Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling

E. L. Hamaker,[1], T. Asparouhov[2], and B. Muthén[2]

1. Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, The Netherlands; 2. Muthén and Muthén, Los Angeles, CA, USA

To be published as Chapter 31 in: *The Handbook of Structural Equation Modeling (2nd edition)*; Rick H. Hoyle (Ed.); Publisher: Guilford Press.

Author Note:

This research is supported by an *ERC Consolidator Grant* awarded to E. L. Hamaker (ERC-2019-COG-865468). Correspondence concerning this manuscript should be addressed to E. L. Hamaker, Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, P.O. Box 80140,3508 TC, Utrecht, The Netherlands. Email: e.l.hamaker@uu.nl.

Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling

Dynamic structural equation modeling (DSEM; Asparouhov, Hamaker, & Muthén, 2018) is an innovative modeling framework that is implemented in Mplus for the analysis of intensive longitudinal data. Such data stem from measurement techniques like experience sampling, daily diaries, and ambulatory assessments, and are characterized by many repeated measures that are typically densely spaced in time (Bolger, Davis, & Rafaeli, 2003; Trull & Ebner-Priemer, 2013; Walls & Schafer, 2006). Since technological innovations, such as smart phones, activity trackers, and other wearable devices, have made it much easier to collect such data, they are now increasingly more often obtained from large samples of cases like individuals, dyads, or households (Hamaker & Wichers, 2017; Mehl & Conner, 2012).

DSEM was developed to exploit the richness of intensive longitudinal data through a combination of three well-known modeling traditions. The core of DSEM is formed by *time series modeling*, which is used to account for dynamic (or lagged) relations within the data of a single case over time. This $N = 1$ technique is combined with *multilevel modeling*, to facilitate the analysis of multiple cases simultaneously, while allowing for quantitative differences between them. Additionally, the *structural equation modeling* component allows for the further analysis of these quantitative differences using path analysis and/or factor analysis. The result is a general framework that encompasses a vast array of models for intensive longitudinal data, and allows for various research questions about dynamics and individual differences therein.

With the current chapter we aim to showcase the flexibility of the DSEM framework. To this end, we take an empirical dataset as our point of departure and present a series of models that can be used to tackle particular challenges associated with it. These data come from a randomized controlled trial in which participants with a history of depression—but currently in remission—were randomly assigned to either a mindfulness training, or a control condition (for details, see Geschwind, Peeters, Drukker, van Os, & Wichers, 2011). Before and after the intervention, these participants were measured 10 times per day (at semi-random time points), for 6 days, using self-reports of emotional states, thoughts, behaviors, and events. In addition to these experience sampling measures, baseline measures were taken prior to each of these intensive measurement episodes.

In the first part, entitled *The DSEM Framework*, we focus on analyzing the intensive longitudinal data from the first episode only, and show how the three modeling traditions—time series modeling, multilevel modeling, and structural equation modeling—contribute to the general DSEM framework. In the second part, entitled *Using DSEM From Pretest-Posttest Data*, we show how to model the pretest-posttest features of the data, which also illustrates more generally how one can deal with multiple groups and/or multiple waves of intensive longitudinal data in DSEM. Our more in-depth focus on particular models and research questions in this chapter prohibits us to cover the entire breadth of DSEM; to somewhat compensate for this, we briefly summarize alternative modeling options in the *Discussion* section, where we also identify avenues for future research. For a more elaborate discussion of the DSEM framework, its assumptions and underlying technicalities (such as the ins and outs of Bayesian estimation), we refer the reader to other DSEM publications (cf. Asparouhov et al., 2018; Asparouhov & Muthén, 2019, 2020; Hamaker, Asparouhov, Brose, Schmiedek, & Muthén, 2018). Throughout the chapter we will present and discuss the results from the empirical data; the Mplus output files of these analyses and additional explanation are made available on an accompanying website (https://ellenhamaker.github.io/DSEM-book-chapter/) for further reference.

1. The DSEM framework

We make use of a running example that consists of two variables: *momentary negative affect*, and *unpleasantness of events that occurred since the previous beep*.[1] These data differ from other bivariate examples in the DSEM literature (e.g., Hamaker et al., 2018), in that even though the two variables were measured at the same time (and they are thus included on the same line in the datafile), they are characterized by an implicit lagged relation between them due to the phrasing: The negative affect measure is about the current moment, whereas the event measure refers to the entire interval between the previous beep and now. Therefore, we argue there is a logical reason for regressing negative affect on unpleasantness of events measured at the same time, in order to investigate whether and how events seem to affect someone's affective state. Note that such implicit lagged relations are not entirely uncommon in intensive longitudinal data: A similar feature arises when obtaining measures of behavior during the day and sleep quality the following night as discussed in Armstrong, Covington, Unick, and Black (2019).

In this section we begin with considering $N = 1$ time series models for these data that are based on analyzing the data for each person separately. Subsequently, we move to multilevel extensions of these models, which are based on analyzing the data of all individuals simultaneously, while allowing for quantitative differences between them. Finally, we add the SEM component, by which we can further model the individual differences in the person-specific parameters.

1.1 Time Series Analysis (For $N = 1$ Data)

Time series analysis is a class of techniques that were developed to handle a large number of repeated measures from a single case (Hamilton, 1994). These techniques have

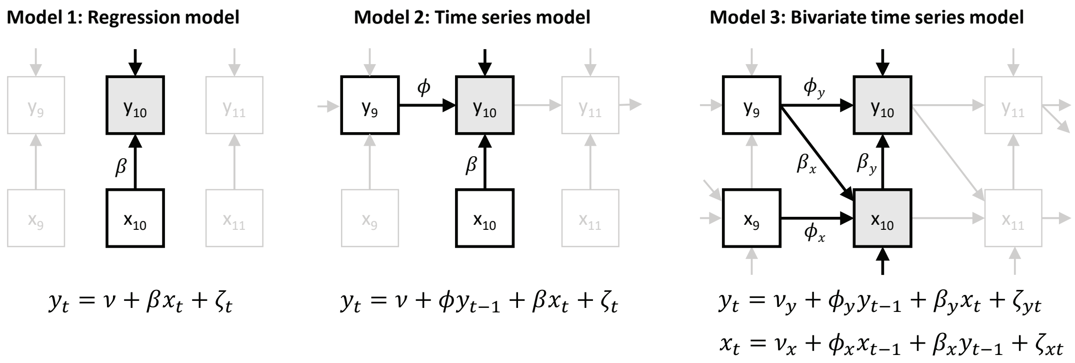
[1] The original variable in the empirical dataset was scaled with zero indicating a neutral event, positive scores indicating a pleasant event, and negative scores indicating an unpleasant event. To ease interpretation, we rescaled this variable by multiplying it by -1.

been very popular in disciplines such as econometrics, meteorology and seismology. In psychology, time series analysis has been recognized as a powerful idiographic approach that allows a researcher to study the patterns of fluctuations within a particular individual over time (Molenaar, 1985; Nesselroade, 2007). In many of the original applications in psychology, the focus has been on the factor structure and the way in which observed variables are related to a smaller number of underlying latent variables (Baldwin, 1946; Cattell, Cattell, & Rhymer, 1947). However, in many of the more recent applications of $N = 1$ time series analysis in psychology, the main focus has been on the dynamic relations between observed variables, specifically on the autoregressive and cross-lagged regressions (Gates & Molenaar, 2012; van der Krieke et al., 2015). Most importantly though, the $N = 1$ approach implies that the data for each individual are analyzed separately from that of others, thereby allowing for a maximal degree of idiosyncracies in the results.

Below we discuss three $N = 1$ models with increasing complexity, using the example of negative affect and unpleasantness of events. We consider the third model the most appropriate and interesting, but include the other two for didactic purposes. Subsequently, we discuss how to deal with missing data and unequal intervals between observations in estimating the time series models. We end with summarizing the empirical results for these models when applied to each individual separately in the dataset.

1.1.1 Three $N = 1$ Models. Let y refer to negative affect, and x to unpleasantness of events. The three single level ($N = 1$) models that we consider are visualized in Figure 1. The focus in these path diagrams is on an arbitrary occasion ($t = 10$) to highlight how each outcome (shaded observation) is predicted; these representations can be generalized by only keeping the bolded parts, and replacing the occasion index 10 by t, and 9 by $t - 1$. The regression equations pertaining to these more general representations are given below the path diagrams.

In Model 1, we begin with regressing negative affect (y_t) on the unpleasantness of events (x_t). Because there are no lagged relations in this model (i.e., the regression includes two variables with the same time index), this is not a dynamic model yet. However, since there is an implicit lagged relation between the outcome and the predictor, the regression coefficient (β) from this model indicates how affect tends to change after a one unit increase in unpleasantness of events. We may expect a positive regression coefficient here, but it is also plausible that this parameter differs across individuals and that some individuals have a stronger affective response than others (Geschwind et al., 2011; Suls, Green, & Hillis, 1998; Wichers et al., 2009).


Model 1: Regression model
Model 2: Time series model
Model 3: Bivariate time series model
y9
y10
y11
x9
x10
x11
β
ϕ
ϕ_y
ϕ_x
β_x
β_y

$$y_t = \nu + \beta x_t + \zeta_t$$

$$y_t = \nu + \phi y_{t-1} + \beta x_t + \zeta_t$$

$$y_t = \nu_y + \phi_y y_{t-1} + \beta_y x_t + \zeta_{yt}$$
$$x_t = \nu_x + \phi_x x_{t-1} + \beta_x y_{t-1} + \zeta_{xt}$$

Figure 1. Three single level $N = 1$ model for time series data. For illustrative purposes, the focus in the path diagrams is on a particular time point (here $t = 10$). The dependent variables are shaded, and all relevant model parts for the prediction of them are bolded. All irrelevant parts are in grey. Below the path diagrams the general regression equations are provided.

In Model 2, we add autoregression to the model through regressing current negative affect (y_t) on preceding negative affect (y_{t-1}); the inclusion of such a lagged relation makes it a time series model. There are both statistical and substantive reasons for including autoregression. From a statistical point of view, we account for autocorrelation in our outcome variable to avoid bias in the parameter estimates. From a substantive perspective, autoregression is a feature with an appealing interpretation, as it captures the tendency of a person not to change much from one occasion to the next. This characteristic has been referred to as *inertia* in the psychological literature during the nineties (Cook, Tyson, White, Gottman, & Murray, 1995; Suls et al., 1998), a concept that was revitalized by Kuppens, Allen, and Sheeber (2010; Koval, Burnett, & Zheng, 2021). It has been described as the degree of carry-over from one occasion to the next, or the level of lingering (Blanke, Neubauer, Houben, Erbas, & Brose, 2021). From a dynamical system's perspective it can be interpreted as regulatory weakness or (inverse) attractor strength, as it quantifies how long it takes a person to return to their equilibrium (i.e., attractor) after being pushed away from it by an external force (Hamaker, 2012; Sosnowska, Kuppens, De Fruyt, & Hofmans, 2019). Individual differences in autoregression strength have been related to an array of individual characteristics, including personality traits, depression, and sex (cf. Koval et al., 2021), although there is also growing concern about the strength and meaningfulness of these relations (Dejonckheere et al., 2019; Wendt et al., 2020).

In Model 3 we consider both negative affect (y_t) and unpleasantness of events (x_t) as outcomes. This allows us to investigate whether the events that a person reports are influenced by their affective states. Such cross-lagged regression from negative affect (y_{t-1}) to subsequent unpleasantness of events (x_t) could reflect that a person's affective states influence their interpretation of events; for instance, when one is feeling happy and content, they may interpret a frown on someone else's face as a sign of thoughtfulness, while in contrast, when one is feeling distressed or angry, they may interpret that same frown as a sign of disapproval. In addition to the effect of one's emotional state on one's interpretation of events, it may also actually shape the events: When one is happy and content, this may be infectious and lead others to respond positively, whereas feelings of distress or anger may result in a negative attitude that more easily triggers negative responses from others. Hence, including the lagged relation from negative affect to unpleasantness of events (β_x) provides more insight into the way these phenomena interact with each other over time. Comparing the standardized cross-regressions will form a way to investigate to what extent they are affected by each other.

Model 3 is closely related to what is known in the time series literature as the first-order vector autoregressive (VAR[1]) model. However, the current model deviates from the standard VAR(1) model, in that instead of having only lag 1 regressions and correlated residuals, the current model has a lag 0 regression from unpleasantness of events (x_{it}) to negative affect (y_{it}); as a consequence, the residuals of the two variables are not allowed to be correlated, as that would result in an unidentified model.

1.1.2 Unequally Spaced Data and Missing Data. The defining feature of a dynamic model is that it contains lagged relations, that is, regressions between variables at different time points. These are of key interest, because they may provide some insight in how changes in one variable precede changes in another. However, a critical feature of lagged relations is that their size depends on the amount of time between the observations. For instance, autoregressive relations tend to decrease as the interval between subsequent observations increases, whereas cross-lagged regressions tend to be zero for very short intervals, can then increase (or decrease) as the interval increases until some maximum (minimum) is reached, after which they return to zero again (Deboeck & Preacher, 2015; Dorman & Griffin, 2015; Driver, Oud, & Voelkle, 2017; Ryan, Kuiper, & Hamaker, 2018). This phenomenon is known as “the lag problem” (Gollob & Reichardt, 1987), and it implies that the time interval between observations is of critical importance in the interpretation.

In intensive longitudinal data, there are three aspects that may result in unequal intervals between observations. First, there are often at least some missing data, which in this case leads to larger intervals between realized observations. Second, many of the measurement techniques are based on purposely using varying time intervals between the observations, to avoid participants anticipating the next beep and adjusting their behavior towards this (e.g., waiting with starting a new activity, such as getting into the car or calling a friend, to be able to fill out the next questionnaire; Bolger et al., 2003; Mehl & Conner, 2012; Trull & Ebner-Priemer, 2013). Furthermore, when there are multiple self-reports per day, there tends to be a longer gap between the last measurement on one day, and the first measurement the following day.

To investigate the effect of varying intervals, De Haan-Rietdijk, Voelkle, Keijsers, and Hamaker (2017) performed a simulation study in which four methods for estimating a VAR(1) model with unequally spaced data were compared. The most truthful way to handle unequal intervals is through the use of continuous time models, which include exact information on the interval length directly in the model (for discussions see Oravecz, Tuerlinckx, & Vandekerckhove, 2011; Driver et al., 2017). An alternative approach that showed to perform about equally well, but that remains in the realm of discrete time modeling, is based on adding missing data in between realized observations as a way to account for the length of the time interval between them. This approach can be described as converting the unequal-interval problem into a missing-data problem. The latter is then tackled with a discrete time Kalman filter approach, which is known to perform well in case of data missing at random (Asparouhov et al., 2018; Harvey, 1989; Kalman, 1960).

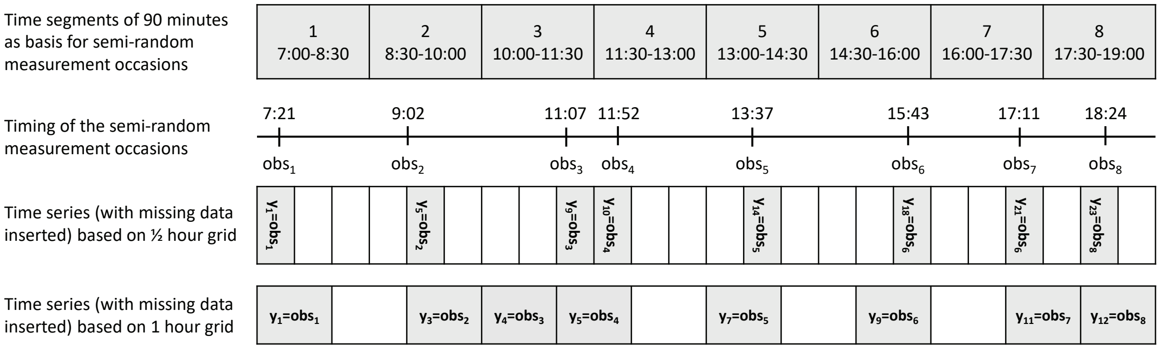


Figure 2. Procedure for handling unequal time intervals between observations. Eight observations were made at random time points within 90 minute blocks. These are then positioned in discrete time, using a particular time grid. Two examples are given: a half-hour grid and a one-hour grid. Shaded segments indicate an observations was positioned within this segment, whereas non-shaded segments indicate missing data that are added to the series. See main text for further explanation.

Figure 2 contains a hypothetical illustration of this procedure. As is typical in experience sampling, the observations are made at random time points within segments of—in this case—90 minutes (see De Haan-Rietdijk et al., 2017). Subsequently, two different time grids are shown, based on creating segments of half an hour, or segments of one hour. Each observation is included in the segment where it took place; when no observation was made within a particular segment, this becomes a missing value in the restructured time series. In some instances, two observations may fall into the same segment of the time grid that is used; this is the case for observations 3 and 4 when using the one-hour grid. In DSEM in Mplus, one of these will then be moved to an adjacent segment, as is also illustrated in Figure 2 (see Appendix A of Asparouhov et al. (2018) for details). This procedure tends to work quite well up to 80% of missing data. Yet, it is important to realize that—in general—lagged parameters (and residual variances) are a function of interval length (Deboeck & Preacher, 2015; Dorman & Griffin, 2015; Driver et al., 2017; Ryan et al., 2018). Hence, the results should always be interpreted with respect to the time grid that was used.

1.1.3 Empirical Illustration: Part 1. The empirical data we use come from 129 participants. In Figure 3 the data from three participants is shown: on the left, the sequences of the two variables are shown, with negative affect in black, and unpleasantness of events in grey; on the right, the histograms for these variables are shown. It shows there is quite some diversity across individuals in the amount and patterns of variability over time. While some individuals are characterized by a somewhat symmetric distribution, there are also individuals that have very skewed distributions, with many observations at the floor or within the lowest region of the scale. Our modeling approach is actually based on the assumption that the residuals are normally distributed, which may not be entirely compatible with these data; we will elaborate on possible alternatives in the discussion.

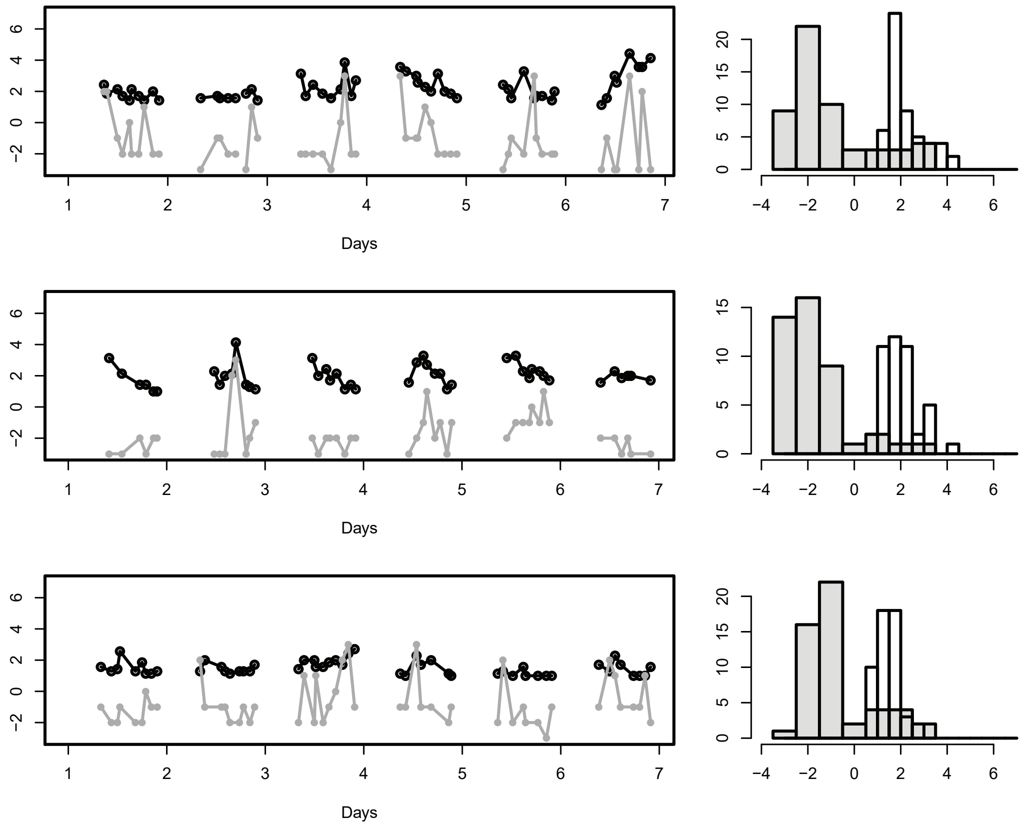


Figure 3. Sequences of three individuals for negative affect (in black), and unpleasantness of events (in grey), and histograms for these variables (black open bars for negative affect, and grey filled for unpleasantness of events).

We perform our DSEM analyses with Mplus v8.6, which is based on Bayesian estimation with non-informative (i.e., flat) priors (for details on Bayesian estimation and DSEM see Asparouhov et al., 2018; for details on model specification, number of iterations, convergence, computation time, etc., see the accompanying website: https://ellenhamaker.github.io/DSEM-book-chapter/). We use the Monte Carlo option in Mplus to analyze the data for each person separately and then average the individual results across all 129 persons. Prior to analysis, we standardize the data per person, as this makes the cross-regression parameters easier to interpret in terms of an effect size, and to compare their relative size to each other. In Table 1 we report the average (across 129 participants) point estimates, the standard deviation (across the participants) of these individual point estimates, and the percentage of individuals whose 95% credibility interval (CI) did not contain zero. The latter implies that, based on an individual's $N = 1$ analysis, there is evidence in the data that, for this person, the parameter differs from zero.

Table 1

Averaged Results For $N = 1$ Models

	Model 1		Model 2		Model 3	
Parameter	$\bar{\hat{\theta}}$ (SD$_{\hat{\theta}}$)	%	$\bar{\hat{\theta}}$ (SD$_{\hat{\theta}}$)	%	$\bar{\hat{\theta}}$ (SD$_{\hat{\theta}}$)	%
β_y	0.307 (0.174)	0.527	0.267 (0.161)	50.4	0.250 (0.174)	38.0
ϕ_y			0.362 (0.276)	49.6	0.364 (0.174)	49.6
β_x					0.101 (0.241)	12.4
ϕ_x					0.038 (0.232)	6.2

Note: Results for the $N = 1$ Models 1, 2, and 3 averaged across 129 participants, including: the cross-regression from unpleasantness of events to negative affect (β_y); the autoregression of negative affect (ϕ_y); the cross-regression from negative affect to unpleasantness of events (β_x); and the autoregression of unpleasantness of events (β_x). $\bar{\hat{\theta}}$ represents the average (across participants) standardized parameter estimate; SD$_{\hat{\theta}}$ represents the standard deviation (across participants) of this estimate; % indicates the percentage of participants whose 95% credibility interval does not cover zero.

If we focus on the results of Model 3, this shows that there is evidence for reciprocal effects between negative affect and unpleasantness of events for some individuals, and that there seem to be three times as many participants whose negative affect is affected by the unpleasantness of events (i.e., β_y), than vice versa (i.e., β_x). Moreover, unpleasantness of events seems on average determined more by prior negative affect (as quantified by β_x), than by prior unpleasantness of events (as quantified by ϕ_x).

While these analyses give some insight in the within-person dynamics and individual differences therein, it does not allow us to investigate how the individual differences (for instance, in means and cross-regressions) are related to each other or to other person characteristics. To obtain more insight in these patterns of individual differences, we turn to a multilevel approach.

1.2 Combining Time Series Modeling with Multilevel Modeling

Multilevel modeling is based on analyzing clustered data, using a model for the within-cluster variation and a model for the between-cluster variation (Raudenbush & Bryk, 2002). Although we have to impose the same model at the within-level for every person, multilevel analysis allows for quantitative differences between individuals in their parameters. Such individual differences are referred to as random effects, and are bounded by a distribution. This implies that, in contrast to the replicated time series analysis presented above where the individual parameters could take on any value, in the multilevel approach they are restricted to come from—for instance—a multivariate normal distribution. Note, however, that the effect of such a distribution becomes weaker as the sample size (at the within level) increases. The random effects can be further investigated with a model at the between level.

1.2.1 Three Dynamic Multilevel Models.

Fundamental to the multilevel approach in DSEM is that the observed variables for individual i at occasion t (i.e., negative affect y_{it} and unpleasantness of events x_{it}), are decomposed into a person mean (i.e., $y_i^{(b)}$ and $x_i^{(b)}$), and a temporal deviation from that mean (i.e., $y_{it}^{(w)}$ and $x_{it}^{(w)}$); this is visualized at the top left of Figure 4. The latter components are then further modeled at the within level using a time series model to account for the dynamic relations within a person over time. The individual means only contain between-person variance, and can be further modeled at the between level. We consider three models that are based on combining time series analysis with multilevel analysis. Their analytical expressions are presented in Table 2, as Models 1, 2 and 3.

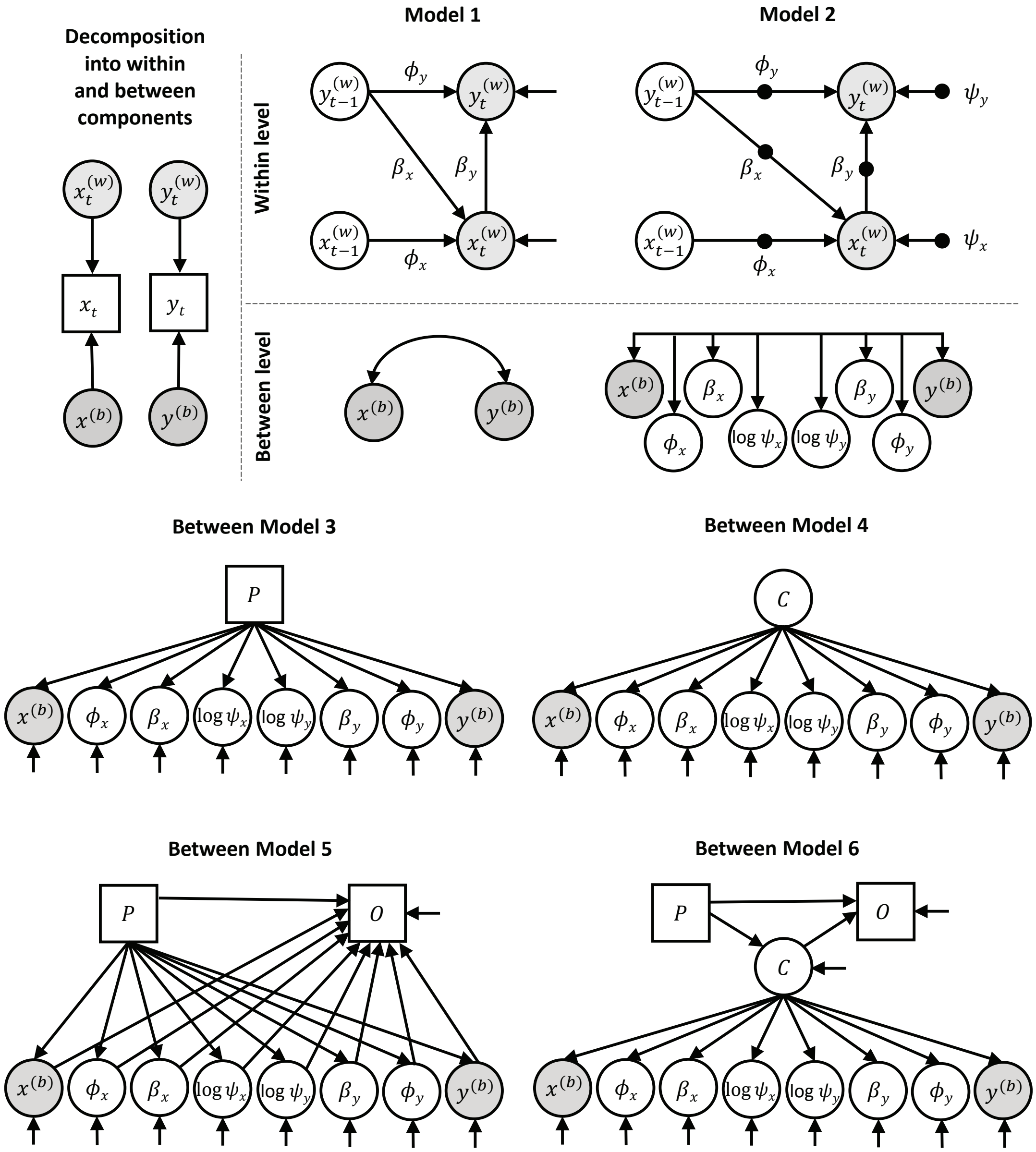


Figure 4. Six multilevel dynamic structural equation models. Observed time-varying variables (represented by squares) are decomposed into a within-person component that varies over time, and a between-person component that is invariant over time (represented by circles). Model 1 has fixed parameters at the within level, meaning every person gets the same regression coefficients and residual variances. Models 2-6 have random slopes and random residual variances, represented by filled circles at the within level that correspond to the open circles at the between level.

Table 2

Six Dynamic Structural Equation Models

Model	Within Expression	Between Expression
1	$x^{(w)}_{it} = \phi_x x^{(w)}_{it-1} + \beta_x y^{(w)}_{it-1} + \zeta_{xit}$ $y^{(w)}_{it} = \phi_y y^{(w)}_{it-1} + \beta_y x^{(w)}_{it} + \zeta_{yit}$	$x^{(b)}_{i} = \gamma_{x0} + u_{x0i}$ $y^{(b)}_{i} = \gamma_{y0} + u_{y0i}$
2	$x^{(w)}_{it} = \phi_{xi} x^{(w)}_{it-1} + \beta_{xi} y^{(w)}_{it-1} + \zeta_{xit}$ $y^{(w)}_{it} = \phi_{yi} y^{(w)}_{it-1} + \beta_{yi} x^{(w)}_{it} + \zeta_{yit}$	$x^{(b)}_{i} = \gamma_{x0} + u_{x0i}$ $y^{(b)}_{i} = \gamma_{y0} + u_{y0i}$ $\phi_{xi} = \gamma_{x1} + u_{x1i}$ $\phi_{yi} = \gamma_{y1} + u_{y1i}$ $\beta_{xi} = \gamma_{x2} + u_{x2i}$ $\beta_{yi} = \gamma_{y2} + u_{y2i}$ $log(\psi_{xi}) = \gamma_{x3} + u_{x3i}$ $log(\psi_{yi}) = \gamma_{y3} + u_{y3i}$
3	As Model 2	$x^{(b)}_{i} = \gamma_{x00} + \gamma_{x01} P_i + u_{x0i}$ $y^{(b)}_{i} = \gamma_{y00} + \gamma_{y01} P_i + u_{y0i}$ $\phi_{xi} = \gamma_{x10} + \gamma_{x11} P_i + u_{x1i}$ $\phi_{yi} = \gamma_{y10} + \gamma_{y11} P_i + u_{y1i}$ $\beta_{xi} = \gamma_{x20} + \gamma_{x21} P_i + u_{x2i}$ $\beta_{yi} = \gamma_{y20} + \gamma_{y21} P_i + u_{y2i}$ $log(\psi_{xi}) = \gamma_{x30} + \gamma_{x31} P_i + u_{x3i}$ $log(\psi_{yi}) = \gamma_{y30} + \gamma_{y31} P_i + u_{y3i}$
4	As Model 2	As Model 3, but with latent variable C_i instead of observed variable P_i
5	As Model 2	As Model 3 but with the addition:

Six Dynamic Structural Equation Models (continued)

Model	Within Expression	Between expression
		$O_i = \tau_0 + \tau_1 P_i + \tau_2 x_i^{(b)} + \tau_3 y_i^{(b)}$ $+\tau_4 \phi_{xi} + \tau_5 \phi_{yi} + \tau_6 \beta_{xi} + \tau_7 \beta_{yi}$ $+\tau_8 log(\psi_{xi}) + \tau_9 log(\psi_{xi}) + \epsilon_i$
6	As Model 2	As Model 4, but with P_i and O_i: $O_i = \tau_0 + \tau_1 P_i + \tau_2 C_i + \epsilon_{Oi}$ $C_i = \tau_3 P_i + \epsilon_{Ci}$

Note: Six dynamic structural equation models for two variables x and y: Model 1, 2, and 3 are based on combining time series modeling with multilevel modeling; Models 4, 5, and 6 also include a structural equation modeling component. Model 1 combines random means with fixed slopes and fixed residual variances. All other models have random means, slopes and residual variances. In Model 2 these are correlated. Model 3 includes an observed predictor P_i for the random effects. Model 4 contains a latent variable C_i with the random effects as indicators. Model 5 includes an observed predictor P_i and an observed outcome O_i of the random effects. Model 6 is as Model 4, but includes an observed predictor P_i and observed outcome O_i of the latent variable C_i as well.

Multilevel Model 1 uses the bivariate time series model from the individual analysis (i.e., $N = 1$ Model 3) as the within level model (see Figure 4). The model has fixed slopes, which implies that every person is characterized by the same autoregressive and cross-regressive parameters.[2] Furthermore, the residual variances are also the same for each person in this model. Hence, the only source of individual differences here are the two random effects that stem from the decomposition, that is, the individual mean of negative

[2] The terms random effect and fixed effect are used quite differently in different disciplines; see https://statmodeling.stat.columbia.edu/2005/01/25/why_i_dont_use/

affect ($y_i^{(b)}$) and the individual mean of unpleasantness of events ($x_i^{(b)}$). These scores can be interpreted as trait scores, or as a person's setpoint or equilibrium: When there are no external influences, the person will return to these values over time. Since these components do not vary over time, they only exist at the between level, where they are allowed to be correlated, as shown in Figure 4.

Our previous $N = 1$ analyses, however, suggested that there may be quite some variation across individuals for the autoregressive and cross-regressive parameters. Therefore, in Model 2 we allow for regression parameters and residual variances to be individual specific. The latter are included to represent that individuals may be differently affected by external and internal factors (Jongerling, Laurenceau, & Hamaker, 2015). In the top right panel of Figure 4 the random slopes and residual variances at the within level are represented by filled circles; these random effects become (latent) variables represented by open circles at the between level. In Model 2 all eight random effects (i.e., two means, four regression parameters, and two residual variances) are correlated with each other. The residual variances are log transformed to ensure that the individual variances are never negative.

Model 3 is based on the same within level model as Model 2, while at the between level an observed predictor P_i is included for the random effects, as shown in Figure 4 (Between Model 3). Including such observed level 2 predictors—also known as time-invariant predictors or baseline covariates—is quite common in multilevel analysis. When a random slope is predicted by a between level variable, this is sometimes referred to as a cross-level interaction, as it is based on an interaction between a within level variable and a between level variable. Furthermore, we can also investigate whether individual differences in the residual variances are predictable by the between level covariate. For interpretation purposes, it is helpful to grand mean center between level predictors like P_i; that way, when regressing random effects on these predictors, the intercepts (e.g., γ_{x00} and γ_{x10} in Table 2) can be interpreted as the mean or average of a random effect.

The models described here are closely related to the multilevel VAR(1) models discussed in Hamaker et al. (2018). However, as was already discussed when presenting the $N = 1$ models, a critical difference is that the current models contain a lag 0 regression, and as a consequence the residuals of the two variables are not allowed to be correlated. Hence, in contrast to the models considered in Hamaker et al. (2018) where the random residual variances were combined with random residual covariance, which required the introduction of a separate latent variable, the current models with random residual variances (i.e., Models 2 and 3) do not require such an additional latent variable to capture individual differences in the commonness of the error structure.

1.2.2 Empirical Illustration: Part 2. As a first descriptive of the empirical data from a multilevel perspective, we consider the intraclass correlations of both variables. The intraclass correlation can be expressed as the between level variance divided by the total variance of a variable, and it thus represents the proportion of total variance that stems from stable, trait-like between-person differences. For negative affect, the intraclass correlation is 0.455 (CI=(0.392, 0.523)), meaning that about half of the observed variance is due to stable between-person differences, while the other half is due to fluctuations within individuals over time. For unpleasantness of events, we find an intraclass correlation of 0.092 (CI=(0.070, 0.121)), meaning that the variation in this variable is mostly due to within-person fluctuations over time.

In comparing the parameter estimates that are obtained with these multilevel models, we focus on the standardized results again. Standardizing parameters in multilevel models is not common, as there are various variances that can be used for this purpose. Schuurman, Ferrer, de Boer-Sonnenschein, and Hamaker (2016) argue that standardization of person-specific parameters in multilevel models—such as the cross-regression β_{xi} and β_{yi} in Models 2 and 3 above—should be done using the person-specific within-person variances of the associated variables, reasoning that this most closely corresponds to standardization of parameters as it is done in $N = 1$ analysis. This has been implemented in Mplus, where the parameters are standardized per person within each iteration of the Bayesian MCMC algorithm, resulting in a posterior distribution for each standardized parameter per person (cf. Asparouhov et al., 2018).

The average, individually standardized, parameter estimates of the three multilevel models discussed above are presented in Table 3. It shows that overall, these parameters and CIs are pretty stable across the three models, which is what we would expect. When comparing these average slope estimates to the ones obtained in the replicated time series approach based on analyzing the data of each individual separately (see Model 3, Table 1), we also see that the average cross-regressions are quite similar, but that the average autoregressive parameters from the replicated time series approach are clearly lower than the average autoregressions obtained with these multilevel models; the latter is in line with Nickel's bias (Asparouhov et al., 2018).

Table 3

Parameter Estimates For Dynamic Multilevel Models

Parameter	Model 1	Model 2	Model 3
β_y	0.221 (0.199, 0.243)	0.232 (0.210, 0.252)	0.218 (0.197, 0.239)
ϕ_y	0.534 (0.512, 0.557)	0.470 (0.443, 0.495)	0.476 (0.450, 0.502)
β_x	0.108 (0.077, 0.141)	0.102 (0.071, 0.137)	0.106 (0.074, 0.139)
ϕ_x	0.154 (0.116, 0.194)	0.129 (0.089, 0.166)	0.131 (0.090, 0.172)

Note: Averaged individually standardized estimates (and their credibility interval) for: the cross-regressive parameter from unpleasantness of events to negative affect (β_y), the autoregressive parameter for negative affect (ϕ_y), the cross-lagged parameter from negative affect to unpleasantness of events (β_x), and the autoregressive parameter for unpleasantness of events (ϕ_x). Model 1 has fixed slopes and residual variances; Model 2 has random slopes and residual variances that are allowed to be correlated; Model 3 includes a single observed predictor for these random effects.

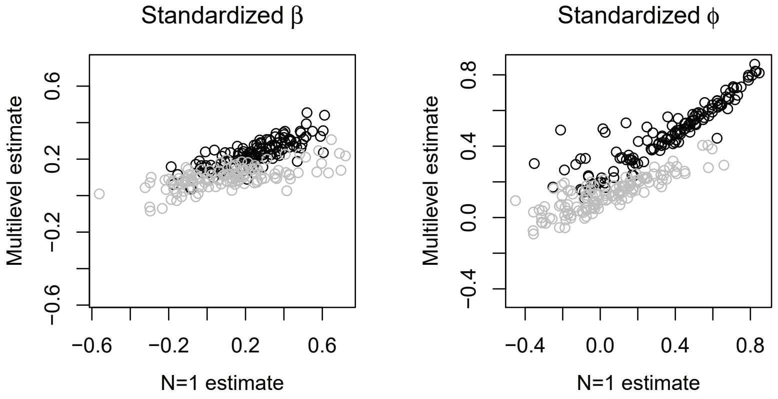


Figure 5. Standardized parameters from the multilevel model and the $N = 1$ analyses plotted against each other. Left panel contains the standardized cross-regressions β_{yi} (in black; from unpleasantness of events to negative affect) and β_{xi} (in grey); right panel contains the standardized autoregressions ϕ_{yi} (in black; for negative affect) and ϕ_{xi} (in grey; for unpleasantness of events).

To gain more insight in how the replicated time series approach and the multilevel approach deviate, we have plotted the individual standardized cross-regression parameters and autoregressions from multilevel Model 2 against these parameters from the replicated time series approach in Figure 5. This clearly shows that the multilevel approach is based on "borrowing strength" across cases: The variability of the multilevel estimates is much smaller than that of the replicated time series approach, as the estimates are pulled towards the grand mean. This shrinkage is determined by the uncertainty of the parameter estimates, and it can be seen that especially some of the more extreme values (e.g., the negative autoregressive parameters from the individual analyses) are pulled quite strongly towards the grand mean in the multilevel analysis (for an explicit discussion of the degree of shrinkage as a function of the reliability of estimates in the context of frequentist multilevel modeling, see Chapter 3 of Raudenbush & Bryk, 2002). Yet, the correlations between the estimates from these two approaches are considerable: It is 0.79 for the standardized β_{yi}, 0.66 for the standardized β_{xi}, 0.91 for ϕ_{yi}, and 0.84 for ϕ_{xi}.

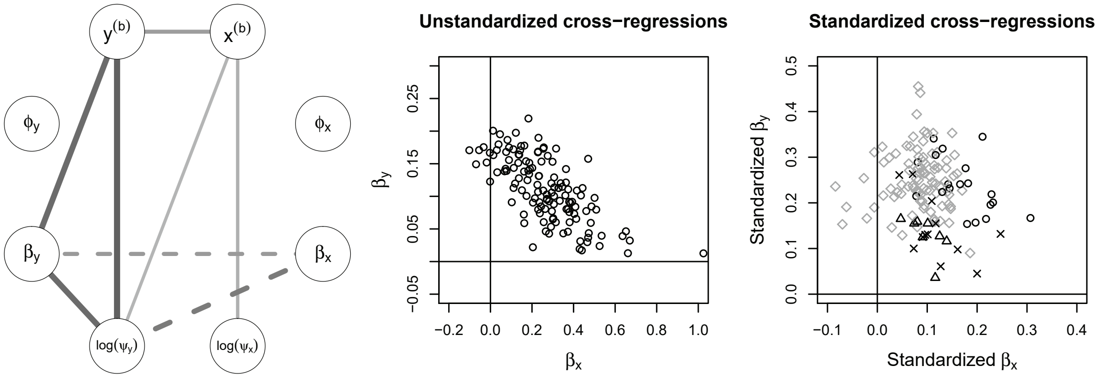


Figure 6. Between-person results for Model 2. Left panel represents the correlations between the individual's mean on negative affect ($y^{(b)}$), the individual's mean on unpleasantness of events ($x^{(b)}$), the autoregressive parameters for both (ϕ_{yi} and ϕ_{xi}), the cross-regressions (β_{yi} and β_{xi}), and the log of the residual variances of each ($log(\psi_{yi})$ and $log(\psi_{xi})$). Positive correlations are represented as solid lines, negative correlation as dashed lines; thickness of the lines indicates strength of the correlation. Middle panel shows the relation between the two random cross-regressions. Right panel shows the relation between the individually standardized cross-regressions. A distinction is made between four groups based on whether the credibility intervals of the two parameters contained zero or not: a) both CIs did not include zero (black circles); b) only the CI for the standardized β_{xi} did not include zero (black crosses); c) only the CI for the standardized β_{yi} did not include zero (grey diamonds); and d) both CIs contained zero (black triangles).

In Model 2 there are eight random effects, so there are $(8 \times 7)/2 = 28$ correlations between them. Of these, eight have a CI that does not contain zero, whereas 20 CIs cover zero. This is visualized in the left panel of Figure 6, where negative correlations are represented by dashed lines, positive correlations are represented by solid lines, and the thickness of the connections indicates the size of the correlations. The largest correlation here is 0.692 (between the mean of negative affect ($y^{(b)}$), and the residual variance of negative affect ($log(\psi_{yi})$), and the smallest correlation is 0.324 (between the mean of unpleasantness of events ($x^{(b)}$), and the residual variance of negative affect). The results suggest that individuals who are high on average on negative affect tend to have large residual variances, and a high cross-regressive parameter from unpleasantness of events to negative affect (β_{yi}). This could be interpreted as meaning these individuals are more responsive to changes in measured events, but also to unmeasured factors (Jongerling et al., 2015).

Somewhat puzzling perhaps, is the negative correlation between the two cross-regressive coefficients (β_{yi} and β_{xi}). We have plotted these individual slopes of the 129 participants against each other in the middle panel of Figure 6. It shows that individuals with a large cross-regressive parameter from unpleasantness of events to negative affect (β_{yi}), tend to have a cross-regressive parameter from negative affect to unpleasantness of events (β_{xi}) that is close to zero, and vice versa. These are unstandardized parameters, however, which means they are scale dependent, and they are actually inversely related to the individual's variability in y and x.[3] We have therefore also plotted the individually standardized cross-regressions in Figure 6 (see right panel). This shows that the relatively strong negative correlation we found for the unstandardized cross-regression coefficients, disappears when focusing on the standardized cross-regressions. Which of these relations should be considered more interesting from a substantive point of view, is open for debate and may depend on the context and purpose.

We may also choose to take an even further individually oriented perspective, by considering whether the individual CIs contained zero or not, either for the standardized or the unstandardized coefficients. We show this for the standardized parameters, which divides the 129 participants into four groups: 19 individuals have CIs for both standardized cross-regressions that do not contain zero (black circles in the right panel of Figure 6); 90

[3] If we assume the autoregressions are zero, we have: $\beta_y = cor(x_t, y_t)\frac{sd(y)}{sd(x)}$ and $\beta_x = cor(x_t, y_{t-1})\frac{sd(x)}{sd(y)}$ for this model; this shows that the two unstandardized parameters are inversely related to the variability in the two variables.

individuals (represented as grey diamonds) have a CI for their standardized β_{yi} that does not contain zero, while the CI for their standardized β_{xi} does; 12 individuals (represented by black crosses) have a CI for their standardized β_{xi} that does not contain zero, while the CI for their standardized β_{yi} does; and 8 individuals (represented as black triangles), whose CIs for both standardized cross-regressions contained zero. Hence, there were 109 out of 129 individuals who showed evidence for a spill-over effect from unpleasantness of events to negative affect, and 21 individuals who showed evidence for spill-over from their negative affect to unpleasantness of evens.

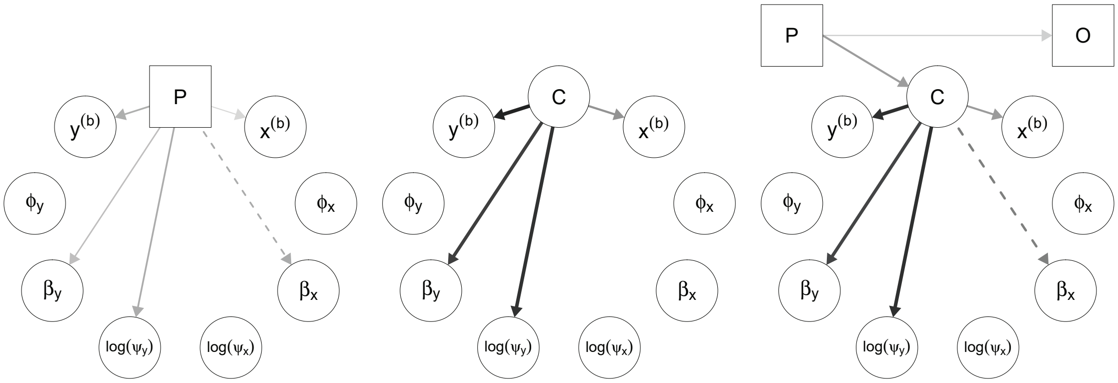


Figure 7. Between-person results for Models 3, 4 and 6. Left panel shows the standardized regression coefficients from Model 3, where the random effects are regressed on an observed predictor. Middle panel shows standardized factor loadings from Model 4, in which the random effects are indicators of a single latent variable. Right panel shows standardized factor loadings and regression coefficients for the observed predictor, latent variable and observed outcome. Positive parameters are represented with solid lines, negative parameters with dashed lines; thickness of the lines indicates the size of the parameter.

In Model 3 we include a baseline measurement of depression as a predictor (P_i) of the eight (unstandardized) random effects, and find that five regression coefficients have a CI that does not include zero. The left panel of Figure 7 is a visualization of the standardized regression parameters from this model, where again the thickness indicates the size, solid lines imply a positive parameter, and dashed lines a negative one. It shows that individuals high on baseline depression (P_i) tend to have a higher average for negative affect ($y_i^{(b)}$), and a higher average for unpleasant events ($x_i^{(b)}$). Furthermore, individuals high on baseline depression also tend to respond with more change in their negative affect after a one unit change in unpleasantness of events (β_{yi}), and their residual variance for negative affect also tends to be larger ($log(\psi_{yi})$). Finally, individuals with a higher score on baseline depression tend to have a lower unstandardized cross-regression from negative affect to unpleasantness of events (β_{xi}), which means that their level of negative affect does not tend to spill-over into the unpleasantness of the events they experience. Note however that, again, these relations may be very different when considering individually standardized cross-regressions, and it is not obvious which of these should be preferred.

1.3 Combining Time Series Modeling and Multilevel Modeling with SEM

We can further model the random effects—which include the individual means, slopes, and variances—at the between level, using an SEM approach. This implies that we can specify latent variable models and/or path models, in which we include time-invariant observed variables, and the random effects. Below, we consider three examples of this.

1.3.1 Three Full DSEM Models. The between level model for the three final models we consider are also included in Figure 4, and their expressions are presented in Table 2 as Models 4, 5 and 6. In Model 4, rather than using an observed variable P_i to predict the random effects, we specify a latent variable (or factor) C_i to account for what the random effects have in common. Models 5 and 6 can be thought of as path analysis (or mediation models), in which there is an observed predictor P_i and an observed outcome O_i. In Model 5, the effect of the predictor on the outcome is partly mediated by the random effects. In contrast, in Model 6 the indirect effect is through the common factor C_i, rather than through all eight random effects. The latter model, which combines path analysis and factor analysis, is therefore simpler than Model 5 in terms of the number of parameters that need to be estimated.

1.3.2 Empirical Illustration: Part 3. The results obtained at the between level for Model 4 are visualized in the middle panel of Figure 7. It shows that especially the mean negative affect ($y_i^{(b)}$), the effect of unpleasantness of events on negative affect (β_{yi}), and the residual variance of negative affect ($log(\psi_{yi})$) are largely determined by the underlying factor (C_i), as the standardized factor loadings for these three indicators lie between 0.757 and 0.856. Additionally, the mean for unpleasantness of events ($x_i^{(b)}$) is also related to this underlying factor, but less strongly than the first three indicators. This pattern is somewhat similar to the pattern observed for Model 3 with the observed baseline predictor (see left panel of Figure 7), although the current model has stronger relations, and there is no connection with the slope for negative affect on subsequent unpleasantness of events (β_{xi}).

Models 5 and 6 are based on including both an observed baseline variable, here the depression score prior to the first experience sampling episode, and an observed distal outcome, here the depression score after the experience sampling episode (prior to the second episode). While we assume there may be a direct effect of the baseline measure on the distal outcome, we investigate whether there are also indirect effects through the random effects, that is, the means, autoregression, cross-regressions and residual variances. In Model 5, we therefore have one direct effect and eight indirect effects. When estimating this model, we encounter some problems that seem to imply that the model we are trying to estimate is too complex for the data.[4] While further steps could be taken—such as fixing certain regression coefficients to zero, or specifying more informative priors—we do not pursue with this model here.

Instead, we move to Model 6, which is a simpler model in that it is based on extracting a common source of variance from the random effects (i.e., a latent variable, like Model 4), and using this to model an indirect effect. The results for this model are

[4] Specifically, we found that the trace plots of some of the parameters showed eruptions of extreme values; see https://ellenhamaker.github.io/DSEM-book-chapter/ for more details.

visualized in the right panel of Figure 7. We find a direct effect of the observed baseline covariate on the distal outcome (standardized regression coefficient is 0.184, CI=(0.036, 0.323)). Furthermore, while there is an effect of the baseline covariate on the common factor (standardized regression coefficient is 0.386, CI=(0.255, 0.500)), there is no evidence that the common factor affects the distal outcome, and therefore we conclude there is no indirect effect from prior depression to later depression through the common factor of the random effects. What is notable about this model in comparison to Model 4, is that the effect of negative affect on subsequent unpleasantness of events (i.e., the random slope β_{xi}) is now an indicator of the common factor with a standardized factor loading of -0.508 (CI=(-0.940, -0.068)). Hence, including the baseline predictor as a covariate of the factor somewhat changes the character of the common factor.

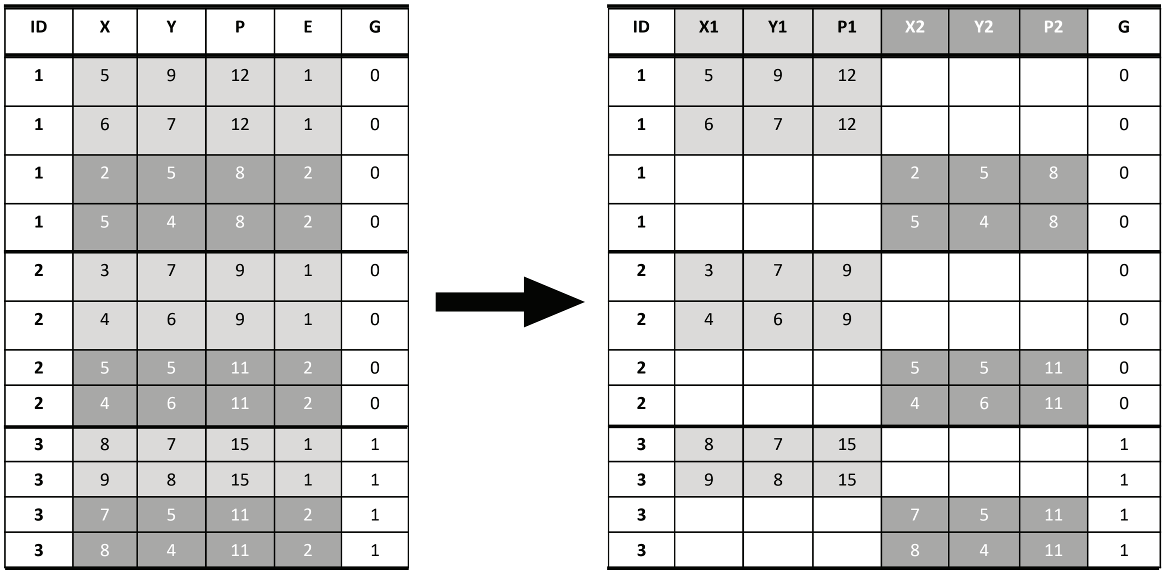
2. Using DSEM For Pretest-Posttest Data

Thus far we have discussed models that can be used when there is a single episode of intensive longitudinal measures obtained from a single group. However, the data that we are using actually have a more complex structure in that after this initial episode, individuals were randomly assigned to either a treatment or a control condition, and a second episode of intensive longitudinal measures was obtained after the treatment period. Hence, we have a pretest-posttest design with experience sampling data.

There are three basic questions of interest in a pretest-posttest design. First, we want to ensure that there are no initial differences between the groups on the pretest. Second, we want to know whether there is an effect of time, which we can investigate by looking at whether the control group changes from pretest to posttest. Finally and most importantly, we want to see whether the treatment has an effect, and thus whether there are differences between the groups on the posttest. With the current dataset, each of these questions can be posed with respect to the means; for instance, we may hypothesize that the mean of negative affect decreases as a result of treatment. But we can also consider effects on the autoregressive and cross-regressive parameters, and the residual variances; for instance, we may hypothesize that the carryover of negative affect and the spill-over of events into negative affect are reduced by treatment.

We can now distinguish between two factors in the design. First, group G_i is a between-person factor that can be represented by a dummy variable in our analyses. Second, episode E_{it} is a within-person factor, and while we could also use a dummy variable, that would have severe drawbacks for our analysis. Suppose we would use a dummy that represents the posttest episode; this implies that the within level predictors during the posttest episode are not centered with the individual's means from that episode, but with the means of the pretest episode. As a result the regression coefficient of the dummy would not represent the actual change in mean between the pretest and the posttest, and it would become hard—if not impossible—to actually determine this change based on the parameter esitmates. Moreover, this approach would not allow us to investigate changes in autoregression, cross-regression, or residual variances.

To avoid these issues, we restructure the data such that a variable that was measured during both episodes is now represented by two separate variables: One that contains the observations that were made during the first episode, and another that contains the observations made during the second episode. This is illustrated in Figure 8. It shows that the variables in the restructured datafile that represent observations associated with the first episode (i.e., $x1_{it}$, $y1_{it}$, and—if included—the baseline measure for this episode, $p1_i$) have missing values for the time points that fall in the second episode, while the variables that represent observations associated with the second episode (i.e., $x2_{it}$, $y2_{it}$, and—if included—the baseline measure for this episode, $p2_i$) contain missing values for the time points from the first episode. In this way, we get separate variables for each episode and these are each decomposed into a within and a between component. Subsequently, the within-person components can then be modeled for each episode separately, allowing for different slopes and residual variances in each episode.



ID	X	Y	P	E	G
1	5	9	12	1	0
1	6	7	12	1	0
1	2	5	8	2	0
1	5	4	8	2	0
2	3	7	9	1	0
2	4	6	9	1	0
2	5	5	11	2	0
2	4	6	11	2	0
3	8	7	15	1	1
3	9	8	15	1	1
3	7	5	11	2	1
3	8	4	11	2	1

ID	X1	Y1	P1	X2	Y2	P2	G
1	5	9	12				0
1	6	7	12				0
1				2	5	8	0
1				5	4	8	0
2	3	7	9				0
2	4	6	9				0
2				5	5	11	0
2				4	6	11	0
3	8	7	15				1
3	9	8	15				1
3				7	5	11	1
3				8	4	11	1

Figure 8. Restructured data from a pretest-posttest design with two episodes of intensive longitudinal measurements. Variables in the original datafile depicted on the left include: identifier for cluster (e.g., person; ID); two variables from the intensive measurements (X and Y), a baseline covariate measured prior to every intensive measurement episode (P); identifier for episode (E); identifier for treatment group (G). Variables in the restructured datafile on the right include: ID and G as in the original datafile; two variables from the intensive measurements during the first episode (X1 and Y1), and a baseline covariate that is measured prior to the first episode (P1); two variables from the intensive measurements during the second episode (X2 and Y2), and a baseline covariate that is measured prior to the second episode (P2). Blank cells correspond to missing data.

Here we focus on a pretest-posttest DSEM analysis with negative affect (y_{it}) as the outcome, and unpleasantness of events (x_{it}) as its predictor. Our ultimate interest is in whether treatment has an effect on the mean level of negative affect ($y_{it}^{(b)}$), the inertia in negative affect (ϕ_i), the sensitivity of negative affect to unpleasantness of events (β_i), and the sensitivity of negative affect to other, unmeasured sources (ψ_i). Each of these four random effects is estimated for each individual during both episodes, as well as the mean

Table 4

Prettest-Posttest Model For Two Groups

Equation	Expression	Description
1	$y1_{it}^{(w)} = \phi_{1i} y1_{it-1}^{(w)} + \beta_{1i} x1_{it}^{(w)} + \zeta_{xit}$	Within model for negative affect (episode 1)
2	$y1_{i}^{(b)} = \gamma_{00} + \gamma_{01} G_i + u_{0i}$	Mean of negative affect (episode 1)
3	$\phi_{1i} = \gamma_{10} + \gamma_{11} G_i + u_{1i}$	Autoregression for negative affect (episode 1)
4	$\beta_{1i} = \gamma_{20} + \gamma_{21} G_i + u_{2i}$	Cross-regression from events to affect (episode 1)
5	$log(\psi_{1i}) = \gamma_{30} + \gamma_{31} G_i + u_{3i}$	Log residual variance (episode 1)
6	$x1_{it}^{(b)} = \gamma_{40} + \gamma_{41} G_i + u_{4i}$	Mean of predictor events (episode 1)
7	$y2_{it}^{(w)} = \phi_{2i} y2_{it-1}^{(w)} + \beta_{2i} x2_{it}^{(w)} + \zeta_{yit}$	Within model for negative affect (episode 2)
8	$\Delta y_{i}^{(b)} = \gamma_{50} + \gamma_{51} G_i + u_{5i}$	Change in mean of negative affect
9	$\Delta \phi_i = \gamma_{60} + \gamma_{61} G_i + u_{6i}$	Change in autoregression of negative affect
10	$\Delta \beta_i = \gamma_{70} + \gamma_{71} G_i + u_{7i}$	Change in cross-regression from events to affect
11	$\Delta log(\psi_i) = \gamma_{80} + \gamma_{81} G_i + u_{8i}$	Change in log of residual variance
12	$\Delta x_{it}^{(b)} = \gamma_{90} + \gamma_{91} G_i + u_{9i}$	Change in mean of predictor events

Note: The first six equations are for episode 1, the latter six for episode 2. Equations 1 and 7 are within level expressions. Equations 2-6 are between level equations to determine whether there were initial differences between the two groups on: the mean for the outcome variable (captured by the regression parameter γ_{10}); the autoregression (γ_{11}); the cross-regression (γ_{21}); the log residual variance (γ_{31}); and the mean for the predictor (γ_{41}). Equations 8-12 are between level equations used to model the changes in: the mean ($\Delta y_i^{(b)} = y2_i^{(b)} - y1_i^{(b)}$); the autoregression ($\Delta \phi_i = \phi_{2i} - \phi_{1i}$); the cross-regression ($\Delta \beta_i = \beta_{2i} - \beta_{1i}$); and the log of the residual variance ($\Delta log(\psi_i) = log(\psi_{2i}) - log(\psi_{1i})$). The intercepts in these expressions (i.e., γ_{50} to γ_{90}) capture changes in the reference group (when $G_i = 0$); if these parameters are different from zero, this implies a change (on average) due to time. The regression coefficients for the dummy variable G_i (i.e., γ_{51} to γ_{91}) capture differential change between the two groups; hence, if these are different from zero, this represent a treatment effect.

level of unpleasantness of events ($x_i^{(b)}$). At the between level, the change of each of these parameters across the two episodes is regressed on the grouping variable (G_i), to determine whether the average change in these parameters is different across the treatment groups. The analytical expressions included in Table 4 show that: the regression coefficients for G_i during the first episode (γ_{01} to γ_{41}) indicate whether there are initial group differences; the intercepts during the second episode (γ_{50} to γ_{90}) indicate whether there is an effect of time; and the regression coefficients for G_i during the second episode (γ_{51} to γ_{91}) indicate whether there is a treatment effect.

When applying this model to the empirical dataset, we find that there are no initial differences between the two groups, as all the CIs of γ_{01} to γ_{41} contain zero. This is actually in line with what one would expect based on random assignment of participants to the treatment groups. Regarding the changes in the parameters, we see evidence for a change in the log residual variance of negative affect ($\Delta log(\psi)$ as a result of time: When regressing the change in this parameter on group, the intercept (γ_{90}) was estimated to be -0.267 (CI=(-0.508, -0.029)). This can be interpreted as meaning there is less residual variance in negative affect during the second episode when compared to the first episode. Furthermore, we also find evidence that three other parameters were affected by treatment. First, we find a negative effect of group on the change in mean of negative affect ($\gamma_{51} = -0.286$, CI=(-0.480, -0.096)), which implies there is, on average, a decrease in the mean level of negative affect among individuals who received treatment. Second, there is also a negative effect of group on the change in average unpleasantness of events ($\gamma_{91} = -0.264$, CI=(-0.421, -0.106), which implies that people indicate to experience less unpleasantness of events after treatment. Third, there is a negative effect of group on the change in the autoregressive parameter ($\gamma_{61} = -0.132$, CI=(-0.247, -0.007), which implies that the carry-over or inertia in negative affect is reduced due to treatment. Taken together, this would imply that after the mindfulness training, individuals experience lower levels of negative affect, lower levels of unpleasantness of events, and that they also tend to recover from perturbations to their negative affect and return to their equilibrium more quickly.

3. Discussion

In this chapter we have shown how time series modeling, multilevel modeling and structural equation modeling are combined in the DSEM framework for the analysis of intensive longitudinal data. We decided to focus on an empirical dataset, because it allowed us to highlight some of the fundamental strengths of DSEM and to illustrate how to use the framework to tackle specific research questions. However, this more in-depth treatment of certain DSEM aspects has precluded us from a more broad presentation of the diverse modeling opportunities offered by DSEM and other modeling strategies that exist for intensive longitudinal data. To point the reader to these alternatives, below we provide a brief overview with references. Furthermore, we discuss some of the most pressing unresolved issues for which future research is needed.

3.1 Other Modeling Options

In this chapter, we have been able to present only a few of the many possibilities that currently exist for modeling the dynamics of intensive longitudinal data. There are diverse flexible Bayesian packages like WinBUGS, jags, and stan, that allow researchers to build their own DSEM models. Furthermore, there are various R-packages that have been developed for the analysis of intensive longitudinal data. These include: *ctsem* (Driver et al., 2017) for continuous and discrete time modeling of $N = 1$ data and multilevel data; *mlVAR* (Epskamp, Deserno, & Bringmann, 2017), which estimates a multilevel first order vector autoregressive model; *gimme* (Gates & Molenaar, 2012), which is based on replicated $N = 1$ analyses that are then combined in a bottom-up approach; and *dynr* (Ou, Hunter, & Chow, 2019) for $N = 1$ regime-switching models.

The DSEM framework as it is implemented in Mplus also includes alternative modeling options that were not covered in the current chapter. First, it is possible to include latent variables in the time series model that is specified at the within level. This implies that we can study underlying constructs that are measured with multiple indicators, but it is also possible to have a single indicator, and use a latent variable model to separate the underlying process from measurement error (Schuurman & Hamaker, 2019), or to specify moving average terms (Asparouhov et al., 2018).

Second, there is an option to include time-varying predictors at the within level and model the lagged relations between the residuals rather than between the within-person components themselves. This is referred to as residual DSEM (RDSEM; Asparouhov et al., 2018; Asparouhov & Muthén, 2020), and can be of interest when there are, for instance, increasing or decreasing trends over time that vary across individuals. Other such trends could be cycles or repetitive patterns due to a circadian rhythm, a day-of-the-week effect, or a monthly cycle (Liu & West, 2016; Ram et al., 2005).

Third, the full DSEM framework contains an additional time-varying component that allows for random effects of time. When there are time series data from multiple individuals, these observations can be thought of as being clustered within individuals, but also within time points. This cross-classification allows for the study of changes over time of the observed and latent variables, but also of the structural parameters, such as autoregressions, cross-regressions, or factor loadings (Asparouhov et al., 2018).

Fourth, DSEM also allows for the analysis of categorical observed data through the use of a probit link function; this is based on specifying a continuous latent response variable behind the categorical observed variable (Asparouhov et al., 2018; Asparouhov & Muthén, 2019). Alternatively, when there is a strong floor effect for (some of) the participants (e.g., see the data in Figure 3), it may be useful to consider a two-part (semicontinuous) modeling approach, as suggested by Olsen and Schafer (2001). The two-part approach splits a variable into a binary variable that indicates whether the original variable has a value greater than the floor, and a continuous variable that represents the value above the floor. When the binary variable indicates that the original variable is at the floor value, the continuous variable is given a missing data flag. Two-part modeling is available in DSEM, where it converts a univariate model with a variable with a floor effect into a bivariate DSEM model with a categorical and a continuous variable.

Finally, although at this point not part of the released Mplus version yet, the DSEM framework has also been extended to account for regime-switching (Asparouhov, Hamaker, & Muthén, 2017). This implies that individuals may switch between distinct states that are each characterized by different means, variances, and dynamics. This matches well with certain descriptions of psychopathology, but may also prove useful in analyzing the data of healthy participants (Hamaker, Grasman, & Kamphuis, 2016).

3.2 Unresolved Issues

With the host of modeling opportunities that DSEM has to offer, a series of questions arises regarding: a) the underlying assumptions and the consequences of violating them; b) how to build and evaluate a model; and c) how to interpret the results and put them to use in practice. As this is still a relatively new research area, many of these questions have only been partly answered at best. Below, we elaborate on what we consider the most pressing issues in this field, in the hope that future research will soon bring more clarity on them.

All the analyses performed in this chapter are based on the assumption that the observed data are continuous rather than categorical; moreover, it is assumed that there is variability at both levels, and that the residuals at each level are multivariate normally distributed. However, as the data in Figure 3 already showed, assumptions regarding the distribution of observed variables are likely to be violated in practice. Especially when measuring variables such as symptoms or negative affect items in the general population, there tend to be many individuals with a skewed distribution, and a larger portion of observations at or near the floor. This forms a violation of the underlying assumptions. Alternatively, we could analyze the data as categorical or use two-part modeling as described above, or develop multilevel discrete-valued time series models. However, at this point the actual consequences of such violations—and hence, the actual need for alternatives to overcome them—are unclear, and require further simulation research. Related to this, research is needed to determine whether non-normal within level residuals flag a problem (e.g., biased estimates, CIs that are too wide or too narrow), or that there are no serious consequences. The assumption of multivariate normality at the between level is imposed by the prior that is used for the random effects. However, this may not be a real concern with enough data, as this prior will be overruled by the data when the time series are long enough. However, what can be considered "long enough" remains an open question at this point, and is likely to also depend on the complexity of the model (e.g., number of observed variables, number of random effects), and violations of other model assumptions.

Another major challenge is how to evaluate a model. In contrast to SEM, where every model that is specified is nested under the saturated model, when doing any of the DSEM analyses, there is no such thing as a saturated model. The reason for this is that the repeated measures are not independent, and therefore, not only is the lag zero (concurrent) covariance structure of interest, but also the covariance structures at all other possible lags contain information about dependencies that we try to account for with the model. In the time series literature, model fit is therefore often evaluated in two ways. First, the residuals of a model are obtained, to determine whether there is any autocorrelation left in them; if there is, this implies that the model did not fully account for the temporal dependencies that are present in the data, and it should be further improved. Second, the appropriateness of a model is often evaluated by considering its forecasts, which can be done using a cross-validation approach (cf. Hyndman & Athanasopoulos, 2021). Note, however, that forecasting is a very specific task, and a model that provides good forecasts does not necessarily provide a good description of the underlying mechanisms; hence, it depends on the goal one has, whether evaluating forecasts is a useful way to determine model fit.

In the absence of measures for overall model fit, we may still revert to model comparison through specifying two or more models that represent rivaling hypotheses, and comparing their appropriateness for the data to each other. This can currently be done using the deviance information criterion (DIC) in Mplus. However, the DIC is not always comparable across models, and it may also be rather unstable, making it a difficult to use measure in practice. Alternatively, one can make use of Bayes factors, although these tend to be very sensitive to the specifications of the priors. More locally, we can determine whether there is evidence for specific parameters in a model, using their CIs like we have shown in the empirical analyses discussed above. If the comparison of nested models involves multiple parameter testing, the Bayesian Wald test can be used instead (Asparouhov & Muthén, 2021). Furthermore, the development of posterior predictive checks in the context of DSEM may also prove beneficial for evaluating local model fit.

Finally, there are major challenges when it comes to interpreting results in a wider research context. Assuming that the goal is to unravel causal mechanisms, the DSEM analysis can be considered a step in between theory development and an actual intervention study, in that we can gain evidence for the theory before investing in an experimental study. However, using DSEM results for causal inferences is not straightforward. First, it is important to realize that lagged parameters are specific to the interval we focus on, and patterns may change (and even reverse) when we consider other interval lengths (Deboeck & Preacher, 2015; Ryan & Hamaker, 2021; Ryan et al., 2018). Second, if we find that a specific cross-regression changed as a result of treatment, this implies that treatment had a causal effect on the parameter, but it does not mean that the parameter itself represents a causal effect. We still need to consider whether (time-varying) confounders may have biased this within-person relation. Third, as our discussion of Model 2 with the random slopes has illustrated, between-level correlations between the unstandardized random effects may be very different from the between-level correlations between standardized random effects. Which of these are more informative may depend on the circumstances, but it is an aspect that researchers need to consider. Furthermore, while unstandardized parameters are informative about expected change for a one unit increase in the predictor, and standardized parameters are informative for proportions of explained variance, neither tells us how much change is possible as a result of an intervention: For instance, even if a specific variable has a small effect, when it is possible to increase this variable by a large amount compared to its "natural" variation, it may actually be a valuable target for an intervention. The latter requires thorough domain knowledge, rather than sophisticated statistics. These issues show that causal inference and reasoning are quite complex in this kind of research, and more research is needed in this area.

3.3 To Conclude

DSEM is a powerful toolbox of well-integrated statistical techniques that allow us to study the dynamics in intensive longitudinal data, investigate individual differences in these, and relate such differences to each other and to other person characteristics. With the stark increase of studies based on intensive longitudinal data, the need for innovative techniques that tap into the richness of these data is also growing. We hope and expect to see a lot of development in this area over the next few years in terms of new techniques and extensions of existing ones, of what should be considered good practice and rules of thumb when doing these kind of analyses, and in how to use these techniques for causal inference.

References

Armstrong, B., Covington, L. B., Unick, G. J., & Black, M. M. (2019). Bidirectional effects of sleep and sedentary behavior among toddlers: A dynamic multilevel modling approach. *Journal of Pediatric Psychology*, *44*, 275-285.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2017). Dynamic latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 257-269.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 359-388.

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 119-142.

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 275-297.

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 1-14.

Baldwin, A. L. (1946). The study of individual personality by means of the intraindividual correlation. *Journal of Personality*, *14*, 151-168.

Blanke, E. S., Neubauer, A. B., Houben, M., Erbas, Y., & Brose, A. (2021). Why do my thoughts feel so bad? getting at the reciprocal effects of rumination and negative affect using dynamic structural equation modeling. *Emotion*.

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

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. D. (1947). P-technique demonstrated in determining psycho-physiological source traits in a normal individual. *Psychometrika*, *12*(4), 267–288.

Cook, J., Tyson, R., White, R. R., Gottman, J. M., & Murray, J. (1995). Mathematics of marital conflict: Qualitative dynamic mathematical modeling of marital interaction. *Journal of Family Psychology*, *9*, 110-130.

Deboeck, P. R., & Preacher, K. J. (2015). No need to be discrete: A method for continuous time mediation analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 1-15.

De Haan-Rietdijk, S., Voelkle, M., Keijsers, L., & Hamaker, E. L. (2017). Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data.

Frontiers in Psychology, *8*, 1849.

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological wellbeing. *Nature Human Behavior*, *3*, 478-491.

Dorman, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods*, *20*, 489–505.

Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software*, *77*, 1-35.

Epskamp, S., Deserno, M. K., & Bringmann, L. F. (2017). mlVAR: Multi-level vector autoregression [computer software manual]. (R package version 0.4). Retrieved from `https://cran.r-project.org/web/packages/mlVAR/mlVAR.pdf`

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *65*, 310-319.

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 Clinincal Psychology*, *79*, 618-628.

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*, 80–92.

Hamaker, E. L. (2012). Why researchers should think "within-person" a paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (p. 43-61). New York, NY: Guilford Publications.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models. *Multivariate Behavioral Research*, *53*, 820-841.

Hamaker, E. L., Grasman, R. P. P. P., & Kamphuis, J. H. (2016). Modeling BAS dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment*, *23*, 436-446.

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.

Hamilton, J. D. (1994). *Time series analysis.* Princeton, NJ: Princeton University Press.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge, UK: University Press.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). Melbourne, Australia: OTexts. OTexts.com/fpp3. Accessed on: May 3, 2021.

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*, *184*, 334-349.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering: Transactions of the ASME Series D*, *82*, 35-45.

Koval, P., Burnett, P. T., & Zheng, Y. (2021). Affect dynamics. In P. Kuppens & C. Waugh (Eds.), (chap. Emotional inertia: On the conservation of emotional momentum).

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*, 984-991.

Liu, Y., & West, S. G. (2016). Weekly cycles in daily report data: An overlooked issue. *Journal of Personality*, *84*, 560-579.

Mehl, M. R., & Conner, T. S. (Eds.). (2012). *Handbook of research methods for studying daily life.* New York, NY: The Guilford Press.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*, 181–202.

Nesselroade, J. R. (2007). Factoring at the individual level: Some matters for the second century of factor analysis. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (p. 249-264). Mahwah, NJ: Lawrence Erlbaum.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, *96*, 730–745.

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2011). A hierarchical latent stochastic difference equation model for affective dynamics. *Psychological Methods*, *16*, 468–490.

Ou, L., Hunter, M., & Chow, S.-M. (2019). Whats for dynr: A package for linear and nonlinear dynamic modeling in R. *The R Journal*, *11*, 91-111.

Ram, N., Chow, S.-M., 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*, 773-790.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ryan, O., & Hamaker, E. L. (2021). Time to intervene: A continuous-time approach to network analysis and centrality. *Psychometrika*, *Published online ahead of print*.

Ryan, O., Kuiper, R. M., & Hamaker, E. L. (2018). A continuous time approach to intensive longitudinal data: The what, why and how. In K. van Montfort, J. Oud, & M. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (p. 27-54). Springer.

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*, 206-221.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive models. *Psychological Methods*, *24*, 70-91.

Sosnowska, J., Kuppens, P., De Fruyt, F., & Hofmans, J. (2019). A dynamic systems approach to personality: The personality dynamics (persdyn) model. *Personality and Individual Differences*, *144*, 11-18.

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*, 127-136.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151-176.

van der Krieke, L., Emerencia, A. C., Bos, E. H., Rosmalen, J., Riese, H., Aiello, M., ... de Jonge, P. (2015). Ecological momentary assessmnets and automated time series analysis to promote tailored health care: A proof-of-principle study. *JMIR Res Protoc.*, *4*, e100.

Walls, T. A., & Schafer, J. L. (Eds.). (2006). *Models for intesnive longitudinal data*. New York, NY: Oxford University Press.

Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Woods, W. C., Denissen, J. J. A., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, *34*, 1060-1072.

Wichers, M. C., Barge-Schaapveld, D. G. C. M., 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*, 923-931.