STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 3
ISSN: 1070-5511 | E-ISSN: 1532-8007

Measurement Model Misspecification in Dynamic Structural Equation Models: Power, Reliability, and Other Considerations

Hyungeun Oh, Michael D. Hunter, and Sy-Miin Chow

The Pennsylvania State University

ABSTRACT
Dynamic Structural Equation Models (DSEMs) integrate multilevel modeling, time series analysis, and structural equation modeling within a Bayesian estimation framework, offering a versatile tool for analyzing intensive longitudinal data (ILD). However, the impact of measurement structure misspecification in DSEMs, especially under varying reliability conditions and model complexities, remains underexplored. Our Monte Carlo simulation revealed that omitting measurement errors when present led to severe biases in dynamic parameters regardless of reliability conditions, though power remained high. Increasing the number of participants and time points ameliorated but did not eliminate all biases. A single-indicator DSEMs with a measurement structure using composite scores showed similar performance to multiple indicators DSEMs. Empirical applications showed discrepancies in dynamic parameters based on the number of indicators and measurement structures used. Leveraging these findings, we provide design recommendations, functions for extending reliability indices from single-indicator to multiple-indicator models, and guidelines for power evaluations under different reliability conditions.

KEYWORDS
Dynamic structural equation models; intensive longitudinal data; measurement error; within-person dynamic process

1. Introduction

With technological developments such as smartphones and wearable devices, studies using intensive longitudinal data (ILD) have increased considerably in the social and behavioral sciences in the last two decades (Hamaker & Wichers, 2017). Researchers are now better equipped than ever to gather data with more time points, more closely spaced in time, and less invasively with data collection methods such as experience sampling (Scollon et al., 2003), ambulatory assessment (Fahrenberg et al., 2007), daily diaries (Bolger et al., 2003), and ecological momentary assessment (Smyth & Smyth, 2003). That is, unlike traditional longitudinal panel data with a small number of time points, ILD has more observations per person (e.g., two measurement occasions per day for a month). This characteristic of ILD offers a lens to investigate within-person dynamic processes and corresponding between-person differences, rather than focusing on mean changes as is common with panel data (Hamaker & Wichers, 2017; Ram & Gerstorf, 2009; Wang et al., 2012).

However, the very characteristics that define ILD's unique strengths also introduce challenges. The need for frequent participant involvement, typically involving multiple items, can be burdensome (Diener & Tay, 2014), and may exacerbate the effects of measurement errors. For instance, the repetitive nature of ILD collection often necessitates the use of fewer items to mitigate respondent burden, thereby leading to less redundancy between items (Stigler, 1990), and higher susceptibility to errors that can recur across measurement occasions (Cronbach & Furby, 1970). The use of measurements with sound psychometric properties for capturing both within- and between-person variations is thus crucial.

Research designs involving ILD often deal with latent variables using multiple items, such as affect or mood, necessitating the measurement of underlying constructs that are not directly observable, making the consideration of measurement error essential. However, many psychological applications modeling ILD did not consider measurement and thus may have measurement model misspecification (Hamaker et al., 2018; McNeish et al., 2021). The effects of measurement error are also ignored in most of the recent work on dynamic network models, which aim to provide a deeper understanding of the intricate processes and connections within various systems (Park et al., 2020; Ellington & Baruník, 2020). This omission is largely due to the characteristics of popular time series models such as vector autoregressive models, which account for dynamic or process noise but overlook measurement noise. Process noise affects future values of the manifest variables through the variables' current and previous values, hence the name *process* or *dynamic* noise, while measurement errors are independent across time.

In longitudinal panel data, the exclusion of the measurement model for outcome variables reduces the precision of parameter estimates, especially outside ILD contexts (Kuhfeld & Soland, 2022). Although ILD studies often use composite scores to represent latent variables, this can compromise reliability and validity (McNeish et al., 2021), when not employing dynamic modeling that integrates a measurement model with multiple indicators, such as a dynamic factor model (Molenaar, 1985). Schuurman and Hamaker (2019) highlighted the benefits of incorporating measurement error using a single indicator in dynamic models, contrasting with approaches that neglect measurement error. However, they did not evaluate reliability, suggesting a research avenue, such as if within-person reliability in ILD is high, the difference between models assuming latent variables and those that do not may be negligible. This could imply potential model simplification without losing accuracy when data reliability is high. Additionally, existing literature lacks a comparison of model performance between single-indicator measurement models and dynamic factor models with the same structural relations.

Dynamic Structural Equation Models (DSEMs) are a powerful tool for analyzing ILD, integrating techniques from multilevel modeling, time series analysis, and structural equation modeling, and using Bayesian estimation (Asparouhov et al., 2018; McNeish & Hamaker, 2020). While many applications of DSEMs focus primarily or even exclusively on structural relationships between variables, they often overlook measurement models and differences in the quality of indicators for the variables. The main purpose of this study is to evaluate the impact of omitting measurement error in DSEMs. We aim to provide insights into effective sample size and power planning, considering different measurement models. Through a Monte Carlo simulation study, we evaluate the effects of omitting measurement error in DSEMs when they are indeed present across the conditions with different numbers of manifest indicators, effect sizes (within- and between-person reliability), and when multiple indicators are aggregated into a single composite score.

The remainder of this paper is organized as follows. Initially, we introduce the general DSEM framework. This is followed by an exploration of three special cases of DSEMs that serve as the basis of our evaluations. Next, we proceed with a discussion of how reliability may be quantified in ILD studies. Subsequently, we conduct a Monte Carlo simulation study to evaluate DSEM performance under various measurement conditions. We then present an empirical application that examines the impact of measurement error specifications on affective states and state personality data, demonstrating how key conclusions about individual dynamics are sensitive to these varied specifications. Lastly, we conclude with a discussion on the findings and directions for future research.

1.1. Motivating DSEM Variations

We will start by introducing the general DSEM framework adopted in this article, which distinctly divides the total variability observed in multivariate time series data into within-person and between-person components. Firstly, the total variability observed in the multiple multivariate time series data for individual i at time t can be decomposed into two parts (within & between) as detailed in Equation (1):

$$\boldsymbol{y}_{i,t} = \boldsymbol{y}_{i,t}^{W} + \boldsymbol{y}_{i}^{B}, \tag{1}$$

where the observed data vector $\boldsymbol{y}_{i,t} = (y_{1it}, y_{2it}, \ldots y_{pit})'$ is a p-dimensional vector, where p is the number of indicators. And $y_{j,i,t} (j = 1, 2, \ldots, p)$ is the observed value of the jth indicator for individual i at time t. The within-person component, $\boldsymbol{y}_{i,t}^{W}$, represents the deviations of individual i at time t from their corresponding individual-specific means. This component captures the transient, state-like changes that vary over time within an individual due to external influences or internal processes. In contrast, the between-person component, $\boldsymbol{y}_{i}^{B} = (y_{1i}^{B}, y_{2i}^{B}, \ldots y_{pi}^{B})'$, represents the individual-specific latent mean vector of indicators across measurement occasions for individual i. This component is stable across time and captures the traits of individuals, reflecting higher or lower stable scores on the indicators relative to others. After decomposing total variability into within and between variability, then we describe the within-level measurement model for the time-specific indicator vector $\boldsymbol{y}_{i,t}^{W}$. It is modeled as follows:

$$\begin{aligned} \boldsymbol{y}_{i,t}^{W} &= \boldsymbol{L}^{W}\boldsymbol{f}_{i,t} + \boldsymbol{\omega}_{i,t}, \quad \boldsymbol{\omega}_{it} \sim N(\boldsymbol{0}, \Sigma_{\omega}) \\ \boldsymbol{f}_{i,t} &= \Phi_{i}\boldsymbol{f}_{i,(t-1)} + \boldsymbol{\epsilon}_{i,t}, \quad \boldsymbol{\epsilon}_{it} \sim N(\boldsymbol{0}, \Sigma_{\epsilon}) \end{aligned} \tag{2}$$

where $\boldsymbol{f}_{i,t} = (f_{1it}, \ldots, f_{qit})'$ denotes the vector of q latent time-specific factors for individual i at time t, each factor f_{kit} (for $k = 1, 2, \ldots, q$) represents a within-person factor (i.e., latent state), measured by the indicators $y_{1it}^{W}, y_{2it}^{W}, \ldots, y_{pit}^{W}$. $\boldsymbol{L}^{\boldsymbol{W}} = (\lambda_{jki})_{p \times q}$ is the factor loading matrix for the kth latent factor of individual i, mapping it to the jth indicator. $\boldsymbol{\omega}_{\boldsymbol{it}} = (\omega_{1it}, \omega_{2it}, \ldots \omega_{pit})'$ represents the latent within-level measurement errors. It is assumed to follow a normal distribution, $\boldsymbol{\omega}_{it} \sim N(\boldsymbol{0}, \Sigma_{\omega})$. Then, the dynamic processes among within-factor (f_{kit}) are described as Φ_{i}, reflecting the dynamic relationship with itself and other latent variables at the previous occasion. This is also interpreted as individual carry-over effects between occasions of latent factors itself (i.e., Autoregression coefficients; AR) and spill-over effects between different latent factors (i.e., Cross-lagged regression coefficients; CR). Lastly, $\boldsymbol{\epsilon}_{it}$ is the vector of within residual variances (i.e., process noise, innovation, or dynamic errors), which reflects the reactivity of the dynamic processes for individual i at time t. We assume $\boldsymbol{\epsilon}_{it}$ follows a multivariate normal distribution, $\boldsymbol{\epsilon}_{it} \sim N(\boldsymbol{0}, \Sigma_{\epsilon})$.

Subsequently, we delineate the between-level measurement model for the individual-specific indicator vector $\boldsymbol{y}_{i}^{B}$. It is modeled as follows:

$$\boldsymbol{y}_{i}^{B} = \boldsymbol{v}_{i} = \boldsymbol{\alpha} + \boldsymbol{L}^{B}\boldsymbol{fb}_{i} + \boldsymbol{\psi}_{i}, \ \boldsymbol{fb}_{i} \sim N(\boldsymbol{0}, \Sigma_{\omega^{B}}), \ \boldsymbol{\psi}_{i} \sim N(\boldsymbol{0}, \Sigma_{\psi}) \tag{3}$$

where $\boldsymbol{fb}_i$ represents a between-factor (i.e., latent trait) vector encompassing scores of the within-level factors, which is assumed to follow a multivariate normal distribution, $\boldsymbol{fb}_i \sim MVN(\boldsymbol{0}, \Sigma_{\omega^B})$. Although Σ_{ω^B} can accommodate the same or a different number of latent factors at the between-level compared to the within-level, this scenario was not considered in this paper. The between-factor loadings are represented as $\boldsymbol{L^B} = (\lambda_{jk}^B)_{p\times q}$, where the matrix maps the kth between latent factor to the jth latent item intercepts. The vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)'$ represents the fixed effect intercept vector of $\boldsymbol{y}_i^B$, and $\boldsymbol{\psi}_i = (\psi_{1i}, \psi_{2i}, ..., \psi_{pi})'$ is the vector of random effects components of each intercept, with each component, such as ψ_{1i}, assumed to be normally distributed, $\boldsymbol{\psi}_i \sim \mathcal{N}(\boldsymbol{0}, \Sigma_\psi)$.

Next, we introduce three different bivariate variations of DSEM with the same structural model, but alternative ways of handling measurement error. The names of the models corresponding to the numbering of those introduced in the preceding section are as follows: (1) the multilevel first-order Vector Autoregressive model (ML-VAR(1)), (2) the multilevel first-order Measurement Error Vector Autoregressive model (ML-MEVAR(1)), and (3) the Multilevel first-order Dynamic Factor Vector Autoregressive model (ML-DFVAR(1)). Thorough introductions and illustrations of related variations can be found elsewhere (Asparouhov et al., 2018; Li et al., 2022; McNeish et al., 2021; Oh & Jahng, 2023; Schuurman & Hamaker, 2019).

1.1.1. The Multilevel First-Order Vector Autoregressive Model (ML-VAR(1))

The first variation of DSEM considered is a multilevel extension of the VAR model, a dynamic model widely adopted in the network modeling literature that captures over-time dynamics of a set of variables only at the manifest indicators level (Epskamp et al., 2018; Hamaker et al., 2015; Wang et al., 2012; Li et al., 2022; Park et al., 2020; Wang et al., 2012). For VAR, the general formulations previously described can be simplified by imposing specific constraints. Specifically, for the VAR model, both $\boldsymbol{L^W}$ and $\boldsymbol{L^B}$ should be $\mathbf{I}$ matrices, and Σ_ω and Σ_{ω^B} should also be set to zero. These constraints facilitate a straightforward adaptation of the general model to explain the dynamics in VAR configurations. Specifically, we considered a bivariate ML-VAR(1) model (see Figure 1a) expressed as:

$$\boldsymbol{y}_{i,t}^W = \Phi_i \boldsymbol{y}_{i,(t-1)}^W + \boldsymbol{\epsilon}_{i,t}, \quad \boldsymbol{\epsilon}_{i,t} \sim N(\boldsymbol{0}, \Sigma_\epsilon) \tag{4}$$

$$\boldsymbol{y}_i^B = \boldsymbol{\nu}_i = \boldsymbol{\alpha} + \boldsymbol{\psi}_i, \quad \boldsymbol{\psi}_i \sim N(\boldsymbol{0}, \Sigma_\psi) \tag{5}$$

By distinguishing the variance from the manifest observations y_{jit} $(j = 1, 2)$ for person i at time point t and variable j, we can get within components $\boldsymbol{y}_{i,t}^W$ and between components $\boldsymbol{y}_i^B$. As shown in Equations (4), the structural relationship between variables is modeled in within components, here as a 2×2 coefficient matrix in which $\phi_{11,i}$, and $\phi_{22,i}$ represent person-specific AR parameters, and $\phi_{21,i}$ and $\phi_{12,i}$ represent person-specific CR parameters. Process noises are described as $\epsilon_{j,i,t}$ $(j = 1, 2)$, where Σ_ϵ is the process noise covariance matrix with variances $\sigma_{\epsilon 1}^2$ and $\sigma_{\epsilon 2}^2$, respectively. As distinct from the broader DSEM framework presented elsewhere (Asparouhov et al., 2018), here we restrict the process noise variances to be person-invariant to reduce the complexity and computational burden associated with the DSEM models for the small to moderate sample size configurations considered in this article. In addition, consistent with standard VAR conventions, all dynamics are hypothesized to unfold at the manifest variable level, with only process noises, but not measurement errors in the system. That is, the influence of previous process disturbances at time $t-1$ $(\epsilon_{j,i,t-1})$ continues to be carried forward to $y_{1,i,t}$ and $y_{2,i,t}$ through the auto- and cross-regression effects.

By employing the between components $(\boldsymbol{y}_i^B)$, a source of between-person differences, the person-specific intercepts that may be understood as typical or trait levels of the dynamic processes are modeled. As shown in Equation (5), α_j and $\psi_{j,i}$ represent the grand mean of the person-specific intercept and random effect associated with the jth person-specific intercept for person i, for manifest process $y_{j,i,t}$. This model posits that the deviations of $y_{1,i,t}$ and $y_{2,i,t}$ from their person-specific intercepts, $\nu_{1,i}$ and $\nu_{2,i}$ relate to the deviations of $y_{1,i,t-1}$ and $y_{2,i,t-1}$ from the same intercepts at the previous time point.

Thus, in this model, the within-person deviations from the person-specific intercepts are postulated to follow VAR(1) relations, but between-person differences exist in such within-person dynamics through interindividual differences in AR and CR coefficients: $\phi_{11,i}$, $\phi_{12,i}$, $\phi_{21,i}$, and $\phi_{22,i}$. In this regard, the *between-level model* that allows for the examination of individual differences in the variables is expressed as (see Figure 1a):

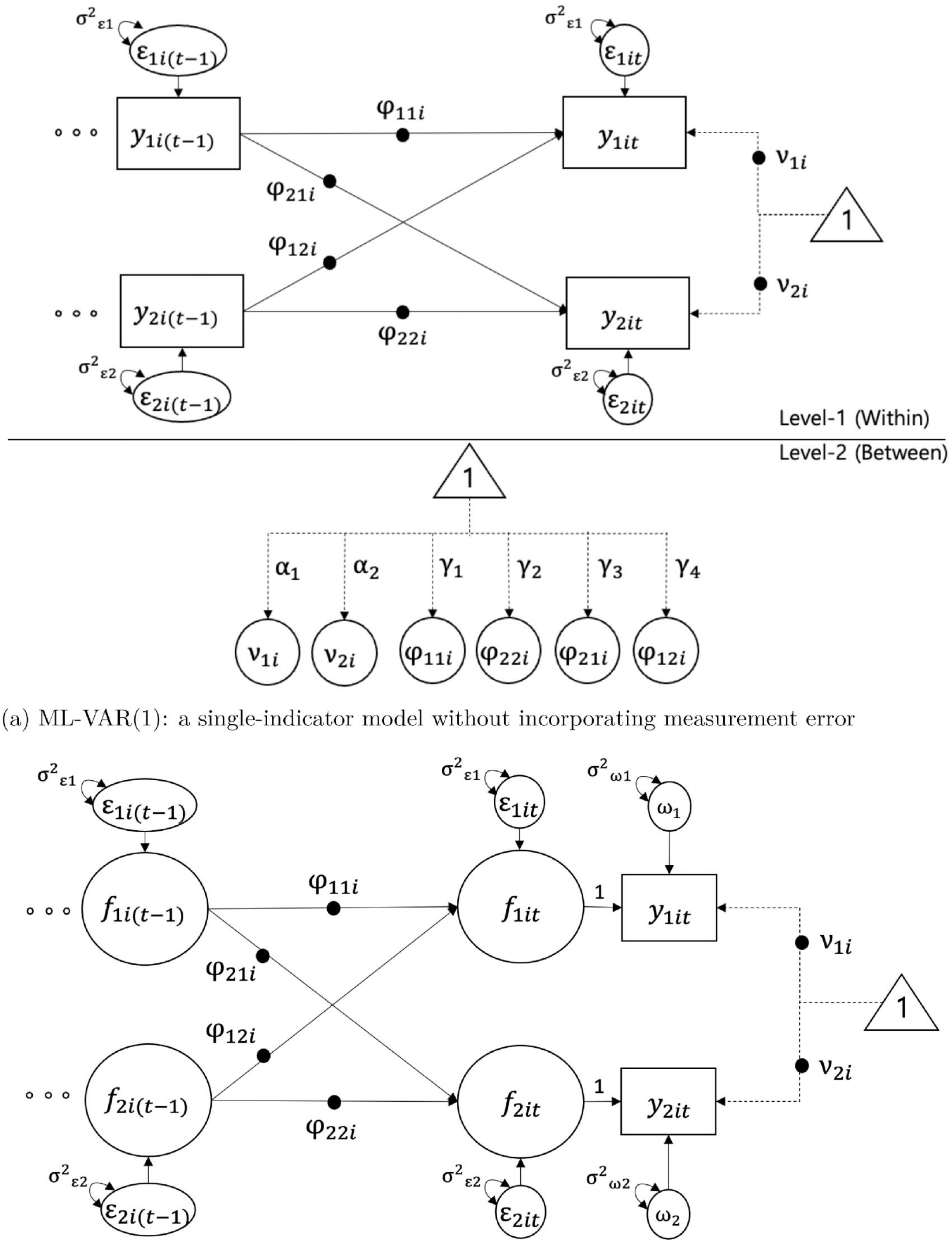
$$\begin{aligned} &\phi_{11i} = \gamma_1 + u_{1i}, \quad \phi_{22i} = \gamma_2 + u_{2i}, \quad \phi_{12i} = \gamma_3 + u_{3i}, \\ &\phi_{21i} = \gamma_4 + u_{4i}, \quad \nu_{1i} = \alpha_1 + \psi_{1i}, \quad \nu_{2i} = \alpha_2 + \psi_{2i} \end{aligned} \tag{6}$$

where γ_1–γ_4 represent the fixed effects, or grand mean of the AR and CR parameters respectively, with corresponding random effect components, u_{1i}–u_{4i}.

1.1.2. The Multilevel First-Order Measurement Error Vector Autoregressive Model (ML-MEVAR(1))

Conventional SEM cannot identify a measurement model with a single indicator due to insufficient information to simultaneously estimate true scores and error variances. However, in the context of time series analysis, repeated measurements over time provide sufficient information to identify both true scores and error variance through a single indicator. That is, measurement error variances are distinguishable from latent process noises because the influence of the latter carries forward in time, whereas measurement noise affects one occasion and that occasion only without propagating forward (Hamilton, 2020). We detail how these differences between measurement and process noises may be specified within the ML-MEVAR(1), a special case of the DSEM model summarized in Equations (1–3).

The ML-MEVAR(1) can be obtained as a special case of the DSEM model by setting the matrices $\boldsymbol{L^W}$ and $\boldsymbol{L^B}$ as identity matrices $\mathbf{I}$, and $\boldsymbol{\omega_i^B}$ to zero. In this context, the



(a) ML-VAR(1): a single-indicator model without incorporating measurement error

(b) ML-MEVAR(1): a measurement model with a single-indicator

Figure 1. ML-VAR(1) and ML-MEVAR(1) model.

second DSEM variation considered in this study is a multilevel, multivariate extension of the model in Equations (4–5), denoted herein as the multilevel Measurement Error Vector Autoregressive (1) model (i.e., ML-MEVAR(1)), expressed (see Figure 1b) as:

$$\boldsymbol{y}^W_{i,t} = \boldsymbol{I}\boldsymbol{f}_{i,t} + \boldsymbol{\omega}_{i,t}, \quad \boldsymbol{\omega}_{i,t} \sim N(\boldsymbol{0}, \Sigma_\omega), \tag{7}$$

$$\boldsymbol{f}_{i,t} = \Phi_i \boldsymbol{f}_{i,(t-1)} + \boldsymbol{\epsilon}_{i,t}, \quad \boldsymbol{\epsilon}_{i,t} \sim N(\boldsymbol{0}, \Sigma_\epsilon) \tag{8}$$

For model identification, Φ_i cannot be a null matrix for all individuals; Σ_ϵ is a diagonal matrix, and Σ_ω may be a full or diagonal symmetric matrix.

As indicated in the measurement model (Equation 7), each within components $\boldsymbol{y}^W_{i,t}$ (j = 1,2) has a factor loading of 1 for its corresponding latent factor ($\boldsymbol{f}_{i,t}$). Both the ML-VAR(1) and ML-MEVAR(1) models assume the same between-person differences in parameters. The between-person model of MEVAR is the same as Equations (5) and (6).

A key difference between them is the inclusion of measurement error terms, $\boldsymbol{\omega}_{\boldsymbol{i},\boldsymbol{t}}$, in the ML-MEVAR(1), as specified in the measurement model in Equation 7. That is, the measurement model considers the within-components by postulating a latent factor with a loading of 1 to account for measurement error. This allows for an analysis of the structural relationships among the factors, as seen in Equation 8, distinguishing it from the VAR model.

1.1.3. The Multilevel Dynamic Factor Vector Autoregressive Model (ML-DFVAR(1))

The third DSEM variation considered in the present article is a multilevel dynamic factor VAR model in which six manifest indicators are used to define two latent factors (with 3 indicators each), and VAR(1) relations are assumed to characterize the dynamics of the latent factors. Other variations from this class of dynamic factor models (Molenaar, 1985) used to describe the relations in multivariate time series data can be found elsewhere (Gates et al., 2023; Nesselroade et al., 2002).

The "within" and "between" portions of the DFVAR are expressed as:

$$\begin{aligned} \boldsymbol{y}_{i,t}^{W} &= \boldsymbol{L}^{W}\boldsymbol{f}_{i,t} + \boldsymbol{\omega}_{i,t}, \quad \boldsymbol{\omega}_{i,t} \sim N(\boldsymbol{0}, \Sigma_{\omega}) \\ \boldsymbol{f}_{i,t} &= \Phi_i \boldsymbol{f}_{i,(t-1)} + \boldsymbol{\epsilon}_{i,t}, \quad \boldsymbol{\epsilon}_{i,t} \sim N(\boldsymbol{0}, \Sigma_{\epsilon}) \end{aligned} \tag{9}$$

$$\begin{aligned} &\boldsymbol{y}_{i}^{B} = \boldsymbol{v}_i = \boldsymbol{\alpha} + \boldsymbol{L}^{B}\boldsymbol{fb}_i + \boldsymbol{\psi}_i, \quad \boldsymbol{fb}_i \sim N(\boldsymbol{0}, \Sigma_{\boldsymbol{\omega}^{B}}), \\ &\boldsymbol{\psi}_i \sim N(\boldsymbol{0}, \Sigma_{\psi}) \end{aligned} \tag{10}$$

In Equation 9, $\boldsymbol{y}_{i,t}^{W}$ represents the six-dimensional vector of within components assumed to load onto two-dimensional latent vector $\boldsymbol{f}_{i,t}$ from the general formulation. $\boldsymbol{L}^{W}$ is the 6×2 factor loadings matrix for of individual $\boldsymbol{i}$. Consequently, Equation 10 models a six-dimensional random intercept, $\boldsymbol{v}_i$. These person-specific intercepts are depicted as random variables, marked by small black dots, as shown in Figure 2a. The model specifications and assumptions for the DFVAR model correspond to those described in Equation 2 for the within model and Equation 3 for the between model. As with the ML-MEVAR(1) model, for the measurement and process noise-related parameters to be uniquely distinguishable from each other, the measurement error covariance matrix, Σ_{ω}, has to be diagonal in structure; the process noise covariance matrix, $\boldsymbol{\epsilon}_{i,t}$, can assume a full or diagonal symmetric structure; and not all individuals' Φ_i can be null matrices.

In summary, the three DSEM models share the same conceptual framework, modeling identical structural relationships. They differ only in how they handle measurement errors and the between measurement models in the DFVAR model. Previously, we discussed these DSEM variations in bivariate cases and generalized formats. Building on this, we now focus on reliability, maintaining a general approach for multivariate analysis and setting the stage for evaluating reliability across different model configurations. Thus, we will explore how reliability can be quantified within ILD using the concepts of within- and between-reliability across these three DSEM models. Hereafter, the three multilevel bivariate models will be referred to as VAR, MEVAR, and DFVAR.

1.2. Reliability in ILD

In the context of longitudinal studies, the reliability of measurement is a critical concern, with previous work highlighting key factors such as the number of measurement occasions and measurement error variance. Previous work on reliability in longitudinal contexts focused on growth curve models. To name just a few of the important contributions to representing reliability in this longitudinal context, Willett (1989) derived a coefficient for reliability in the growth rate itself; Laenen et al. (2009) derived both time-average and time-specific reliability coefficients; and Marcoulides (2019) reviewed and specified several additional longitudinal reliability metrics. In the context of multilevel vector autoregression models, Schuurman and Hamaker (2019) proposed explicit indices for evaluating both within- and between-person reliability. In this article, we capitalize on Schuurman and Hamaker's reliability indices to investigate the consequences of measurement misspecification as related to variations in within- and between-person reliability levels.

As shown earlier, observed variables can be disaggregated into between and within components. The within components can be further divided into true within-person variance and measurement error variance, thus decomposing the total variance into three parts (Schuurman & Hamaker, 2019). The variance of the random intercept term (inter-individual variance) is denoted by $\Sigma_{\boldsymbol{\psi}}$, the expected within-person fluctuations variance by $\mathbb{E}[\mathrm{Var}(\boldsymbol{f}_{\boldsymbol{i}})] = \Sigma_{f_{it}}$, and the expected measurement error variance within individuals by $\Sigma_{\boldsymbol{\omega}}$. Therefore, the total variance is expressed in Equation 11:

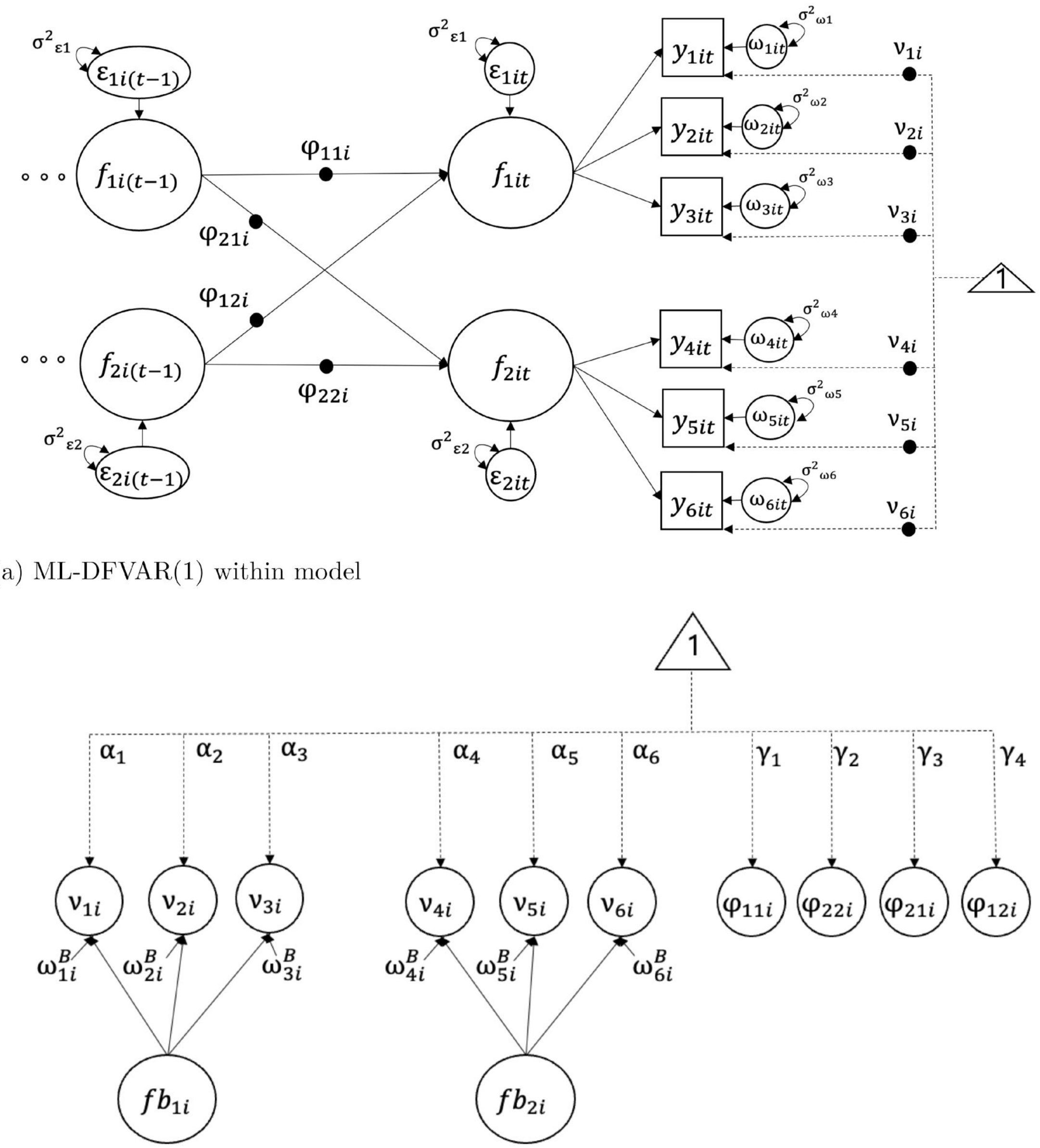
$$V(\boldsymbol{y}) = \Sigma_{\boldsymbol{\psi}} + \Sigma_{f_{it}} + \Sigma_{\boldsymbol{\omega}} \tag{11}$$

For the $\Sigma_{f_{it}}$ we use the expected value of the asymptotic variance of the within factors as given by Equation 12.

$$\Sigma_{f_{it}} = (\boldsymbol{I} - \Phi_{\boldsymbol{i}} \otimes \Phi_{\boldsymbol{i}})^{-1}\mathrm{vec}(\Sigma_{\boldsymbol{\epsilon}_{\boldsymbol{i}}}) \tag{12}$$

where $\boldsymbol{I}$ is an identity matrix, $\otimes$ indicates the Kronecker product, function vec() transforms a matrix into a column vector, and mat() transforms a vector into a matrix (cf., Kim & Nelson, 1999). This unique variance corresponds to the diagonal entries in the individual's covariance matrix $\Sigma_{f_{it}}$, encapsulating the variances and covariances of f for the individual. Even in a VAR model, this matrix serves the same purpose as in the MEVAR, capturing structural relationships, but without latent modeling in VAR. By changing the notation from $\Sigma_{f_{it}}$ to $\mathbb{E}[\mathrm{Var}(y)]$, it effectively captures all variances and covariances as specified in Equation (4), highlighting the similarity between the two processes despite the lack of latent modeling in VAR.

Extending the derivations by Schuurman and Hamaker (2019) to allow for the inclusion of a measurement structure, we can define reliability for models VAR and MEVAR to allow for the inclusion of a measurement structure separating between variability (e.g., $\Sigma_{\boldsymbol{\psi}}$), measurement error variances (e.g., $\Sigma_{\boldsymbol{\omega}}$), and true within variances (e.g., $\Sigma_{f_{it}}$). The total variance of an observed multivariate y for an "average" person as defined by Equations (2) and (3).



(a) ML-DFVAR(1) within model

(b) ML-DFVAR(1) between model

Figure 2. ML-DFVAR(1): measurement models with multiple indicators.

$$\begin{aligned}\text{Prototypical Var}(\boldsymbol{y}) &= \text{Between-vars} + \text{Prototypical Within-vars}\\ &= (\boldsymbol{L}^B\Sigma_\psi\boldsymbol{L}^{BT} + \Sigma_{\omega^B}) + (\boldsymbol{L}^W\Sigma_{f_{it}}\boldsymbol{L}^{WT})\end{aligned} \tag{13}$$

Also, note that the between factor variance impacts Σ_ψ and not $\Sigma_{f_{it}}$. The decomposition of variance as presented in Equation (1.2) permits the estimation of a spectrum of reliabilities for our measurements of the observed variable y. For instance, an aggregate reliability of our y measures that encompasses both intra-individual and inter-individual variances can be derived by Equation (14):

$$\left(\Sigma_\psi + (\boldsymbol{L}^W\Sigma_{f_{it}}\boldsymbol{L}^{WT})\right)\left(\Sigma_\psi + (\boldsymbol{L}^W\Sigma_{f_{it}}\boldsymbol{L}^{WT}) + \Sigma_\omega\right)^{-1} \tag{14}$$

Nevertheless, such reliability combines the variances caused by two disparate true scores: the stable trait scores Σ_ψ and the true within-person variability $\Sigma_{f_{it}}$. Consequently, while observed scores may reliably measure specific traits, they might not accurately reflect an individual's true variations over time, and vice versa. This conflation implies that conventional reliability calculations do not clearly indicate how suitable these observed scores are for specific inferences. Additionally, it overlooks that the consistency of these scores in capturing within-person fluctuations can vary between individuals. Therefore, it is more beneficial to separately evaluate the reliability of trait scores and within-person variances.

1.2.1. Reliability for Between-Person Differences

The estimation of between-person reliability is focused on the trait scores, which are the systematic components of interest in variability. It should be noted that, for simplicity, the random variances and covariances are not permitted to covary among themselves or with other random variables such as the random trait scores or regression coefficients.

The between-person reliability of the variable y can be derived as follows:

$$rel_B(\boldsymbol{y}) = \Sigma_\psi \left(\Sigma_\psi + (\boldsymbol{L}^W \Sigma_{f_{it}} \boldsymbol{L}^{WT}) + \Sigma_\omega\right)^{-1} \tag{15}$$

Subsequently, the between-person reliability may be diminished if the deviations of the trait scores over time are substantial. In essence, similar to conventional reliability assessments, the within-person variability observed in scores is not the primary focus and is deemed as noise in the context of this reliability measure. This between-person reliability signifies the dependability of measurements that one could anticipate, on average, if the measurement of trait scores was of primary interest, analogous to a single measurement per participant in a cross-sectional investigation.

1.2.2. Reliabilities for Within-Person Dynamics

The dynamics of each individual's experience can differ significantly, which may affect the reliability of their respective score assessments. By examining the variances unique to each individual as outlined in Equation (12), the reliability of within-person fluctuations can be assessed separately for each individual. It is important to note that for any individual, the between-person variance component is omitted, hence the term Σ_ψ is not included. The reliability of an individual's observed scores is then calculated by Equation (16):

$$rel_W(\boldsymbol{y_i}) = (\boldsymbol{L}^W \Sigma_{f_{it}} \boldsymbol{L}^{WT})\left((\boldsymbol{L}^W \Sigma_{f_{it}} \boldsymbol{L}^{WT}) + \Sigma_\omega\right)^{-1} \tag{16}$$

This equation illustrates that the reliability can vary among individuals due to differences in the autoregressive or cross-lagged associations, or when the variability of their new inputs, denoted by $\Sigma_{f_{it}}$, differs. Similarly, variations in measurement error, represented by Σ_ω, also contribute to these differences in reliability.

2. Simulation Study

2.1. Simulation Designs

This simulation study aimed to investigate the consequences of omitting measurement errors in three DSEM variations, focusing on point estimates, standard errors, model complexities, statistical power, sample size, time points, and reliability. We sought to address the following research questions: (1) What are the consequences of omitting measurement errors when they are indeed present in low-reliability conditions? (2) How do these effects change across varying levels of reliability? (3) How do measurement model performances differ between models using a single indicator and those utilizing multiple indicators? and (4) How do the effects differ across various sample size configurations in terms of N and T?

To address these questions, we performed a simulation study to demonstrate the practical model performances under different conditions with 100 Monte Carlo replications for each condition. Four factors were manipulated, including sample size ($N = 200, 500$), number of repeated measurement occasions ($T = 60, 150$), the extent of reliability (low vs. high), and number of indicators in the model (single vs. multiple). This resulted in $2 \times 2 \times 2 \times 2 = 16$ conditions. Since the minimum recommended sample size for modeling random intercepts, AR, and process noise starts at $N = 200$, $T = 100$ (Asparouhov et al., 2018; Schultzberg & Muthén, 2018). We chose $N = 200$ and $T = 60$ and extended to $N = 500$ and $T = 150$ to examine the effects of sample size and time points. This design balanced the total data length, ensuring equivalency with $N200 \times T150 = N500 \times T60$.

For the construction of different levels of between and within reliability, we manipulated reliability by altering the variance of the intercept as the source of between variability, and measurement error variance as the source of within variability, excluding true within-variability. For the low-reliability condition, within-reliability was set to 0.6 and between-reliability to 0.5 by setting the intercept variance to 4 and the measurement error variance to 3. For the high-reliability condition, within-reliability was set to 0.9 and between-reliability to 0.7 by setting the intercept variance to 2.5 and the measurement error variance to 0.4. High and low-reliability thresholds were based on conventional interpretations, with 0.9 considered high and 0.6 considered low, similar to Cronbach's alpha standards. The choice of parameter values, such as the intercept variance of 2.5 for high-reliability, was based on empirical ILD analysis (Oh & Jahng, 2023).

Data generation was conducted in R 4.3.1, and model fitting was executed in Mplus 8.4 using the MplusAutomation package in R (R Core Team, 2023; Muthén & Muthén, 2017; Hallquist & Wiley, 2018). To address the first two research questions, we generated data based on the MEVAR model specified in Equation 7 as the true model with varying sample sizes, time points, and reliability conditions. We then fitted the correctly specified MEVAR or the misspecified VAR model, omitting measurement errors. For the third research question, we generated data using the DFVAR model specified in Equation 9 with 6 manifest variables. To mirror the previous discussions on cross-level measurement invariance in multilevel factor models (Ryu, 2014) and the dynamic SEM framework (Asparouhov et al., 2018), we imposed equality constraints between the within-level factor loadings ($\boldsymbol{L}^W$) and the between-level factor loadings ($\boldsymbol{L}^B$) in both data generation and model fitting. We emphasize, however, that these invariance constraints are neither strictly required for model identification nor do they meaningfully affect the key results, and may be relaxed if theoretical or empirical considerations warrant. We fitted the data to the true model (DFVAR) and two incorrect single-indicator models (VAR & MEVAR) using composite scores, which were averages across the item subset for each factor. Composite scores fitted to a single-indicator measurement model are referred to as MEVARC, and those without a measurement model as VARC. Detailed code for data generation and model fitting can be found in the Online Supplementary Materials.

As a consequence, we fitted the MEVAR, VAR, DFVAR, MEVARC, VARC model for the generated data set using two MCMC chains of 300,000 iterations, of which the first half was discarded as burn-in. Model performances were evaluated using relative biases, root-mean-square errors (RMSEs), 95% coverage rate, standard deviation biases, and power for the parameter estimates. The convergence of MCMC chains was monitored using the trace plot for each parameter and the Gelman–Rubin potential scale reduction factor (PSRF) less than 1.1 (Brooks & Gelman, 1998; Gelman & Rubin, 1992), which is the Mplus default setting, with a value close to 1 indicating convergence. In addition, we monitored the effective sample size (ESS) for each parameter, aiming for an ESS greater than 400 as recommended by Zitzmann and Hecht (2019). It is important to note that a higher number of iterations is typically needed to accomplish this target ESS level for variance parameters such as process noise or measurement error variance. Following preliminary inspection of the diagnostic information, we decided to employ 300,000 MCMC iterations on two separate chains for each Monte Carlo replication of the simulation study. Notably, since Mplus does not automatically report ESS in its default settings, researchers need to use the `SAVEDATA: BPARAMETERS = mcmc_samples.dat` option to save MCMC samples for calculating ESS. In the following section, we organize our simulation results in the order of the research questions.

2.2. Simulation Results

2.2.1. Consequences of Omitting Measurement Errors under Low Reliability

As elaborated earlier, widely used, standard form of VAR models only includes process noises but not measurement errors, whereas the other two models, MEVAR and DFVAR, do incorporate them. This distinction naturally leads to the question of how model performance might differ when measurement errors are not considered. We first address the question by focusing on the relative biases in low-reliability conditions. Relative biases of the parameters are illustrated in Figure 3, which serves to determine whether a parameter is overestimated or underestimated. The conventional threshold for an acceptable level of relative bias, denoted as ±0.1, is also represented by dashed lines in Figure 3.

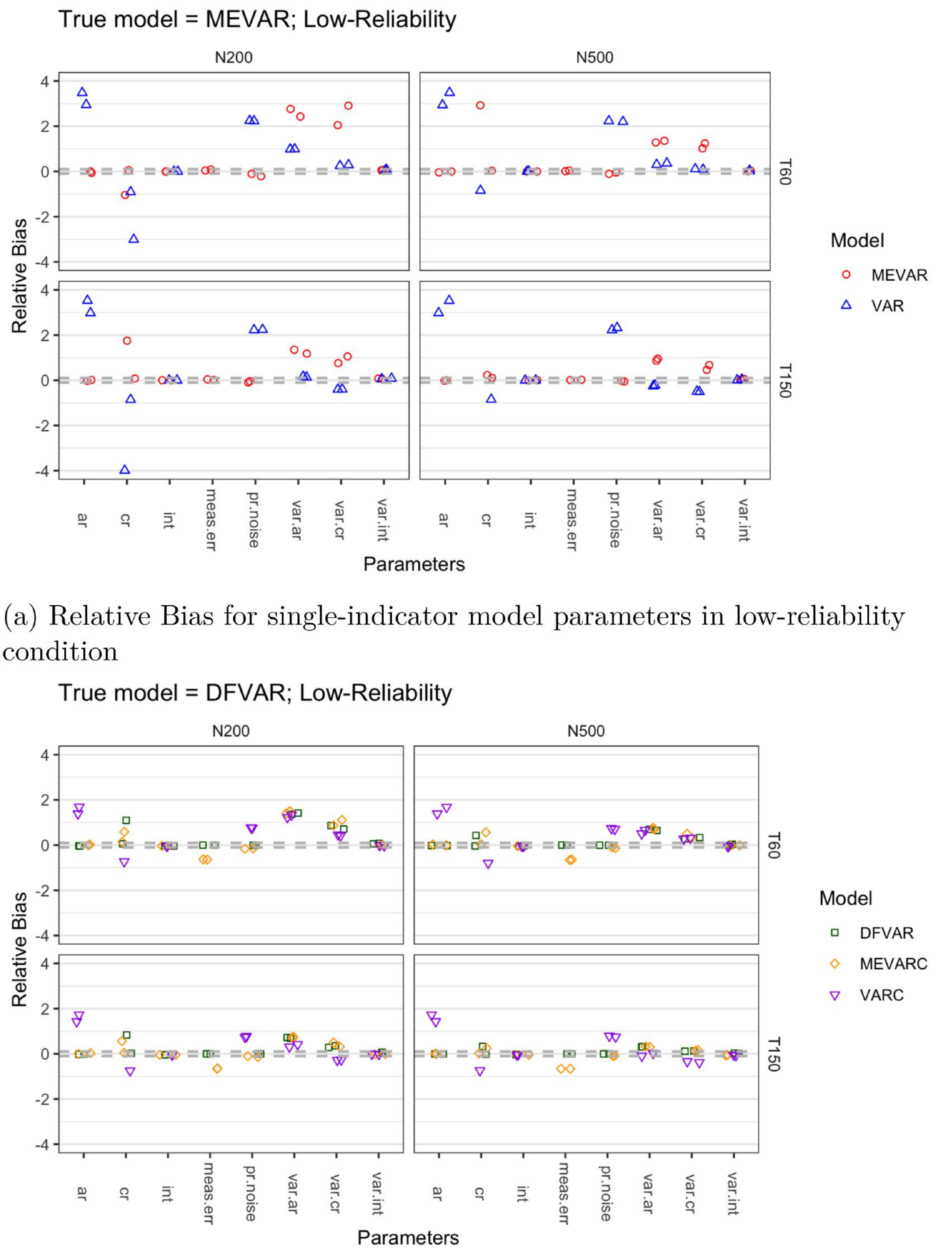
As shown in Figure 3a, when the true data generating model was MEVAR but the VAR model was fitted, significant positive relative biases (exceeding 3–4) were observed in the AR and process noise parameters compared to the correctly specified MEVAR model. The MEVAR model maintained an acceptable level of relative bias for most parameters including measurement error, while the VAR model could not estimate measurement error. Intercept parameters (e.g., int) and random effect variance parameters (e.g., var.ar, var.int) were relatively unaffected by the omission of measurement errors. However, CR showed large relative biases in all conditions due to small true values (e.g., 0.042, −0.002), which, when used in the relative bias formula ($(\hat{\theta} - \theta) \ / \ \theta$), resulted in disproportionately large bias values even for small differences. Figure 3b revealed a similar pattern, where the VARC model, representing results from fitting the VAR model to composite scores from multiple indicators generated using the DFVAR model, showed greater positive relative biases in dynamic parameters (i.e., ar, cr, and process noise) compared to the DFVAR model.

In summary, the omission of measurement errors when they were indeed present yielded a pronounced overestimation of the AR, CR, and process noise variance parameters. Based on these relative bias evaluations, it can be concluded that in data abundant with measurement errors, models incorporating a measurement framework outperform models without accounting for measurement error. This pattern is evident regardless of whether the true model is a single indicator (Figure 3a) or multiple indicators (Figure 3b). Regarding power, excluding the cross-lagged effects with very small true values, it surpassed 0.8 across all conditions. Importantly, satisfactory power levels were attained even in scenarios with relatively few data lengths, such as N = 200 and T = 60, under conditions of low reliability. Under the low-reliability conditions, we observed substantial relative biases due to the omission of measurement errors. Corresponding RMSEs, biases, and coverage rates also exhibited significant gaps. Under these low-reliability conditions, such differences are anticipated. Accordingly, in the subsequent section, we will investigate the extent to which these effects diminish under high-reliability conditions and provide a more detailed analysis of the observed patterns.

2.2.2. Effects of Measurement Error Omission across Reliability Conditions

To elucidate the effects of measurement error omission under different reliability conditions, we generated data under $N = 200$, 500 and $T = 60$, 150 under high within (0.9) and between (0.7) reliability conditions. In this study, we did not generate data with perfect reliability, as it is unrealistic for ILD. Instead, we established a high-reliability condition to simulate a more realistic and achievable level of reliability, allowing us to assess model performance under feasible conditions.

The results for high-reliability conditions are shown in Figure 4. Despite increased data reliability, the VAR and VARC models still showed relative bias in AR and process noise compared to the MEVAR and DFVAR models. However, the magnitude of this relative bias decreased as reliability increased. For process noise, the relative bias dropped from 2 to 3 in low-reliability to below 1 in high-reliability condition, though ideally, a relative bias below 0.1 would indicate an accurate estimate. Similarly, the relative bias in AR decreased: from approximately 4 in single-indicator models and 2 in multiple-indicator models in the low-reliability condition to around 1 and 0.5, respectively in high-reliability. Nevertheless, a pattern of overestimation persisted across both parameters, regardless of whether the model was single-indicator or multiple indicators when we omit the measurement error. Moreover, as shown in

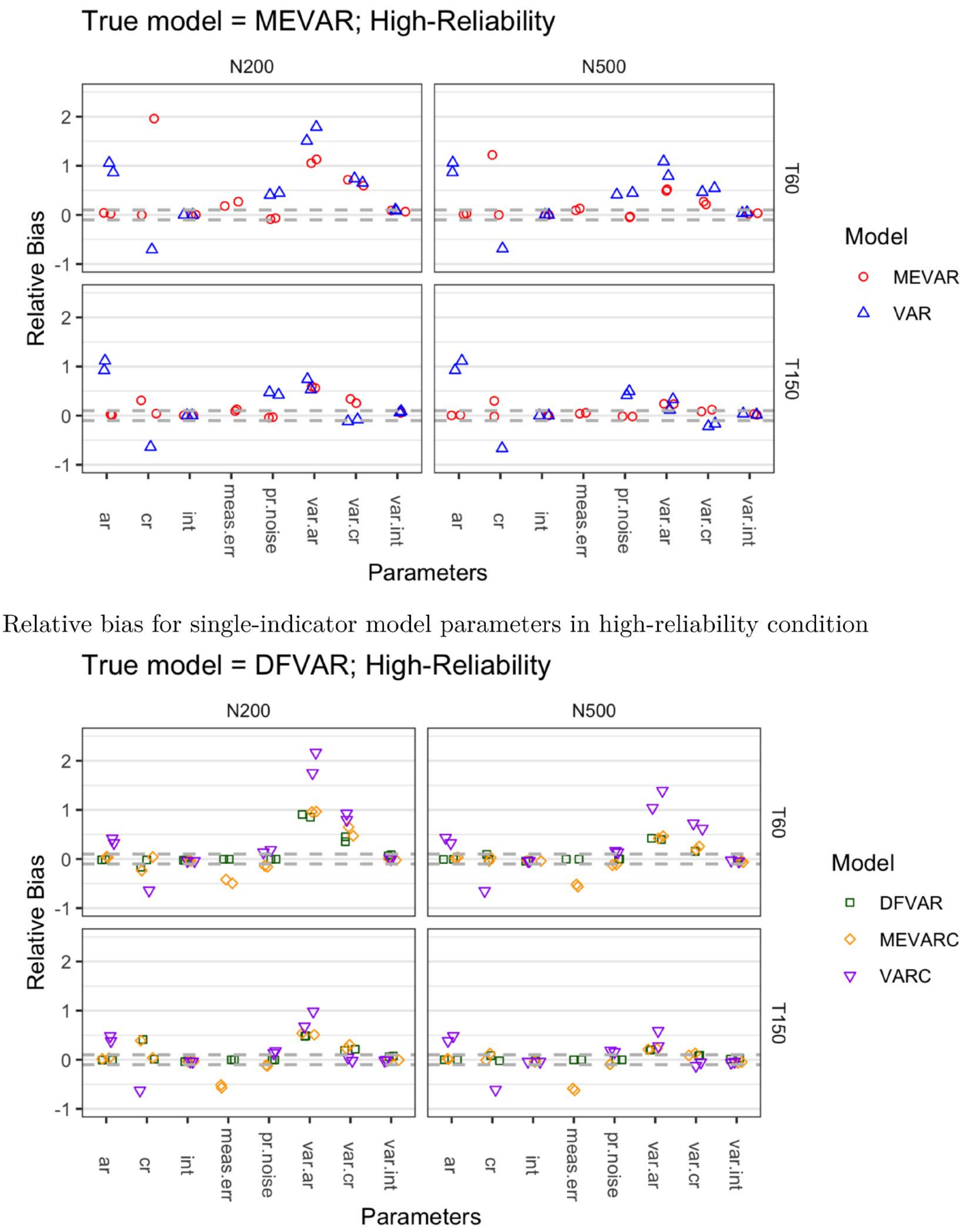


(a) Relative Bias for single-indicator model parameters in low-reliability condition

(b) Relative Bias for multiple indicators model parameters in low-reliability condition

Figure 3. Relative bias of single-indicator and multiple indicators model parameters in low-reliability conditions. The dashed line marks the ±0.1 conventional threshold of acceptable relative bias. ar = autoregression, indicating lagged effect of a variable on itself. cr = cross-regression, representing lagged effects from other variables. int = intercept, representing the latent average of the dynamic variable over the measurement occasions. meas.error = measurement error variances, reflecting residual variances with no impact on dynamics; pr.noise = process noise, capturing unmeasured stochastic influences on dynamics. var.ar, var.cr, and var.int = variances of the ar, cr, and intercept. In the legend, MEVAR = fitting MEVAR using MEVAR as the true model, VAR = fitting VAR using MEVAR as the true model. DFVAR = fitting DFVAR using DFVAR as the true model, MEVARC = fitting MEVAR to composite scores computed across multiple items generated using the DFVAR as the true model, VARC = fitting VAR using DFVAR as the true model.

Figure 4, when the true model is specified, the relative bias for the dynamic parameters of interest (e.g., AR, CR, and process noise) was below 0.1 in all cases, except when the true value of the CR parameters was very close to zero (-0.002). In this context, as seen in Figure 4b, the MEVARC model accounting for measurement error showed relative bias close to 0.1 for the dynamic parameters, even though the true model was the DFVAR. This low relative bias for



(a) Relative bias for single-indicator model parameters in high-reliability condition

(b) Relative bias for multiple indicators model parameters in high-reliability condition

Figure 4. Relative bias of single-indicator and multiple indicator models parameters in high-reliability conditions.

the dynamic parameters was observed despite the large relative bias for the measurement error variance estimates. Although the true model showed a large relative bias for between-person parameters – such as the variance of AR and CR – we can conclude this inflation was due to division by a near-zero denominator (true value = 0.01), because their absolute bias was small. As data reliability increased, the gap between models with and without measurement error narrowed, with the MEVAR and VAR showing a bigger discrepancy than the DFVAR and VARC. In high-reliability conditions, the power of the parameters remained similar to those in low-reliability conditions. Most conditions achieved a power above 0.8 for all parameters except for the CR parameters with near-zero true values, even with smaller sample sizes and without accounting for measurement error in the model.

Regarding other simulation results than relative bias, such as RMSE, 95% Coverage rate, and SD bias, we posit that if discrepancies emerge in high-reliability conditions, they are likely to be mirrored in low-reliability conditions as well, thus we will primarily focus on the results of high-reliability conditions. Upon examining the RMSE, a similar pattern to relative bias is noticeable, particularly in the context of point estimate results. The RMSE for the VAR model remains higher than that for the MEVAR, notable in AR and process noise. The performance gap between models gets smaller in AR, with the VARC model continuing to exhibit suboptimal performance. With respect to the 95% coverage rates, neither the MEVAR model for single-item data nor the DFVAR model for multiple-item data achieved satisfactory coverage rates for all parameters. Specifically, the variances of AR and CR showed poor coverage in low sample size conditions. In misspecified models, many parameters exhibited poor coverage. This shortfall prompts a closer investigation into whether Bayesian credible intervals (CIs) consistently miss true values due to overly narrow intervals or deviations from the true model parameters, as indicated by an examination of SD bias.[1] The results of SD bias are shown in Figure 5. A negative SD bias indicates consistently narrower Bayesian CIs, while a positive SD bias indicates overestimated posterior standard errors. In both Figure 5a and b, the positive SD bias in the variance of AR suggests an overestimated standard error, implying the true value is further from the estimated mean. However, the SD bias for this between-person parameter demonstrated improved estimation accuracy with increased sample size, implying that although uncertainty around the estimates is accurately quantified, the point estimates themselves remain biased. The misspecified model displays a positive SD bias, indicating a significant distance between true values and estimates. For the MEVARC model, the positive SD bias in measurement error, process noise, and AR variance suggests a similar discrepancy. However, since this does not affect the actual point estimate bias, it suggests that the MEVARC model likely has broader CIs.

2.2.3. Effects of Measurement Error Omission and Misspecification of Measurement Model Structure

While we tried to gain insights into the effects of measurement error omission, we fitted different models with and without measurement error parameters to data characterized by various levels of reliability. As observed in Figures 3 and 4, both the MEVAR and DFVAR models significantly outperformed the VAR and VARC models, achieving a satisfactory level of relative bias.

In this section, we are trying to change the viewpoint from examining performance differences between models with and without measurement errors to directly comparing the two models accounting for measurement errors. However, direct comparison between these two models is challenging; fitting data generated from a true model with a single item into a DFVAR model is unfeasible, whereas fitting a MEVAR model with composite scores derived from data generated by a true model with multiple items is viable. Hence, we focus on the results that data generated using the ML-DFVAR(1) model with three indicators per VAR process with high and low-reliability conditions, and fitted to ML-DFVAR(1), ML-MEVAR(1) using composite scores (MEVARC; denoted in Figure 1b), and ML-VAR(1) using composite scores (VARC).

In our study, we observed a consistent pattern of underestimation of measurement error variance in the MEVARC model compared to the DFVAR model. In both low- and high-reliability conditions, as depicted in Figures 3b and 4b, the MEVARC model exhibits an underestimation of measurement error variance compared to the DFVAR model. This underestimation likely stems from the aggregation of item-specific measurement errors into composite scores, which average out these errors, reducing their distinctive magnitude. This aggregation process obscured the true extent of measurement errors and their variances, leading to this underestimation across both high and low-reliability conditions. However, aside from the measurement error variance, both MEVARC and DFVAR models show negligible relative bias in other parameters in varying reliability conditions, indicating that the averaging process has minimal impact on estimating these interest parameters and their performance remains similar. Moreover, it is also noted that the performances of the two models (DFVAR & MEVARC) are substantially different from the results of the VARC models.

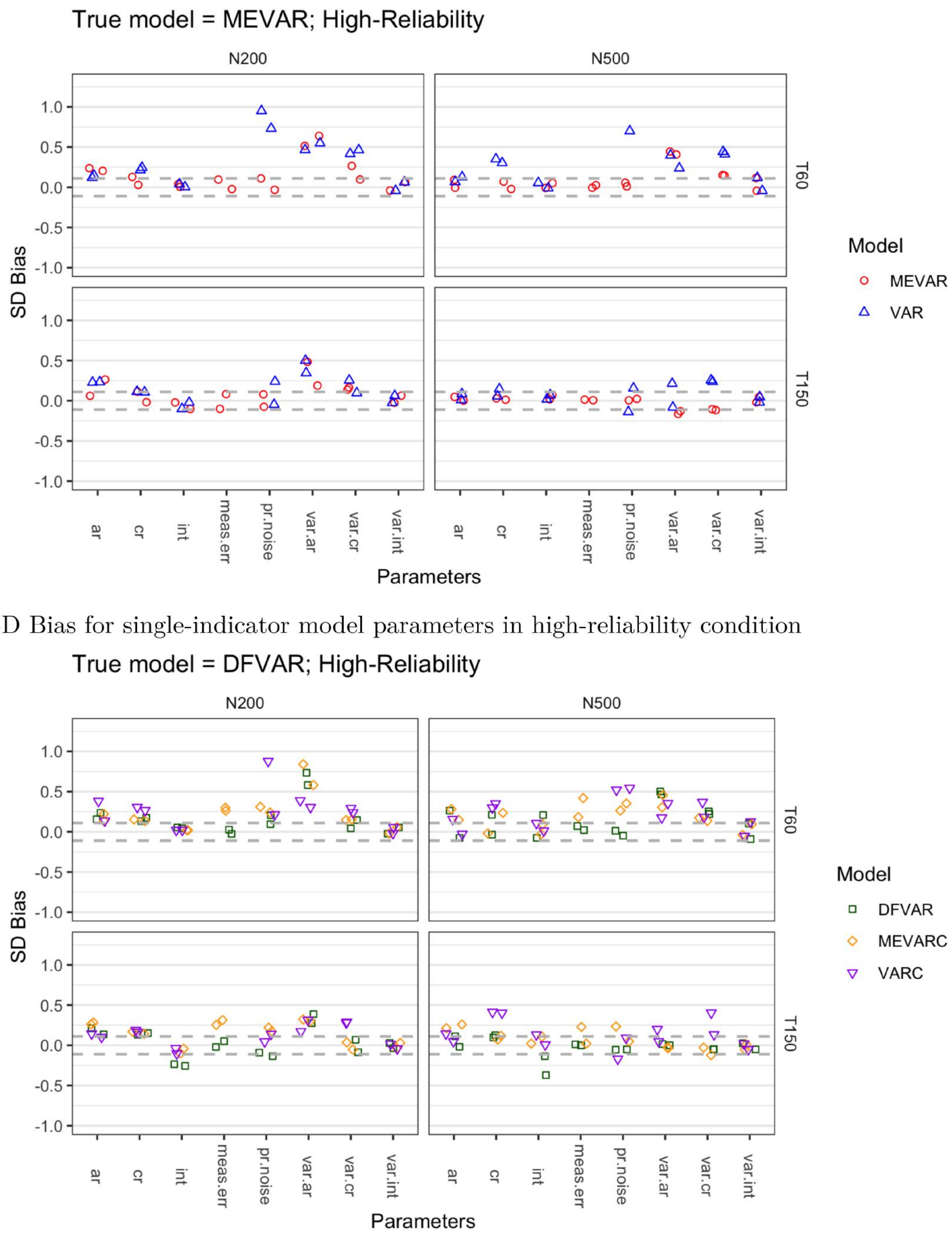
2.2.4. Effects of Different Sample Size Configurations

The impact of sample size (N) and the number of time points (T) is always a matter of interest to researchers related to sample size planning. The effects of measurement model misspecification with different configurations of N and T are depicted in Figures 3 and 4 by comparing the first row ($T = 60$) with the second row ($T = 150$). From Figures 3a and 4a, the conclusions concerning variations in relative biases across models remained regardless of changes in N or T. In other words, in single-indicator models, the extent of overestimation in parameters such as the AR and process noise variance parameters remained regardless of the increase in N or T. However, Figure 3b indicates a slight improvement in the performance of all models when either N or T increased. Notably, an increase in N particularly aids in reducing the bias in the variances of the AR and CR parameters (random effect). In contrast, an increase in T contributed more clearly to the AR, and especially the CR parameters. As shown in Figures 3 and 4, the effect of N looks greater than that of T. Conversely, the effect of T looks greater than that of N with respect to the SD bias, as represented in Figure 5.

2.3. Conclusion

In our findings, dynamic parameters were consistently overestimated where measurement errors were not properly considered, even as sample size and time points increased. In general, as sample size increased, it had notable effects on the variance of AR and CR, while an increase in time points

[1] The SD bias is calculated using the formula:
$\text{SD bias} = \frac{\text{PosteriorSD} - \text{EmpiricalMonteCarloSD}}{\text{EmpiricalMonteCarloSD}}$,
where *Posterior SD* is the standard deviation of the parameter estimates from the posterior distribution, and *Empirical Monte Carlo SD* is the standard deviation of the parameter estimates obtained through empirical Monte Carlo simulations.



(a) SD Bias for single-indicator model parameters in high-reliability condition

(b) SD Bias for multiple indicators model parameters in high-reliability condition

Figure 5. SD bias of single-indicator and multiple indicator models parameters in high-reliability conditions.

significantly affected dynamic parameters. Enhancing the reliability of measurements significantly improved model performance and reduced discrepancies between models with and without explicit measurement error consideration, but it is not a panacea for overestimation. Power remained robust, exceeding 0.8 across all conditions—even in low-reliability settings with smaller sample sizes and fewer time points, except for CR due to having small true values. In simulations, using composite scores to fit MEVARC to data generated by DFVAR showed promise. This approach effectively averaged out measurement error effects, providing satisfactory estimates for most parameters, except for standard error estimates and coverage rates. Given the complexity and computational demands of the DFVAR model, the MEVARC approach offers a simpler and more computationally manageable alternative.

3. Empirical Data Analysis

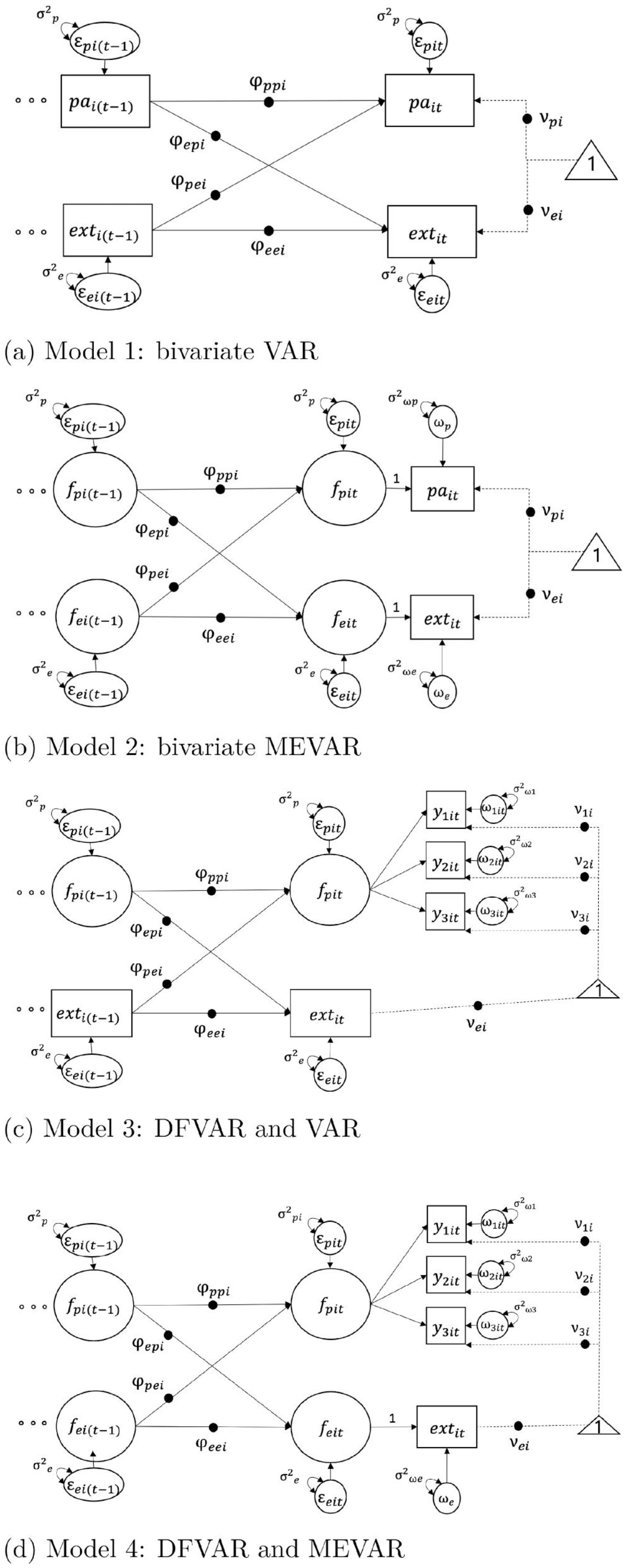
In this section, we provide an empirical example to illustrate the application of four different models in the DSEMs framework. The simulation study showed that both measurement models using single-indicator or multiple indicators outperformed the models without incorporating measurement error both in low & high-reliability conditions. Therefore, to investigate whether these patterns are also evident when analyzing actual data, we examined the bivariate dynamics by considering four different models having the same structural parameters of interest but varied in the method of specifying the variables.

3.1. Intensive Longitudinal Data for Illustration

ILD used in this application were from the Affective Dynamics and Individual Differences (ADID; Emotions and Dynamic Systems Laboratory, 2010) study, published in part in previous studies (Chow & Zhang, 2013; Gates et al., 2023; Hutton & Chow, 2014; You et al., 2020). The total sample consisted of 217 participants, ages 18–86 years, who provided ratings five times per day over one month, accumulating 150 measurement occasions (i.e., time points) per individual. The dataset captured the moment-to-moment fluctuations in positive affect (PA) and state personality of participants. PA was measured using 10 items from the *Positive Affect and Negative Affect Schedule* (PANAS; Watson et al., 1988) rated from 1 (very slight or no experience) to 5 (extreme experience). Extraversion (EXT) was measured using three items from the revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992), reflecting self-perceived tendencies toward being Passive or Active, Unenergetic or Energetic, and Dominant or Submissive, relative to previous self-reports. Following previous studies (Chow & Zhang, 2013; You et al., 2020), three item parcels for PA and a single composite score for extraversion were used, aggregated into two equally spaced time blocks per day. Studying the interrelationships between state EXT and PA is intriguing due to the dynamic nature of state EXT, offering insights into how variations in personality traits influence momentary happiness (Lucas & Baird, 2004). This investigation illuminates the intricate interaction between personality and emotion, pivotal for understanding the temporal dynamics of well-being (Diener et al., 1999). Such understanding significantly enriches affective science, highlighting the role of personality states in the regulation and experience of emotions.

3.2. Data Analysis

We considered four bivariate models as shown in Figure 6. Model 1 (VAR) specifies both variables without measurement errors. Model 2 (MEVAR) employs single-indicator measurement models for both variables. Model 3 (DFVAR+VAR) uses a multiple-indicator measurement model for one variable and no measurement error for the other. Model 4 (DFVAR+MEVAR) combines a multiple-indicator model for one variable and a single-indicator measurement model for the other. We examined the dynamic relationships between PA and EXT in all models. Model estimation was conducted in Mplus Version 8.4 using Bayesian Markov Chain Monte Carlo with two chains, set to run a minimum of 2,000 and up to 120,000 iterations per chain. Convergence was assessed with PSR ($\hat{R} \leq 1.10$) (Gelman et al., 2014), and trace plots were examined for checking non-convergence. Mplus discarded the first half of iterations as burn-in, using the latter half for posterior assessments. Parameter estimates were derived from the median of the posterior distribution.



(a) Model 1: bivariate VAR

(b) Model 2: bivariate MEVAR

(c) Model 3: DFVAR and VAR

(d) Model 4: DFVAR and MEVAR

Figure 6. Four different DSEMs analyzing dynamic process and their individual differences between positive affect (PA) and extraversion (EXT).

Table 1. Estimates of parameters for bivariate VAR (model 1), bivariate MEVAR (model 2), DFVAR-VAR (model 3), and DFVAR-MEVAR (model 4).

		Model 1		Model 2		Model 3		Model 4	
	Parameters	Est	Sd	Est	Sd	Est	Sd	Est	Sd
ν_{pi}	α_{p1}	2.571*	0.033*	2.573*	0.034*	2.633*	0.032*	2.626*	0.033*
	$\sigma^2_{u_{pv1}}$	0.232*	0.027*	0.208*	0.025*	0.198*	0.024*	0.197*	0.025*
	α_{p2}	–	–	–	–	2.685*	0.035*	2.677*	0.035*
	$\sigma^2_{u_{pv2}}$	–	–	–	–	0.198*	0.024*	0.197*	0.025*
	α_{p3}	–	–	–	–	2.447*	0.036*	2.438*	0.036*
	$\sigma^2_{u_{pv3}}$	–	–	–	–	0.198*	0.024*	0.197*	0.025*
ν_{ei}	α_e	2.817*	0.031*	2.810*	0.031*	2.828*	0.032*	2.810*	0.030*
	$\sigma^2_{u_{e\nu}}$	0.180*	0.019*	0.158*	0.019*	0.175*	0.020*	0.158*	0.020*
ϕ_{ppi}	γ_1	−0.477*	0.022*	0.458*	0.061*	0.262*	0.021*	0.385*	0.038*
	$\sigma^2_{u_1}$	0.066*	0.010*	0.110*	0.028*	0.050*	0.008*	0.101*	0.023*
ϕ_{eei}	γ_2	−0.251*	0.024*	0.107*	0.070*	−0.020	0.016	0.133*	0.062*
	$\sigma^2_{u_2}$	0.081*	0.011*	0.395*	0.066*	0.011*	0.002*	0.385*	0.060*
ϕ_{epi}	γ_3	0.413*	0.022*	0.026	0.048*	0.063*	0.018*	−0.001	0.033
	$\sigma^2_{u_3}$	0.087*	0.011*	0.116*	0.036*	0.033*	0.006*	0.072*	0.016*
ϕ_{pei}	γ_4	0.341*	0.019*	−0.284*	0.070*	0.277	0.119*	−0.272*	0.070*
	$\sigma^2_{u_4}$	0.067*	0.008*	0.271*	0.073*	0.056*	0.008*	0.351*	0.079*
σ^2_p	γ_5	0.726*	0.087*	0.477*	0.040*	0.555*	0.040*	0.641*	0.014*
σ^2_e	γ_6	0.942*	0.123*	0.252*	0.028*	0.977*	0.014*	0.221*	0.018*
M.E	$\sigma^2_{\upsilon_{p1}}$	–	–	0.177*	0.044*	0.198*	0.006*	0.198*	0.006*
	$\sigma^2_{\upsilon_{p2}}$	–	–	–	–	0.211*	0.006*	0.209*	0.005*
	$\sigma^2_{\upsilon_{p3}}$	–	–	–	–	0.448*	0.007*	0.449*	0.007*
	$\sigma^2_{\upsilon_e}$	–	–	0.616*	0.025*	–	–	0.657*	0.019*

Note. Estimates (Est) are the median of the Bayesian Posterior Distribution and Standard deviations (Sd) are represented. In the variable index, p represents positive affect(PA), and e represents Extraversion(EXT). ν_{pi} and ν_{ei} represent random intercepts for PA and EXT, respectively. ϕ_{ppi} and ϕ_{eei} indicate random autoregression. ϕ_{epi} and ϕ_{pei} are cross-lagged effects. σ^2_p and σ^2_e represent residual variance for PA and EXT. Random effects have the estimates of the mean and variance of the effects. $\sigma^2_{\omega p}$ and $\sigma^2_{\omega e}$ represent the variance of within-person measurement error for PA and EXT. In model 3 and model 4, $\sigma^2_{\omega p}$ and $\sigma^2_{\omega e}$ values are calculated as the sum of within residual variances for three items over sum of within factor loadings of each 3 items for PA and EXT, respectively. In the same way, the α_p and α_e, which represent the grand mean intercept, were calculated as the mean of each item's mean in model 3 and model 4.

Before discussing the priors, note that the target parameters in this application included two random intercepts, two random AR parameters, two random CR effects, two process noise, and two measurement error variances. For the fixed effects of the intercepts, AR, CR, and process noise, we assigned weak informative priors with large variances ($N(2, 100)$, $N(0.5, 1)$, $N(0, 0.5)$, $N(0, 10)$). For the random effects of the intercepts, AR, CR, and process noise, we set inverse-Wishart priors ($IW(2.5, -7)$, $IW(0.3, -7)$, $IW(0.05, -7)$, $IW(0.07, -3)$), considering the six random variables in each model.[2] Lastly, we assigned an inverse gamma prior ($IG(-1, 0)$) for the two measurement errors, which is the default prior.

3.3. Results

The point estimates of the fixed effects, variances of the random effects, and standard deviations of each parameter across the four models are presented in Table 1. Descriptions for each parameter are provided in the table notes. The focus is on the dynamic parameters among models with the same structural relationships, not on capturing measurement error itself. The analysis revealed significant differences in the estimated dynamic parameters among the models. All four models exhibited similar grand means and random variances for the random intercepts, but the dynamic parameters differed markedly. Specifically, models without a measurement model for at least one variable (Models 1 and 3) showed substantial differences compared to those with a measurement model for both variables (Models 2 and 4). This divergence was particularly notable in the AR and CR values, with some even differing in sign. Additionally, process noise (σ^2_p, σ^2_e) and measurement errors ($\sigma^2_{\omega p}, \sigma^2_{\omega e}$) were most pronounced in Model 1, which omitted measurement errors, with Model 3 showing similar patterns. For detailed values, refer to Table 1.

Upon closer inspection of the dynamic parameters, particularly the AR for the EXT variable (γ_2), Models 1 and 3 exhibited quite different coefficient values compared to Models 2 and 4. This discrepancy highlights how a measurement model can lead to significantly different outcomes. For the AR of the PA variable (γ_1), only Model 1's values differed significantly from Models 2, 3, and 4. This can be attributed to Model 3 specifying the PA variable alone within a measurement model, resulting in AR values similar to Models 2 and 4, while Model 1 remains distinct. Conceptually, positive AR values for EXT and PA are more plausible (Larsen & Ketelaar, 1991). This pattern is also evident in the CR, especially from EXT to PA (γ_4). Models 1 and 3 showed positive coefficients, whereas Models 2 and 4 demonstrated negative ones, indicating self-regulation mechanisms. In the reverse CR effect from PA to EXT (γ_3), Models 1 and 3 exhibited significant differences, while Models 2 and 4 did not, suggesting a lack of influence in this direction. For detailed values, refer to the corresponding table.

Consistent with the simulation study results, we found that omitting measurement models led to notably different AR and CR parameter values, resulting in distinct interpretations of the relationships between PA and state EXT. When neither variable, or only one, was specified under a measurement model, the results differed significantly,

[2] The default prior for variance and covariance parameters in Mplus is $IW(0, -\text{df} - 1)$, where df denotes the number of random variables.

sometimes even showing opposite signs. Using composite scores attenuated the AR and CR magnitudes, but the fundamental conclusions about the PA and EXT relationship remained consistent with latent variable findings. The pattern of estimated values showed similarity within pairs, leading to a divergence between Models 1 and 3 and Models 2 and 4, consistent with the simulation study. In short, Models 2 and 4 are preferred over Models 1 and 3.

It is interesting to note the within and between reliability of each variable. Unlike conventional reliability calculations, which can be performed at the data level regardless of the model, between and within reliability in ILD can be assessed after fitting the model. In this paper, reliability from Models 2 and 4 is reported using Equations (15) and (16). The between-reliability for PA and EXT is 0.205 and 0.153 in Model 2, and 0.157 and 0.152 in Model 4. The within-reliability for PA and EXT is 0.781 and 0.293 in Model 2, and 0.731 and 0.255 in Model 4. The low between-reliability values for PA and EXT indicate limited differentiation between individuals, while the low within-reliability for EXT suggests high variability and low stability within individuals over time, meaning that whenever individuals deviate from their person-specific intercepts in EXT, they tended to revert back to these intercept values more quickly than PA.

4. Discussion

In ILD, data collection trade-offs are inherent. While it captures within-process dynamics, frequent measurements over short periods can burden participants and introduce measurement errors. Furthermore, the variance source of ILD comes from both between-person (inter-individual differences) and within-person (dynamic processes, including errors) levels. This paper examines Monte Carlo simulations under various conditions: presence or absence of measurement models, data reliability, various sample sizes and time points, as well as discrepancies between generated and analyzed models.

Monte Carlo simulations from this study reveal considerable biases in dynamic parameters when measurement errors are overlooked, an issue frequently encountered in psychological research that often deals with latent constructs. These findings highlight the importance of integrating measurement models in DSEM. Although factors such as ILD reliability, sample size, and time points help reduce bias, they do not eliminate the substantial bias caused by measurement errors. Our results show that MEVAR models are more suitable than VAR models for single-item data. For multiple-item data, both DFVAR and MEVAR with composite scores (MEVARC) were as effective as VAR with composite scores (VARC). This suggests that MEVARC is a viable, simpler alternative to DFVAR, offering parsimony and computational efficiency without compromising performance.

An empirical validation using a bivariate model supported the simulation results, showing consistent biases when measurement errors were not addressed. Discrepancies were observed based on whether measurement models were applied to both versus at least one or neither variable. While the empirical study aligns with the simulation outcomes, the unknown true values and true model specifications limit the accuracy and validity of the results, even with measurement models included (Rhemtulla et al., 2018). Future research should refine model specifications and improve measurement validity, especially for constructs like state positive affect and extraversion. Consistent with previous studies (McNiel et al., 2010; Larsen & Ketelaar, 1991), the structural relationships elucidated by measurement models better match observed data, offering more plausible interpretations. However, including measurement models does not always enhance accuracy. As shown in Table 1, Bayesian estimates for some parameters in models accounting for measurement error have slightly larger posterior standard deviations, likely due to the bias-variance trade-off in more complex models (Ledgerwood & Shrout, 2011).

While a single-indicator measurement model simplifies analysis, it has notable limitations. Primarily, it cannot estimate between residual variances, unlike a multiple-indicator model (ω_i^B) term in Equation (3). Additionally, it assumes measurement invariance over time, rather than evaluating it. Conversely, a multiple-indicator model allows for the evaluation of measurement invariance using a cross-classified model, with one factor representing the latent construct and the other representing time (McNeish et al., 2021). If the time factor variance is insignificant, it suggests no measurement differences over time. These aspects are not discussed in this paper but warrant further exploration to enhance the robustness and interpretability of measurement models in psychological research.

However, the multiple-indicator model has drawbacks, such as increased model complexity and longer computational times. In our simulations, DFVAR took around 2 hours per replication, compared to 30 minutes for MEVAR and 20 minutes for VAR under the same conditions. The choice between multiple-indicator and single-indicator approaches should be guided by the theoretical framework. If the latent variable supports a multiple-indicator model, DFVAR is preferable, with MEVARC as a viable alternative using composite scores. Otherwise, MEVAR may suffice.

4.1. Limitations and Future Directions

Our evaluations of measurement errors are by no means extensive, and several future extensions warrant closer inspection. Firstly, this study's exploration of the DSEM framework is limited to continuous indicators. Future research should accommodate ordinal or categorical items and adapt multilevel reliability estimation for binary or ordinal ILD. Additionally, extending the DSEM to categorical variables (Asparouhov et al., 2018) or outcomes (Savord, 2023) is a potential future direction. Also, considering the nature of ILD, these data may exhibit irregularly spaced observations. Therefore, extending to Continuous Time (CT) models could more accurately capture these dynamics. Furthermore, missing observations are likely frequent, and this study did not investigate the influence of missing data patterns and non-normality on parameter estimates and model fit. Given their potential significant impact, these aspects deserve further exploration in future research (Ji et al., 2018, 2020).

Moreover, the simulation study can include a broader range of conditions to better capture the diverse configurations within DSEM applications. For example, we did not consider person-specific process noise variances due to estimation and convergence issues with the sample sizes used in this study. Our Monte Carlo simulation results indicated that process noise variances were biased when measurement errors were incorrectly omitted. With person-specific variances, we anticipated even less stability in estimates and more severe spill-over biases to the AR and CR parameters. However, other parameters were well recovered even with $N = 200$ and $T = 60$, smaller than the recommended sample sizes for DSEM models with person-specific process noise variances (Schultzberg & Muthén, 2018; Asparouhov et al., 2018). We recommend that researchers use streamlined versions of the DSEM when their N and T are too small to fit the full DSEM model.

Additionally, the finding that power was adequately recovered with smaller sample sizes than previously recommended suggests a promising direction for future research. Exploring power and model performance with sample sizes smaller than $N = 200$ and $T = 60$ could provide valuable insights. Previous studies have shown that single-indicator models can estimate random measurement error parameters, but their feasibility depends heavily on the magnitude of the AR coefficient and is prone to under-identification challenges (Schuurman & Hamaker, 2019). Moreover, according to previous research (Asparouhov et al., 2018), to estimate subject-specific parameters for AR, CR, process noise, and measurement, at least $T \geq 200$ per person is required. Therefore, examining the necessary sample size configurations for sufficiently powered DSEM variations is critical for future ILD planning.

In light of our findings, it is crucial to address the discrepancies in reliability conditions between simulation studies and empirical data. The predefined factor structure of the DSEM model, while useful, may not fully capture the complex dynamics of real-world data, leading to lower observed reliability. Future research should critically evaluate and adjust the factor structures used in simulations to better reflect empirical data. Exploring alternative models, such as differential equation models, could enhance our understanding of ILD dynamics and improve reliability. Ensuring that ILD models are both theoretically and empirically robust is essential for accurately capturing real-world data complexities.

Lastly, this study was limited to bivariate models. Other studies using dynamic network models (Park et al., 2020; Ellington & Baruník, 2020) have employed higher-dimensional VAR structures. Investigating the feasibility and performance of higher-dimensional dynamic network models would be valuable. Incorporating more variables might amplify discrepancies between VAR and MEVAR models, as the distorting elements of reliability among variables will be included. Given the reported consequences of omitting measurement errors, we encourage future researchers to incorporate measurement errors into their dynamic network models.

In conclusion, this study underscores the importance of incorporating measurement models in ILD analysis, especially in psychological research where latent constructs are common. Our Monte Carlo simulations and empirical applications demonstrate that the misspecification of measurement models can significantly bias dynamic parameters across various reliability conditions, sample sizes, and time points, despite high power. Our results advocate the MEVAR model over the DFVAR model due to its comparable performance and greater computational efficiency. Additionally, we provide functions for extending reliability indices and power evaluations. These findings highlight the necessity of proper measurement model specification to accurately capture ILD dynamics, regardless of reliability.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling*, *25*, 359–388. https://doi.org/10.1080/10705511.2017.1406803

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455. https://doi.org/10.1080/10618600.1998.10474787

Chow, S.-M., & Zhang, G. (2013). Nonlinear regime-switching state-space (RSSS) models. *Psychometrika*, *78*, 740–768. https://doi.org/10.1007/s11336-013-9330-8

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources, Inc.

Cronbach, L., & Furby, L. (1970). How we should measure" change": Or should we? *Psychological Bulletin*, *74*, 68–80. https://doi.org/10.1037/h0029382

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin*, *125*, 276–302. https://doi.org/10.1037/0033-2909.125.2.276

Diener, E., & Tay, L. (2014). Review of the day reconstruction method (DRM). *Social Indicators Research*, *116*, 255–267. https://doi.org/10.1007/s11205-013-0279-x

Ellington, M., & Baruník, J. (2020). *Dynamic networks in large financial and economic systems*. Available at SSRN 3651134.

Emotions and Dynamic Systems Laboratory. (2010). *The affective dynamics and individual differences (ADID) study: Developing non-stationary and network-based methods for modeling the perception and physiology of emotions* [Unpublished doctoral dissertation]. University of North Carolina at Chapel Hill.

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*, 453–480. https://doi.org/10.1080/00273171.2018.1454823

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment-monitoring behavior in daily life settings. *European Journal of Psychological Assessment*, *23*, 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Gates, K. M., Chow, S., & Molenaar, P. C. (2023). *Intensive longitudinal analysis of human processes*. CRC Press.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472. https://doi.org/10.1214/ss/1177011136

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An r package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*, 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*, *53*, 820–841. https://doi.org/10.1080/00273171.2018.1446819

Hamaker, E. L., Ceulemans, E., Grasman, R., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*, 316–322. https://doi.org/10.1177/1754073915590619

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*, 10–15. https://doi.org/10.1177/0963721416666518

Hamilton, J. D. (2020). *Time series analysis*. Princeton University Press.

Hutton, R. S., & Chow, S.-M. (2014). Longitudinal multi-trait-state-method model using ordinal data. *Multivariate Behavioral Research*, *49*, 269–282. https://doi.org/10.1080/00273171.2014.903832

Ji, L., Chen, M., Oravecz, Z., Cummings, E. M., Lu, Z.-H., & Chow, S.-M. (2020). A Bayesian vector autoregressive model with nonignorable missingness in dependent variables and covariates: Development, evaluation, and application to family processes. *Structural Equation Modeling*, *27*, 442–467. https://doi.org/10.1080/10705511.2019.1623681

Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling*, *25*, 715–736. https://doi.org/10.1080/10705511.2017.1417046

Kim, C.-J., & Nelson, C. R. (1999). *State-space models with regime switching: Classical and Gibbs-sampling approaches with applications*. MIT Press Books.

Kuhfeld, M., & Soland, J. (2022). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*, *27*, 234–260. https://doi.org/10.1037/met0000367

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009). Reliability of a longitudinal sequence of scale ratings. *Psychometrika*, *74*, 49–64. https://doi.org/10.1007/s11336-008-9079-7

Larsen, R. J., & Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *Journal of Personality and Social Psychology*, *61*, 132–140. https://doi.org/10.1037/0022-3514.61.1.132

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*, 1174–1188. https://doi.org/10.1037/a0024776

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement*, *49*, 587–602. https://doi.org/10.1177/001316448904900309

You, D., Hunter, M., Chen, M., & Chow, S.-M. (2020). A diagnostic procedure for detecting outliers in linear state–space models. *Multivariate Behavioral Research*, *55*, 231–255. https://doi.org/10.1080/00273171.2019.1627659

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling*, *26*, 646–661. https://doi.org/10.1080/10705511.2018.1545232

Li, Y., Wood, J., Ji, L., Chow, S.-M., & Oravecz, Z. (2022). Fitting multilevel vector autoregressive models in Stan, JAGS, and Mplus. *Structural Equation Modeling*, *29*, 452–475. https://doi.org/10.1080/10705511.2021.1911657

Lucas, R. E., & Baird, B. M. (2004). Extraversion and emotional reactivity. *Journal of Personality and Social Psychology*, *86*, 473–485. https://doi.org/10.1037/0022-3514.86.3.473

Marcoulides, K. M. (2019). Reliability estimation in longitudinal studies using latent growth curve modeling. *Measurement: Interdisciplinary Research and Perspectives*, *17*, 67–77. https://doi.org/10.1080/15366367.2018.1522169

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*, 610–635. http://doi.apa.org/getdoi.cfm?https://doi.org/10.1037/met0000250

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling*, *28*, 807–822. https://doi.org/10.1080/10705511.2021.1915788

McNiel, J. M., Lowman, J. C., & Fleeson, W. (2010). The effect of state extraversion on four types of affect. *European Journal of Personality*, *24*, 18–35. https://doi.org/10.1002/per.738

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202. https://doi.org/10.1007/BF02294246

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.) [Computer software manual]. Muthén & Muthén.

Nesselroade, J. R., McArdle, J. J., Aggen, S. H., & Meyers, J. M. (2002). Dynamic factor analysis models for representing process in multivariate time-series. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 235–265). Lawrence Erlbaum Associates.

Oh, H., & Jahng, S. (2023). Incorporating measurement error in the dynamic structural equation modeling using a single indicator or multiple indicators. *Structural Equation Modeling*, *30*, 501–514. https://doi.org/10.1080/10705511.2022.2103703

Park, J. J., Chow, S.-M., Fisher, Z. F., & Molenaar, P. (2020). Affect and personality: Ramifications of modeling (non-) directionality in dynamic network models. *European Journal of Psychological Assessment*, *36*, 1009–1023. https://doi.org/10.1027/1015-5759/a000612

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software manual]. https://www.R-project.org/

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*, 778–791. https://doi.org/10.1037/a0017915

Rhemtulla, M., van Bork, R., & Borsboom, D. (2018). *WorseThanMeasurementError*. Open Science Framework. https://osf.io/42s8b. https://doi.org/10.31219/osf.io/42s8b

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *The British Journal of Mathematical and Statistical Psychology*, *67*, 172–194. https://doi.org/10.1111/bmsp.12014

Savord, A. (2023). *Evaluation of univariate and multivariate dynamic structural equation models with categorical outcomes* [Unpublished doctoral dissertation]. Arizona State University.

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*, 495–515. https://doi.org/10.1080/10705511.2017.1392862

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, *24*, 70–91. http://doi.apa.org/getdoi.cfm?https://doi.org/10.1037/met0000188

Scollon, C. N., Kim-Prieto, C., & Scollon, C. N. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*, 5–34. https://doi.org/10.1023/A:1023605205115

Smyth, J. M., & Smyth, J. M. (2003). Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies*, *4*, 35–52. https://doi.org/10.1023/A:1023657221954

Stigler, S. M. (1990). *The history of statistics*. Harvard University Press.

Wang, L. P., Hamaker, E. L., & Bergeman, C. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*, 567–581. https://doi.org/10.1037/a0029317

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063