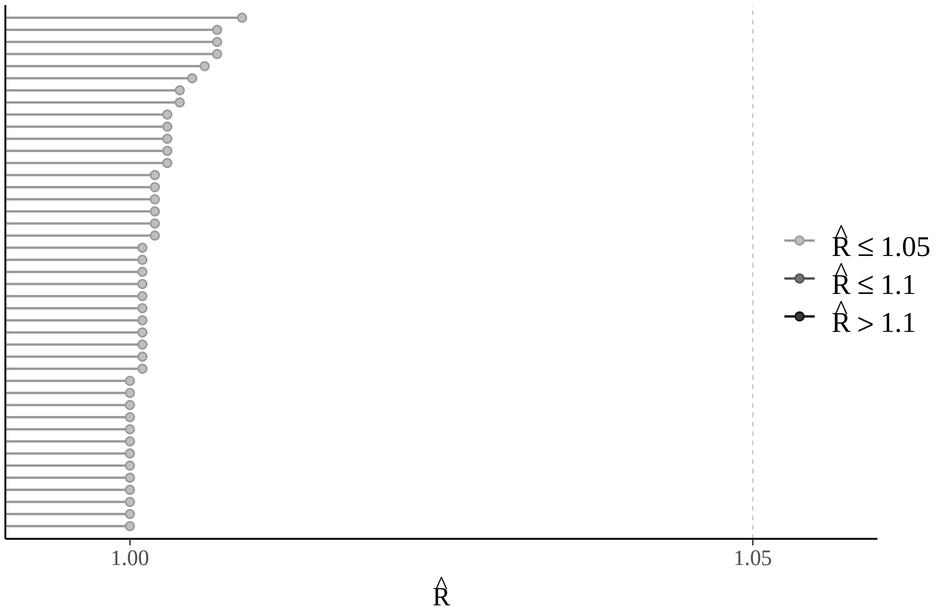
Note. $\mathscr{M}_{1b}$: Model 1b. $\hat{R}$: Gelman-Rubin Statistic.

Figure C.3: Gelman-Rubin Statistics of $\mathscr{M}_2$



Note. $\mathscr{M}_2$: Model 2. $\hat{R}$: Gelman-Rubin Statistic.

Figure C.4: Gelman-Rubin Statistics of $\mathscr{M}_{2b}$



Note. $\mathscr{M}_{2b}$: Model 2b. $\hat{R}$: Gelman-Rubin Statistic.

C.4 Results with Stationary Sample

We tested stationarity with the Kwiatkowski-Phillips-Schmidt-Shin test to study the trend-stationarity in the data (Kwiatkowski et al., 1992). The results of these tests reduced our sample size to 451 individuals once individuals with at least one nonstationary time series were excluded. We repeated the analyses with the reduced sample (see Tables C.1 and C.2). The results did not differ to a large extent in comparison with the analyses with the whole sample reported in the article. Here we present the results with the stationary sample for completeness.

Table C.1: Unstandardized Estimates of the Key Parameters of the Models using the Situation Variable Event

Parameter	$\mathscr{M}_1$ Est. [95% C.I.]	$\mathscr{M}_{1b}$ Est. [95% C.I.]
	Between-level	
Eg-Ev Interaction Effect β_{111}	-0.15 [-0.22,-0.07]	-0.14 [-0.23,-0.04]
En-Ev Interaction Effect β_{121}	-0.17 [-0.24,-0.1]	-0.16 [-0.24,-0.07]
Ch-Ev Interaction Effect β_{131}	-0.23 [-0.31,-0.16]	-0.24 [-0.34,-0.15]
Opt-Eg-Ev Interaction Effect $\beta_{OPT,1}$		-0.07 [-0.36,0.21]
Opt-En-Ev Interaction Effect $\beta_{OPT,2}$		-0.1 [-0.38,0.19]
Opt-Ch-Ev Interaction Effect $\beta_{OPT,3}$		0.01 [-0.3,0.31]
AR Effect Mean $E(\varphi)$	0.29 [0.27,0.31]	0.29 [0.27,0.32]
Eg-Ev Effect Residual Variance $Var(\omega_{EG,11})$	23.76 [13.42,36.58]	23.87 [13.48,37.16]
En-Ev Effect Residual Variance $Var(\omega_{EN,21})$	35.72 [23.66,50.67]	35.7 [23.68,51.06]
Ch-Ev Effect Residual Variance $Var(\omega_{CH,31})$	47.44 [33.59,63.73]	47.64 [33.72,64.53]
AR Effect Variance $Var(\varphi)$	0.030 [0.024,0.038]	0.031 [0.024,0.038]
	Model Fit Information	
Number of Free Parameters	34	43
DIC	1280670.22	1282051.36
pD	99514.70	99594.22

Note. $\mathscr{M}_1$: Model 1, $\mathscr{M}_{1b}$: Model 1b, Est.: Unstandardized estimate, C.I.: Credibility interval, Eg: Energetic, En: Enthusiastic, Ch: Cheerful, Ev: Event, DIC: Deviance information criterion, pD: Estimated number of parameters.

Table C.2: Unstandardized Estimates of the Key Parameters of the Models using the Situation Variable Alone

Parameter	$\mathscr{M}_2$ Est. [95% C.I.]	$\mathscr{M}_{2b}$ Est. [95% C.I.]
	Between-level	
Eg-Al Interaction Effect β_{111}	-0.08 [-0.13,-0.04]	-0.06 [-0.12,0]
En-Al Interaction Effect β_{121}	-0.11 [-0.15,-0.07]	-0.1 [-0.15,-0.05]
Ch-Al Interaction Effect β_{131}	-0.09 [-0.13,-0.05]	-0.08 [-0.13,-0.03]
Opt-Eg-Al Interaction Effect $\beta_{OPT,1}$		-0.11 [-0.3,0.07]
Opt-En-Al Interaction Effect $\beta_{OPT,2}$		-0.08 [-0.25,0.08]
Opt-Ch-Al Interaction Effect $\beta_{OPT,3}$		-0.04 [-0.2,0.13]
AR Effect Mean $E(\varphi)$	0.3 [0.28,0.32]	0.3 [0.28,0.32]
Eg-Al Effect Residual Variance $Var(\omega_{EG,11})$	16.56 [12.23,21.75]	16.27 [11.82,21.61]
En-Al Effect Residual Variance $Var(\omega_{EN,21})$	12.57 [8.91,16.95]	12.41 [8.74,16.67]
Ch-Al Effect Residual Variance $Var(\omega_{CH,31})$	9.77 [6.53,13.79]	9.67 [6.3,13.72]
AR Effect Variance $Var(\varphi)$	0.030 [0.024,0.037]	0.030 [0.024,0.038]
	Model Fit Information	
Number of Free Parameters	34	43
DIC	1356724.47	1358088.51
pD	105463.61	105473.08

Note. $\mathscr{M}_2$: Model 2, $\mathscr{M}_{2b}$: Model 2b, Est.: Unstandardized estimate, C.I.: Credibility interval, Eg: Energetic, En: Enthusiastic, Ch: Cheerful, Al: Alone, DIC: Deviance information criterion, pD: Estimated number of parameters.

Appendix D

In this appendix, we present additional Figures to support the results presented in the Chapter 4. As mentioned in Chapter 4, the time-varying dynamic partial credit model (TV-DPCM) integrates the partial credit model (PCM; Masters, 2016) and the time-varying autoregressive model (TV-AR; Bringmann et al., 2017) into one model. In particular, Figures D.1 through D.6 complement the results of the simulation study. These figures summarize the recovery and accuracy statistics for the latent state dispositions, the attractor, the variance of the innovation, the process variance. Then, Figures D.7 through D.9 present diagnostic plots for the fitted model to the items of self-esteem during phases 1 and 2. Lastly, Table D.1 and Figures D.10 through D.12 show the results from fitting the model to the items of self-esteem during the whole duration of the study.

Figure D.1: Mean Absolute Bias of the Threshold Parameters per Condition.

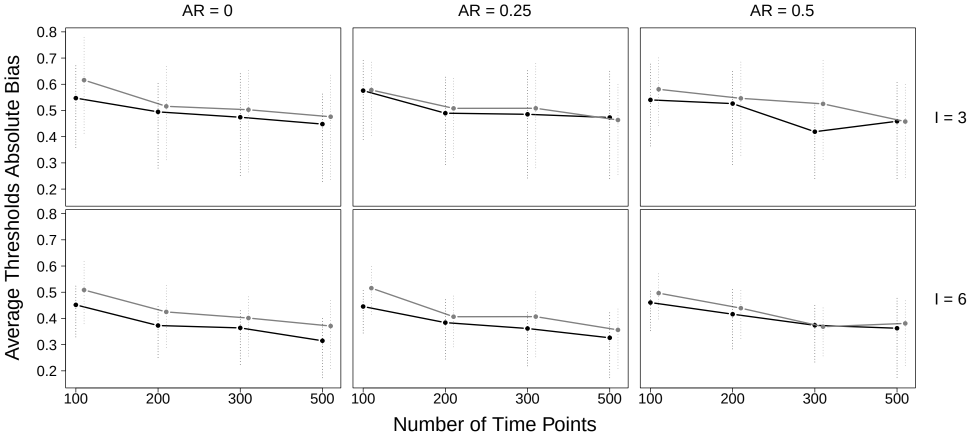


Figure D.2: Mean RMSE of the Threshold Parameters per Condition.

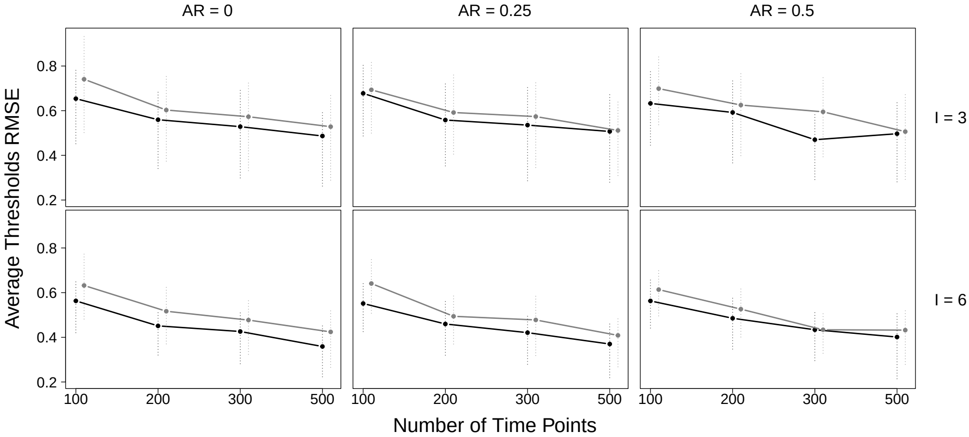
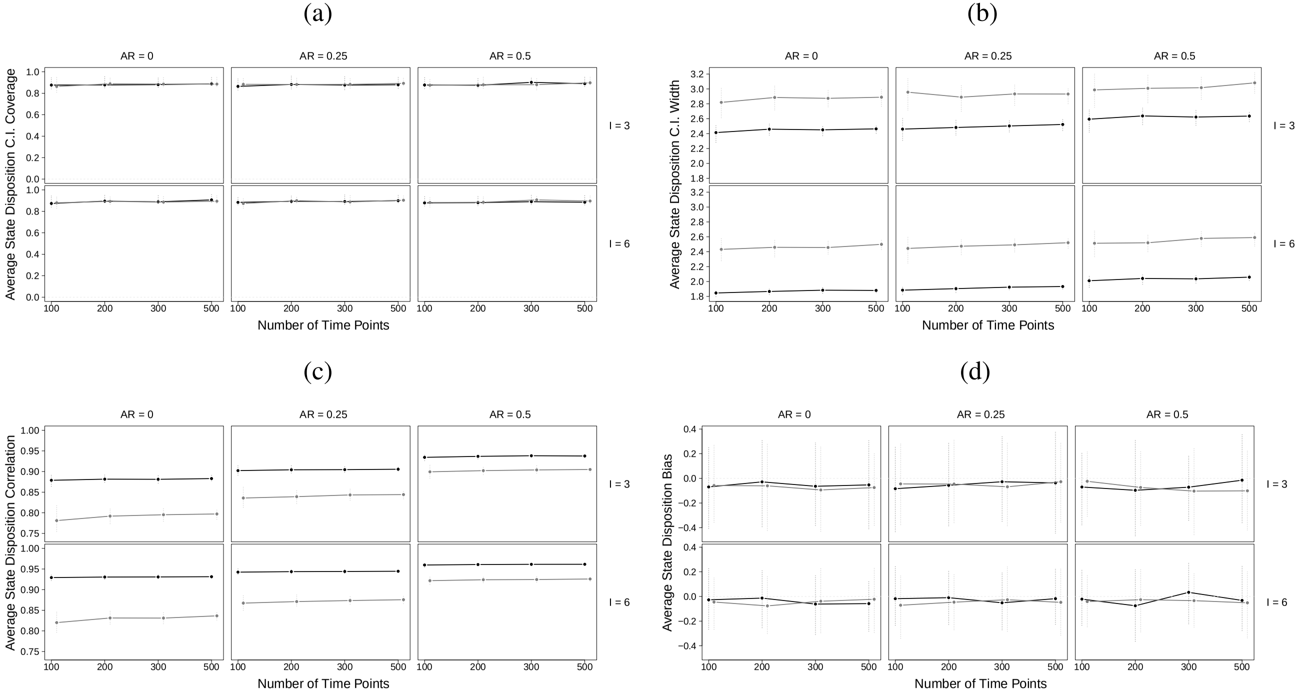
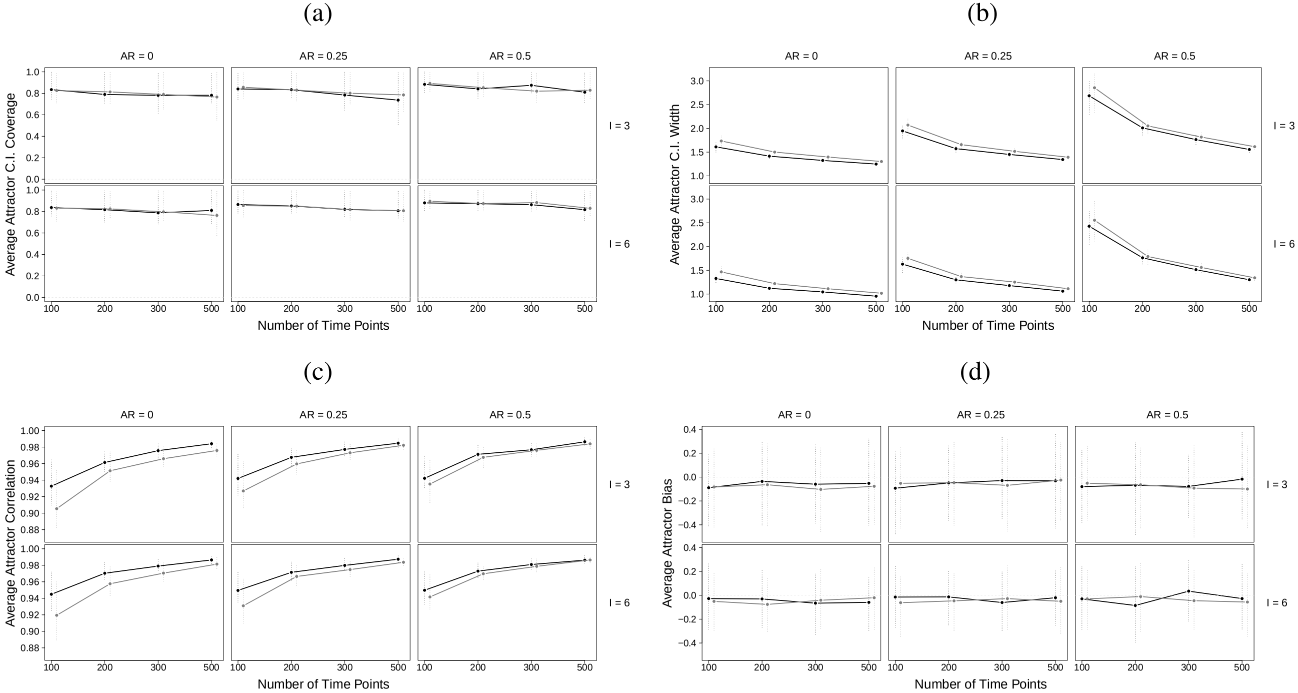


Figure D.3: Parameter Recovery and Accuracy Statistics of the Latent State Disposition Parameters.



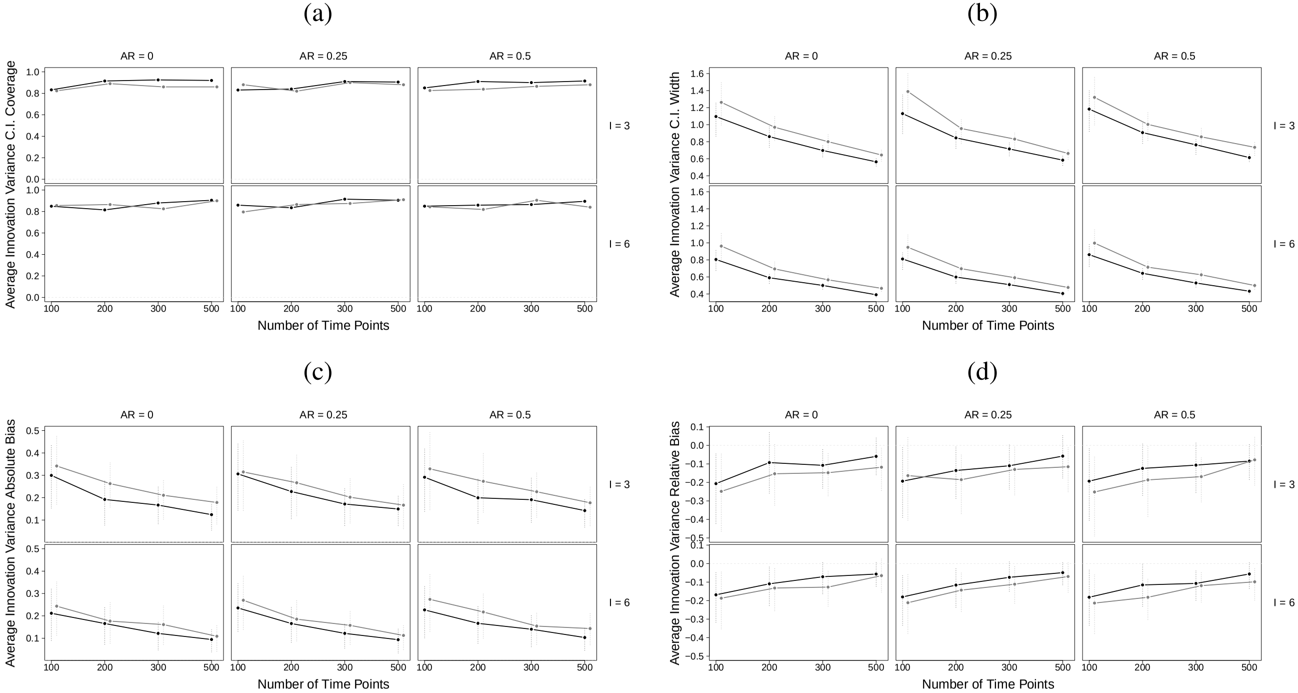
Note. The black lines represent the conditions where there were no missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average correlation between the true and the estimated thresholds, and (D) average bias per condition.

Figure D.4: Parameter Recovery and Accuracy Statistics of the Attractor.



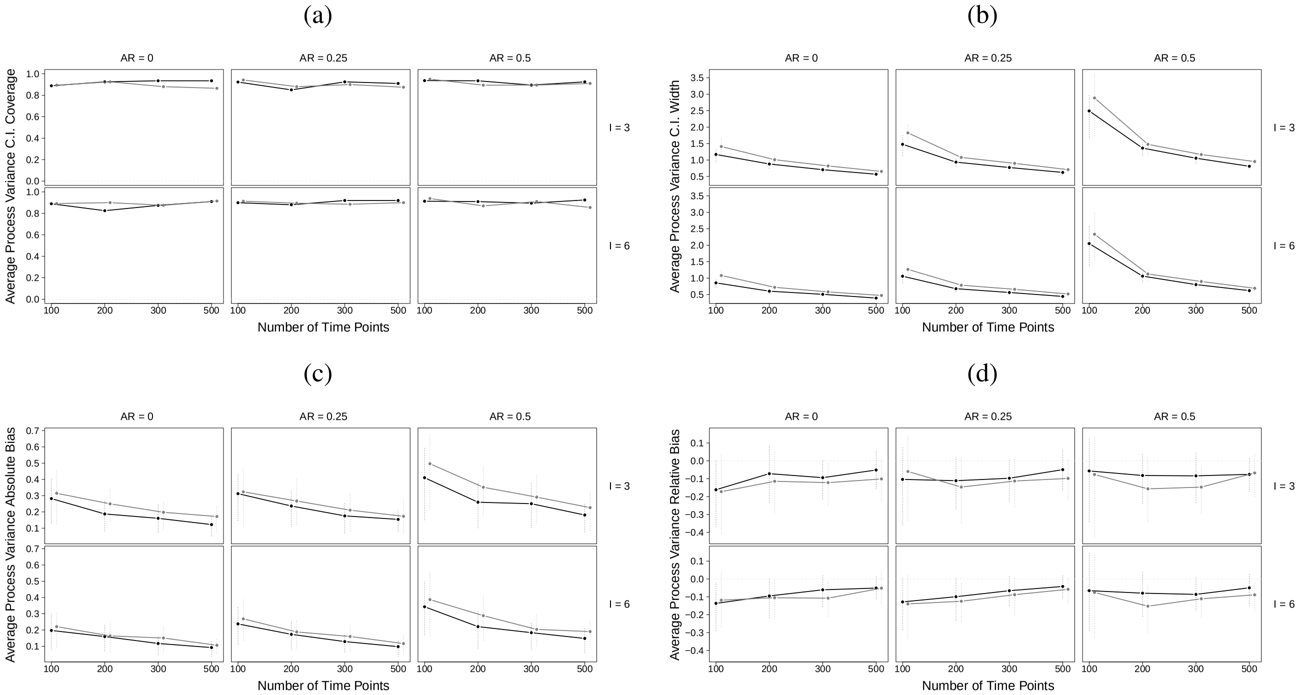
Note. The black lines represent the conditions where there were no missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average correlation between the true and the estimated thresholds, and (D) average bias per condition.

Figure D.5: Parameter Recovery and Accuracy Statistics of the Variance of the Innovations.



Note. The black lines represent the conditions where there was not missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average absolute bias, and (D) average relative bias per condition.

Figure D.6: Parameter Recovery and Accuracy Statistics of the Process Variance.



Note. The black lines represent the conditions where there was not missing data and the gray lines represent the conditions with 30% missing values. The vertical dotted lines around the dots represent the interquartile range per condition. (A) Coverage proportion, (B) average width of the credibility interval, (C) average absolute bias, and (D) average relative bias per condition.

Figure D.7: Gelman-Rubin Statistics for the Estimated Parameters of the TV-DPCM

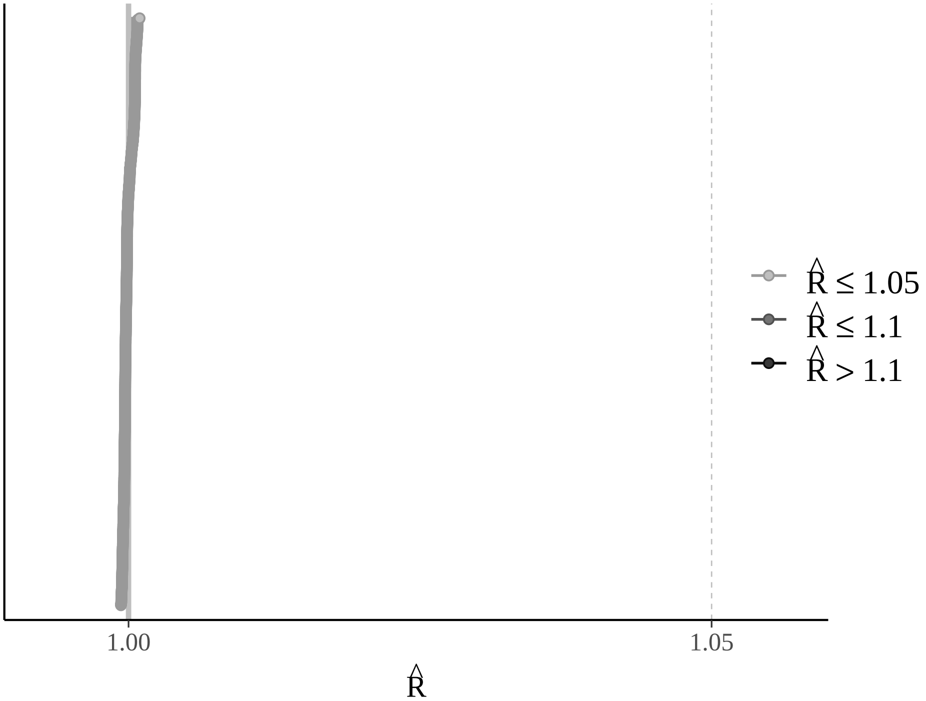


Figure D.8: Traceplots of Selected Parameters

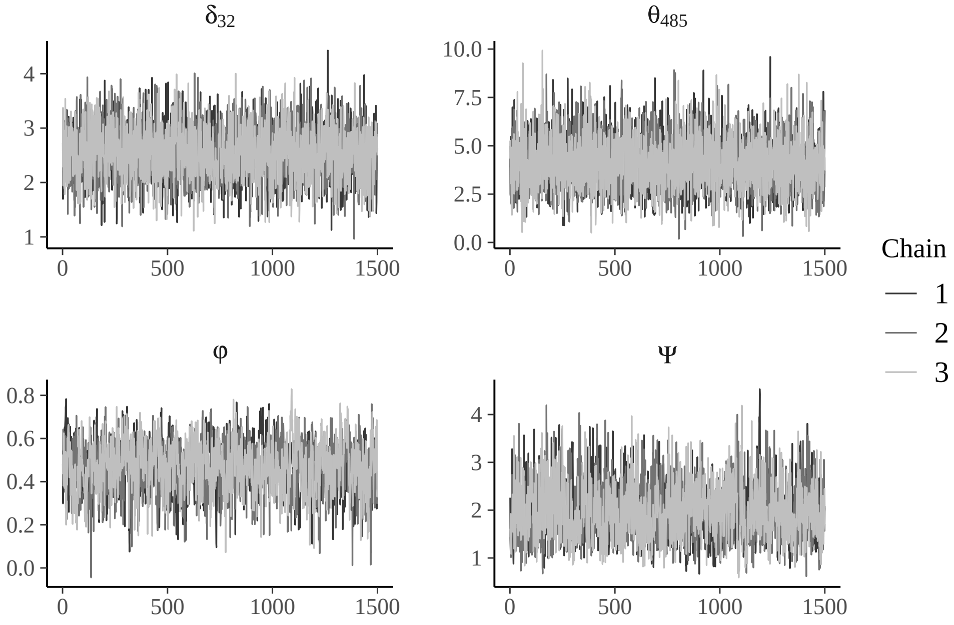


Figure D.9: Autocorrelation Plots of Selected Parameters

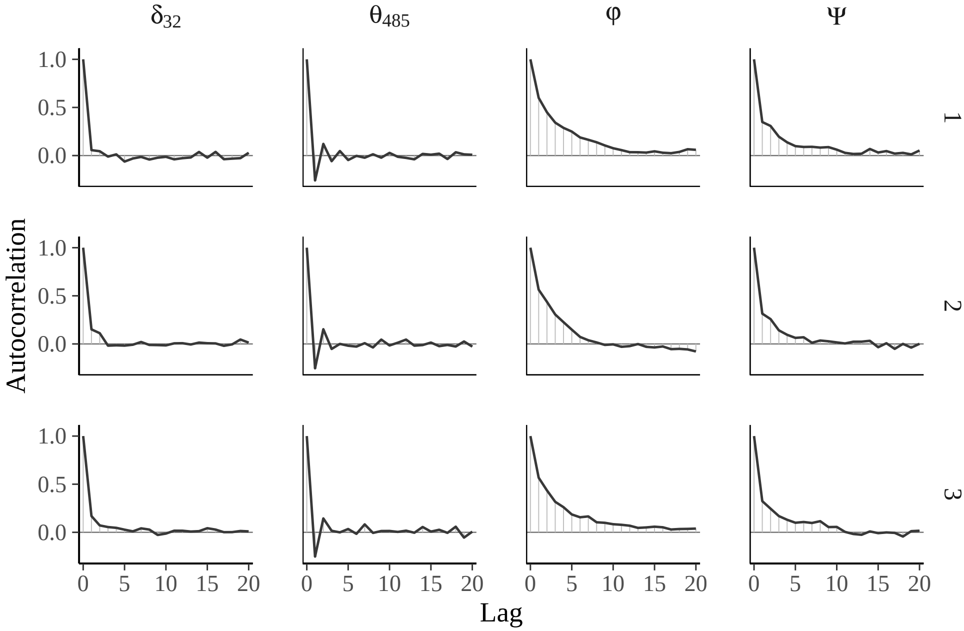
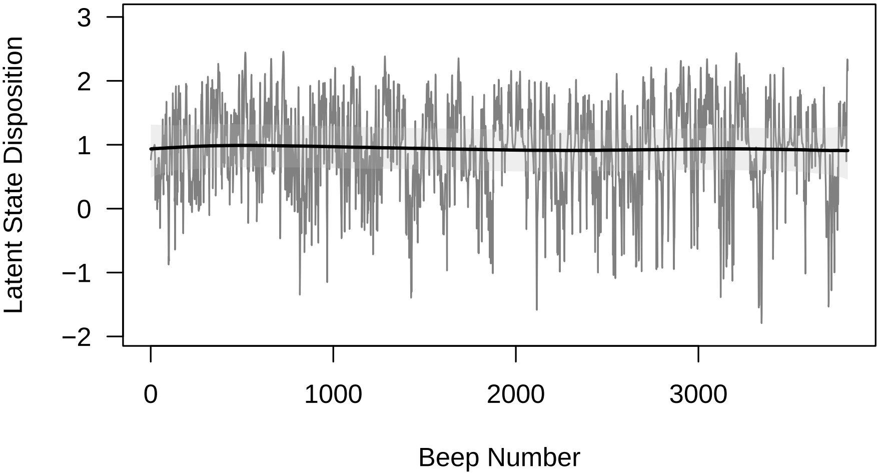


Table D.1: Estimated Parameters of the TV-DPCM Fitted to the Items of Self-Esteem during Phases 1 through 5

	Median	SD	C.I.	ESS
$\hat{\delta}_{11}$	-2.93	0.43	$(-3.76, -2.06)$	1273
$\hat{\delta}_{12}$	0.75	0.40	$(-0.04, 1.54)$	979
$\hat{\delta}_{21}$	-2.27	0.43	$(-3.1, -1.43)$	1211
$\hat{\delta}_{22}$	-0.93	0.40	$(-1.7, -0.15)$	1131
$\hat{\delta}_{31}$	0.78	0.40	$(-0.01, 1.55)$	1026
$\hat{\delta}_{32}$	3.42	0.42	$(2.6, 4.25)$	986
$\hat{\varphi}$	0.71	0.03	$(0.65, 0.77)$	693
$\hat{\Psi}$	3.44	0.44	$(2.65, 4.4)$	988
$\hat{\sigma}^2$	7.03	0.76	$(5.71, 8.71)$	1299

Note. C.I. = 95% central credible interval.

Figure D.10: Estimated Latent Dynamic Process



Note. The estimated latent state dispositions for each beep (observed and missing) including phases 1 through 5 are represented with the gray line. The trend of the dynamic process or attractor is represented with the black line alongside with its 95% credibility interval band in light gray.

Figure D.11: Item Characteristic Functions for the Items of Self-Esteem during Phases 1 through 5

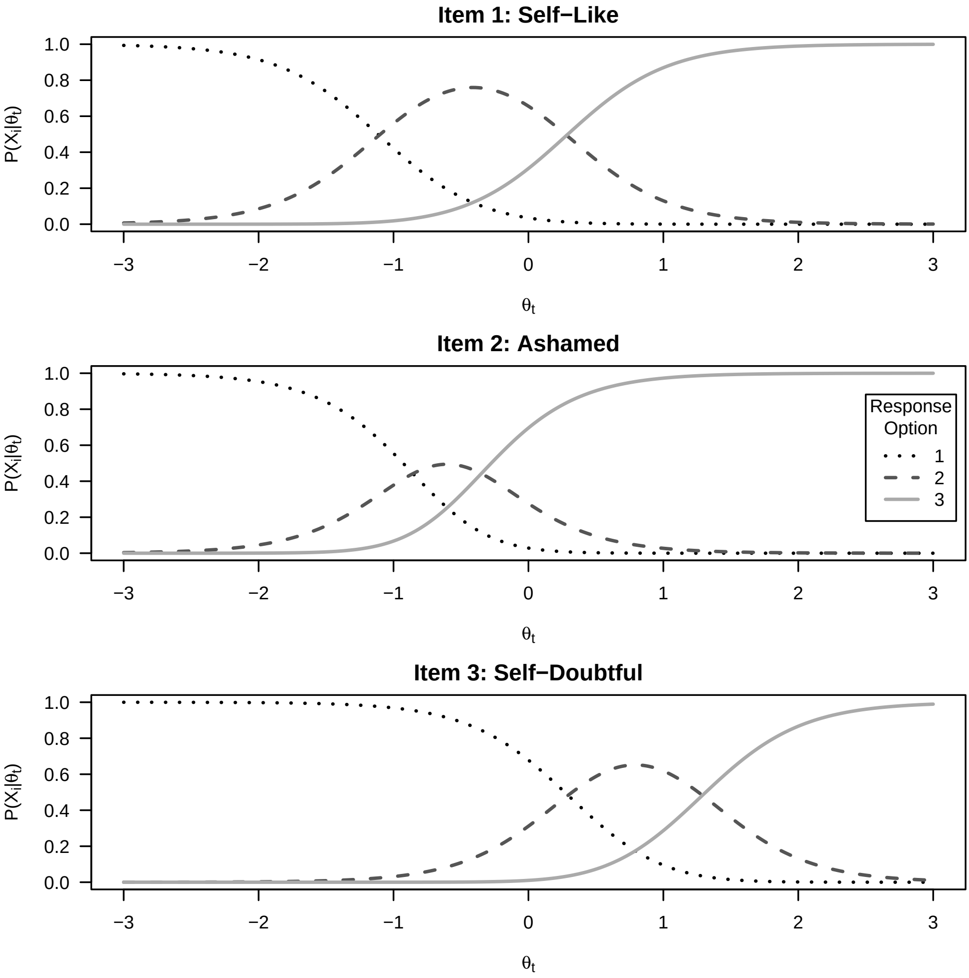
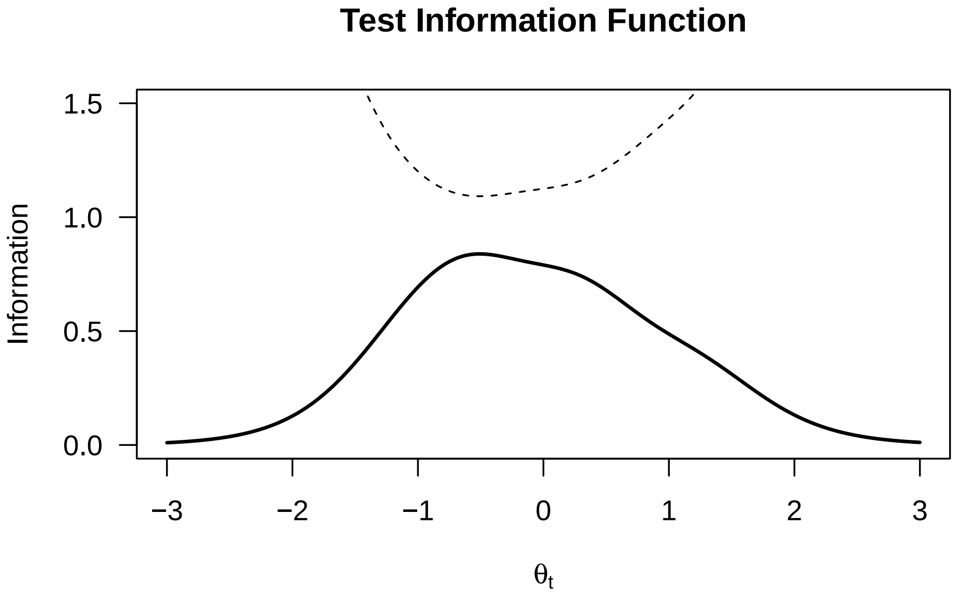


Figure D.12: Test Information Function when Analyzing Phases 1 through 5



Samenvatting

Het gebruik van smartphones en andere elektronische apparaten heeft het verzamelen van data voor het bestuderen van psychologische dynamiek vergemakkelijkt. Tegenwoordig is het relatief eenvoudig om proefpersonen gedurende weken of zelfs maanden drie of meer keer per dag een korte vragenlijst te laten invullen. Dit resulteert in tientallen metingen per persoon. Dit soort data staan bekend als intensieve longitudinale data of tijdreeks data. Meten in de psychologie is echter verre van eenvoudig, en het meten van psychologische processen over tijd, voegt nog een extra moeilijkheidsgraad toe. Dit maakt het bijzonder belangrijk om duidelijkheid te verschaffen over wat precies wordt gemeten.

In dit proefschrift behandel ik de vraag “wat meten we?” vanuit twee hoofdonderwerpen. Een eerste onderwerp is het onderscheid tussen “traits” en “states”. Deze twee begrippen komen onvermijdelijk in beeld bij intensieve longitudinale data omdat ze traditioneel gebruikt worden om respectievelijk de stabiliteit en de variabiliteit van psychologische constructen te beschrijven. Toch zijn er in de literatuur verschillende definities van “traits” en “states” te vinden (ziet Endler & Magnusson, 1976; Hamaker et al., 2007; Nezlek, 2007; Steyer et al., 1999). Echter, duidelijke definities van wat wij onder “traits” en “states” verstaan is zeer belangrijk voor het gebied van dynamische psychologische processen om data en theorie beter met elkaar in verband te brengen en te begrijpen wat er gemeten wordt. Ten tweede besteed ik bijzondere aandacht aan het begrip meetfout. Telkens wanneer wij iets meten, gaat onze meting gepaard met meetfouten. Begrijpen in welke mate onze metingen meetfouten vertonen houdt rechtstreeks verband met het begrijpen van wat wij meten en hoe betrouwbaar onze metingen zijn. Meetfouten zijn vooral belangrijk bij psychologische metingen omdat de meeste constructen in de psychologie niet direct waarneembaar zijn. Ondanks het belang ervan is de meetfout echter zelden een thema in intensief longitudinaal onderzoek (Schuurman & Hamaker, 2019; Schuurman et al., 2015).

Om deze lacune in de literatuur op te vullen, houd ik in onze modellen wel expliciet rekening met meetfouten in intensief longitudinaal onderzoek op basis van de latente state-trait theorie (LST; Steyer et al., 2015; Steyer et al., 1999) en de itemresponstheorie (IRT; Embretson & Reise, 2013).

Dit proefschrift bestaat in totaal uit 6 hoofdstukken, waaronder een algemene inleiding, vier onderzoekspapers en één algemene discussie. In de algemene inleiding presenteer ik de theoretische achtergrond met betrekking tot (a) het belang van het bestuderen van psychologische dynamiek, (b) het onderscheid tussen "traits" en "states", (c) meetfouten, en (d) een schets van het proefschrift. Nu volgt een korte samenvatting van elk van de volgende hoofdstukken, te beginnen met hoofdstuk 2.

Hoofdstuk 2

In dit hoofdstuk hebben we onderzocht hoe we "traits" en "states" kunnen onderscheiden in intensieve longitudinale gegevens met drie populaire longitudinale structural equation modellen (SEM): Het multistate-singletrait model (MSST; Steyer et al., 2015; Steyer et al., 1999), het common and unique trait state model (CUTS; Hamaker et al., 2017), en het trait-state-occasion model (TSO; Eid et al., 2017). Aangezien deze modellen werden ontwikkeld om longitudinale gegevens met een klein aantal tijdspunten te analyseren, werden de modellen geherformuleerd als multilevel SEM modellen om de analyse van intensieve longitudinale gegevens te vergemakkelijken. De modellen werden getest in een simulatiestudie om hun prestaties in termen van convergentie en schatting van de parameters onder verschillende omstandigheden te beoordelen. In het algemeen waren de prestaties van de multilevel versies van de modellen bevredigend en presteerde de multilevel versie van de TSO het best in alle omstandigheden. Bovendien heben we geïllustreerd hoe de resultaten van deze modellen kunnen worden ingepast en geïnterpreteerd aan de hand van de analyse van gegevens uit het project HoeGekIsNL (van der Krieke et al., 2017; van der Krieke et al., 2016). In dit hoofdstuk hebben we ook besproken hoe deze modellen zich verhouden tot andere modellen en raamwerken die gebruikt worden om intensieve longitudinale data te bestuderen, zoals dynamische factoranalyse benaderingen en het dynamische SEM.

Hoofdstuk 3

In dit hoofdstuk hebben we een uitbreiding voorgesteld van de multilevel versie van het TSO-model dat we het mixed-effect trait-state-occasion model (ME-TSO) noem.

Deze uitbreiding is gebaseerd op de random en fixed situation benadering van de LST theorie (Geiser et al., 2015b) en het dynamic structural equation modeling framework (Asparouhov et al., 2018). Het doel van deze uitbreiding was om het model algemener te maken en beter geschikt voor de analyse van intensieve longitudinale data. Ten eerste kan het TSO-model, door de integratie van de "random en fixed" situatiebenadering, nu rekening houden met de context van de situatie en de interactie tussen de persoon en de situatie identificeren. In een intensieve longitudinale studie kunnen we bijvoorbeeld geïnteresseerd zijn in het effect van het alleen zijn versus het verkeren in een sociale situatie op het positieve affect van individuen. Dit effect kan worden bestudeerd met het voorgestelde ME-TSO-model. Bovendien erkennen we in de ME-TSO de heterogeniteit van de psychologische dynamiek, daarom hebben we toegestaan dat het autoregressieve effect varieert tussen de individuen, wat betekent dat voor elke proefpersoon een ander autoregressief effect wordt geschat. Bijgevolg mochten de variantiecoëfficiënten, die essentieel zijn voor de interpretatie van LST-modellen, ook variëren tussen individuen. Om te laten zien hoe de ME-TSO kan worden ingepast en hoe de resultaten kunnen worden geïnterpreteerd, hebben we ook gegevens gebruikt van het HoeGekIsNL project (van der Krieke et al., 2017; van der Krieke et al., 2016).

Hoofdstuk 4

De meeste statistische modellen en raamwerken die zijn ontwikkeld om psychologische dynamische processen te analyseren, vereisen continue data. Psychologische onderzoekers gebruiken echter niet altijd continue schalen om constructen te meten. Juist het gebruik van geordende categorische schalen, zoals Likert-schalen, is zeer populair in intensief longitudinaal onderzoek (Vachon et al., 2019). Om deze reden en in navolging van het idee van meetmodellen voor intensieve longitudinale gegevens, hebben we in dit hoofdstuk een IRT-model voorgesteld voor de analyse van psychologische tijdreeksen. Wij noemen dit model het time-varying partial credit model (TV-DPCM). IRT modellen zijn niet lineaire modellen die speciaal zijn ontwikkeld voor het analyseren van schalen met dichotome en polytome items. Dit raamwerk maakt het mogelijk om de interactie tussen de personen en de items te bestuderen, en de eigenschappen en kwaliteit van de gebruikte items en schalen te analyseren. Het doel van het TV-DPCM is het analyseren van psychologische tijdreeksen van één proefpersoon wanneer een set van likert-schaal items wordt gebruikt om een construct te meten zoals negatief affect. Het TV-DPCM integreert het partiële credit model (Masters, 2016) en het tijdsvariërende autoregressieve model (Bringmann et al., 2017).

Dit betekent dat het model kan omgaan met niet-stationaire tijdreeksen wanneer het latente dynamische proces een niet-lineaire trend volgt. We hebben de prestaties van het TV-DPCM beoordeeld in een simulatiestudie en het model geïllustreerd met een empirisch voorbeeld. Bovendien, door gebruik te maken van bijkomende kenmerken van IRT zoals de item karakteristieke functies, de item informatie functies, en de test informatie functie, laten we zien hoe het TV-DPCM kan worden gebruikt om een uitgebreide analyse te maken van de schalen in intensief longitudinaal onderzoek. Door bijvoorbeeld de items die “eigenwaarde" meten te analyseren, tonen we aan dat deze set items informatiever was bij het meten van lagere niveaus van eigenwaarde, maar dat de deelnemer meestal gemiddeld een hoog niveau van eigenwaarde ervoer. Dit wees erop dat meer items die hoge niveaus van eigenwaarde meten nodig waren om de eigenwaarde van de deelnemer accuraat te kunnen meten.

Hoofdstuk 5

Bij de toepassing van nieuwe modellen kan het voor onderzoekers onduidelijk zijn hoe zij de passing (goodness-of-fit) van het model moeten beoordelen. In dit hoofdstuk hebben we dit probleem aangepakt voor het TV-DPCM methode. Aangezien het TV-DPCM binnen het Bayesiaanse raamwerk is geïmplementeerd, hebben we ons geconcentreerd op de implementatie van de posterior predictive model checking methode (PPMC; Gelman & Rubin, 1992; Rubin, 1984) voor het beoordelen van de goodness-of-fit van het TV-DPCM. De PPMC-methode bestaat uit het vergelijken van kenmerken van de waargenomen gegevens met kenmerken van “gerepliceerde” gegevens op basis van het gepaste model. Als de verschillen tussen de waargenomen gegevens en de “gerepliceerde” gegevens te groot zijn, dan zijn er aanwijzingen voor modelmisfit. Om de PPMC-methode voor het TV-DPCM te implementeren, hebben we teststatistieken en discrepantiemetingen gebruikt en aangepast die gebruikt zijn om de goodness-of-fit van andere populaire IRT-modellen te beoordelen (Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011). We hebben deze teststatistieken en discrepantiematen getest met verschillende passende en niet passende gesimuleerde data. De resultaten van deze analyses toonden aan onder welke omstandigheden deze goodness-of-fit maten doeltreffend waren om modelmisfit vast te stellen.

Hoofdstuk 6

Tenslotte presenteer ik in dit laatste hoofdstuk een algemene discussie over dit proefschrift. Eerst belicht ik de belangrijkste bevindingen van ons onderzoek in de voorgaande hoofdstukken. Daarna bespreek ik het belang van duidelijke definities van wat we verstaan onder “traits” en “states” om meer duidelijkheid te hebben over wat we meten in intensieve longitudinale data. Vanuit een theoretisch oogpunt denk ik dat de LST-theorie kan bijdragen aan een beter begrip van metingen in intensief longitudinaal onderzoek. Ik benadruk ook dat IRT nog niet vaak is toegepast in intensieve longitudinale settings, maar beragumnteer ook de voordelen van een IRT benaderingen in dit veld. Bijvoorbeeld, zoals aangetoond in dit proefschrift, IRT benaderingen maken het mogelijk om de eigenschappen en kwaliteit van de items en de schalen die gebruikt worden in intensief longitudinaal onderzoek diepgaand te bestuderen. Tenslotte erken ik dat onderzoekers nog voor veel uitdagingen kunnen komen te staan die niet kunnen worden aangepakt met de statistische modellen die in dit proefschrift worden gepresenteerd. Desalniettemin legt dit onderzoek een aantal grondslagen voor het omgaan met meetfouten in de context van psychologische dynamische processen.

Summary

The use of smartphones and other electronic devices has facilitated the collection of data to study psychological dynamics. Nowadays, it is relatively simple to get participants to complete a short questionnaire three or more times a day for weeks or even months, resulting in dozens of measurements from the same individual. This kind of data is known as intensive longitudinal data or time series data. However, measurement in psychology is far from easy, and the fact that we now measure psychological processes over time adds another layer of difficulty. This makes it especially important to have clarity on what exactly is being measured.

In this thesis, to address the question “what are we measuring?”, I focus on two main topics. One first topic is the distinction between “traits” and “states”. These two concepts inevitably come to mind when dealing with intensive longitudinal data, as they have traditionally been used to describe the stability and the variability of psychological constructs, respectively. Yet, several definitions of traits and states can be found in the literature (see Endler & Magnusson, 1976; Hamaker et al., 2007; Nezlek, 2007; Steyer et al., 1999). I consider that having clear definitions of what we understand for traits and states is important for the field of psychological dynamics in order to better relate the data with the theory and to understand what we are measuring. Then, I give special attention to the concept of measurement error. Whenever we measure something, our measurement is typically affected by sources other than the target of our measure. This ‘noise’ in the data is commonly referred to as measurement error. Understanding to what extent our measurements are affected by measurement error is directly associated with understanding what we are measuring. Measurement error is specially important in psychological measurement because most of what is measured in psychology is non observable. Hence, measurement error tends to be considered a major threat. However, in spite of its importance, measurement error has rarely been addressed in intensive longitudinal research (Schuurman & Hamaker, 2019; Schuurman et al., 2015). Hence, to fill this gap in the literature, I account for measurement

error in intensive longitudinal data based on the latent state-trait theory (LST; Steyer et al., 2015; Steyer et al., 1999) and the item response theory (IRT; Embretson & Reise, 2013).

This thesis is composed by a total of 6 chapters including a general introduction, four chapters that are published articles or manuscripts of articles submitted for review or in preparation, and one general discussion. The general introduction presents the theoretical background regarding (a) the importance of studying psychological dynamics, (b) the distinction between traits and states, (c) measurement error, and (d) an outline of the dissertation. Next, I also present a brief summary of each of the following chapters, starting from Chapter 2.

Chapter 2

In this chapter, we studied how to distinguish traits and states in intensive longitudinal data with three popular longitudinal structural equation models: The multistate-singletrait model (MSST; Steyer et al., 2015; Steyer et al., 1999), the common and unique trait state model (CUTS; Hamaker et al., 2017), and the trait-state-occasion model (TSO; Eid et al., 2017). Given that these models were developed to analyze longitudinal data with a few number of time points, the models were reformulated as multilevel structural equation models to facilitate the analysis of intensive longitudinal data. The models were tested in a simulation study to assess their performance in terms of convergence and recovery of the parameters under different conditions. In general, the performance of the multilevel versions of the models was satisfactory and the multilevel version of the TSO performed the best across all conditions. Moreover, we illustrated how to fit and interpret the results of these models by analyzing data from the project HowNutsAreTheDutch (van der Krieke et al., 2017; van der Krieke et al., 2016). In this chapter, we also discussed how these models are related to other models and frameworks used to study intensive longitudinal data, such as dynamic factor analysis approaches and the dynamic structural equation modeling framework.

Chapter 3

In this chapter, we proposed an extension of the multilevel version of the TSO model that we refer to as the mixed-effect trait-state-occasion model (ME-TSO). This extension is based on the random and fixed situation approach of the LST theory (Geiser et al., 2015b) and the dynamic structural equation modeling framework (Asparouhov et al., 2018). The purpose of this extension was to make the model more general and better suited to analyze intensive longitudinal data. First, by incorporating the random and fixed situation approach, the ME-TSO model can take the context of the situation into account and identify the interaction between the person and the situation. For example, in an intensive longitudinal study, we might be interested in how being alone versus being in a social situation has an effect in the positive affect of the individuals. This effect can be studied with the proposed ME-TSO model. Furthermore, in the ME-TSO, we acknowledge the heterogeneity of psychological dynamics. For this reason, we allowed the autoregressive effect to vary randomly across individuals, meaning that a different autoregressive effect is estimated for each subject. As a consequence, the variance coefficients, which are key for the interpretation of LST models, were also allowed to vary across individuals. To show how to fit the ME-TSO and interpret its results, we also used data from the HowNutsAreTheDutch project (van der Krieke et al., 2017; van der Krieke et al., 2016).

Chapter 4

Most of the statistical models and frameworks developed to analyze psychological dynamics require continuous data. However, psychological researchers not always use continuous scales to measure the constructs of interest. In fact, using ordered categorical scales like Likert scales is highly popular in intensive longitudinal research (Vachon et al., 2019). For this reason and following the idea of measurement models for intensive longitudinal data, in this chapter we proposed an IRT model for the analysis of psychological time series. We refer to this model as the time-varying partial credit model (TV-DPCM). In a nutshell, IRT models are non linear models that have been especially developed to analyze scales with dichotomous and polytomous items. This framework allows studying the interaction between the persons and the items, and analyzing the properties and quality of the items and scales used. The purpose of the TV-DPCM is to allow analyzing psychological time series of one subject when a set of likert-scale items was used to measure a construct such as negative affect. The TV-DPCM integrates the partial credit model (Masters, 2016) and the time-varying autoregressive model (Bringmann et al., 2017). This means that the model can handle non-stationary time series when the latent dynamic process follows a non linear trend. We assessed the performance of the TV-DPCM in a simulation study and illustrated the model with an empirical example. Moreover, by using additional features of IRT such as the item characteristic function, the item information function, and the test information function, we show how the TV-DPCM can be used to make a comprehensive analysis of the scales used in intensive longitudinal research. For example, by analyzing the items of self-esteem, we showed that this set of items was more informative at measuring lower levels of self-esteem, but the participant mostly experienced medium and high levels of self-esteem. This indicated that more items measuring high levels of self-esteem were needed to accurately measure the participant's self-esteem.

Chapter 5

When applying new models, practitioners might find that it is unclear how they are supposed to assess the goodness-of-fit of the model. In this chapter, we tackled this issue for the TV-DPCM. Given that the TV-DPCM was implemented within the Bayesian framework, we focused on implementing the posterior predictive model checking method (PPMC; Gelman & Rubin, 1992; Rubin, 1984) for assessing the goodness-of-fit of the TV-DPCM. In a nutshell, the PPMC method consists of comparing features of the observed data with features of replicated data based on the fitted model. If the differences between the observed data and the replicated data are too large, then one infers that there is evidence of model misfit. To implement the PPMC method for the TV-DPCM, we used and adapted test statistics and discrepancy measures that have been used to assess the goodness-of-fit of other popular IRT models (Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011). We tested these test statistics and discrepancy measures with a handful of fitting and misfitting simulated data. The results from these analyses showed under which conditions these goodness-of-fit measures were effective in determining model misfit.

Chapter 6

In this final chapter, I present a general discussion of this thesis. First, I highlight the main findings of my research across the previous chapters. Then, I further discuss the importance of having clear definitions of what we understand for traits and states, to have more clarity about what we measure in intensive longitudinal data. Personally, I think that from a theoretical point of view, the LST theory can contribute to a better understanding of measurements in intensive longitudinal research. I also highlight that IRT has remained relatively unexplored in intensive longitudinal settings, but there is much that IRT approaches can offer to the field. For example, as shown in this thesis, IRT approaches allow a thorough understanding of the properties and quality of the items and scales used in intensive longitudinal research. Finally, I acknowledge that there are still many challenges that researchers may face that cannot be addressed with the statistical models presented in this thesis. Nonetheless, this research sets some foundations for accounting for measurement error in the field of psychological dynamics.

Resumen

El uso de teléfonos inteligentes y otros dispositivos electrónicos ha facilitado la recolección de datos para estudiar procesos psicológicos dinámicos. Hoy en día, es relativamente fácil hacer que los participantes completen un breve cuestionario tres o más veces al día durante semanas o incluso meses, lo que da como resultado docenas de mediciones del mismo individuo. Este tipo de datos se conoce como datos longitudinales intensivos o datos de series tiempo. Sin embargo, la medición en psicología no es un proceso sencillo y el hecho de que ahora midamos procesos psicológicos a lo largo del tiempo agrega otro nivel de dificultad. Por lo tanto, es especialmente importante tener claridad sobre qué estamos midiendo exactamente.

En esta tesis, para responder la pregunta“¿qué estamos midiendo?", nos enfocamos en dos temas. En primer lugar consideramos la distinción entre“rasgos" y “estados". Estos dos conceptos necesariamente tienen que ser discutidos cuando se trabaja con datos longitudinales intensivos, ya que tradicionalmente se han utilizado para describir la estabilidad y la variabilidad de los constructos psicológicos. Teniendo en cuenta que se han propuesto múltiples definiciones de estos conceptos en la literatura (ver Endler & Magnusson, 1976; Hamaker et al., 2007; Nezlek, 2007; Steyer et al., 1999), consideramos que es importante clarificar que es lo que entendemos por rasgos y estados para el estudio de procesos psicológicos dinámicos con el fin de poder relacionar mejor los datos con la teoría y comprender mejor lo que estamos midiendo. En segundo lugar, nos enfocamos en el concepto de “error de medición”. Cada vez que medimos algo, nuestras mediciones se suelen ver afectadas por fuentes distintas de lo que realmente queremos medir. Este ‘ruido’ en los datos se conoce comúnmente como error de medición. De esta manera, comprender hasta qué punto nuestras mediciones se ven afectadas por el error de medición está directamente asociado con comprender qué estamos midiendo. El error de medición es especialmente importante en la medición en psicología porque la mayor parte de los fenómenos psicológicos que intentamos medir no son observables. Por lo tanto, el error de medición

tiende a considerarse una amenaza importante a la validez de nuestros resultados. Sin embargo, a pesar de su importancia, el error de medición rara vez es tenido en cuenta en las investigaciones con datos longitudinales intensivos (Schuurman & Hamaker, 2019; Schuurman et al., 2015). Así, para estudiar el error de medición en datos longitudinales intensivos, proponemos diferentes modelos basados en la teoría del estado-rasgo latente (LST; Steyer et al., 2015; Steyer et al., 1999) y la teoría de respuesta al ítem (IRT; Embretson & Reise, 2013).

Esta tesis está compuesta por un total de 6 capítulos que incluyen una introducción general, cuatro capítulos que son artículos publicados o manuscritos de artículos presentados para revisión de pares o en preparación, y una discusión general. La introducción general presenta los antecedentes teóricos con respecto a (a) la importancia de estudiar los procesos psicológicos dinámicos, (b) la distinción entre rasgos y estados, (c) el error de medición y (d) un resumen de la esta tesis. A continuación, también presentamos un breve resumen de cada uno de los siguientes capítulos, comenzando por el Capítulo 2.

Capítulo 2

En este capítulo, estudiamos cómo distinguir rasgos y estados en datos longitudinales intensivos con tres modelos de ecuaciones estructurales longitudinales: El modelo de multiestado y rasgo único (MSST; Steyer et al., 2015; Steyer et al., 1999), el modelo de rasgo y estado - común y único (CUTS; Hamaker et al., 2017), y el modelo rasgo-estado-ocasión (TSO; Eid et al., 2017). Debido a que estos modelos se desarrollaron para analizar datos longitudinales con un número reducido de medidas repetidas, los modelos se reformularon como modelos de ecuaciones estructurales multinivel para facilitar el análisis de datos longitudinales intensivos. Los modelos fueron probados en un estudio de simulación para evaluar su desempeño en términos de la convergencia del algoritmo y la precisión de la recuperación de los parámetros bajo diferentes condiciones. En general, las versiones multinivel de los modelos tuvieron un desempeño satisfactorio y la versión multinivel del TSO fue la que se desempeñó mejor en todas las condiciones. Adicionalmente, demostramos cómo ajustar e interpretar los resultados de estos modelos mediante el análisis de datos del proyecto HowNutsAreTheDutch (van der Krieke et al., 2017; van der Krieke et al., 2016). En este capítulo, también discutimos cómo estos modelos se relacionan con otros modelos y enfoques estadísticos utilizados para estudiar datos longitudinales intensivos, como por ejemplo, el análisis factorial dinámico y los modelos de ecuaciones estructurales dinámicos.

Capítulo 3

En este capítulo, propusimos una extensión de la versión multinivel del modelo TSO a la que nos referimos como modelo de efectos mixto de rasgo-estado-ocasión (ME-TSO). Esta extensión se propuso con base en el modelo de situación aleatoria y fija de la teoría del estado-rasgo latente (Geiser et al., 2015b) y en los modelos de ecuaciones estructurales dinámicos (Asparouhov et al., 2018). El propósito de esta extensión era hacer que el modelo fuera más flexible y más adecuado para analizar datos longitudinales intensivos. En primer lugar, al incorporar el modelo de situación fija y aleatoria, el ME-TSO puede tener en cuenta el contexto de la situación e identificar como interactuan los efectos de la persona y la situación. Por ejemplo, en un estudio longitudinal intensivo, nos podría interesar estudiar el efecto que tiene en el afecto positivo de una persona el hecho de que la persona se encuentre sola o acompaãda, para esto, sugerimos utilizar el ME-TSO. Adicionalmente, en el ME-TSO, reconocemos la heterogeneidad de los procesos psicológicos dinámicos, es decir, que no todas las personas de comportan y reaccionan de la misma manera. Por esta razón, en el ME-TSO, permitimos que el efecto autorregresivo varíe aleatoriamente entre los individuos, lo que significa que se estima un efecto autorregresivo diferente para cada sujeto. Por consiguiente, los coeficientes de varianza, que son esenciales para la interpretación de los modelos LST, también varían entre los individuos. Para mostrar cómo ajustar el ME-TSO e interpretar sus resultados, también usamos datos del proyecto HowNutsAreTheDutch (van der Krieke et al., 2017; van der Krieke et al., 2016).

Capítulo 4

La mayoría de los modelos y enfoques estadísticos desarrollados para analizar los procesos psicológicos dinámicos requieren datos continuos. Sin embargo, en psicología, no siempre se utilizan escalas continuas para medir los constructos de interés. De hecho, el uso de escalas categóricas ordenadas como por ejemplo las escalas Likert es muy popular en la investigación longitudinal intensiva (Vachon et al., 2019). Por esto, en este capítulo propusimos un modelo de IRT para el análisis de series de tiempo en psicología. Este modelo lo nombramos como el modelo dinámico de crédito parcial con tiempo variable (TV-DPCM). En pocas palabras, los modelos TRI son modelos no lineales que se han desarrollado especialmente para analizar escalas con ítems dicotómicos y politómicos. Este enfoque permite estudiar la interacción entre las personas y los ítems, y analizar las propiedades y calidad de los ítems y escalas utilizadas. El propósito del TV-DPCM es permitir el análisis de series de tiempo en psicología de un sujeto cuando se ha utilizado un conjunto de ítems tipo Likert para medir un constructo como por ejemplo el afecto negativo. El TV-DPCM integra el modelo de crédito parcial (Masters, 2016) y el modelo autorregresivo con tiempo variable (Bringmann et al., 2017). Esto significa que el modelo se puede utilizar cuando las series de tiempo no son estacionarias, siendo que el proceso dinámico latente sigue una tendencia no lineal. En este capítulo, evaluamos el desempeño del TV-DPCM en un estudio de simulación y demostramos como utilizarlo con un ejemplo empírico. Además, mediante el uso de propiedades de la IRT, como la función característica del ítem, la función de información del ítem y la función de información del test, mostramos cómo se puede usar el TV-DPCM para realizar un análisis integral de las escalas utilizadas en la investigación longitudinal intensiva. Por ejemplo, al analizar los ítems de autoestima de los datos empíricos, mostramos que esta escala era más informativa para medir niveles relativamente bajos de autoestima, pero el participante en la mayoría del tiempo experimentó niveles medios y altos de autoestima. En consecuencia, estos resultados sugirieron que se necesitaban más ítems que midieran niveles medios y altos de autoestima para poder medir con precisión la autoestima del participante.

Capítulo 5

Al utilizar modelos estadísticos nuevos, los investigadores pueden encontrar que no es claro cuales son los procedimientos para evaluar la bondad de ajuste del modelo seleccionado. En este capítulo, abordamos esta problemática para el TV-DPCM por medio de el método de evaluación predictiva a posteriori del modelo (PPMC; Gelman & Rubin, 1992; Rubin, 1984), dado que el TV-DPCM se implementó dentro del enfoque bayesiano. En pocas palabras, el método PPMC consiste en comparar las características de los datos observados con las características de los datos replicados con base en el modelo ajustado. Si las diferencias entre los datos observados y los datos replicados son demasiado grandes, se infiere que existe evidencia de un desajuste del modelo. Para implementar el método PPMC para el TV-DPCM, usamos y adaptamos estadísticos de prueba y medidas de discrepancia que se han usado para evaluar la bondad de ajuste de otros modelos de IRT (Li et al., 2017; Sinharay et al., 2006; Zhu & Stone, 2011). En nuestro estudio, simulamos bases de datos con base en el TV-DPCM incluyendo violaciones a algunos supuestos del modelo para examinar los diferentes estadísticos de prueba y medidas de discrepancia. Los resultados de estos análisis mostraron en qué condiciones estas medidas de bondad de ajuste fueron efectivas para indentificar el desajuste del modelo.

Capítulo 6

En este capítulo final, presentamos las principales conclusiones y trabajo a futuro que se derivan de esta tesis. En primer lugar, destacamos los principales hallazgos de nuestra investigación en los capítulos anteriores. Luego, analizamos más a fondo la importancia de tener definiciones claras de lo que entendemos por rasgos y estados, para tener más claridad sobre lo que medimos en datos longitudinales intensivos. Personalmente, creemos que desde un punto de vista teórico, la teoría LST puede contribuir a una mejor comprensión de las medidas en la investigación longitudinal intensiva. También destacamos que IRT ha permanecido relativamente inexplorado en entornos longitudinales intensivos, pero hay mucho que los enfoques de IRT pueden ofrecer al campo de estudio. Por ejemplo, como se muestra en esta tesis, los enfoques de IRT permiten estudiar a profundidad las propiedades y la calidad de los ítems y las escalas que se utilizan en la investigación longitudinal intensiva. Finalmente, reconocemos que aún existen muchos desafíos que los investigadores pueden enfrentar y que no pueden solucionarse con los modelos estadísticos presentados en esta tesis. No obstante, esta investigación sienta algunas bases para tener en cuenta el error de medición en el estudio de procesos psicológicos dinámicos.

Acknowledgements

Six years ago, I would have never imagined that I would be graduating from a Ph.D. and become a doctor at just 30 years old. Yet, this is the path that I ended up walking. Fortunately, I did not walk this (sometimes difficult) path alone, as I found many helping hands along the way. These words are for all the persons who, in one way or another, helped me to reach this goal.

First, I am extremely grateful to my supervisors. I could not have wished for a better team to work with. Laura, you have been a great mentor throughout these years. You have been strict and demanding but also supportive and understanding. Whenever I needed you, you were there for me. You not only taught me how to do research in intensive longitudinal data but also that doing science while being kind and empathetic is also possible and that that is what scientists should aim for. Overall, more than a supervisor, you are a very dear friend to me. Jorge, working with you has been delightful. You cannot imagine how much I admire you and your work. About 90% of my programming skills in R, I owe them to you. You certainly left a mark on me and my career. Also, I really like that you are rigorous, flexible, and open-minded at the same time. I am also very thankful that you kept your commitment as a supervisor even after you moved all the way to Japan. Rob, I very much appreciate your role in my supervisory team. Even though we did not meet very regularly, you were always there when we needed you. You helped us to see the big picture and you were also very supportive, understanding, and flexible. I sincerely hope that I get to work again with each of you at some point in the future.

Over the years, I had also several collaborators for my publications, some of which are included in this thesis. Peter, thank you so much for sharing the data from hoegekisnl with us for my first and second publications and for taking the time to read and make suggestions for the manuscripts. Sandip, working with you has been a very enriching experience for me. I have learned a lot from you during our meetings, which were quite hard to schedule as we were on three different continents. Sara and

Vanessa, I am very thankful that you invited me to collaborate with you. I very much enjoyed doing the analyses for your projects. Also, I want to thank Ellen Hamaker, Christian Geiser, Michael Eid, and Rolf Steyer, who were very kind in answering my questions via email on several occasions. By doing this you helped us to move forward. Lastly, I want to thank Markus Eronen who was a supervisor in the shadows. I know that you got to read several of my manuscripts and that you gave suggestions on the structure and writing. For this, I am very thankful.

To my paranymphs, Anna and Pieter, I am greatly grateful as well. Thank you both for taking this role. Anna, I met you in the middle of the Covid pandemic, when you were about to start your Ph.D. under Laura's supervision. I am honored to be your colleague. You are a very nice person, an excellent researcher, and I deeply admire how disciplined you are. You certainly have a bright future and you made me very happy when you accepted the role of my paranymph. Pieter, you were our (Angélica's and my) first friend in the Netherlands. You were very welcoming and open to us. Your board games and your tea ceremonies helped us build this beautiful friendship. Even after you moved back to Amsterdam and after six years, you are still in our hearts. Thank you for these many years of friendship and for being by my side while reaching the biggest milestone of my academic career so far.

Next, to my colleagues and friends in the LaBlab (Laura Bringmann's intensive longitudinal data lab); Jannis, Marie, Timon, Marmar, Arnout, and Anna (once again); thank you for being amazing. I very much enjoyed sharing my work with you and hearing about your projects. I feel so proud of all of you and your successes fill me with joy. Working by your side really helped me feel that I belonged somewhere and took away the feelings of loneliness.

I am also thankful to my current and former colleagues at the department of Psychometrics and Statistics at the University of Groningen; Anja, Merle, Joyce, Max, Marvin, Jasmine, Juan, Mali, Sarahanne, Lieke, Daniela, Nitin, Susan, Marieke, Casper, Don, Henk, Edith, and Karin. Working with you was a very nice experience. I learned a lot from all of you, not only because many of you were my teachers during the research master, but also because of the interesting discussions during the research meetings and the casual conversations we had. I will miss the nice environment of the department. I feel that in this group, we are genuinely happy about the successes of each other, which is something that should be kept up. In particular, I feel very

grateful to Anja. It was a pleasure being your office mate for all these years. You are an amazing researcher and a really good friend.

Also, walking this path would not have been possible without the good foundations that I was lucky to have, which is why I also want to thank Olga Rodriguez and Erika Arias, who were my mentors in Colombia. Profe Olga, desde cuarto semestre en el pregrado, yo ya sabía que quería dedicarme a la psicometría, sin embargo, haber tenido la fortuna de que fueras mi profesora de psicometría y supervisora de práctica y trabajo de grado, fue lo que me dió absoluta certeza de que este era mi camino. Agradezco mucho todas las enseñanzas y el apoyo que me has dado durante y después del pregrado. También agradezco la amistad que me has brindado y que siempre nos has recibido a mi y a Angélica con los brazos abiertos cuando vamos a Colombia. Erika, a ti también te agradezco mucho por todas las enseñanzas. Por tí no siento más que profunda admiración, por tu saber, tu disciplina, y tu amabilidad. También fuiste una excelente mentora durante el pregrado en la clase de evaluación y durante la práctica en el SAP, y tu apoyo y enseñanzas son lo que me ha permitido llegar a donde estoy. A las dos las aprecio y admiro mucho y espero que esta amistad pueda perdurar por muchos años más.

To all the friends I made in the Netherlands and who have been part of my life during the last 6 years. Gonneke, Frank, Edwina, Roy, Julian, Max, Januschka, Timo, and Deepti, thank you so much for being there for me for the good and the bad times. Thank you for the tasty dinners, for the board games nights, for helping me rehearse important presentations, for listening to me trying to explain my (sometimes over-complicated) work, and for bearing with my conversations about manga and video games. Also, to my Latin-American classmates and friends from the research master, Anton and Daniela, fue genial haberlos conocido, ustedes ayudaron a que todo el cambio que implica venir a un país desconocido a estudiar fuera menos traumático y más ameno. Espero, que podamos volver a coincidir en cualquier lugar del mundo. También a todos los colombianos que sin imaginarlo vine a conocer en Groningen y que se volvieron cercanos amigos para hoy y el mañana, Miguel, Mayerli, Jorgito, Gonzalo, Dina, y Esteban, gracias! Que chimba haberlos conocido y haber compartido tanto tiempo de calidad con todos ustedes. Por último, Ana (y Koen), por coincidencia, esta aventura la empezamos juntos y siempre has estado ahí (especialmente para Angélica). Gracias por recibirnos tantas veces en Amsterdam y por acompañarnos en los momentos importantes. In general, because of all of you, the day I leave the Netherlands, I will be leaving a part of my heart behind. I will really miss you all.

A todas mis amistades en Colombia, los de psicometría, mis compañeros de pregrado, los muchachos de la Drinko, y a quienes conozco desde el colegio, gracias por su apoyo por todos estos años, por siempre sacar tiempo para mi cuando voy a Colombia, por celebrar conmigo mis logros, y por su amistad incondicional. Me llena de alegría que a pesar de la distancia, la amistad ha perdurado. Estar en contacto con ustedes me ha ayudado a conservar mi cordura o lo que queda de ella.

También le agradezco con todo mi corazón a mi familia que me apoyado y acompañado en esta travesia y en la aventura que esta por empezar. En particular a mis padres y mi hermana, gracias por siempre creer en mi, por siempre motivarme a seguir mis sueños, por asesorarme para tomar las mejores decisiones, y en general por ser una constante en mi vida. Sé que siempre puedo contar con ustedes y ustedes siempre pueden contar conmigo. Los amo.

A la familia de mi esposa, Alvaro, Claudia, Daniela, y Clara, muchas gracias por todo el apoyo que nos han brindado a mi y a Angélica durante estos años. Gracias por visitarnos y recibirnos con los brazos abiertos cuando los hemos visitado. Gracias por darnos animos a siempre seguir adelante a pesar de las dificultades.

Por último, estoy infinitamente agradecido con mi esposa Angélica. Sin tí, nada de esto habría sido posible, tú eres una de las principales razones por la cual esta aventura comenzó y mira hasta donde hemos llegado. Soy privilegiado de tenerte a mi lado y de que hayamos recorrido este camino juntos. Tú me has dado fuerza y motivación cuando me ha faltado y sin importar las adversidades, has estado ahí para ayudarme a superarlas. La admiración y el respeto que siento por tí son inimaginablemente enigmantes. Si tuviera que recorrer esta camino una y otra vez, volvería a hacerlo, sin dudarlo, contigo a mi lado. Te amo.

Y por insistencia de mi esposa, también le doy gracias al bodoque feo que tengo por mascota. Pinche Brownie con su gordura y fantochería me hace querer aplastarlo cada vez que lo veo y para ser honesto le trajo mucha alegría a mi vida durante la pandemia y el doctorado.

About the author

Sebastián Castro Álvarez was born on September 1st, 1992, in Bogotá, Colombia. He started his undergraduate studies in psychology in 2010 at the Universidad Nacional de Colombia. As part of his studies, he did a one-year internship in psychometrics at the Laboratorio de Psicometría. In this internship, he studied advanced statistical methods applied in psychometrics, he did statistical consultancy for bachelor and master theses, and he did psychological assessment for children in the Servicio de Atención Psicológica of the Universidad Nacional de Colombia. After obtaining his Bachelor's degree in 2015, he worked in the development and scoring of the test of knowledge for the contest for the selection of notaries in Colombia. In September 2016, he started his Research Master in "Behavioural and Social Sciences" in the "Psychometrics and Statistics" specialization at the University of Groningen. To do his master's studies, he obtained funding with the loan-scholarship program from COLFUTURO, and the Holland Scholarship, which is financed by the Dutch Ministry of Education, Culture and Science. For his master thesis project, he worked in collaboration with prof. dr. Jorge N. Tendeiro and dr. Laura F. Bringmann to study how to differentiate between traits and states in intensive longitudinal data by using longitudinal structural equation models. In 2018, he started his Ph.D. at the Department of Psychometrics and Statistics at the Faculty of Behavioural and Social Sciences of the University of Groningen, which was funded thanks to the Ph.D. scholarship programme of the same university. Sebastián did his Ph.D. under the supervision of dr. Laura F. Bringmann, prof. dr. Jorge N. Tendeiro, and prof. dr. Rob R. Meijer. As of January 2023, he is going to work as a postdoctoral researcher at the Department of Ecology of the University of California, Davis under the mentorship of dr. Siewi Liu.

References

Adolf, J. K., Voelkle, M. C., Brose, A., & Schmiedek, F. (2017). Capturing Context-Related Change in Emotional Dynamics via Fixed Moderated Time Series Analysis. *Multivariate Behavioral Research*, *52*(4), 499–531. https://doi.org/10.1080/00273171.2017.1321978

Allen, B. P., & Potkay, C. R. (1981). On the arbitrary distinction between states and traits. *Journal of Personality and Social Psychology*, *41*(5), 916–928. https://doi.org/10.1037/0022-3514.41.5.916

Allen, B. P., & Potkay, C. R. (1983). Just as arbitrary as ever: Comments on Zuckerman's rejoinder. *Journal of Personality and Social Psychology*, *44*(5), 1087–1089. https://doi.org/10.1037/0022-3514.44.5.1087

Alston, W. P. (1975). Traits, Consistency and Conceptual Alternatives for Personality Theory. *Journal for the Theory of Social Behaviour*, *5*(1), 17–48. https://doi.org/10.1111/j.1468-5914.1975.tb00341.x

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2017). Dynamic Latent Class Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 257–269. https://doi.org/10.1080/10705511.2016.1253479

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using Mplus: Technical implementation.* (tech. rep.). Technical Report. Los Angeles: Muthén & Muthén. http://www.statmodel.com/download/bayes2.pdf

Beltz, A. M., Beekman, C., Molenaar, P. C. M., & Buss, K. A. (2013). Mapping Temporal Dynamics in Social Interactions With Unified Structural Equation Modeling: A Description and Demonstration Revealing Time-Dependent Sex Differences in Play Behavior. *Applied Developmental Science*, *17*(3), 152–168. https://doi.org/10.1080/10888691.2013.805953

Beltz, A. M., & Gates, K. M. (2017). Network Mapping with GIMME. *Multivariate Behavioral Research*, *52*(6), 789–804. https://doi.org/10.1080/00273171.2017.1373014

Betancourt, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. https://doi.org/10.48550/arxiv.1701.02434

Borsboom, D., & Cramer, A. O. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Bos, F. M., Schoevers, R. A., & aan het Rot, M. (2015). Experience sampling and ecological momentary assessment studies in psychopharmacology: A systematic review. https://doi.org/10.1016/j.euroneuro.2015.08.008

Brennan, R. L. (2005). Generalizability Theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Bringmann, L. F., & Eronen, M. I. (2018). Don’t blame the model: Reconsidering the network approach to psychopathology. *Psychological Review*, *125*(4), 606–615. https://doi.org/10.1037/rev0000108

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological methods*, *22*(3), 409–425. https://doi.org/10.1037/met0000085

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing Temporal Emotion Dynamics Using Networks. *Assessment*, *23*(4), 425–435. https://doi.org/10.1177/1073191116645909

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A Network Approach to Psychopathology: New Insights into Clinical Longitudinal Data (G. A. de Erausquin, Ed.). *PLoS ONE*, *8*(4), 1–13. https://doi.org/10.1371/journal.pone.0060188

Burke, P. A., Kraut, R. E., & Dworkin, R. H. (1984). Traits, consistency, and self-schemata: What do our methods measure? *Journal of Personality and Social Psychology*, *47*(3), 568–579. https://doi.org/10.1037/0022-3514.47.3.568

Cai, L. (2010). A Two-Tier Full-Information Item Factor Analysis Model with Applications. *Psychometrika*, *75*(4), 581–612. https://doi.org/10.1007/s11336-010-9178-0

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Castro-Alvarez, S. (2019). LST_Analyses. https://github.com/secastroal/LST_Analyses

Castro-Alvarez, S., Bringmann, L. F., Meijer, R. R., & Tendeiro, J. N. (2022a). *A Time-Varying Dynamic Partial Credit Model to Analyze Polytomous and Multivariate Time Series Data*. https://doi.org/10.31234/osf.io/udnbt

Castro-Alvarez, S., Sinharay, S., Bringmann, L. F., Meijer, R. R., & Tendeiro, J. N. (2022b). *Posterior Predictive Model Checking Methods for the Time-Varying Dynamic Partial Credit Model.*

Castro-Alvarez, S., Tendeiro, J. N., de Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022c). Mixed-Effects Trait-State-Occasion Model: Studying the Psychometric Properties and the Person–Situation Interactions of Psychological Dynamics. *Structural Equation Modeling*, *29*(3), 438–451. https://doi.org/10.1080/10705511.2021.1961587

Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022d). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods*, *27*(1), 17–43. https://doi.org/10.1037/met0000393

Cattell, R. B. (1963). The structuring of change by P-technique and incremental R-technique. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 167–198). The University of Wisconsin Press.

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. Springer US. https://doi.org/10.1007/978-1-4684-2262-7

Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of states and traits: Dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology*, *54*(4), 541–557. https://doi.org/10.1037/0022-3514.54.4.541

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction* (6th ed.). Chapman; Hall/CRC. https://doi.org/10.4324/9780203491683

Chatfield, C. (2004). *The Analysis of Time Series : An Introduction.* (6th ed.). Chapman; Hall/CRC. https://doi.org/10.4324/9780203491683

Chow, S. M., Hamaker, E. L., Fujita, F., & Boker, S. M. (2009). Representing time-varying cyclic dynamics using multiple-subject state-space models. *British*

Journal of Mathematical and Statistical Psychology, *62*(3), 683–716. https://doi.org/10.1348/000711008X384080

Chow, S. M., Ho, M. h. R., Hamaker, E. L., & Dolan, C. V. (2010). Equivalence and differences between structural equation modeling and state-space modeling techniques. *Structural Equation Modeling*, *17*(2), 303–332. https://doi.org/10.1080/10705511003661553

Chow, S. M., Zu, J., Shifren, K., & Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research*, *46*(2), 303–339. https://doi.org/10.1080/00273171.2011.563697

Chow, S.-M., Hamagani, F., & Nesselroade, J. R. (2007). Age differences in dynamical emotion-cognition linkages. *Psychology and aging*, *22*(4), 765–780. https://doi.org/10.1037/0882-7974.22.4.765

Cole, D. A., Jacquez, F. M., Truss, A. E., Pineda, A. Q., Weitlauf, A. S., Tilghman-Osborne, C. E., Felton, J. W., & Maxwell, M. A. (2009). Gender differences in the longitudinal structure of cognitive diatheses for depression in children and adolescents. *Journal of Clinical Psychology*, *65*(12), 1312–1326. https://doi.org/10.1002/jclp.20631

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and Conceptual Problems With Longitudinal Trait-State Models: Introducing a Trait-State-Occasion Model. *Psychological Methods*, *10*(1), 3–20. https://doi.org/10.1037/1082-989X.10.1.3

Conway, C. C., Hopwood, C. J., Morey, L. C., & Skodol, A. E. (2018). Borderline personality disorder is equally trait-like and state-like over ten years in adult psychiatric patients. *Journal of Abnormal Psychology*, *127*(6), 590–601. https://doi.org/10.1037/abn0000364

Courvoisier, D. S., Eid, M., Lischetzke, T., & Schreiber, W. H. (2010). Psychometric properties of a computerized mobile phone method for assessing mood in daily life. *Emotion (Washington, D.C.)*, *10*(1), 115–24. https://doi.org/10.1037/a0017813

Courvoisier, D. S., Eid, M., & Nussbeck, F. W. (2007). Mixture distribution latent state-trait analysis: Basic ideas and applications. *Psychological Methods*, *12*(1), 80–104. https://doi.org/10.1037/1082-989X.12.1.80

Cramer, A. O. J., Van Der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Measurable Like Temperature or Mereological like Flocking? on the Nature of Personality Traits.

European Journal of Personality, *26*(4), 451–459. https://doi.org/10.1002/per.1879

Crayen, C., Eid, M., Lischetzke, T., & Vermunt, J. K. (2017). A continuous-time mixture latent- state-trait Markov model for experience sampling data: Application and evaluation. *European Journal of Psychological Assessment*, *33*(4), 296–311. https://doi.org/10.1027/1015-5759/a000418

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2016). Get Over It! A Multilevel Threshold Autoregressive Model for State-dependent Affect Regulation. *Psychometrika*, *81*(1), 217–241. https://doi.org/10.1007/s11336-014-9417-x

de Haan-Rietdijk, S., Kuppens, P., Bergeman, C. S., Sheeber, L. B., Allen, N. B., & Hamaker, E. L. (2017a). On the Use of Mixed Markov Models for Intensive Longitudinal Data. *Multivariate Behavioral Research*, *52*(6), 747–767. https://doi.org/10.1080/00273171.2017.1370364

de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017b). Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology*, *8*(OCT), 1849. https://doi.org/10.3389/fpsyg.2017.01849

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. *Journal of Educational and Behavioral Statistics*, *41*(6), 628–649.

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*(2), 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Donoghue, J. R., & Isham, S. P. (1998). A Comparison of Procedures to Detect Item Parameter Drift. *Applied Psychological Measurement*, *22*(1), 33–51. https://doi.org/10.1177/01466216980221002

Ebner-Priemer, U. W., Houben, M., Santangelo, P., Kleindienst, N., Tuerlinckx, F., Oravecz, Z., Verleysen, G., Van Deun, K., Bohus, M., & Kuppens, P. (2015). Unraveling affective dysregulation in borderline personality disorder: a theoretical model and empirical evidence. *Journal of abnormal psychology*, *124*(1), 186–198. https://doi.org/10.1037/abn0000021

Eid, M., Courvoisier, D. S., & Lischetzke, T. (2012). Structural equation modeling of ambulatory assessment data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 384–406). Guilford Press. http://psycnet.apa.org/record/2012-05165-021

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the Definition of Latent-State-Trait Models With Autoregressive Effects. *European Journal of Psychological Assessment*, *33*(4), 285–295. https://doi.org/10.1027/1015-5759/a000435

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, 107319112095710. https://doi.org/10.1177/1073191120957102

Elmer, T., Geschwind, N., Peeters, F., Wichers, M., & Bringmann, L. (2020). Getting stuck in social isolation: Solitude inertia and depressive symptoms. *Journal of Abnormal Psychology*, *129*(7), 713–723. https://doi.org/10.1037/abn0000588

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781410605269

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin*, *83*(5), 956–974. https://doi.org/10.1037/0033-2909.83.5.956

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network Psychometrics. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The wiley handbook of psychometric testing* (pp. 953–986). Wiley. https://doi.org/10.1002/9781118489772.ch30

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, *37*(7), 1097–1126. https://doi.org/10.1037/0022-3514.37.7.1097

Epstein, S. (1981). &quot;The stability of behavior: II. Implications for psychological research&quot;: Reply to Lieberman. *American Psychologist*, *36*(6), 696–697. https://doi.org/10.1037/0003-066X.36.6.696

Fan, J., & Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer New York. https://doi.org/10.1007/978-0-387-69395-8

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967–984. https://doi.org/10.1037/0022-3514.74.4.967

Fox, J.-P. (2010). *Bayesian Item Response Modeling*. Springer New York. https://doi.org/10.1007/978-1-4419-0742-4

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time...or not? lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354–1367. https://doi.org/10.1037/pas0000275

Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomyn, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment*, *29*(9), 1120–1128. https://doi.org/10.1037/pas0000411

Gabry, J., & Mahr, T. (2021). bayesplot: Plotting for Bayesian Models. https://mc-stan.org/bayesplot/

Geiser, C. (2020). *Longitudinal Structural Equation Modeling with Mplus: A Latent State-Trait Perspective*. Guilford Publications.

Geiser, C., Bishop, J., Lockhart, G., Shiffman, S., & Grenard, J. L. (2013). Analyzing latent state-trait and multiple-indicator latent growth curve models as multilevel structural equation models. *Frontiers in Psychology*, *4*, 975. https://doi.org/10.3389/fpsyg.2013.00975

Geiser, C., Götz, T., Preckel, F., & Freund, P. A. (2017). States and Traits: Theory, models, and assessment. *European Journal of Psychological Assessment*, *33*(4), 219–223. https://doi.org/10.1027/1015-5759/a000413

Geiser, C., Keller, B. T., Lockhart, G., Eid, M., Cole, D. A., & Koch, T. (2015a). Distinguishing state variability from trait change in longitudinal data: The role of measurement (non)invariance in latent state-trait analyses. *Behavior Research Methods*, *47*(1), 172–203. https://doi.org/10.3758/s13428-014-0457-z

Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., Servera, M., & Shiffman, S. (2015b). Analyzing person, situation and person × situation interaction effects: Latent state-trait models for the combination of random and fixed situations. *Psychological Methods*, *20*(2), 165–192. https://doi.org/10.1037/met0000026

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*, *17*(2), 255–283. https://doi.org/10.1037/a0026977

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis Third Edition* (3rd ed.). CRC Press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–807. https://www.jstor.org/stable/24306036?seq=1#metadata_info_tab_contents

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: a randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*(5), 618–628. http://psycnet.apa.org/buy/2011-14915-001

Giraitis, L., Kapetanios, G., & Yates, T. (2014). Inference on stochastic time-varying coefficient models. *Journal of Econometrics*, *179*(1), 46–65. https://doi.org/10.1016/j.jeconom.2013.10.009

Glas, C. A., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217–233. https://doi.org/10.1177/0146621603027003003

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation : An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the Frontiers of Modeling Intensive Longitudinal Data: Dynamic Structural Equation Models for the Affective Measurements from the COGITO Study. *Multivariate Behavioral Research*, 1–22. https://doi.org/10.1080/00273171.2018.1446819

Hamaker, E. L., Ceulemans, E., Grasman, R. P., & Tuerlinckx, F. (2015). Modeling Affect Dynamics: State of the Art and Future Challenges. *Emotion Review*, *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

Hamaker, E. L., Schuurman, N. K., & Zijlmans, E. A. O. (2017). Using a Few Snapshots to Distinguish Mountains from Waves: Weak Factorial Invariance in the

Context of Trait-State Research. *Multivariate Behavioral Research*, *52*(1), 47–60. https://doi.org/10.1080/00273171.2016.1251299

Hamaker, E. L. (2005). Conditions for the Equivalence of the Autoregressive Latent Trajectory Model and a Latent Growth Curve Model With Autoregressive Disturbances. *METHODS & RESEARCH*, *33*(3), 404–416. https://doi.org/10.1177/0049124104270220

Hamaker, E. L. (2012). Why researchers should think &quot;within-person&quot;: A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). Guilford Press. http://psycnet.apa.org/record/2012-05165-003

Hamaker, E. L., & Dolan, C. V. (2009). Idiographic Data Analysis: Quantitative Methods—From Simple to Advanced. In J. Valsiner, P. Molenaar, M. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 191–216). Springer US. https://doi.org/10.1007/978-0-387-95922-1{_}9

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in psychology*, *5*, 1492. https://doi.org/10.3389/fpsyg.2014.01492

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated trait–state model. *Journal of Research in Personality*, *41*(2), 295–315. https://doi.org/10.1016/J.JRP.2006.04.003

Hamaker, E. L., & Wichers, M. (2017). No Time Like the Present. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory. *Item Response Theory*. https://doi.org/10.1007/978-94-017-1988-9

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press. https://doi.org/10.1515/9780691218632

Hastie, T., & Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(4), 757–796. http://www.jstor.org/stable/2345993

Hecht, M., Hardt, K., Driver, C. C., & Voelkle, M. C. (2019). Bayesian continuous-time Rasch models. *Psychological Methods*, *24*(4), 516–537. https://doi.org/10.1037/met0000205

Hertzog, C., & Nesselroade, J. R. (1987). Beyond Autoregressive Models: Some Implications of the Trait-State Distinction for the Structural Modeling of Developmental Change. *Child Development*, *58*(1), 93. https://doi.org/10.2307/1130294

Hoover, D. R., Rice, J. A., Colin, O. W., & Yang, L. I. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, *85*(4), 809–822. https://doi.org/10.1093/biomet/85.4.809

Houben, M., Ceulemans, E., & Kuppens, P. (2020). Modeling Intensive Longitudinal Data. In A. G. C. Wright & M. N. Hallquist (Eds.), *The cambridge handbook of research methods in clinical psychology* (pp. 312–326). Cambridge University Press. https://doi.org/10.1017/9781316995808.030

Howard, K. I., & Forehand, G. A. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement*, *22*(4), 731–735. https://doi.org/10.1177/001316446202200407

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., Neale, M. C., Sisk, C. L., & Klump, K. (2016). Test Reliability at the Individual Level. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 532–543. https://doi.org/10.1080/10705511.2016.1148605

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A Multilevel AR(1) Model: Allowing for Inter-Individual Differences in Trait-Scores, Inertia, and Innovation Variance. *Multivariate Behavioral Research*, *50*(3), 334–349. https://doi.org/10.1080/00273171.2014.1003772

Jönsson, K. (2011). Testing stationarity in small- and medium-sized samples when disturbances are serially correlated. *Oxford Bulletin of Economics and Statistics*, *73*(5), 669–690. https://doi.org/10.1111/j.1468-0084.2010.00620.x

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, Transactions of the ASME*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of consulting and clinical psychology*, *63*(1), 52–9. http://www.ncbi.nlm.nih.gov/pubmed/7896990

Kenny, D. A., & Zautra, A. (2001). Trait–state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change.* (pp. 243–263). American Psychological Association. https://doi.org/10.1037/10409-008

Kharratzadeh, M. (2017). Splines In Stan. https://mc-stan.org/users/documentation/case-studies/splines_in_stan.html

Kim, S., & Camilli, G. (2014). An item response theory approach to longitudinal analysis with application to summer setback in preschool language/literacy. *Large-scale Assessments in Education*, *2*(1), 1. https://doi.org/10.1186/2196-0739-2-1

Koestner, R., Bernieri, F., & Zuckerman, M. (1992). Self-Regulation and Consistency between Attitudes, Traits, and Behaviors. *Personality and Social Psychology Bulletin*, *18*(1), 52–59. https://doi.org/10.1177/0146167292181008

Kossakowski, J. J., Groot, P. C., Haslbeck, J. M. B., Borsboom, D., & Wichers, M. (2017). Data from 'Critical Slowing Down as a Personalized Early Warning Signal for Depression'. *Journal of Open Psychology Data*, *5*(1). https://doi.org/10.5334/jopd.29

Kropko, J. (2013). *Dynamic measurement of political phenomena: Item response theory for time-series data* (tech. rep.). Columbia University. https://nanopdf.com/download/item-response-theory-for-time-series-data_pdf#

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition* (2nd ed.). Elsevier Science. https://doi.org/10.1016/B978-0-12-405888-0.09999-2

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional Inertia and Psychological Maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1-3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

LaGrange, B., & Cole, D. A. (2008). An expansion of the trait-state-occasion model: Accounting for shared method variance. *Structural Equation Modeling*, *15*(2), 241–271. https://doi.org/10.1080/10705510801922381

Leybourne, S., & Newbold, P. (1999). On the Size Properties of Phillips–Perron Tests. *Journal of Time Series Analysis*, *20*(1), 51–61. https://doi.org/10.1111/1467-9892.00125

Li, T., Xie, C., & Jiao, H. (2017). Assessing fit of alternative unidimensional polytomous IRT models using posterior predictive model checking. *Psychological Methods*, *22*(2), 397–408. https://doi.org/10.1037/met0000082

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data, Third Edition*. Wiley. https://doi.org/10.1002/9781119482260

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *Journal of Mathematical Psychology*, *55*, 68–83. https://doi.org/10.1016/j.jmp.2010.08.004

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lubin, B., Van Whitlock, R., Reddy, D., & Petren, S. (2001). A comparison of the short and long forms of the Multiple Affect Adjective Check List—Revised (MAACL-R). *Journal of Clinical Psychology*, *57*(3), 411–416. https://doi.org/10.1002/jclp.1023

Masters, G. N. (2016). Partial Credit Model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1: Models* (pp. 109–126). CRC Press. https://doi.org/10.1201/9781315374512

McNeish, D., & Hamaker, E. L. (2020). A Primer on Two-Level Dynamic Structural Equation Models for Intensive Longitudinal Data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in Intensive Longitudinal Data. *https://doi.org/10.1080/10705511.2021.1915788*, *28*(5), 807–822. https://doi.org/10.1080/10705511.2021.1915788

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1–19. https://doi.org/10.3758/s13428-020-01398-0

Meijer, R. R., & Nering, M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Meijer, R. R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11 SUPPL. 3), S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Mischel, W. (1968). *Personality and assessment.* Wiley. https://psycnet.apa.org/record/2003-00025-000

Mischel, W. (2004). Toward an Integrative Science of the Person. *Annual Review of Psychology*, *55*(1), 1–22. https://doi.org/10.1146/annurev.psych.55.042902.130709

Moeller, J., Ivcevic, Z., Brackett, M. A., & White, A. E. (2018). Mixed emotions: Network analyses of intra-individual co-occurrences within and across situations. *Emotion*, *18*(8), 1106–1121. https://doi.org/10.1037/emo0000419

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204{_}1

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202. https://doi.org/10.1007/BF02294246

Molenaar, P. C., Sinclair, K. O., Rovine, M. J., Ram, N., & Corneal, S. E. (2009). Analyzing Developmental Processes on an Individual Level Using Nonstationary Time Series Modeling. *Developmental Psychology*, *45*(1), 260–271. https://doi.org/10.1037/a0014170

Musci, R. J., Masyn, K. E., Benke, K., Maher, B., Uhl, G., & Ialongo, N. S. (2016). The effects of the interplay of genetics and early environmental risk on the course of internalizing symptoms from late childhood through adolescence. *Development and Psychopathology*, *28*(1), 225–237. https://doi.org/10.1017/S0954579415000401

Mushtaq, R. (2011). Augmented Dickey Fuller Test. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.1911068

Muthén, B. O. (1991). Multilevel Factor Analysis of Class and Student Achievement Components. *Journal of Educational Measurement*, *28*(4), 338–354. https://doi.org/10.1111/j.1745-3984.1991.tb00363.x

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

Myin-Germeys, I., & Kuppens, P. (2021). Experience sampling methods, an introduction. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing esm studies* (2nd ed., pp. 7–19). Center for Research on Experience Sampling; Ambulatory Methods Leuven. https://www.kuleuven.be/samenwerking/real/real-book/index.htm

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change:*

Recent advances, unanswered questions, future directions. (pp. 92–105). American Psychological Association. https://doi.org/10.1037/10099-006

Nezlek, J. B. (2007). A multilevel framework for understanding relationships among traits, states, situations and behaviours. *European Journal of Personality*, *21*(6), 789–810. https://doi.org/10.1002/per.640

Nezlek, J. B. (2012). Multilevel modeling analyses of diary-style data. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 357–383). Guilford Press. http://psycnet.apa.org/record/2012-05165-020

Orlando, M., & Thissen, D. (2000). Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models* (144th ed.). Sage.

Pan, J., Ip, E. H., & Dubé, L. (2020). Multilevel Heterogeneous Factor Analysis and Application to Ecological Momentary Assessment. *Psychometrika*, *85*(1), 75–100. https://doi.org/10.1007/s11336-019-09691-4

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346. https://doi.org/10.1093/BIOMET/75.2.335

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425. https://doi.org/10.1016/S1364-6613(02)01964-2

R Core Team. (2022). R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The Questionable Ecological Validity of Ecological Momentary Assessment: Considerations for Design and Analysis. *Research in Human Development*, *14*(3), 253–270. https://doi.org/10.1080/15427609.2017.1340052

Ram, N., Chow, S. Y., Bowles, R. P., Wang, L., Grimm, K., Fujita, F., & Nesselroade, J. R. (2005). Examining interindividual differences in cyclicity of pleasant and unpleasant affects using spectral analysis and item response modeling. *Psychometrika*, *70*(4), 773–790. https://doi.org/10.1007/s11336-001-1270-5

Ram, N., & Gerstorf, D. (2009). Time-Structured and Net Intraindividual Variability: Tools for Examining the Development of Dynamic Characteristics and Processes. *Psychology and Aging*, *24*(4), 778–791. https://doi.org/10.1037/a0017915

Reis, H. T. (2012). Why researchers should think &quot;real-world&quot;: A conceptual rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 3–21). Guilford Press. http://psycnet.apa.org/record/2012-05165-001

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Rijn, P. v., Dolan, C. V., & Molenaar, P. C. M. (2010). State space methods for item response modeling of multisubject time series. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development.* (pp. 125–151). American Psychological Association. https://doi.org/10.1037/12140-008

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*(1), 3–25. https://doi.org/10.1037/0033-2909.126.1.3

Roesch, S. C., Aldridge, A. A., Stocking, S. N., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel Factor Analysis and Structural Equation Modeling of Daily Diary Coping Data: Modeling Trait and State Variation. *Multivariate Behavioral Research*, *45*(5), 767–789. https://doi.org/10.1080/00273171.2010.519276

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rovine, M. J., & Walls, T. A. (2006). Multilevel Autoregressive Modeling of Interindividual Differences in the Stability of a Process. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data2* (pp. 124–147). Oxford University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. https://doi.org/10.1214/aos/1176346785

Samejima, F. (1997). Graded response model. *Handbook of modern item response theory* (pp. 85–100). Springer.

Schultzberg, M., & Muthén, B. (2018). Number of Subjects and Time Points Needed for Multilevel Time-Series Analysis: A Simulation Study of Dynamic Structural Equation Modeling. *Structural Equation Modeling*, *25*(4), 495–515. https://doi.org/10.1080/10705511.2017.1392862

Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological methods*, *21*(2), 206–21. https://doi.org/10.1037/met0000062

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, *24*(1), 70–91. https://doi.org/10.1037/met0000188

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology*, *6*, 1038. https://doi.org/10.3389/fpsyg.2015.01038

Schwarz, N. (2012). Why researchers should think &quot;real-time&quot;: A cognitive rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 22–42). Guilford Press. http://psycnet.apa.org/record/2012-05165-002

Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling Time-Dependent Association in Longitudinal Data: A Lag as Moderator Approach. *Multivariate Behavioral Research*, *47*(5), 697–716. https://doi.org/10.1080/00273171.2012.715557

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, *4*(1), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R examples*. Springer International Publishing. https://doi.org/10.1007/978-3-319-52452-8

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed Beeps and Missing Data. *Social Science Computer Review*, *31*(4), 471–481. https://doi.org/10.1177/0894439313479902

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, *42*(4), 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 429–449. https://doi.org/10.1348/000711005X66888

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Smid, S. C., Depaoli, S., & Van De Schoot, R. (2020). Predicting a Distal Outcome Variable From a Latent Growth Model: ML versus Bayesian Estimation. *Structural Equation Modeling*, *27*(2), 169–191. https://doi.org/10.1080/10705511.2019.1604140

Snippe, E., Nyklíček, I., Schroevers, M. J., & Bos, E. H. (2015). The temporal order of change in daily mindfulness and affect during mindfulness-based stress reduction. *Journal of counseling psychology*, *62*(2), 106–14. https://doi.org/10.1037/cou0000057

Song, H., & Ferrer, E. (2012). Bayesian Estimation of Random Coefficient Dynamic Factor Models. *Multivariate Behavioral Research*, *47*(1), 26–60. https://doi.org/10.1080/00273171.2012.640593

Song, H., & Zhang, Z. (2014). Analyzing Multiple Multivariate Time Series Data Using Multilevel Dynamic Factor Models. *Multivariate Behavioral Research*, *49*(1), 67–77. https://doi.org/10.1080/00273171.2013.851018

Spielberger, C. D., & Sydeman, S. J. (1994). State-trait anxiety inventory and state-trait anger expression inventory. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 292–321). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1037/h0020743

Stan Development Team. (2020). RStan: The R interface to Stan. http://mc-stan.org/

Stan Development Team. (2022). *Stan modeling language users guide and reference manual 2.29*. https://mc-stan.org

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79–98. http://psycnet.apa.org/record/1994-31946-001

Steyer, R., Geiser, C., & Fiege, C. (2012). Latent state-trait models. *Apa handbook of research methods in psychology, vol 3: Data analysis and research publication.* (pp. 291–308). American Psychological Association. https://doi.org/10.1037/13621-014

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A Theory of States and Traits—Revised. *Annual Review of Clinical Psychology*, *11*(1), 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, *13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0.CO;2-A

Steyer, R., & Schmitt, T. (1994). The theory of confounding and its application in causal modeling with latent variables. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 36–67). US: Sage Publications, Inc. http://psycnet.apa.org/record/1996-97111-002

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136. https://doi.org/10.1177/0146167298242002

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Trull, T. J., & Ebner-Priemer, U. (2014). The Role of Ambulatory Assessment in Psychological Science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, *129*(1), 56–63. https://doi.org/10.1037/abn0000473

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12). https://doi.org/10.2196/14475

van der Krieke, L., Blaauw, F. J., Emerencia, A. C., Schenk, H. M., Slaets, J. P., Bos, E. H., de Jonge, P., & Jeronimus, B. F. (2017). Temporal Dynamics of Health and Well-Being: A Crowdsourcing Approach to Momentary Assessments and Automated Generation of Personalized Feedback. *Psychosomatic Medicine*, *79*(2), 213–223. https://doi.org/10.1097/PSY.0000000000000378

van der Krieke, L., Jeronimus, B. F., Blaauw, F. J., Wanders, R. B., Emerencia, A. C., Schenk, H. M., Vos, S. D., Snippe, E., Wichers, M., Wigman, J. T., Bos, E. H., Wardenaar, K. J., & De Jonge, P. (2016). HowNutsAreTheDutch (HoeGekIsNL): A crowdsourcing study of mental symptoms and strengths. *International Journal of Methods in Psychiatric Research*, *25*(2), 123–144. https://doi.org/10.1002/mpr.1495

van der Linden, W. J. (2016). Unidimensional logistic response models. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1: Models* (pp. 13–30). CRC Press. https://doi.org/10.1201/9781315374512

van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2021). Latent Variable Models and Networks: Statistical Equivalence and Testability. *Multivariate Behavioral Research*, *56*(2), 175–198. https://doi.org/10.1080/00273171.2019.1672515

van Roekel, E., Scholte, R. H., Engels, R. C., Goossens, L., & Verhagen, M. (2015). Loneliness in the Daily Lives of Adolescents: An Experience Sampling Study Examining the Effects of Social Contexts. *Journal of Early Adolescence*, *35*(7), 905–930. https://doi.org/10.1177/0272431614547049

Voelkle, M. C., & Oud, J. H. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 103–126. https://doi.org/10.1111/j.2044-8317.2012.02043.x

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An SEM Approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17*(2), 176–192. https://doi.org/10.1037/a0027543

Vogelsmeier, L. V., Vermunt, J. K., Böing-Messing, F., & De Roover, K. (2019a). Continuous-time latent markov factor analysis for exploring measurement model changes across time. *Methodology*, *15*(1), 29–42. https://doi.org/10.1027/1614-2241/A000176

Vogelsmeier, L. V., Vermunt, J. K., Keijsers, L., & De Roover, K. (2020). Latent Markov Latent Trait Analysis for Exploring Measurement Model Changes in Intensive Longitudinal Data. *Evaluation and the Health Professions*, *44*(1), 61–76. https://doi.org/10.1177/0163278720976762

Vogelsmeier, L. V., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019b). Latent Markov Factor Analysis for Exploring Measurement Model Changes in Time-Intensive Longitudinal Studies. *Structural Equation Modeling*, *26*(4), 557–575. https://doi.org/10.1080/10705511.2018.1554445 von Davier, M. (2016). Rasch model. In W. J. van der Linden (Ed.), *Handbook of item response theory: Volume 1: Models* (pp. 31–48). CRC Press. https://doi.org/10.1201/9781315374512

von Neumann, J., Kent, R. H., Bellinson, H. R., & Hart, B. I. (1941). The Mean Square Successive Difference. *The Annals of Mathematical Statistics*, *12*(2), 153–162.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized Adaptive Testing*. Routledge. https://doi.org/10.4324/9781410605931

Walls, T. A., Jung, H., & Schwartz, J. E. (2006). Multilevel Models for Intensive Longitudinal Data. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 3–37). Oxford University Press.

Walls, T. A., & Schafer, J. L. (L. (2006). *Models for intensive longitudinal data*. Oxford University Press.

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of applied measurement*, *9*(4), 387–408.

Wang, X., Berger, J. O., & Burdick, D. S. (2013). Bayesian analysis of dynamic item response models in educational testing. *Annals of Applied Statistics*, *7*(1), 126–153. https://doi.org/10.1214/12-AOAS608

Wichers, M., Peeters, F., Geschwind, N., Jacobs, N., Simons, C. J., Derom, C., Thiery, E., Delespaul, P. H., & van Os, J. (2010). Unveiling patterns of affective responses in daily life may improve outcome prediction in depression: A momentary assessment study. *Journal of Affective Disorders*, *124*(1-2), 191–195. https://doi.org/10.1016/j.jad.2009.11.010

Wichers, M. C., Barge-Schaapveld, D. Q., Nicolson, N. A., Peeters, F., De Vries, M., Mengelers, R., & Van Os, J. (2009). Reduced stress-sensitivity or increased reward experience: The psychological mechanism of response to antidepressant medication. *Neuropsychopharmacology*, *34*(4), 923–931. https://doi.org/10.1038/npp.2008.66

Wichers, M., & Groot, P. C. (2016). Critical Slowing Down as a Personalized Early Warning Signal for Depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116. https://doi.org/10.1159/000441458

Wood, S. N. (2017). *Generalized additive models: An introduction with R, second edition*. https://doi.org/10.1201/9781315370279

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yik, M. S. M., Russell, J. A., & Feldman Barrett, L. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology*, *77*(3), 600–619. https://doi.org/10.1037/0022-3514.77.3.600

Zelenski, J. M., & Larsen, R. J. (2000). The Distribution of Basic Emotions in Everyday Life: A State and Trait Perspective from Experience Sampling Data. *Journal of Research in Personality*, *34*(2), 178–197. https://doi.org/10.1006/jrpe.1999.2275

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using bayesian methods. *Journal of Educational Measurement*, *48*(1), 81–97. https://doi.org/10.1111/j.1745-3984.2011.00132.x

Zuckerman, M. (1983). The distinction between trait and state scales is not arbitrary: Comment on Allen and Potkay’s &quot;On the arbitrary distinction between traits and states.&quot; *Journal of Personality and Social Psychology*, *44*(5), 1083–1086. https://doi.org/10.1037/0022-3514.44.5.1083

Zuckerman, M., Lubin, B., & Rinck, C. M. (1983). Construction of new scales for the Multiple Affect Adjective Check List. *Journal of Behavioral Assessment*, *5*(2), 119–129. https://doi.org/10.1007/BF01321444

To conclude, my experience after four years of Ph.D. can be summarized as follows: