American Psychological Association

Psychological Methods

1082-989X/19/$12.00

2019, Vol. 24, No. 1, 70–91
http://dx.doi.org/10.1037/met0000188

Measurement Error and Person-Specific Reliability in Multilevel Autoregressive Modeling

Noémi K. Schuurman
Tilburg University

Ellen L. Hamaker
Utrecht University Utrecht

Abstract

An increasing number of researchers in psychology are collecting intensive longitudinal data in order to study psychological processes on an intraindividual level. An increasingly popular way to analyze these data is autoregressive time series modeling; either by modeling the repeated measures for a single individual using classic $n = 1$ autoregressive models, or by using multilevel extensions of these models, with the dynamics for each individual modeled at Level 1 and interindividual differences in these dynamics modeled at Level 2. However, while it is widely accepted in psychology that psychological measurements usually contain a certain amount of measurement error, the issue of measurement error is largely neglected in applied psychological (autoregressive) time series modeling: The regular autoregressive model incorporates innovations, or "dynamic errors," but not measurement error. In this article we discuss the concepts of reliability and measurement error in the context of dynamic (VAR(1)) models, and the consequences of disregarding measurement error variance in the data. For this purpose, we present a preliminary model that accounts for measurement error for constructs that are measured with a single indicator. We further discuss how this model could be used to investigate the between-person reliability of the measurements, as well as the (person-specific) within-person reliabilities and any individual differences in these reliabilities. We illustrate the consequences of assuming perfect reliability, the preliminary model, and reliabilities, using an empirical application in which we relate women's general positive affect to their positive affect concerning their romantic relationship.

Translational Abstract

An increasing number of researchers in psychology are collecting data that consist of many repeated measures (say 25 or more) for many individuals. These data can be used to study psychological processes on a within-person level. An increasingly popular way to analyze these data is autoregressive time series modeling; either by modeling the repeated measures for a single individual or for multiple individuals by using multilevel extensions of these models. However, while it is widely accepted in psychology that psychological measurements usually contain a certain amount of measurement error, the issue of measurement error is largely neglected in applied psychological (autoregressive) time series modeling. In this context we discuss the consequences of disregarding measurement error, and present a preliminary model that accounts for measurement error for constructs that are measured with one item. We discuss how this model could be used to investigate the reliability of the measurements across individuals, as well as the reliabilities per individual. We illustrate the consequences of disregarding measurement error, the preliminary model, and reliabilities, using an empirical application in which we relate women's general positive affect to their positive affect concerning their romantic relationship.

Keywords: reliability, intensive longitudinal data, autoregressive modeling, measurement error, time series analysis

In psychology there is an increased attention for modeling within-person, dynamical processes, using intensive longitudinal data. Intensive longitudinal data consist of many repeated measures, say 20 or more, typically for multiple individuals (Walls & Schafer, 2005). Such data are becoming readily available to psychological researchers due to the development of personal devices such as smartphones. As a result, psychological scientists are reaching for new modeling techniques that get the most out of these rich, complex data (Hamaker, Ceulemans, Grasman, & Tuerlinckx, 2015; Hamaker & Wichers, 2017).

A promising approach for analyzing intensive longitudinal data is multilevel vector autoregressive (VAR) modeling. These multilevel models are based on classical VAR models, which are fitted for single subjects (e.g., a person, dyad, country, and so on) for which many repeated measures were taken (Box, Jenkins, Reinsel, & Ljung, 2015; Chatfield, 2004; Hamilton, 1994). In VAR models, multiple variables are regressed on themselves and each other at one or more previous measurement occasions. In psychology,

This article was published Online First September 6, 2018.

Noémi K. Schuurman, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University; Ellen L. Hamaker, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University Utrecht.

An early rendition of this work is part of the dissertation of Noémi K. Schuurman (Schuurman, 2016) and has been presented at APS and IMPS 2017. Ellen L. Hamaker collaborated with Mplus to develop DSEM in Mplus. This study was supported by the Netherlands Organization for Scientific Research (NWO; VIDI Grant 452-10-007).

Correspondence concerning this article should be addressed to Noémi K. Schuurman, Department of Methodology and Statistics, Tilburg University, P.O. Box 80140, 3508 TC Utrecht, the Netherlands. E-mail: n.k.schuurman@uvt.nl

VAR models of Order 1 are most often used (VAR(1) models), in which the variables are regressed on themselves and each other on the nearest previous measurement occasion. VAR modeling makes it possible to investigate how current values of a variable affect values of that variable at the next measurement occasion—the autoregressive effect. The autoregressive effect reflects the amount of carry over across measurement occasions, sometimes also referred to as inertia (Suls, Green, & Hillis, 1998), of the psychological process. In addition, VAR modeling makes it possible to investigate potential reciprocal effects between different variables over time, for example: Does stress affect feelings of depression at the next measurement occasion, do feelings of depression affect stress at the next measurement occasion, or are both the case?

By extending these models to a multilevel model, it is possible to fit these models for multiple individuals simultaneously, taking into account that people may be similar to some extent, and making it easier to generalize results to a larger population than for classical $n = 1$ VAR models (Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016). At the same time, the multilevel model allows for the model parameters to vary across individuals, so that differences between individuals are taken into account. Currently, these models are applied more and more in psychology, especially in the area of affect regulation and dyadic interactions for both $n = 1$ (e.g., Cohn & Tronick, 1989; Madhyastha, Hamaker, & Gottman, 2011; Snippe et al., 2015; Stavrakakis et al., 2015; van der Krieke et al., 2015), and multilevel applications (e.g., Bringmann et al., 2013; De Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2014; Kuppens, Allen, & Sheeber, 2010; Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011; Nezlek & Gable, 2001; Rovine & Walls, 2005; Snippe et al., 2017; Suls et al., 1998; Wang, Hamaker, & Bergeman, 2012).

In most multilevel autoregressive modeling applications in psychology it is implicitly assumed that the observed variables are free of measurement error, that is, perfectly reliable. This is unlikely to be true for most psychological measurements, because most of the measured constructs are not directly observable, and measuring them is complex. In line with this notion, many cross-sectional psychological studies take the reliability of measurements into account by measuring a single construct with multiple exchangeable (or parallel) items, and modeling this measurement structure with latent variable models, such as factor or IRT models. This approach is, on rare occasions, also taken in a multilevel autoregressive modeling context with dynamic (multilevel) factor models (Lodewyckx et al., 2011; Molenaar, 1985; Oravecz & Tuerlinckx, 2011; Song & Ferrer, 2012). However, single-item measures or single variables often play a central role in longitudinal studies (Lucas & Donnellan, 2012). Using multiple items for each construct of interest severely increases the burden on the participants, who may have to fill out these items on a daily or even hourly basis. Moreover, such latent variable models may be considered inappropriate theoretically, for example, because the items cannot be considered exchangeable, such as when each item is considered to play a unique role within a network of items (cf., Borsboom & Cramer, 2013; Borsboom, Mellenbergh, & van Heerden, 2003; Schmittmann et al., 2013), or because the same measurement structure does not apply to each individual in the multilevel model (e.g., there are factorial measurement invariance issues). Regardless of these considerations, however, ignoring the reliability of measurements remains problematic because it leads to substantially different estimated regression effects. For example, it has been shown that for $n = 1$ univariate AR(1) models the autoregressive effects will be estimated much closer to zero compared with the true autoregressive effect if there is measurement error in the data, how much depending on the amount of measurement error variance in the data (Schuurman, Houtveen, & Hamaker, 2015; Staudenmayer & Buonaccorsi, 2005).

The aim of the current work is to emphasize the importance of accounting for potential measurement error variance in dynamic VAR models. For this purpose, we discuss the concepts of reliability and measurement error in the context of (multilevel) VAR(1) models, and we show that disregarding measurement error can severely distort the cross-lagged relations between variables, as well as the autoregressive effects. For this purpose, we discuss an extension of the multilevel VAR model that accounts for measurement error for variables that are measured with a single indicator.[1] By allowing the parameters of this model to vary across persons, it is possible to derive the reliability for our the measurements for each individual, as well as the consequences of assuming perfect reliability.

This also allows us to look at reliability in a new light: While it has occasionally been acknowledged that the reliability of measurements may be different for each person (e.g., as early as 1968 by Lord & Novick, 1968), this is generally disregarded in psychological studies. A notable exception is a recent study by Hu et al. (2016), who create parallel tests to estimate person-specific reliabilities for a replicated time series design. However, by using an extended multilevel VAR approach that accounts for measurement error, we can estimate person-specific reliabilities for the repeated measurements of each individual, as well as estimates for the reliability for measuring between-person differences. Furthermore, we correct for imperfectly reliable measurements during the analysis, potentially even for single indicator measures.

In the remainder, we first discuss reliability as defined in classical test theory and reliability in the context of longitudinal research. After that, we introduce an extended multilevel autoregressive model that incorporates measurement error, we discuss reliability in the context of this model, and use this to show the consequences of disregarding measurement error in a multilevel and multivariate context. After this, we discuss a preliminary implementation of this model in Mplus, with a small simulation study. We fit this model to an empirical data set on the effects of women's general positive affect, and that concerning their romantic relationship, to illustrate the interpretation of the model, accompanying reliability estimates, and consequences of disregarding measurement error. We end with a discussion in which we consider the concept of reliability and measurement error further in light of our findings, and offer suggestions for future work in this area.

Measurement Error and Reliability

Reliability concerns the consistency of measurements. That is, in the hypothetical situation that we would replicate our measure-

[1] This is essentially a dynamic factor model where each indicator loads on a single factor, and is identified by merit of the dynamic process among the factors.

ments of a certain quality of interest while the quality of interest has not changed, perfectly reliable measurements would give the same result for each replication (Cronbach, 1947). In contrast, measurements that are unreliable can result in different scores for each replication. The unreliable part of a score is considered to be due to *random measurement errors*,[2] while the reliable part is what is consistent across replications, and includes the true value of the actual quality of interest and any consistent errors in the measurements (e.g., consistently measuring a person as two pounds heavier than he or she really is). As such, reliable measurements are not necessarily valid, but obtaining reliable measurements is a precondition for obtaining valid measurements.

Although we are interested in reliability in the context of the autoregressive modeling of within-person differences here, the roots of reliability lie in cross-sectional studies of between-person differences. Therefore, we start by discussing the definition of reliability from classical test theory in the context of cross-sectional studies. After that, we discuss reliability in the context of longitudinal (autoregressive modeling) studies of intraindividual differences.

Reliability in Classical Test Theory and Cross-Sectional Studies

Reliability was first defined in the context of classical test theory. As stated previously, reliability concerns the consistency of measurements across replications. A key issue is therefore how to define these "replications." Lazarsfeld (1959; as cited by Lord & Novick, 1968, p. 29) describes the following thought experiment to illustrate what is meant with replications:

> Suppose we ask an individual, Mr. Brown, repeatedly whether he is in favor of the United Nations; suppose further that after each question we "wash his brains" and ask him the same question again. Because Mr. Brown is not certain as to how he feels about the United Nations, he will sometimes give a favorable and sometimes an unfavorable answer. Having gone through this procedure many times, we then compute the proportion of times Mr. Brown was in favor of the United Nations [. . .].

In this example, the proportion of times Mr. Brown was in favor of the United Nations is defined as his "true score," the reliable part of the replicated measurements. That is, in classical test theory the true score θ_i of a specific person i is defined as the expected score over an infinite number of independent replications, such that $\theta_i = \underset{r}{E}[y_i]$, where y_{ri} is the observed score for a certain variable for a specific person i at replication r. The deviations around the true score across the replications are defined as *measurement errors* ϵ_{ri}, such that $y_{ri} = \theta_i + \epsilon_{ri}$.

Although the true score and measurement errors in classical test theory are defined on the level of a specific individual, reliability is defined for the measurements of a *specific population of individuals* (cf., Lord & Novick, 1968; Mellenbergh, 1996). The focus lies on the distribution of the observed and true scores across all individuals in that population. The expected value of the observed scores across the individuals in the population is equal to the expected value of the true scores of these individuals, that is, $\underset{i}{E}[y] = \underset{i}{E}[\theta]$. The variance of the observed scores $V(y)$ is the sum of the variance of the true scores τ^2 and the measurement error variance σ^2_ϵ, that is, $V(y) = \tau^2 + \sigma^2_\epsilon$. The reliability $rel(y)$ of the set of measurements is then defined as the proportion of variance in the observed scores that is due to the variance in the true scores, $rel(y) = \tau^2/V(y) = 1 - \sigma^2_\epsilon/V(y)$. As such, the maximum reliability is equal to 1, indicating that variable y is measured without error in the population, and the minimum reliability is 0, indicating that the measurements consist of only measurement error in the population.

In practice, the true score(s) for any individual are of course unknown, such that in order to be able to take the reliability of our measurements into account, the true scores (or inversely, the measurement errors) have to be estimated from the data. The main idea behind most approaches to estimate or correct for measurement error is the same: Find out what part of the observed scores remains constant across replications (i.e., what part is due to the true trait scores), and what part fluctuates randomly across replications (i.e., what part is due to measurement error). To do this, it is necessary to obtain replicate measurements for the construct of interest, specifically, replications of the same kind as described in the thought experiment cited by Lord and Novick (1968, p. 29; citing Lazarsfeld, 1959). Obtaining such replications is not an easy feat, and the different reliability measures available are in general based on different ideas on how to obtain these replications (Cronbach, 1947).

The most well-known methods for estimating reliability are parallel-test reliability methods, internal consistency methods, and test–retest reliability methods. *Parallel-test reliability* is based on the construction and administration of two or more "parallel tests" to each individual, that is, tests that are constructed to be equivalent (cf., Borsboom, 2003; Cronbach, 1947, 1990; Lord & Novick, 1968). *Internal consistency reliability* is used for composite scores, and circumvents the construction of parallel tests by treating the components from which the scores are composed as the replications of the construct under scrutiny (cf., Borsboom, 2003; Cronbach, 1947, 1990; Lord & Novick, 1968). For instance, for a test score that consists of multiple items, each item may be considered a replicate measurement of the construct, such that the correlations between these items may be used as an indicator of reliability (e.g., as is done for Cronbach's alpha, and latent variable models such as factor models, IRT models Mellenbergh, 1996). Finally, *test–retest reliability* is based on the repeated administration of the same test, that is, each individual fills out the same test multiple times (cf., Borsboom, 2003; Cronbach, 1947; Lord & Novick, 1968).

Reliability for Longitudinal Studies

The reliability measures we discussed above were developed in the context of cross-sectional studies, which investigate interindividual differences: differences between trait scores of individuals (the true scores in that context), which are stable across some relevant length of time. Given that in cross-sectional studies the interest is typically in studying stable differences between subjects, within-subject fluctuations are treated as the measurement error in these reliability estimates, while the stable between person differences are considered to be the reliable part of the data.

However, in longitudinal studies we are usually explicitly interested in within subject fluctuations over time (intraindividual

[2] Throughout the text we will refer to "random measurement errors" as measurement error.

differences in state scores, rather than interindividual differences in trait scores). Hence, it is insufficient for longitudinal studies to only separate the variance of the observed scores into variance due to between-person differences and variance due to within-person differences. For longitudinal studies one also needs to establish what part of the within-person differences is due to measurement error, and what part of the within-person differences is due to variation in true scores over time—that is, due to systematic within-person dynamic processes.

A promising way to do this for intensive longitudinal data, that was recently introduced by Hu et al. (2016), is to create parallel tests, or use an internal consistency approach for the repeated measurements of single subjects, to obtain reliabilities per person (assuming that the measurements are independent). However, creating parallel tests or indicators is difficult (see, e.g., Hu et al., 2016), and not a viable option if some or all variables are measured with a single indicator. Furthermore, we may wish to not only obtain an estimate of the reliability of our repeated measures, but to also correct our parameter estimates for this reliability. That is, we need to not only to estimate, but also account for the reliability of our repeated measures within our longitudinal model.

The additional step of separating the measurement error variance from within-person variance due to a systematic dynamic process, and doing this during the analyses, has gotten attention in the literature on autoregressive modeling in the context of *panel data*. Panel data consist of a few repeated measures (say two to five) for many participants, and are usually analyzed with structural equation models. In some panel data models, measurement error is separated from systematic within-person differences that result from an autoregressive process by using multiple indicators in a factor structure (e.g., Edmondson et al., 2013). The trait-state-error (TSE) model suggested by Kenny and Zautra (1995), also accounts for the reliability of single indicators, and can be seen as an extension of the Quasi-Simplex model (Jöreskog, 1970). In the TSE model, systematic between-person differences ("traits"), are separated from systematic within-person differences result from an autoregressive process ("states"), and measurement error. A downside of the TSE model is—as is the case for many panel data models—that the model ignores potential individual differences in the dynamic processes, while it is unlikely that dynamic processes are the same for each person (Hu et al., 2016; Kenny & Zautra, 1995; Molenaar, 2004). Allowing for such individual differences generally requires intensive longitudinal data, rather than panel data.

An intensive longitudinal data alternative to the TSE model is $n = 1$ (autoregressive) time series modeling, in which models are fitted for each person separately based on such intensive longitudinal data. The advantage of this approach is that the model can be tailored to each person, so that individual differences in dynamics are taken into account, and that there is no between-person variance to account for. Although the time series models used in psychological practice usually do not take measurement error into account, it is possible to do so, even for single items (cf., Schuurman et al., 2015). Downsides of the $n = 1$ approach are, however, that one needs relatively many repeated measures per person to fit these models, and that it is hard to generalize the results for specific individual to a larger population.

However, it is possible to extend the $n = 1$ models, including those that take measurement error into account, to a multilevel setting (or similarly, extend a multilevel VAR model so it incorporates measurement error, or extend the TSE model to incorporate random effects for all parameters of the within-person process). A multilevel approach allows for fitting the model for all individuals at once, and evaluating to what extent the within-person process differ across persons, which makes it easier to generalize results to the population of individuals. Furthermore, by allowing the dynamic process to be different for each person, it is possible to take into account that some persons' responses can be measured more reliably than others.

In the following section we will introduce such a multilevel model, and we will use the model equations to derive various reliabilities, including person-specific reliabilities. We will use these to evaluate what the consequences would be if we fit a VAR(1) model that disregards measurement error to data that contains measurement error. Later on, we will present a preliminary implementation in Mplus to fit this model, and obtain person-specific reliability estimates. We provide a small preliminary simulation study for the performance of the model. Finally, we apply it to an empirical example, illustrating the interpretation of the model and the reliability estimates we obtain in practice.

Measurement Errors in the Multilevel VAR(1) Model

In the following, we will introduce the extended multilevel VAR(1) model that incorporates measurement error, which we refer to as the Measurement Error VAR(1) (MEVAR(1) model).[3] The multilevel MEVAR(1) model consists of two levels. At Level 1, the within-person process for each individual is specified, and at Level 2, the between-person differences in this process across individuals are specified. We will discuss both levels in order, starting with Level 1. We focus on the specification of a bivariate model where the dependent variable is only regressed on its nearest previous measurement occasion (the VAR(1) model), with a measurement model for a single indicator. This model can however be extended further to models that include more dependent variables and predictors, or to use a measurement model with multiple indicators (which would result in a multilevel dynamic factor model Molenaar, 1985; Song & Ferrer, 2012, with the addition of random covariance matrices). A graphical representation of the bivariate MEVAR(1) model is presented in Figure 1.

Level 1 of the MEVAR(1) Model

The Level 1 model is specified with two equations, the measurement equation, and the transition equation (using a state space model representation, cf., Harvey, 1989; Kalman, 1960; Kim & Nelson, 1999). In the measurement equation presented in Equation 1, the observed scores for variable y_1 and y_2 for person i at measurement occasion t are contained in 2×1 vector $\boldsymbol{y}_{ti}$. These observed scores are separated into three 2×1 vectors, $\boldsymbol{\mu}_i$, $\tilde{\boldsymbol{y}}_{ti}$, and $\boldsymbol{\epsilon}_{ti}$, that is,

$$\begin{bmatrix} y_{1ti} \\ y_{2ti} \end{bmatrix} = \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix} + \begin{bmatrix} \tilde{y}_{1ti} \\ \tilde{y}_{2ti} \end{bmatrix} + \begin{bmatrix} \epsilon_{1ti} \\ \epsilon_{2ti} \end{bmatrix} \tag{1}$$

[3] Schuurman et al. (2015) referred to the $n = 1$, univariate version of this model as the AR(1) + WN (AR(1) + White Noise) model. Mplus refers to that univariate model as the MEAR(1) model. We chose to align with the names used in Mplus here.

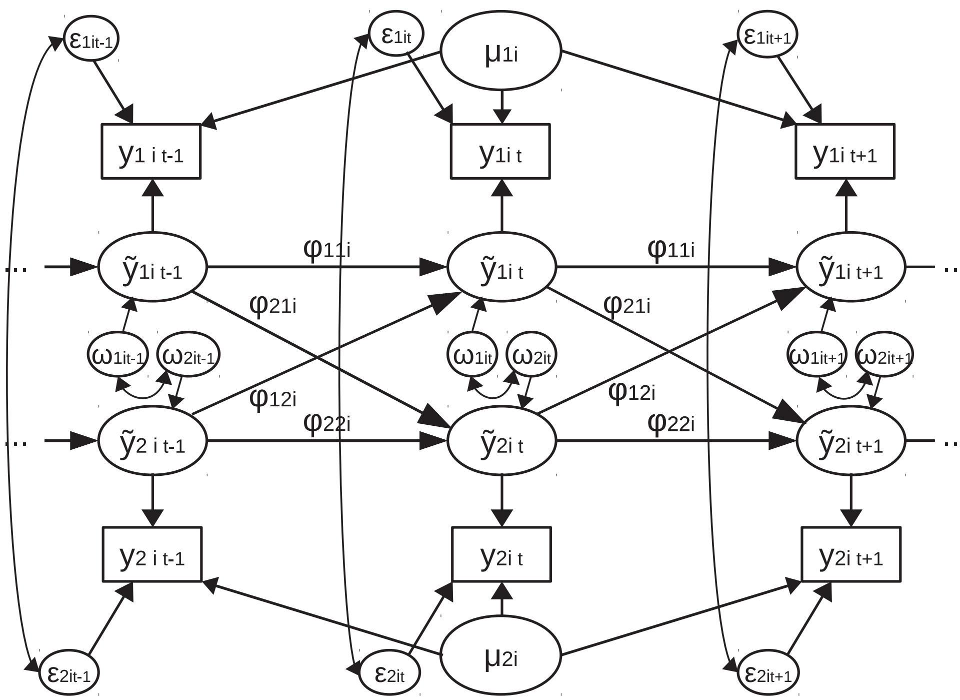

Figure 1. A graphical representation of the multilevel MEVAR(1) model, a VAR(1) model which takes measurement error into account. Measurement error is captured in the terms ϵ_i.

$$\begin{bmatrix}\epsilon_{1ti}\\ \epsilon_{2ti}\end{bmatrix} \sim MvN\left\{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_{\epsilon 11i} & \\ \sigma_{\epsilon 12i} & \sigma^2_{\epsilon 22i}\end{bmatrix}\right\}. \tag{2}$$

The vector $\boldsymbol{\mu}_i$ contains the person-specific means μ_{1i} and μ_{2i} for each variable, for individual i. These means are stable across the repeated measurements for each individual, and therefore reflect a baseline, or "trait" part, for each persons' scores. For example, some persons are on average more extroverted than others. The differences between the trait scores μ_{1i} and μ_{2i} across persons, reflect *systematic, trait-like, between-person differences*.

The vectors $\tilde{\mathbf{y}}_{ti}$ and $\boldsymbol{\epsilon}_{ti}$ in Equation 1 together reflect the within-person fluctuations around the person-specific trait scores: While some persons are on average more extroverted than others, a specific person i may display more or less extroverted behavior across different occasions t. The terms ϵ_{1ti} and ϵ_{2ti} capture *measurement error* for person i at occasion t, and, as shown in Equation 2, are assumed to be serially uncorrelated, and multivariate normally distributed with means equal to zero, and 2 × 2 covariance matrix $\boldsymbol{\Sigma}_{\epsilon i}$.

The terms $\tilde{y}_{1ti}$ and $\tilde{y}_{2ti}$ reflect the deviations from the mean of each variable for person i at occasion t that are due to a *systematic dynamic (autoregressive) process*. The autoregressive process for $\tilde{y}_{1ti}$ and $\tilde{y}_{2ti}$ is further specified in the transition equation as

$$\begin{bmatrix}\tilde{y}_{1ti}\\ \tilde{y}_{2ti}\end{bmatrix} = \begin{bmatrix}\phi_{11i} & \phi_{12,i}\\ \phi_{21,i} & \phi_{22i}\end{bmatrix}\begin{bmatrix}\tilde{y}_{1t-1i}\\ \tilde{y}_{2t-1i}\end{bmatrix} + \begin{bmatrix}\omega_{1ti}\\ \omega_{2ti}\end{bmatrix} \tag{3}$$

$$\begin{bmatrix}\omega_{1ti}\\ \omega_{2ti}\end{bmatrix} \sim MvN\left\{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_{\omega 11i} & \\ \sigma_{\omega 12i} & \sigma^2_{\omega 22i}\end{bmatrix}\right\}. \tag{4}$$

That is, the variables $\tilde{y}_{1ti}$ and $\tilde{y}_{2ti}$ in Equation 3 depend on themselves and each other at the previous measurement occasion, such that they constitute a VAR(1) process. The regression coefficients of the VAR(1) process are gathered in 2 × 2 matrix Φ_i. The relationship between the variables and themselves at the previous measurement occasion for person i is reflected in the autoregressive coefficients ϕ_{11i} for variable $\tilde{y}_{1i}$, and ϕ_{22i} for variable $\tilde{y}_{2i}$. Positive autoregressive coefficients indicate that the score of the current measurement occasion will be similar to that of the previous measurement occasion—the larger the autoregressive coefficient the more similar the scores will be. As such, autoregressive parameters reflect carry-over in a process, which has also been referred to as inertia (Kuppens et al., 2010; Suls et al., 1998). On the other hand, an autoregressive coefficient of zero or near zero indicates that the previous value of a variable does not, or hardly carries over to the next occasion. A negative autoregression coefficient indicates that a relatively high score at the previous measurement occasion is usually followed by a relatively low score at the current measurement occasion, and vice versa. Negative autoregressive effects seem relatively rare in psychological research, but could be expected for processes that concern intake of substances, such as smoking, drinking, and eating (e.g., Rovine & Walls, 2005).

The effects of $\tilde{y}_{1t-1i}$ and $\tilde{y}_{2t-1i}$ on each other at the next occasion is reflected in the cross-lagged coefficients ϕ_{21i}, and ϕ_{12i}. That is, if we study motivation and job satisfaction and their effects on each other over time, the effect of satisfaction on motivation at the next occasion for person i is reflected in ϕ_{msi}, and the effect of motivation on satisfaction in ϕ_{smi}.

The residuals for the transition equation, ω_{1ti} and ω_{2ti}, reflect perturbations of the dynamic process for person i at occasion t, and are referred to as innovations. As can be seen from Equation 4, we assume here that these innovations are normally distributed with means of zero, and covariance matrix $\boldsymbol{\Sigma}_{\omega i}$.

Innovations versus measurement error. The innovations ω_{1ti} and ω_{2ti} and the terms that capture the measurement error ϵ_{1ti} and ϵ_{2ti} are substantially different: The innovations ω_{1ti} and ω_{2ti} represent any unmeasured effects on the observed variables that are carried over from one measurement occasion to the next through the autoregressive and cross-lagged effects. This is visible from Figure 1, where the effect of ω_{t-1i} is passed along through $\tilde{y}_{t-1i}$, and to $\tilde{y}_{ti}$ via the autoregressive effect ϕ_i. Because the innovations affect the system across multiple occasions, the innovations are sometimes also referred to as dynamic errors. Measurement error, on the other hand, is specific to one occasion. Consider, for instance, the classical examples of a measurement error, where someone accidentally checks the wrong answer on a questionnaire, or presses the wrong button during a computer task. The effects of these errors do not carry over to the next measurement occasion, but are specific to that moment. Such occasion-specific effects are not captured by the innovations, but should be modeled separately. By including ϵ_{1ti} and ϵ_{2ti} in the measurement equation, the measurement error is separated from the autoregressive processes for $\tilde{y}_{ti}$, as can be seen from Figure 1, where the measurement error is *not* passed along through arrows to future measurement occasions. As such, the measurement error can be distinguished from the dynamic errors ω_{1ti} and ω_{2ti}. In traditional (multilevel) VAR(1) models the innovations are incorporated in the model, whereas the terms ϵ_{1ti} and ϵ_{2ti} are not, and as such potential measurement error is disregarded.

Some additional remarks about the terms ϵ_{1ti} and ϵ_{2ti} are in place in the current context: Although these terms will capture measurement error present in the data, they may also capture other within-person fluctuations that are specific to occasion t. In fact, they will capture *anything* that affects the variables at one occasion, of which the effect is dissipated before the next measurement occasion. For example, if someone fills out an hourly questionnaire on mood while eating a tasty snack, this may influence that person's mood at that occasion, but this effect on mood may have dissipated before the next measurement occasion an hour later. Then, this effect would end up in the terms ϵ_{ti}, even though it does not reflect an actual error of measurement but a true, occasion-specific, fluctuation in mood. Hence, while we refer to the terms ϵ_{ti} as "measurement error," they actually represent a mix of occasion-specific fluctuations of the true score and measurement errors. Of course, regardless of whether the occasion-specific fluctuations are true fluctuations or measurement error, it is important to take them into account. Furthermore, for these reasons, we also allow for correlations between the measurement error terms of different variables in the model. Given that the measurement error terms may contain "true" occasion-specific fluctuations in the observed scores, these fluctuations may be the result of shared unobserved effects, resulting in correlations between the measurement error terms of different variables. It is important to take these potential correlations into account in the model. Such correlations indicate that the measurement error terms capture more systematic than just "pure" measurement error. We will return to these issues in the Discussion section.

Separating measurement error from innovations. A second important point is that the innovations and the occasion-specific fluctuations are distinguishable in the model *only* by merit of the autoregressive and cross-lagged effects. If there is no time dependency in the data, and the autoregressive and cross-lagged effects are equal to zero for a variable in the model, the measurement error and innovations cannot be distinguished from each other for that variable. Essentially there are no dynamic errors in that case, there is only measurement error—and as a result, the model will no longer be identified. In any given sample, however, it is extremely unlikely that these will be exactly equal to zero, and consistent with this, Schuurman et al. (2015) found for the $n = 1$ case that a Bayesian MEAR(1) model still provided reasonable estimates of the model parameters if the true autoregressive effect was zero. For data where the dynamic effects are very close to zero however, the model may be more difficult to empirically identify and estimate, and hence require more data to get more accurate and precise estimates, than when the dynamic effects are not close to zero (Schuurman et al., 2015). Note however that the results of the simulation study by Schuurman, Houtveen, and Hamaker (2015) are based on $n = 1$ models. An elaborate simulation study for the multilevel case should show how this pans out exactly for the group level parameters of the multilevel model. We discuss a small, preliminary, simulation study on the performance of the multilevel MEVAR(1) model in a later section.

Level 2 of the MEVAR(1) Model

At Level 2 of the multilevel model the individual differences in the dynamic processes of the individuals are modeled. It seems natural that the means, autoregressive, and cross-lagged regressive coefficients differ from person to person, and in most multilevel VAR(1) applications, this is accounted for (e.g., Bringmann et al., 2013; De Haan-Rietdijk et al., 2014; Jongerling, Laurenceau, & Hamaker, 2015; Lodewyckx et al., 2011; Schuurman et al., 2016). As such, we allow the means, autoregressive, and cross-lagged regressive coefficients to vary across persons. We assume that each individual's parameter comes from a common population, with a common probability distribution. Characteristics from this distribution, such as its mean and variance, can be used to make inferences about the between-person differences in the within-person dynamics of the individuals. Specifically, we assume that the means μ_i and the regression parameters ϕ_{11i}, ϕ_{22i}, ϕ_{12i}, and ϕ_{21i} are multivariate normally distributed, with means $\gamma_{\mu1}$, $\gamma_{\mu2}$, $\gamma_{\phi11}$, $\gamma_{\phi12}$, $\gamma_{\phi21}$, and $\gamma_{\phi22}$, and a 6×6 covariance matrix Ψ. The means γ, also referred to as fixed effects, reflect population averages for the individual means (the trait scores), autoregressive, and cross-lagged effects. The personal deviations from the fixed effects are also referred to as random effects, and their variances are included in Ψ. The covariances in Ψ reflect the associations between the person-specific parameters across persons.

In addition to the trait scores and regression effects contained in μ_i and Φ_i respectively, it is also important to consider that the variability of the measurement error and innovations may differ across individuals: The variances and covariances of the innovations and measurement error may differ from person to person (cf., Lord & Novick, 1968). The innovation variance and the measurement error variance may indicate, for example, sensitivity or reactivity to external (unobserved) events, with more sensitive or reactive people having larger variances. Further, each individual may of course experience different external events, also resulting in different innovation or measurement error variances and covariances for different persons. As such, it seems likely that the variances of the innovations and measurement error, and the covariances or correlations between the innovations or measurement error, vary in magnitude across persons, which should be accounted for in the model. Yet (co)variances are usually kept fixed across persons in the multilevel literature, including in multilevel time series applications in psychology (for an exception, see Jongerling et al., 2015). The decision to keep these (co)variance matrices fixed seems be more practically than theoretically motivated, as the software options for doing this are still limited. In our implementation of the model in Mplus v8 we will allow for random covariance matrices for the innovations and for the measurement (details in a later section).[4] As a result, we will obtain estimates for each person's innovation covariance matrix and measurement error covariance matrix. Next to this, we obtain estimates for the average innovation and measurement error variances and covariances in the group of persons (the fixed effects), that is, $\gamma_{\sigma^2_{\omega 11}}$, $\gamma_{\sigma^2_{\omega 22}}$, $\gamma_{\sigma_{\omega 12}}$, $\gamma_{\sigma^2_{\epsilon 11}}$, $\gamma_{\sigma^2_{\epsilon 22}}$, and $\gamma_{\sigma_{\epsilon 12}}$. Furthermore, we obtain estimates for the variance for each of the innovation and measurement error variances and covariances, respectively $\psi^2_{\sigma^2_{\omega 11}}$, $\psi^2_{\sigma^2_{\omega 22}}$, $\psi^2_{\sigma_{\omega 12}}$, $\psi^2_{\sigma^2_{\epsilon 11}}$, $\psi^2_{\sigma^2_{\epsilon 22}}$, and $\psi^2_{\sigma_{\epsilon 12}}$. Such a variance, for instance the variance of the measurement error variances of variable y_1, $\psi^2_{\sigma^2_{\epsilon 11}}$, indicates how much variability there is in the measurement error variances for variable y_1 across persons. The variance of the covariances between the innovations of variable y_1 and y_2, $\psi^2_{\sigma_{\omega 12}}$, for example, indicates how much variance there is in these covariances across persons.

Deriving Reliabilities From the Multilevel MEVAR(1) Model

The model parameters of the MEVAR(1) model are corrected for the reliability of the data, by modeling the measurement error using the terms ϵ_i. We can also use the model parameters to calculate the between-person reliability and the within-person reliabilities for each person. This information can in turn be used to determine the consequences of disregarding measurement error in the VAR(1) model, as will be discussed in more detail in the next section. Below, we will first discuss the composition of the variance of the multilevel MEVAR(1) model: What part of the data is due to between-person variance, what part is due to within-person variation over time, and for the within-person variation—how much is the result of the dynamic process, and how much may be due to measurement error. After this, we discuss how to derive the between-person reliability and the within-person reliabilities for each person based on this.

The reliability for a specific variable can be calculated as the proportion of true score variance to the total variance for that variable, or equivalently, as one minus the proportion of measurement error variance to the total variance. The total variance V(y) for each variable in the MEVAR(1) model, taken over all participants' repeated measures, can be decomposed into three parts: The between-person variance or trait score variance ψ^2_μ (i.e., the variance of the person-specific means), the expected value of the person-specific variances for $\tilde{y}$ (the VAR(1) process captured in the transition equation in Equation 3) $\underset{i}{E}[\tau^2_{\tilde{y}}]$, and the expected value of the person-specific measurement error variances $\gamma_{\sigma^2_\epsilon}$. Hence, we have

$$V(y) = \psi^2_\mu + \underset{i}{E}\left[\tau^2_{\tilde{y}}\right] + \gamma_{\sigma^2_\epsilon}. \tag{5}$$

The terms ψ^2_μ and $\gamma_{\sigma^2_\epsilon}$ are both parameters of the MEVAR(1) model specified in the previous section. The term $\underset{i}{E}[\tau^2_{\tilde{y}}]$ is the expectation taken over all the person-specific variances for variable $\tilde{y}$, where the person-specific variance for this variable for person i is equal to $\tau^2_{\tilde{y}i}$. This person-specific variance $\tau^2_{\tilde{y}i}$ in turn is equal to the diagonal element for that variable of the covariance matrix $\boldsymbol{T}_{\tilde{y}i}$ for person i. The covariance matrix $\boldsymbol{T}_{\tilde{y}i}$ contains the variances and covariances of $\tilde{y}$ for person i, that is, the variances and covariances of the VAR(1) process specified in the transition equation (Equation 3) for person i. This covariance matrix is equal to

$$\boldsymbol{T}_{\tilde{y}i} = mat((I - \Phi_i \otimes \Phi_i)^{-1} vec(\Sigma_{\omega i})), \tag{6}$$

where 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, p. 27).

Based on the variance decomposition in Equation 5, we can calculate various reliabilities for our measurements of variable y. For example, we can determine an overall reliability of our measurements y of both within-person or between-person differences, by calculating $rel(y) = (\psi^2_\mu + \underset{i}{E}[\tau^2_{\tilde{y}}])/V(y)$. However, this reliability blends the variances of two kinds of true scores—that of the trait-scores (ψ^2_μ) and that of the true within-person fluctuations ($\underset{i}{E}[\tau^2_{\tilde{y}}]$). As such, the observed scores might be highly reliable as a measurement of the trait scores, and hardly reliable as a measurement of the true within person fluctuations, or vice versa, but this would not be clear from our calculated reliability—instead it would reflect a mix of these two different reliabilities. As a result, it is not very clear for what kind of inferences the observed scores are suitable or unsuitable based on this calculated reliability. It also does not take into account that the reliability of the observed scores with regard to measuring true within person fluctuations may be different from person to person. It seems therefore more useful to determine the reliability for the trait scores and the within-person fluctuations deviations around these scores separately.

In the following, we will first discuss the reliability for the trait-scores, which indicates the reliability of the measures for making inferences about stable differences between people, akin to the traditional reliability estimates developed for cross-sectional studies. After that, we will discuss reliabilities for the within-person fluctuations: both person-specific reliabilities, and the average reliability for the group as a whole. These indicate the reliability of the measures for making inferences about intraindividual differences, which can be different for each individual. In a following section we will use these reliabilities to determine the consequences of disregarding measurement error.

Reliability for Systematic Between-Person Differences

For the between-person reliability estimate, the true scores we are interested in measuring are the trait scores, that is, the person-

[4] Note, however, that for sake of simplicity, here we do not allow the random variances and covariance to covary among themselves, or with the random trait scores or regression coefficients.

specific means. The variance of the trait scores is captured in parameter ψ^2_μ in the MEVAR(1) model, and is one of the three terms of the total variance presented in Equation 5. We can obtain the between-person reliability of variable y using

$$rel_b(y) = \frac{\psi^2_\mu}{V(y)}. \quad (7)$$

Note that the denominator in this equation contains the total variance, which next to the variance of the trait scores ψ^2_μ, is comprised of both the expectation of the measurement error variance across persons, $\gamma_{\sigma^2_\epsilon}$, as well as the expectation of the variances of the within-person fluctuations that are due to the dynamic process, $\underset{i}{E}[\tau^2_{\tilde{y}}]$. Therefore, the between-person reliability will be low if on average the deviations from the trait scores over time are relatively large. That is, akin to traditional reliability estimates, all within-person fluctuations in the observed scores are not of direct interest here, and are considered noise for this reliability. This between-person reliability indicates the reliability of the measurements one would get on average, if one was interested in measuring the trait scores, and took one measurement for each participant, just like a cross-sectional study.

Reliabilities for Within-Person Fluctuations

Given that the dynamics for each individual may vary substantially from person to person, this may also be the case for the reliability of their scores. Based on the person-specific variances discussed in Equation 6 we can determine the reliability for the within-person fluctuations over time for each individual separately. For the MEVAR(1) model, the total variance of variable y_i for a specific person i equals

$$v(y_i) = \tau^2_{\tilde{y}i} + \sigma^2_{\epsilon i}. \quad (8)$$

Note that for any specific individual, there is no between-person variance, such that the term ψ^2_μ is excluded. Then, the reliability for the observed scores of a specific person i can be determined with

$$rel_w(y_i) = \frac{\tau^2_{\tilde{y}i}}{v(y_i)}. \quad (9)$$

Note that, as can be seen from this equation, that differences between the reliabilities of different individuals can arise when their autoregressive or cross-lagged associations differ or when the variability of their innovations differs, because then the terms $\tau^2_{\tilde{y}i}$ differ; or when the variability of their measurement error differs, because then the terms $\sigma^2_{\epsilon i}$ differ.

It can also be useful to determine what within-person reliability we can expect on average for a member of our group of interest. To do this, we can use the information on the reliabilities for each individual we observed. For example, we can calculate the average person-specific reliability over the individuals, and the accompanying variance of the person-specific reliabilities, to get an impression of the range of person-specific reliabilities in the group. Furthermore, if Level 2 (person level) predictors are available, such as gender or personality traits, these may also be used to obtain predictions for the person-specific reliabilities (e.g., we could use these Level 2 predictors for the person-specific parameters, and based on the predicted person-specific parameters, determine the associated person-specific reliabilities).

Consequences of Disregarding Measurement Error in VAR(1) Modeling

Given that the model parameters and the reliability of each variable can be different for each person at Level 1 of the multilevel model, the consequences of disregarding measurement error (or other occasion-specific fluctuations) in the data can also be different for the estimated parameters each person. In the following, we will therefore discuss the effects of disregarding measurement error for the person-specific parameter estimates of individuals at Level 1 of the model.[5]

For a univariate AR(1) model, it is known that disregarding measurement error in the data results in AR parameter estimates that are pulled toward zero, so that the autoregressive effect ($|\phi\!\!\!/_i|$) for the AR model will be underestimated compared with the autoregressive effect ($|\phi_i|$) of the true AR + WN model (Schuurman et al., 2015; Staudenmayer & Buonaccorsi, 2005). How much the autoregressive effect will be underestimated in such a univariate AR(1) model depends directly on the person-specific reliability $rel_w(y_i)$ that was defined in Equation 9, that is:

$$\phi\!\!\!/_i = rel_w(y_i)\phi_i, \quad (10)$$

where $\phi\!\!\!/_i$ is the expected estimated autoregressive effect in the AR model and ϕ_i is the true autoregressive effect from the MEAR(1) model, such that the discrepancy between $\phi\!\!\!/_i$ and the true ϕ_i is equal to 1 minus the person-specific reliability.[6]

For a VAR(1) model, the effects of disregarding measurement error are more complicated. The difference between the true Φ_i and the estimated matrix of autoregressive and cross-lagged effects in $\hat{\Phi}_i$ that result when measurement error is disregarded depends on the person-specific reliability matrix $\boldsymbol{rel_w}(\boldsymbol{y}_i)$ (pp. 108–109, Buonaccorsi, 2010; Gleser, 1992), which is equal to

$$\boldsymbol{rel_w}(\boldsymbol{y}_i) = \boldsymbol{T}_{\tilde{y}i}(\boldsymbol{\Sigma}_{\epsilon i} + \boldsymbol{T}_{\tilde{y}i})^{-1}. \quad (11)$$

Each element in this reliability matrix is a rather complex function of the covariances and variances of the true scores and measurement error. For instance, in the bivariate case this results in

[5] Note that because the fixed effects in the multilevel model equal the expectation of the person-specific parameters, the effect of disregarding measurement error on the fixed effects depends on the effects of disregarding measurement error for each person-specific parameter. Hence, the discrepancy in the fixed effects can be determined by taking the expectation of the discrepancies for the person-specific parameters, and thus would be a blend of all the discrepancies in the Level 1 parameters.

[6] Note that $\phi\!\!\!/_i$ is also equal to the first order autocorrelation of both models. The first order autocorrelations of the AR(1) model and the MEAR(1) model are equal, but are not equal for the zeroth or later orders than order 1, with the MEAR(1) model having slower decaying autocorrelations.

$$\boldsymbol{rel_w}(\boldsymbol{y}_i) = \begin{bmatrix} \frac{\tau^2_{\tilde{y}11i}(\tau^2_{\tilde{y}22i} + \sigma^2_{\epsilon 22i}) - \tau_{12i}(\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})}{(\tau^2_{\tilde{y}11i} + \sigma^2_{\epsilon 11i})(\tau^2_{\tilde{y}22i} + \sigma^2_{\epsilon 22i}) - (\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})^2} & \frac{\tau_{\tilde{y}12i}(\tau^2_{\tilde{y}22i} + \sigma^2_{\epsilon 22i}) - \tau^2_{\tilde{y}22i}(\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})}{(\tau^2_{\tilde{y}11i} + \sigma^2_{\epsilon 11i})(\tau^2_{\tilde{y}22i} + \sigma^2_{\epsilon 22i}) - (\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})^2} \\ \frac{\tau_{\tilde{y}12i}(\tau^2_{\tilde{y}11i} + \sigma^2_{\epsilon 11i}) - \tau^2_{\tilde{y}11i}(\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})}{(\tau^2_{\tilde{y}11i} + \sigma^2_{\epsilon 11i})(\tau^2_{\tilde{y}22i} + \sigma^2_{\epsilon 22i}) - (\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})^2} & \frac{\tau^2_{\tilde{y}22i}(\tau^2_{\tilde{y}11i} + \sigma^2_{\epsilon 11i}) - \tau_{\tilde{y}12i}(\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})}{(\tau^2_{\tilde{y}11i} + \sigma^2_{\epsilon 11i})(\tau^2_{\tilde{y}22i} + \sigma^2_{\epsilon 22i}) - (\tau_{\tilde{y}12i} + \sigma_{\epsilon 12i})^2} \end{bmatrix}. \quad (12)$$

The diagonal elements of $\boldsymbol{rel_w}(\boldsymbol{y}_i)$ are related to the person-specific reliabilities $rel_w(y_i)$ for each person: Specifically, if the correlations between the true scores are zero, and the correlations between the measurement error is zero, the diagonals are exactly equal to the reliabilities $rel_w(y_i)$.

The relationship between the expected matrix of autoregressive and cross-lagged effects $\hat{\Phi}_i$ for the VAR(1) model, and the matrix Φ_i for the MEVAR(1) can be expressed as

$$\hat{\Phi}_i = \Phi_i \boldsymbol{rel_w}(\boldsymbol{y}_i), \quad (13)$$

which results in the following for the bivariate case,

$$\hat{\Phi}_i = \begin{bmatrix} \phi_{11}r_{11} + \phi_{12}r_{21} & \phi_{11}r_{21} + \phi_{12}r_{22} \\ \phi_{21}r_{11} + \phi_{22}r_{21} & \phi_{21}r_{12} + \phi_{22}r_{22} \end{bmatrix}, \quad (14)$$

where r_{pq} indicates an element of the reliability matrix. It is important to note from Equations 11 to 14, that the resulting difference between the true parameters from the MEVAR(1) model and the estimated parameters from the VAR(1) model partly depends on the other regression parameters: For instance, the distortion of $\hat{\phi_{11}}$ compared with ϕ_{11} depends on r_{11}, but it also depends on the product of ϕ_{12} and r_{12}, such that the larger ϕ_{12} and r_{12}, the larger the discrepancy between the true ϕ_{11} and the estimate $\hat{\phi_{11}}$ from the VAR(1) model may become. As a result, the more variables that are included in the model, the more complicated and pronounced these discrepancies can become, because the more variables are included, the more distorting terms will be included in each element of $\hat{\Phi}_i$ (e.g., in a 3×3 model the discrepancy between $\hat{\phi_{11}}$ and ϕ_{11} will not only depend on r_{11} and the product ϕ_{12} and r_{12}, but also on the product of ϕ_{13} and r_{31}). Furthermore, note that even if a variable is measured without error, the estimated VAR(1) autoregressive effect for that variable (e.g.) may still differ notably from the true autoregressive effect from the MEVAR(1) model, as a result of measurement error in other variables in the model.

One may further observe from Equations 11 to 14 that the impact of disregarding measurement error on estimates of the autoregressive and cross-lagged effects will depend roughly on two aspects: The person-specific reliability for each variable, and the correlations between the measurement error of the different variables in the model. More specifically, the lower the reliabilities of the variables, the more pronounced the distortion of the autoregressive and cross-lagged effects will be when the VAR(1) model is erroneously fitted to the data. The stronger the (either positive or negative) correlations between the measurement error, the more easily spurious effects will arise.

Disregarding measurement error can have various effects on the estimated autoregressive and cross-lagged effects: They may be underestimated or they may be overestimated, to the extent that effects may "disappear" or even switch signs, and spurious effects may arise. In Figure 2 we present four examples of the effects of disregarding measurement error on the estimated autoregressive and cross-lagged effects in a VAR(1) model, using four pairs of network graphs. The top row of Figure 2 shows network graphs of four true data-generating MEVAR(1) models, and the bottom row shows the corresponding graphs for the results of a regular VAR(1) model that disregards measurement error.[7] Comparing the top and bottom row of network graphs shows that many arrows are thinner for the bottom VAR(1) graphs than for the top MEVAR(1) graphs, indicating that many autoregressive and cross-lagged effects in the VAR(1) model are underestimated compared to the true MEVAR(1) effects. These autoregressive and cross-lagged parameters are pushed toward zero. The extent of the underestimation is different for each variable, depending on the reliability of that variable.[8] Further note, comparing the true MEVAR(1) graphs and the VAR(1) graphs, that many small spurious effects (thin arrows not present in the true model) are introduced in the VAR(1) graphs. Graphs A1 and A2 illustrate that it is possible to find large spurious effects, in this case a large positive spurious effect between Variable 1 and 2. Note that this spurious effect in the VAR(1) graph is actually the strongest effect in the model. Graphs B1 and B2 illustrate that relatively large negative spurious effects may arise as well as a result of disregarded measurement error, here between Variable 1 and 5. Graphs C1 and C2 illustrate how a true large effect of the MEVAR(1) may become underestimated to the point that it "disappears" or even changes signs when fitting a VAR(1) model (i.e., the relationship between Variable 4 and 2 changes signs from .3 in graph C1 to −.01 in graph C2). Finally, Graphs D1 and D2 illustrate that associations may also be overestimated (i.e., for the relationships between Variable 4 and 2), and that many of the effects of disregarding measurement error, that were separately illustrated in Graphs A–C, may occur together. These examples show that disregarding measurement error by using the VAR(1) model instead of the MEVAR(1) model can result in very different estimated autoregressive and cross-lagged effects, and hence different conclusion about the dynamic process under study.

Preliminary Implementation of the MEVAR(1) Model in Mplus

Next to determining the consequences of disregarding measurement error based on the MEVAR(1) model, it is also possible to fit the MEVAR(1) model to account for measurement error in our VAR(1) model, and to estimate the reliabilities discussed in the previous sections. We provide a preliminary implementation of the model here, making use of the Bayesian DSEM modeling in Mplus

[7] Note that these graphs are based on Equation 13—there was no data simulated—such that sampling error is not an issue in these graphs. The reliabilities for the examples range in between .85 to approximately .5 (for the generating values for all the relevant parameters, see Appendix B and C). These reliabilities are similar to those found in an $n = 1$ empirical example by Schuurman et al. (2015; ranging from .5 to .7), those found by Hu et al. (2016) of on average .7 to .75 for the PANAS, and to the results of the empirical example of the current work, which are presented in a later section. These reliabilities may seem low, but this is not necessarily surprising. We will discuss this issue further in the Discussion section.

[8] Note that if one is interested in calculating network statistics for these graphs, such as centrality/betweeenness, that these characteristics can be seriously distorted for the regular VAR(1) model as a result of this.

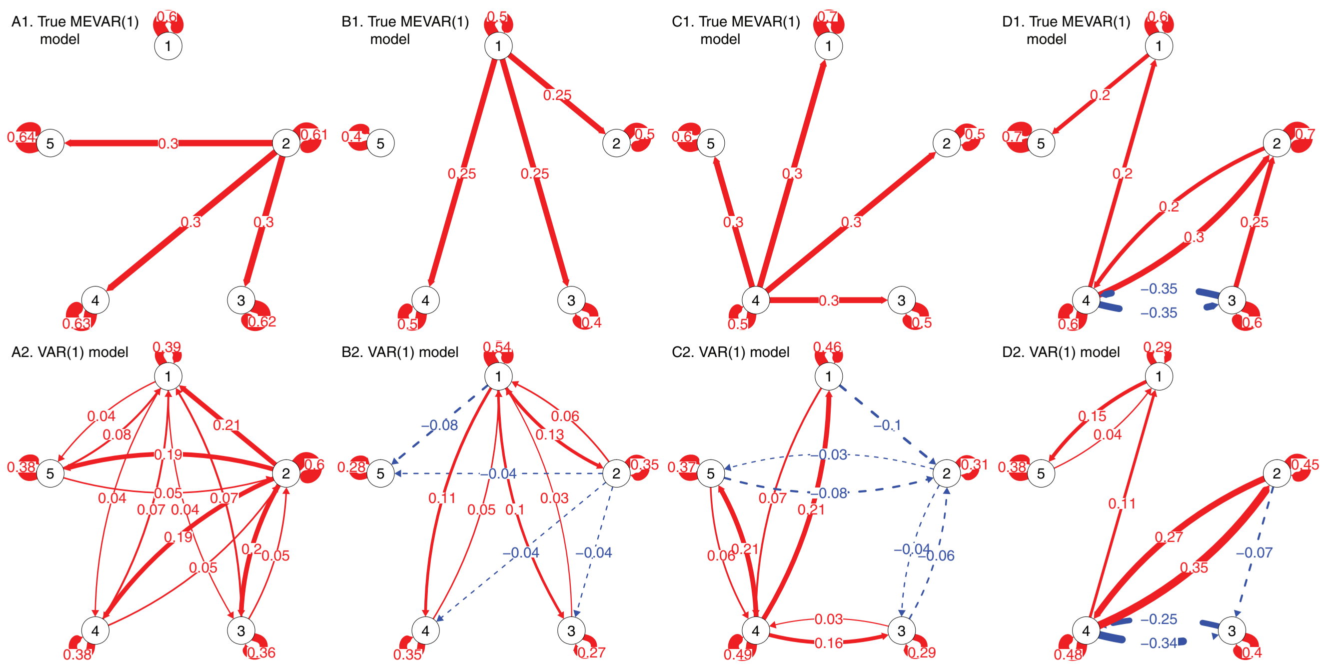

Figure 2. Four pairs of graphs that provide examples of the potential effects of disregarding measurement error (using a regular VAR(1) model) on the regression coefficients. Graphs on the top row (A1, B1, C1, D1) represent the data-generating MEVAR(1) model, while the accompanying graphs on the bottom row (A2, B2, C2, D2) represent the parameter estimates that would be obtained using a VAR(1) model, disregarding measurement error. In these graphs, circles (nodes) represent the measured variables, and arrows between the nodes (edges) represent the relationships between these variables. Edges that point from one variable into itself represent autoregressive effects, and edges that point from one variable to another variable represent cross-lagged effects. Red/solid arrows indicate positive effects, and blue/dashed arrows indicate negative effects. The larger the absolute value of a regression coefficient, the thicker the arrow. Regression coefficients smaller than |.03| in the VAR(1) graphs are suppressed for sake of clarity. See the online article for the color version of this figure.

v8 (Muthen & Muthen, 2017).[9] An advantage of the Bayesian approach is that one gains the flexibility to fit complex multilevel models, as a result of the Bayesian MCMC estimation procedures. This for example allows us to fit a full bivariate multilevel MEVAR(1) model including random means, regression parameters, and covariance matrices. These Bayesian procedures also make it relatively easy to estimate new quantities based on the estimated model parameters, as well as uncertainty intervals for these estimated quantities. This will be especially useful for obtaining the between-person and person-specific reliability estimates. For an introduction to Bayesian statistics and MCMC estimation procedures, we refer the reader to Hoijtink, Klugkist, and Boelen (2008) and Gelman, Carlin, Stern, and Rubin (2003).

We specify the model as described in the previous sections, and make use of Mplus' default prior distributions. Annotated Mplus code for the MEVAR(1) model is provided in Appendix A. Random covariance matrices are modeled in Mplus by specifying a latent variable (z) on which the residuals for Variable 1 and 2 load, with the factor loadings fixed to 1 (cf., Hamaker, Asparouhov, Brose, Schmiedek, & Muthén, 2018). That is, to obtain random measurement error covariance matrices the following is specified:

$$\begin{bmatrix}\epsilon_{1ti}\\ \epsilon_{2ti}\end{bmatrix} = \begin{bmatrix}1\\ 1\end{bmatrix} z_{ti} + \begin{bmatrix}\upsilon_1\\ \upsilon_2\end{bmatrix} \tag{15}$$

Note that as a result of this specification, for a specific individual the variance of the measurement error ϵ_{1t} is then equal to the variance of latent variable z plus the variance of υ_1 ($\sigma_z^2 + \sigma_{\upsilon_1}^2$), and the variance of ϵ_{2t} is then equal to the variance of latent variable z plus the variance of υ_2 ($\sigma_z^2 + \sigma_{\upsilon_2}^2$). The covariance between the innovations is equal to the variance of z (σ_z^2). Mplus assumes normal distributions for the logs of the variances of z, υ_1, and υ_2 across persons, allowing in this way for random variances and covariances across persons. We use this latent variable approach to obtain random measurement error covariance matrices, as well as random innovation covariance matrices. Note however, that in using this latent variable approach, depending on the factor loadings, the covariances will either be restricted to be positive (both

[9] We opt for Mplus because of their flexible implementation of random covariance matrices. Code for an implementation in open source software OpenBUGS/WinBUGS/Jags that implements the model with random covariance matrices for a bivariate model is available from Schuurman (2016). Further, it may be possible to program a similar implementation as that in Mplus—or an alternative implementation of the model with random covariance matrices—in Stan, which is Bayesian open source software based on Hamiltonian sampling (Carpenter et al., 2016).

factor loadings equal to one), or negative (one factor loading equal to −1). That is, one has to decide a priori whether the residual correlations are likely to be positive, or negative, in this implementation of random covariance matrices in Mplus (cf., Hamaker et al., 2018). This is arguably a disadvantage of this random covariance matrix implementation. An advantage of this implementation is, however, that it readily generalizes to larger multivariate models than bivariate models, and (for instance) to models that would include covariates for the random variances and covariances.

We performed a small proof of principle simulation study to provide an indication of the estimability of the model. We evaluate the convergence of the model by inspecting the mixing of the chains and the Gelman-Rubin statistics for the parameters (Gelman & Rubin, 1992). We fitted the model using two chains of each 90,000 iterations, of which half was discarded as burn-in. We simulated data according to the MEVAR(1) model for different sample sizes using Mplus, with 100 replications for each condition: For a sample size of 100 time points and subjects, 100 time points and 200 subjects, and 200 time points and 100 subjects. Depending on the number of subjects and time points, running 100 iterations for two chains took approximately between 10 s to 20 s. We evaluate the performance of the model based on the bias, mean square errors, and coverage rates for the 95% credible intervals for each parameter. The true values and simulation results for the fixed effects and the variances for the random effects are presented in Table 1, these results for the covariances of the random effects are available upon request for sake of sparsity.

For the smallest sample size of 100 time points and subjects, we encountered some issues in the model estimation: For a sample size of 100 subjects and repeated measures, 80 replications could be completed; 20 aborted during estimation because the sampler ventured into nonpositive definite territory for the covariance matrix of the random effects. For a sample size of 100 subjects and 200 repeated measures, 99 replications were completed, and for 200 subjects and 100 time points 96 replications were completed. The results show that for the completed replications the fixed effects and the (co)variances of the random covariance matrices are estimated quite well: Bias and *MSE* are small, and decrease when sample size increases, and coverage rates are overall in the area around .95. We do see some notable bias for the fixed effects of Variable 2: For this variable the measurement error variance is slightly underestimated, and the innovation error variance is slightly overestimated, and the autoregressive effect is therefore also slightly underestimated. For some cases we see that this bias decreases more slowly as sample size increases than the precision of the estimates increases, resulting in lower coverage rates for larger sample sizes. Overall, the fixed effects, and the

Table 1
Bias, Mean Square Error (MSE), and Coverage Rates (CR) for the 95% Credible Intervals, for the MEVAR(1) Model for Different Numbers of Repeated Measures (T) and Subjects (N)

	Bias			*MSE*			CR		
	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 100$	$n = 200$
Parameter = true	$t = 100$	$t = 200$	$t = 100$	$t = 100$	$t = 200$	$t = 100$	$t = 100$	$t = 200$	$t = 100$
$\gamma_{\mu 1} = 5$	.004	.002	.004	.004	.004	.002	.91	.94	.96
$\gamma_{\mu 2} = 5$	.001	−.016	.002	.003	.003	.002	.95	.98	.96
$\gamma\phi_{11} = .5$	−.021	−.011	−.004	.002	.001	.001	.98	.97	.98
$\gamma\phi_{22} = .5$	−.047	−.030	−.041	.004	.002	.003	.89	.91	.80
$\gamma\phi_{12} = .2$	−.006	−.004	−.011	.001	.001	.001	.98	.99	.98
$\gamma\phi_{21} = 0$	.021	.010	.019	.002	.002	.001	.94	.97	.91
$log\gamma_{\sigma^2_{\omega 11}} = -4.1$	.045	.031	−.011	.053	.045	.036	1	.96	.96
$log\gamma_{\sigma^2_{\omega 22}} = -3.2$	.135	.123	.157	.055	.042	.042	.94	.97	.92
$log\gamma_{\sigma^2_{\omega 12}} = -3.9$	.071	.076	.084	.029	.027	.020	.96	.95	.92
$log\gamma_{\sigma^2_{\epsilon 11}} = -2$	−.025	−.003	−.011	.007	.007	.003	.98	.93	.97
$log\gamma_{\sigma^2_{\epsilon 22}} = -2.8$	−.127	−.078	−.128	.051	.032	.034	.93	.94	.92
$log\gamma_{\sigma^2_{\epsilon 12}} = -3.7$	.051	.043	.058	.027	.021	.012	.93	.97	.97
$\psi^2_{\mu 1} = .3$	.035	.037	.016	.004	.004	.001	.89	.86	.94
$\psi^2_{\mu 2} = .3$	.059	.053	.025	.006	.006	.002	.83	.78	.84
$\psi^2_{\phi 11} = .01$	.012	.009	.006	.000	.000	.000	.83	.70	.85
$\psi^2_{\phi 22} = .01$	.018	.012	.010	.000	.000	.000	.75	.71	.69
$\psi^2_{\phi 12} = .01$	.010	.006	.004	.000	.000	.000	.61	.76	.92
$\psi^2_{\phi 21} = .01$	.021	.011	.011	.001	.000	.000	.60	.73	.72
$log\psi^2_{\sigma^2_{\omega 11}} = .5$	−.131	−.111	−.133	.037	.036	.034	1	.99	1
$log\psi^2_{\sigma^2_{\omega 22}} = 1.4$	−.141	−.141	−.154	.095	.063	.061	.96	.97	.90
$log\psi^2_{\sigma_{\omega 12}} = 1.1$	−.143	−.113	−.136	.054	.040	.038	.95	.97	.91
$log\psi^2_{\sigma^2_{\epsilon 11}} = .5$	.031	.020	.020	.008	.006	.003	.98	.98	.99
$log\psi^2_{\sigma^2_{\epsilon 22}} = 1$	.099	.084	.095	.061	.048	.030	.98	.99	.97
$log\psi^2_{\sigma_{\epsilon 12}} = 1.3$	−.112	−.127	−.126	.051	.053	.033	.95	.95	.96

variances of the random covariances matrices, and the covariances between the random effects are recovered well, even for smaller sample sizes.

However, the variances of the random regression coefficients, and to a lesser extent the variances of the random means, are overestimated. This pattern of overestimation of the variances of the random effects is similar to what is known to happen for smaller sample sizes for the regular multilevel VAR(1) model, as well as for multilevel dynamic factor models, for certain specifications of the Inverse-Wishart prior distribution for the covariance matrix of the random effects (Schuurman et al., 2016; Song & Ferrer, 2012). For variances that are quite close to zero, such as those for the regression coefficients, the Inverse-Wishart prior distribution can be informative, resulting in biased estimates for smaller sample sizes, or equivalently, for more complex models, when the data does not completely dominate the prior distribution. It might be possible to improve the estimates of the variances for smaller sample sizes here by using data-informed prior distributions, as suggested by Schuurman et al. (2016). This would require preestimates of the variances, for example, obtained by first estimating a model with uniform priors for the variances of the random parameters (fixing covariances to zero), or from prior information such as previous research or expert knowledge. For the current implementation of the MEVAR(1) model the results for the variances of the regression coefficients should be taken with a grain of salt, especially for smaller numbers of subjects. When the number of subjects increases to 200, the estimates for these variances improve considerably, but are still suboptimal.

An important direction for future research is therefore to investigate if and how the estimates of the variances of the random effects can be improved by specifying better informed priors, including how suitable preestimates of the variances of the random effects could be obtained for this model. Such prior distributions may also help keep the Bayesian sampler in positive definite territory, such that more replications can be completed in each condition. Note, however, that for the current implementation the fixed effects, the covariances of the random effects, and the variances random effects for the random covariance matrices are recovered well.

Furthermore, this implementation will still be more accurate than those of a regular VAR(1) model that disregards measurement error, if measurement error is present in the data. To illustrate this we also fitted a regular multilevel VAR(1) model to data generated from the MEVAR(1) model. As can be seen from the results in Table 2, performance for the VAR(1) is very poor, resulting in large bias and low coverage rates for the regression coefficients and residual variances. The effects of disregarding measurement error in the VAR(1) models depend on the true regression parameters and measurement error (co)variances, and as such may differ from situation to situation, as discussed in the previous section.

Obtaining Reliability Estimates From the Bayesian MEVAR(1) Model

Obtaining the reliability estimates while fitting the MEVAR(1) model can be done by making use of the Bayesian MCMC samples from the posteriors of the estimated parameters. We save these samples in Mplus, import them into R, and perform calculations on these samples to obtain the posteriors for the reliabilities.

Specifically, for the person-specific reliabilities we first need to obtain the measurement error covariance matrices for each person, based on the modeled logs of the variances of z, υ_1 and υ_2. That is,

Table 2

Bias, Mean Square Error (MSE), and Coverage Rates (CR) for the 95% Credible Intervals, for the VAR(1) Model, Where the True Model is the MEVAR(1) Model, for Different Numbers of Repeated Measures (T) and Subjects (N)

	Bias			*MSE*			CR		
	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 100$	$n = 200$	$n = 100$	$n = 100$	$n = 200$
Parameter = true	$t = 100$	$t = 200$	$t = 100$	$t = 100$	$t = 200$	$t = 100$	$t = 100$	$t = 200$	$t = 100$
$\gamma_{\mu 1} = 5$	.002	.001	.004	.003	.004	.002	.94	.93	.96
$\gamma_{\mu 2} = 5$	.003	−.015	.000	.003	.003	.002	.97	.99	.94
$\gamma\phi_{11} = .5$	−.331	−.328	−.330	.110	.108	.109	0	0	0
$\gamma\phi_{22} = .5$	−.240	−.238	−.238	.058	.058	.057	.06	0	0
$\gamma\phi_{12} = .2$	−.055	−.050	−.054	.003	.003	.003	.10	.10	0
$\gamma\phi_{21} = 0$	.006	.008	.006	.000	.000	.000	.92	.90	.90
$log\gamma_{\sigma^2_{\omega 11}} = -4.1$	2.25	2.26	2.25	5.07	5.12	5.05	0	0	0
$log\gamma_{\sigma^2_{\omega 22}} = -3.2$	1.06	1.07	1.06	1.12	1.14	1.12	0	0	0
$log\gamma_{\sigma^2_{\omega 12}} = -3.9$	1.18	1.19	1.18	1.40	1.43	1.41	0	0	0
$\psi^2_{\mu 1} = .3$	.045	.037	.018	.005	.004	.001	.85	.86	.92
$\psi^2_{\mu 2} = .3$	.054	.053	.025	.005	.006	.002	.84	.78	.82
$\psi^2_{\phi 11} = .01$	.009	.008	.007	.000	.000	.000	.29	.26	.20
$\psi^2_{\phi 22} = .01$	.016	.015	.014	.000	.000	.000	.05	0	.00
$\psi^2_{\phi 12} = .01$	.005	.005	.003	.000	.000	.000	.77	.68	.82
$\psi^2_{\phi 21} = .01$	−.005	−.006	−.006	.000	.000	.000	.49	.24	0
$log\psi^2_{\sigma^2_{\omega 11}} = .5$	−.056	−.111	−.059	.008	.008	.006	.89	.86	.81
$log\psi^2_{\sigma^2_{\omega 22}} = 1.4$	−.533	−.141	−.553	.304	.311	.314	.18	.14	0
$log\psi^2_{\sigma^2_{\omega 12}} = 1.1$	−.395	−.113	−.415	.171	.202	.038	.181	.20	.06

for each iteration of the MCMC procedure we first take the exponents of $log(\sigma^2_{\upsilon 1})$, $log(\sigma^2_{\upsilon 2})$, and $log(\sigma^2_z)$. Then we calculate the variances of the measurement error by adding σ^2_z to $\sigma^2_{\upsilon 1}$, and to $\sigma^2_{\upsilon 1}$, in each iteration. The covariance between the measurement error is equal to σ^2_z. We use the same approach to obtain the innovation covariance matrices. After this, in each iteration of the MCMC procedure, we calculate the person-specific covariance matrix $\boldsymbol{T}_{\tilde{y}i}$ for each person i using Equation 6, based on the sampled person-specific regression parameters and the sampled person-specific innovation covariance matrices. From the relevant diagonal element of this covariance matrix, we obtain the variance $\tau^2_{\tilde{y}i}$ for our variable of interest, which we need to estimate the person-specific total variance $v(y_i)$. Based on this estimate and the estimated measurement error variance $\sigma^2_{\epsilon i}$, we then calculate the total variance $v(y_i)$ for each person using Equation 8 in each iteration of the MCMC procedure. Finally, in each iteration we calculate the person-specific reliability making use of Equation 9, based on the estimate we obtained for the total variance $v(y_i)$ and the estimated person-specific variance $\tau^2_{\tilde{y}i}$. This results in an estimate of the person-specific reliability for each iteration of the MCMC procedure for each person, which together result in a posterior distribution of the person-specific reliability for each person. Based on these posterior distributions we can determine a point estimate and credible intervals for each person's reliability.

To calculate the average person-specific reliability across persons, we calculate the mean across the person-specific reliabilities in each iteration of the MCMC procedure. Note that this mean is essentially a sample mean, and that the credible interval for this estimate thus disregards sampling variability; it takes into account the variability across the iterations of the MCMC procedure, but does not take into account that the current selection of subjects is just one possible sample from a larger population. As such, care should be taken when generalizing the average within-person reliability to other samples, as the credible intervals will be too small for this purpose. The larger the number of subjects in the sample, the smaller this discrepancy will be.

For the between-person reliability of a specific variable, we make use of Equation 5 to 7. However, before we can use these equations, we need to obtain the fixed effects for the measurement error variance $\gamma_{\epsilon 1}$ and $\gamma_{\epsilon 2}$, in each iteration of the MCMC procedure, based on the fixed effects for $log(\sigma^2_{\upsilon 1i})$, $log(\sigma^2_{\upsilon 2i})$, and $log(\sigma^2_{zi})$. Note that that if $log(X)$ is normally distributed with mean μ and variance σ_2, X is lognormally distributed, with mean $e^{\mu+\sigma^2/2}$ and variance $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$. Making use of this property, we transform the fixed effects for $log(\sigma^2_{\upsilon 1i})$, $log(\sigma^2_{\upsilon 2i})$, and $log(\sigma^2 zi)$ to a regular scale in each iteration. Afterward, we calculate the fixed effects for the covariances matrices in each iteration, by adding the samples of the resulting fixed effect for σ^2_z to the fixed effect $\sigma^2_{\upsilon 1}$, and to the fixed effect of $\sigma^2_{\upsilon 2}$. The fixed effect for the covariance between the innovations or measurement error is equal to the fixed effect for σ^2_z. Next, we calculate matrix $\boldsymbol{T}_{\tilde{y}i}$ for each person i using Equation 6, based on the estimated person-specific regression parameters and the estimated person-specific innovations covariance matrices. From the relevant diagonal element of each person-specific covariance matrix, we obtain the person-specific variances $\tau^2_{\tilde{y}i}$, which we need to estimate $\underset{i}{E}[\tau^2_{\tilde{y}}]$. To obtain an estimate of $\underset{i}{E}[\tau^2_{\tilde{y}}]$ we calculate the average person-specific variance across persons in each iteration. We then calculate the total variance using Equation 5 in each iteration, based on the estimate we obtained for $\underset{i}{E}[\tau^2_{\tilde{y}}]$, the estimated variance of the person-specific means ψ^2_μ, and the fixed effect for the measurement error variance $\gamma_{\sigma^2_\epsilon}$. Finally, in each iteration we calculate the between-person reliability making use of Equation 7, based on the estimate we obtained for the total variance $V(y)$ and the estimated variance of the person-specific means ψ^2_μ. This results in a sample of the between-person reliability for each iteration of the MCMC procedure, which together form a posterior distribution. Based on this posterior distribution we determine a point estimate for the between-person reliability and the credible interval for the between-person reliability.

Empirical Application on Dyadic Affect Data

In this empirical application we use a data set on affect measurements from a daily diary study including 191 couples, including both university students and couples from the local community (Ferrer, Steele, & Hsieh, 2012; Ferrer & Widaman, 2008). We focus on two measures of positive affect for the women from these couples: (a) daily relationship positive affect (RelPA), that is, the positive affect (PA) each woman experienced specifically about her romantic relationship that day; and (b) general positive affect (GenPA), the PA each woman experienced generally that day. RelPA was measured with nine 5-point Likert scale items (1 = *very slightly or not at all* to 5 = *extremely*), for which the participants indicated to what extent they felt the following ways about their relationship that day: "emotionally intimate," "trusted," "committed," "physically intimate," "free," "loved," "happy," "loving," and "socially supported." GenPA was measured with the PANAS (Watson, Clark, & Tellegen, 1988). The participants reported on their RelPA and GenPA at the end of each day, for approximately 60 to 90 days. For each woman, daily average scores were calculated for both types of PA.

Here, we will investigate how GenPA and RelPA influence each other within women. Specifically, we want to know if (a) the temporal evaluation of one's relationship spreads to other areas in daily life; (b) general affective tone colors the evaluation of one's relationship; or (c) both are the case. The multilevel MEVAR(1) approach allows us to study this question by establishing associations between GenPA and RelPA, and identifying individual differences in these associations.

We fit the model (as specified in Equation 1 to 4) for the RelPA and GenPA, making use of the DSEM module of Mplus Version 8 (Muthen & Muthen, 2017) as described previously (see Appendix A for the Mplus code). In the following, we first present the parameter estimates for the MEVAR(1) model and the regular VAR(1) model. After that we discuss the estimated between-person and person-specific reliabilities for the MEVAR(1) model.

Results for the Dynamics of General and Relationship Positive Affect

The point estimates of the fixed effects and the variances of the random effects, and their 95% credible intervals (CI) are presented in Table 3 for the MEVAR(1) and the regular VAR(1) model. We discuss the autoregressive and cross-lagged effects for the MEVAR(1) and VAR(1) model first, followed by the estimated trait scores, and measurement error and innovation covariance matrices for the MEVAR(1) model.

Table 3
Parameter Estimates for the Bivariate Multilevel MEVAR(1) and VAR(1) Model for Women in a Relationship, Modeling the Relationship Between Daily Relationship and General Positive Affect

Women	Fixed effects		Var. random effects	
Parameter	MEVAR(1) [95% CI]	VAR(1) [95% CI]	MEVAR(1) [95% CI]	VAR(1) [95% CI]
μ_g	2.7 [2.6, 2.7]	2.6 [2.6, 2.7]	.36 [.29, .46]	.38 [.31, .47]
μ_r	3.5 [3.4, 3.6]	3.5 [3.4, 3.6]	.44 [.36, .55]	.48 [.39, .60]
ϕ_g	.75 [.69, .80]	.31 [.28, .34]	.01 [.01, .02]	.03 [.02, .04]
ϕ_r	.59 [.53, .64]	.37 [.34, .40]	.03 [.02, .05]	.03 [.03, .05]
ϕ_{rg}	.07 [.02, .13]	.02 [.00, .04]	.02 [.01, .04]	.01 [.00, .01]
ϕ_{gr}	−.03 [−.07, .00]	.04 [.02, .07]	.01 [.00, .02]	.01 [.01, .02]
$\sigma^2_{\omega g}$	.06 [.05, .08]	.30 [.27, .33]	.00 [.00, .01]	.04 [.03, .05]
$\sigma^2_{\omega r}$	.12 [.10, .15]	.28 [.25, .32]	.02 [.01, .07]	.06 [.04, .11]
$\sigma_{\omega gr}$	.04 [.03, .05]	.10 [.08, .12]	.00 [.00, .01]	.01 [.01, .03]
$\rho_{\omega gr}$	.33 [.15, .59]	.29 [.16, .48]		
$\sigma^2_{\epsilon g}$	.22 [.19, .24]		.02 [.02, .04]	
$\sigma^2_{\epsilon r}$	.14 [.12, .17]		.02 [.01, .05]	
$\sigma_{\epsilon gr}$	.05 [.04, .07]		.01 [.00, .01]	
$\rho_{\omega gr}$	.26 [.11, .53]			
rel_{wg}	.42 [.38, .45]			
rel_{wr}	.58 [.54, .62]			
rel_{bg}	.50 [.39, .57]			
rel_{br}	.54 [.40, .61]			

Note. Parameter estimates for the fixed effects (group means) and the variances of the random effects (group variances) are presented for the person-specific means (μ_g, μ_r), autoregression effects (ϕ_g, ϕ_r), cross-lagged effects (ϕ_{gr}, ϕ_{rg}), the innovation variances ($\sigma^2_{\omega g}$, $\sigma^2_{\omega s}$) and covariance ($\sigma_{\omega gs}$) and correlation ($\rho_{\omega gs}$), the measurement error variances ($\sigma^2_{\epsilon g}$, $\sigma^2_{\epsilon s}$) and covariance ($\sigma_{\epsilon gs}$) and correlation ($\rho_{\epsilon gs}$), the average person-specific reliabilities (rel_{wg}, rel_{wr}), and the between-person reliabilities (rel_{bg}, rel_{br}).

Autoregressive effects. The estimated means and variances for the autoregressive effects indicate that for most women the autoregressive effects of general and RelPA are expected to be positive. For RelPA we would expect that the autoregressive coefficients for 95% of the women range from about .25 to .9. For GenPA we find ranges from about .55 to .95. This indicates that for the most, if not all, women we expect there to be carry-over present in the regulation of both types of positive affect. In other words, if the GenPA or the positive affect about the relationship of a woman is perturbed—for instance resulting in a relatively high PA—the positive affect will linger some time above the average level of PA. However, the same holds when she experiences a relatively low PA: In this case PA will linger for some time below baseline levels due to the autoregressive effect. When we compare these estimated autoregressive effects of the MEVAR(1) model with those of the VAR(1) model, we find that the inertia is estimated to be much weaker for the latter, with average effects of .3 (95% CI [.28, .34]) and .4 (95% CI [.34, .40]) for GenPA and RelPA, respectively, compared with average effects of .75 (95% CI [.69, .80]) and .59 (95% CI [.53, .64]) for the MEVAR(1) model. In fact, the credible intervals for the fixed autoregressive effects for the VAR(1) and MEVAR(1) model do not even overlap, such that inferences about the strength of the carry-over are markedly different for the two models.

Cross-lagged effects. For the MEVAR(1) model, the estimate of the average cross-lagged effect of GenPA at the previous day on current RelPA is equal to −.03 (95% CI [−.07, .00]), which indicates there is no evidence that RelPA positively influences GenPA the next day on average. There is some variance around this effect (.01; 95% CI [.00, .02]), indicating that this cross-lagged effect is expected to range from about −.2 to .15 across women. There is evidence that GenPA colors women's experienced RelPA the following day on average, although the effect may be small: The estimated average cross-lagged effect was .07 (95% CI [.02, .13]) across women, with a variance of .02 (95% CI [.01, .04]). Based on these results, we would expect that the cross-lagged effects of RelPA on GenPA for 95% of women would lie in between approximately −.2 and .35. For the VAR(1) model we find no clear evidence for an effect of GenPA on RelPA on average, but we do find evidence for an effect of RelPA on GenPA—the opposite of what we find for the MEVAR(1) model. That is, in this example, we reach different conclusions for both the autoregressive effects and for the cross-lagged effects in the VAR(1) and MEVAR(1) model.

Trait Scores and Innovation and Measurement Error Covariance Matrices

Based on the estimated means and variances of the traits scores (see Table 3), we find that women on average feel moderately positive to quite positive about their relationship, although there is considerable variance in this across women. Across women, most would be expected to have a trait score between approximately 2.2, indicating they on average feel slightly positive; and 4.8, indicating they feel very positive about their relationship. The average experienced GenPA is estimated to be a bit lower on average (see Table 3), and based on the estimated means and variances of the trait scores across women, we would expect most women to have a trait score between approximately 1.5 (low GenPA) and 3.8 (moderately high GenPA).

When we inspect the estimated correlations between the random trait scores and the regression parameters for the MEVAR(1) model, we find that the traits scores for GenPA and RelPA are positively correlated (.59; 95% CI [.48, .69]): Women that generally have a high mean level of GenPA, also tend to have a relatively high mean level of RelPA. Further, we find that the autoregressive effects for RelPA are negatively correlated with the cross-lagged effects of GenPA on RelPa (−.51; 95% CI [−.74, −.21]). That is, for women with relatively consistent scores for their relationship positive affect, their general PA tends to color their positive affect about their relationship less. We find no evidence for correlations between the remaining random parameters.

Finally, when we inspect the variances and correlations of the innovations and the measurement errors, we find that the estimated average variances lie within a range of .05 to .2, and the variances around these average variances range from .003 to .02 (see Table 3). The average correlation between the innovations of general and RelPA is .33 (95% CI [.15, .59]). This indicates that there is a considerable part of the concurrent association between RelPA and GenPA that cannot be explained by the experienced PA at the previous occasion, that seems to be due to unobserved influences of which the effects are passed along across multiple measurement occasions. The average correlation between the measurement error of GenPA and RelPA is .26 (95% CI [.11, .53]). This indicates that there is also a considerable part of the concurrent association between RelPA and GenPA that seems to be due to unobserved, occasion-specific effects. Note that this correlation indicates that part of the occasion-specific effects are not "pure" measurement errors, but something more systematic.

Reliabilities for Relationship and General Positive Affect

In the following we will discuss two types of reliability for RelPA and GenPA: The between-person reliability, and the person-specific (within-person) reliabilities.

Between-person reliability. We estimated the *between-person reliability*, that is, the proportion of variance that is due to stable differences across women, for RelPA and GenPA based on Equation 7. We found that for both GenPA and RelPA about half of the variance in the observed scores across all women and repeated measures is estimated to be due to systematic differences between women, while the other half is due to differences within women (rel_{bg} = .5; 95% CI [.39, .57]; rel_{br} = .54; 95% CI [.4, .61]).

Within-person reliabilities. In the previous subsection we found considerable variation across women in the regression parameters, and in the variances and correlations for the innovations and measurement error. As a result, the reliabilities for general and RelPA will also differ from woman to woman. Figure 3 contains plots of the posterior distributions, trimmed at their respective 95% CI, for each woman's person-specific reliability. The pink (left) distributions are the posterior distributions for GenPA, and the blue (right) distributions are the posterior distributions for RelPA, and the dots in each distribution represent the median reliability. The posterior distributions are ordered on the estimated median reliability for RelPA (i.e., with the woman with the lowest reliability on the top-right, and the highest on the bottom-left). As can

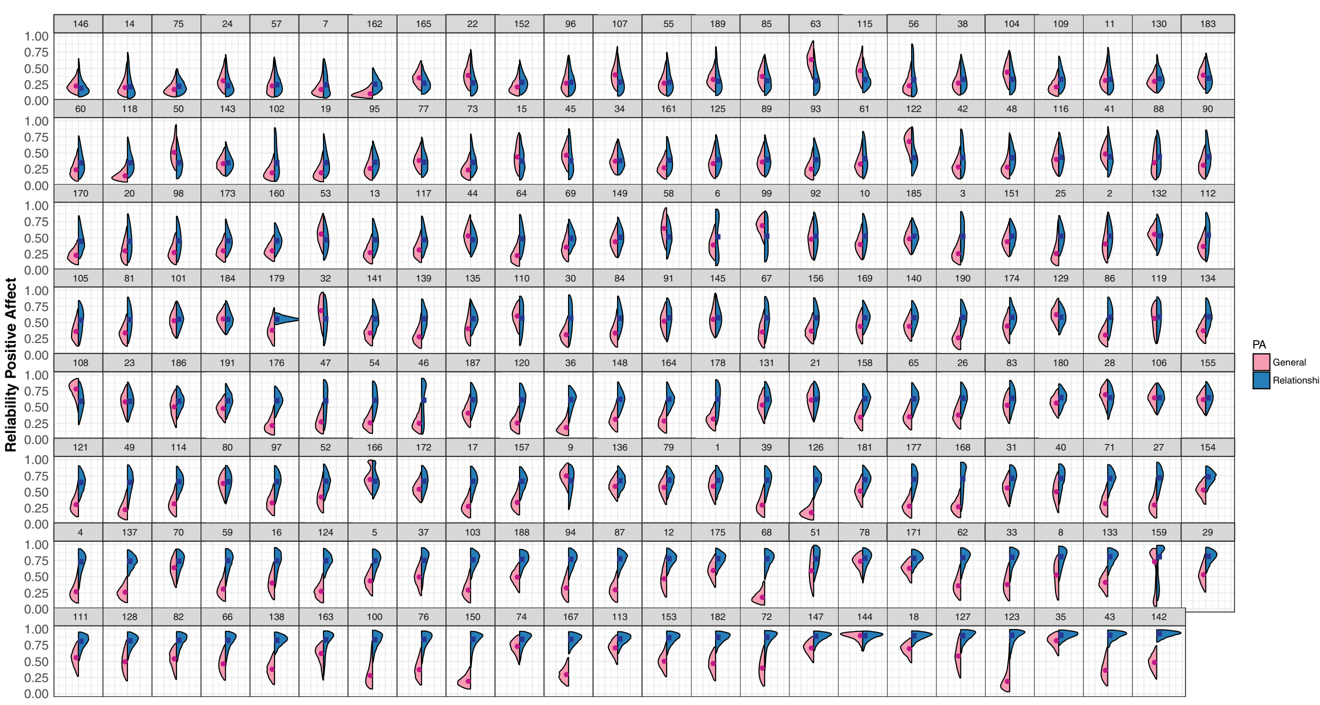

Figure 3. Plots of the posterior distributions for the reliability of relationship positive affect (blue/right), and general positive affect (pink/left) for each woman. The dot in each distribution represents the median for that distribution. The tails of the posteriors are trimmed at their 95% credible intervals. The posteriors are ordered across women based on their median reliability for relationship positive affect, from low to high (top left to bottom right). See the online article for the color version of this figure.

be seen from this figure, there is quite some variation in the estimated reliabilities across women. The lowest and highest estimated reliabilities for RelPA were approximately .20 and .93, and .11 and .90 for GenPA. Overall, the reliabilities for RelPA seem to be estimated slightly higher than those of GenPA. It can also be seen from Figure 3 that there is a fair amount of uncertainty about the person-specific reliabilities, that is, they have wide CIs. Still, there is evidence that the reliabilities of the PA measurements are likely to be lower than .8 for many women (i.e., most of the posterior distributions' mass lies below a reliability of .8). In other words, a considerable part of the variation in the observations for most women can be ascribed to measurement error or other occasion specific fluctuations. This is also reflected in the average person-specific reliabilities (see also Table 3) of .42 for GenPA (95% CI [.38, .45]), and of .58 for RelPA (95% CI [.54, .62]).

Finally, we investigated whether there is an association across women between the reliabilities for RelPA and GenPA. A scatter plot of the point estimates of the person-specific reliabilities is shown in Figure 4. We find evidence for a positive relationship between the reliabilities of GenPA and RelPA, with a correlation of .38 (95% CI [.26, .49]). This indicates that women that have a relatively high reliability for GenPA, also tend to have a high reliability for RelPA.

Discussion

The variance of intensive longitudinal data for a certain variable generally is a mix of between person variation, within-person variation due to dynamic processes, and occasion-specific, random within-person variation which includes measurement errors. Regular VAR(1) models disregard this last source of variance, essentially assuming perfectly reliable measurements. We have shown that if such variance is present in the data, but is disregarded, this can result in severely distorted autoregressive and cross-lagged effects. It is therefore important to consider the reliability of intensive longitudinal measurements, and account for this in our dynamic VAR(1) models.

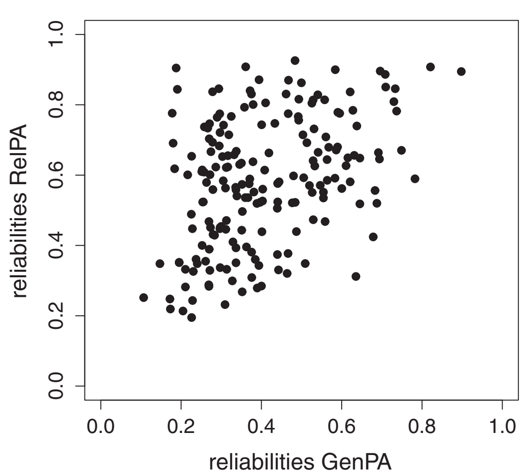

Figure 4. Scatter plots of the point estimates (medians of the posterior distributions) of the person-specific reliabilities for general positive affect (GenPA) in relationship positive affect (RelPA) across women.

The MEVAR(1) model we presented in this work separates the three sources of variation, while also taking into account that there may be differences in the sources of within-person variation across persons. Our proof of principle simulation study indicates that the model can recover most of the model parameters reasonably well, but for smaller numbers of participants the variances of the random regression coefficients are overestimated. A more extensive simulation study across a wide range of sample sizes and parameter values is needed however to determine the performance of the model more generally, and is an important direction for future research. Another clear topic for future research is to see if it is possible to improve the estimates of the variances of the random effects by specifying different priors (e.g., by using the methods suggested in Schuurman et al., 2016). Similarly, an important line of research, both for this and other multilevel models, concerns the investigation of the optimal way to account for random covariance matrices. Allowing for random covariance matrices is nontrivial, and other approaches than we used here may be possible, for example, by sampling random covariances matrices from an Inverse-Wishart distribution. It would also be worthwhile to further investigate what the consequences are of disregarding the differences across persons in the innovation covariance matrices and measurement error covariance matrices, as is often the case in practice (cf., Jongerling et al., 2015, for an exposition on the consequences of disregarding such differences in the innovation variances in a univariate model).

An alternative way to account for measurement error in the multilevel VAR context, if there are multiple items available that measure the same construct, is to make use of the internal consistency approach using a dynamic factor model (Molenaar, 1985; Song & Ferrer, 2012). This would require, however, that the items function to some extent as parallel tests for the latent variable of interest, *for every individual in the multilevel model*, which may be difficult to achieve (but see Hu et al., 2016). Further, in some cases using multiple indicators for each construct may severely increase the burden on the participants, especially if measurements are taken frequently and with short intervals as is the case in many intensive longitudinal studies that use experience sampling. As such, more research on the feasibility of the multilevel dynamic factor model, next to the MEVAR(1) model, is welcome.

By making use of a dynamic multilevel modeling approach, it is possible to evaluate the reliabilities of a measurement instrument for between-person differences in a specific population of individuals, but also person-specific reliabilities for within-person fluctuations. As noted previously, reliability is defined as specific to a certain population and the measurement instrument (Mellenbergh, 1996). That is, reliability estimates for one population (e.g., men) cannot simply be generalized to another (e.g., women). In the case of reliability for within-person psychological processes, each person may have a unique psychological process, and as such may be considered a single subpopulation (within a larger population of individuals). For example, in the empirical example we saw that people differ in their levels of inertia for both their general positive affect and their positive affect concerning their relationship. Furthermore, one can imagine that some people experience more, or are more easily affected by, external events than other people, or some persons may take more care in filling out self-report measures than others. From this perspective, it seems not very informative to state one reliability estimate for all individuals. By taking a multilevel dynamic modeling approach, we can take this into account.

Concerning the concepts of measurement error and reliability, it is important to note again here that the measurement error terms in the multilevel MEVAR(1) model do not only contain variance that is due to measurement errors—they also contain any "true" occasion-specific fluctuations in the construct of interest. As such, reliability estimates based on this measurement error variance can in practice at most provide an estimate of the *lower bound* of the reliability of the observed scores (as is the case for all other reliability estimates in psychology discussed; see also Borsboom, 2003; Guttman, 1945; Ten Berge & Sočan, 2004). To separate true occasion specific fluctuations from measurement errors further, one option may be to also make use of internal consistency reliability measures in the model, by including multiple indicators in a dynamic factor model; if an occasion-specific fluctuation occurs in all indicators at the same time, this may be indicative of a true occasion specific fluctuation rather than a measurement error (see Edmondson et al., 2013, for an example of this in the context of panel modeling). Again, however, this would require that the indicators function to some extent as parallel tests, which is nontrivial to assume. In practice, a combination of the single indicator model described here, together with the use factor structures where reasonable, may prove to be a fruitful approach.

Regardless, however, of whether the "unreliable part" of the data is a result of real fluctuations in the construct, or of measurement errors, it is important to take this type of occasion-specific variation in the data into account. An important contribution of the current work is demonstrating that disregarding this type of variation, as is the case in regular (multilevel) VAR(1) models, results in severely distorted estimated autoregressive and cross-lagged effects. Depending on the reliabilities of the variables and the correlations among the occasion-specific fluctuations, the regression parameters may be under- or overestimated, may switch signs, and spurious cross-lagged relationships may emerge. The lower the person-specific reliability of a specific variable, the more severe these biasing effects will be. An important question for future research is therefore how to test whether a MEVAR(1) or VAR(1) is best suited for the data. Schuurman et al. (2015) show for the $n = 1$ case, that the AIC, BIC, and DIC are not suitable for this purpose. It may be possible however to design a test based on the patterns of autocorrelation decay of the models, or to make use of other information criteria, or Bayes factors.

Another important question is what kind of reliabilities we may expect in practice when the observed scores do include measurement error. Empirical examples that provide estimates of person-specific reliabilities are still very rare. In an $n = 1$ example by Schuurman et al. (2015) about the daily mood of eight women, reliabilities ranged from about .5 to .7. In the empirical application of the current work on general and relationship PA we find similar results, with average reliabilities of .4 and .6 across women. Hu et al. (2016) find slightly higher average within-person reliabilities of about .7 or .75 for the PANAS, using a parallel test and internal consistency approach. The reliabilities found may seem somewhat low, but this result is less surprising for the MEVAR(1) model, considering that ϵ_{ti} could also include true occasion-specific fluctuations in the variable of interest in addition to measurement error. In the empirical example, we found a residual correlation between the measurement error, which implies that at least part of what is classified as measurement error may in fact be something more systematic. Such fluctuations in the observed scores may occur as a result of a wide range of both internal and external influences, including for example the weather, hormone levels, getting a phone call or e-mail, eating a snack, hearing a certain song, and so on, as long as the resulting fluctuations are specific to a single measurement occasion.

It is an interesting notion, that depending on the nature of these true effects, what is classified as measurement error variance may not only depend on the variables and population of interest, but also on the frequency of measurements. If the occasion-specific effects are truly unique to the measurement occasion, as would be the case for truly random errors of measurement, we would expect the variance of the measurement errors to be stable regardless of the spacing or frequency of measurements we take. In practice it might also be possible, however, that part of what is classified as occasion-specific effects are true effects on the variable of interest, of which the effect dissipated before the next measurement occasion.[10] If this is the case, different intervals of measurement may result in different measurement error variances: For example, if one measures once every minute, these effects most likely carry over to the next measurement occasion through the autoregressive effect and will be part of the innovations, but if one measures once every hour, or once a day, or even once a week, such effects may not carry over to the next measurement occasion and become part of the error term ω_{ti} instead. Then, depending on the construct of interest and the frequency of measurements, the proportion of variance due to occasion-specific fluctuations may be considerable. With a different frequency of measurement, the proportion of measurement error variance may be different too.

If this is the case, it may be possible in practice to decrease the proportion of such "true" occasion-specific fluctuations in the model, by measuring more frequently so that the intervals between measurements become smaller, such that these fluctuations essentially become classified as innovations rather than measurement error. However, this is speculative, and will surely not always be practically possible, for instance, because it increases the burden on participants too much. Furthermore, the variance due to truly occasion specific effects, such as actual random measurement errors, will remain regardless of the measurement interval. Of course, to some extent, classical measurement errors such as making a mistake filling out a questionnaire, may be circumvented by designing better measurement instruments or improving the measurement conditions.

In practice, however, it seems likely that a proportion of the variance in our measurements will be the result of effects of which the influence is specific to one measurement occasion, including measurement error, next to effects that are carried over to multiple measurement occasions. Therefore, it would be prudent to account for both dynamic error and measurement error in VAR models, while allowing for interindividual differences in the reliabilities of the measurements of multiple individuals. The multilevel MEVAR(1) model discussed here may provide a relatively flexible environment for accomplishing these two goals. We hope that this work may contribute to the dynamic modeling field by raising awareness of both the

[10] This would imply that different external influences may affect the variable, and carry over at different rates.

consequences of disregarding measurement error, and the possibilities for accounting for measurement error, in dynamic VAR models.

References

Borsboom, D. (2003). *Conceptual issues in psychological measurement*. Enschede, the Netherlands: PrintPartners Ipskamp.

Borsboom, D., & Cramer, A. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9,* 91–121.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. Hoboken, NJ: Wiley.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., . . . Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8,* e60188. http://dx.doi.org/10.1371/journal.pone.0060188

Buonaccorsi, J. P. (2010). *Measurement error, models, methods, and applications*. Boca Raton, FL: Chapman & Hall/CRC.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 20,* 1–37.

Chatfield, C. (2004). *The analysis of time series: An introduction*. Boca Raton, FL: Chapman & Hall/CRC.

Cohn, J. F., & Tronick, E. (1989). Specificity of infants' response to mothers' affective behavior. *Adolescent Psychiatry, 28,* 242–248.

Cronbach, L. J. (1947). Test "reliability" its meaning and determination. *Psychometrika, 12,* 1–16.

Cronbach, L. J. (1990). *Essentials of psychological testing*. New York, NY: Harper & Row.

De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2014). Get over it! a multilevel threshold autoregressive model for state-dependent affect regulation. *Psychometrika,* 1–25.

Edmondson, D., Shaffer, J., Chaplin, W., Burg, M., Stone, A., & Schwartz, J. (2013). Trait anxiety and trait anger measured by ecological momentary assessment and their correspondence with traditional trait questionnaires. *Journal of Research in Personality, 47,* 843–852. http://dx.doi.org/10.1016/j.jrp.2013.08.005

Ferrer, E., Steele, J. S., & Hsieh, F. (2012). Analyzing the dynamics of affective dyadic interactions using patterns of intra- and interindividual variability. *Multivariate Behavioral Research, 47,* 136–171.

Ferrer, E., & Widaman, K. F. (2008). Dynamic factor analysis of dyadic affective processes with inter-group differences. In N. Card, J. Selig, & T. D. Little (Eds.), *Modeling dyadic and interdependent data in the developmental and behavioral sciences* (pp. 107–137). Hillsdale, NJ: Psychology Press.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7,* 457–511.

Gleser, L. J. (1992). The importance of assessing measurement reliability in multivariate regression. *Journal of the American Statistical Association, 87,* 696–707.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate behavioral research,* 1–22.

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

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: Cambridge University Press.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer.

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., . . . Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling, 23,* 532–543. http://dx.doi.org/10.1080/10705511.2016.1148605

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research, 50,* 334–349. http://dx.doi.org/10.1080/00273171.2014.1003772

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *ETS Research Report Series, 1970*(2), i–45. http://dx.doi.org/10.1002/j.2333-8504.1970.tb00599.x

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering, 82,* 35–45.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of consulting and clinical psychology, 63,* 52. http://dx.doi.org/10.1037/0022-006X.63.1.52

Kim, C.-J., & Nelson, C. R. (1999). *State–space models with regime switching*. Cambridge, MA: The MIT Press.

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

Lazarsfeld, P. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science*. New York, NY: McGraw-Hill.

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *Journal of Mathematical Psychology, 55,* 68–83.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston, MA: Addison Wesley.

Lucas, R., & Donnellan, M. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research, 105,* 323–331. http://dx.doi.org/10.1007/s11205-011-9783-z

Madhyastha, T., Hamaker, E. L., & Gottman, J. (2011). Investigating spousal influence using moment-to-moment affect data from marital conflict. *Journal of Family Psychology, 25,* 292–300.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response theory models. *Psychological Methods, 1,* 293–299.

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

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2,* 201–218.

Muthen, L. K., & Muthen, B. O. (2017). Mplus users guide (8th ed.) [Computer software manual]. Los Angeles, CA: Author.

Nezlek, J. B., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day-to-day psychological adjustment. *Personality and Social Psychology Bulletin, 27,* 1692–1704.

Oravecz, Z., & Tuerlinckx, F. (2011). The linear mixed model and the hierarchical Ornstein–Uhlenbeck model: Some equivalences and differences. *British Journal of Mathematical and Statistical Psychology, 64,* 134–160. http://dx.doi.org/10.1348/000711010x498621

Rovine, M. J., & Walls, T. A. (2005). A multilevel autoregressive model to describe interindividual differences in the stability of a process. In J. L. Schafer & T. A. Walls (Eds.), *Models for intensive longitudinal data* (pp. 124–147). New York, NY: Oxford University Press.

Schmittmann, V., Cramer, A., Waldorp, L., Epskamp, S., Kievit, R., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31,* 43–53. http://dx.doi.org/10.1016/j.newideapsych.2011.02.007

Schuurman, N. K. (2016). *Multilevel autoregressive modeling in psychology: Snags and solutions* (Unpublished doctoral dissertation). Utrecht University, Utrecht, the Netherlands.

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. http://dx.doi.org/10.1037/met0000062

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of Inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research, 51,* 185–206. http://dx.doi.org/10.1080/00273171.2015.1065398

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in $n = 1$ psychological autoregressive modeling. *Frontiers in Psychology, 6,* 1038. http://dx.doi.org/10.3389/fpsyg.2015.01038

Snippe, E., Bos, E. H., Ploeg, K. M., van der Sanderman, R., Fleer, J., & Schroevers, M. J. (2015). Time-series analysis of daily changes in mindfulness, repetitive thinking, and depressive symptoms during mindfulness-based treatment. *Mindfulness, 6,* 1053–1062. http://dx.doi.org/10.1007/s12671-014-0354-7

Snippe, E., Viechtbauer, W., Geschwind, N., Klippel, A., De Jonge, P., & Wichers, M. (2017). The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials. *Scientific Reports, 7,* 10.

Song, H., & Ferrer, E. (2012). Bayesian estimation of random coefficient dynamic factor models. *Multivariate Behavioral Research, 47,* 26–60.

Staudenmayer, J., & Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association, 100,* 841–852.

Stavrakakis, N., Booij, S. H., Roest, A. M., Jonge, P., de Oldehinkel, A. J., & Bos, E. H. (2015). Temporal dynamics of physical activity and affect in depressed and nondepressed individuals. *Health Psychology, 34,* 1268.

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

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625. http://dx.doi.org/10.1007/BF02289858

van der Krieke, L., Emerencia, A. C., Bos, E. H., Rosmalen, J., Riese, H., Aiello, M., . . . Jonge, P. (2015). Ecological momentary assessments and automated time series analysis to promote tailored health care: A proof-of-principle study. *JMIR Research Protocols*. Advance online publication. http://dx.doi.org/10.2196/resprot.4000

Walls, T. A., & Schafer, J. L. (2005). *Models for intensive longitudinal data*. New York, NY: Oxford University Press.

Wang, L., Hamaker, E. L., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods, 17,* 567–581.

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.

Appendix A

Mplus Code for the MEVAR(1) Model

Mplus Code

```
DATA: FILE = datafile.dat;
ANA1YSIS: TYPE IS TWOLEVEL RANDOM; estimator=bayes; fbiter=(120000);
          proc=2; bseed = 1 2;
SAVEDATA: bparameters=bp.dat; save = fs(1000); file=fs.dat;
          factors= all; data imputation: thin=100;
VARIABLE: names = clus pag pas;
              MISSING ARE *;
          CLUSTER = clus;

MODEL:

%WITHIN%
!making latent variables ytilde (here f1 and f2)
 f1 by pag@1(&1);
 f2 by pas@1(&1);
 [f1@0];
 [f2@0];
!random regression coefficients
 phi12 | f1 on f2&1;
 phi11 | f1 on f1&1;
```

(*Appendices continue*)

```
 phi21 | f2 on f1&1;
 phi22 | f2 on f2&1;

!random measurement error variances and covariances
 logv1 | pag;
 logv2 | pas;
 Z BY pag@1 pas@1;
 logvZ | Z;

!random innovation variances and covariances
 logv3 | f1;
 logv4 | f2;
 W BY f1@1 f2@1;
 logvW | W;

%BETWEEN%
!fixed effects and variances for random means
 [pag*6.5];
 pag*3;
 [pas*5];
 pas*5;

!fixed effects and variances for random regression coefficients
 [phi11*.5];
 phi11*.01;
 [phi22*.5];
 phi22*.01;
 [phi12*0];
 phi12*.01;
 [phi21*0];
 phi21*.01;

!fixed effects and variances for random measurement error (co)variances
 [logv1*0];
 [logv2*0];
 logv1*.15;
 logv2*.15;
 logvZ*.05;
 [logvZ*.25];

!fixed effects and variances for innovation (co)variances
 [logv3*0];
 [logv4*0];
 logv3*.15;
 logv4*.15;
 logvW*.05;
 [logvW*.25];

! correlations between the random means and regression coefficients
! no correlation between random (co)variances and other parameters for simplicity
pag pas phi11 phi22 phi12 phi21 with pag*0 pas*0 phi11*0 phi22*0 phi12*0 phi21*0;

PLOT:TYPE IS PLOT1 PLOT2 PLOT3;
 OUTPUT:TECH1 TECH8;
```

(*Appendices continue*)

Appendix B

Parameter Values Used for Generating Graphs A and B in Figure 2

	Graph A	Graph B
$\boldsymbol{\Phi}_i$	$\begin{bmatrix} .6 & .0 & .0 & .0 & .0 \\ .0 & .61 & .3 & .3 & .3 \\ .0 & .0 & .62 & .0 & .0 \\ .0 & .0 & .0 & .63 & .0 \\ .0 & .0 & .0 & .0 & .64 \end{bmatrix}$	$\begin{bmatrix} .5 & .25 & .25 & .25 & .0 \\ .0 & .5 & .0 & .0 & .0 \\ .0 & .0 & .4 & .0 & .0 \\ .0 & .0 & .0 & .5 & .0 \\ .0 & .0 & .0 & .0 & .4 \end{bmatrix}$
$\boldsymbol{\Sigma}_{\omega i}$	$\begin{bmatrix} .5 & -.15 & -.1 & -.1 & -.1 \\ -.15 & .5 & .2 & .2 & .21 \\ -.1 & .2 & .5 & .2 & .2 \\ -.1 & .2 & .2 & .5 & .15 \\ -.1 & .2 & .2 & .15 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & .25 & .25 & .25 & .25 \\ .25 & .5 & .1 & .1 & .1 \\ .25 & .1 & .5 & .05 & .05 \\ .25 & .1 & .05 & .5 & .05 \\ .25 & .1 & .05 & .05 & .5 \end{bmatrix}$
$\boldsymbol{\Sigma}_{\epsilon i}$	$\begin{bmatrix} .5 & -.28 & -.25 & -.25 & -.25 \\ -.28 & .5 & .15 & .15 & .17 \\ -.25 & .15 & .5 & .15 & .1 \\ -.25 & .15 & .15 & .5 & .1 \\ -.25 & .17 & .1 & .1 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & .25 & .25 & .25 & .25 \\ .25 & .5 & .1 & .1 & .1 \\ .25 & .1 & .5 & .05 & .05 \\ .25 & .1 & .05 & .5 & .05 \\ .25 & .1 & .05 & .05 & .5 \end{bmatrix}$
$\boldsymbol{rel}_w$	$\begin{bmatrix} .65 & -.04 & .06 & .07 & .07 \\ .17 & .87 & .02 & .02 & .01 \\ .12 & .08 & .59 & .01 & .03 \\ .11 & .09 & .01 & .60 & .00 \\ .12 & .08 & .04 & .00 & .60 \end{bmatrix}$	$\begin{bmatrix} .93 & -.04 & -.08 & -.06 & -.08 \\ .11 & .71 & -.07 & -.07 & -.08 \\ .08 & -.05 & .68 & -.05 & -.04 \\ .09 & -.05 & -.06 & .69 & -.05 \\ .06 & -.06 & -.04 & -.04 & .70 \end{bmatrix}$
$\hat{\boldsymbol{\Phi}}_i$	$\begin{bmatrix} .39 & -.02 & .04 & .04 & .04 \\ .21 & .60 & .20 & .19 & .19 \\ .07 & .05 & .36 & .01 & .02 \\ .07 & .05 & .00 & .38 & .00 \\ .08 & .05 & .02 & .00 & .38 \end{bmatrix}$	$\begin{bmatrix} .54 & .13 & .10 & .11 & -.08 \\ .06 & .35 & -.04 & -.04 & -.04 \\ .03 & -.02 & .27 & -.02 & -.02 \\ .05 & -.03 & -.03 & .35 & -.02 \\ .02 & -.02 & -.02 & -.02 & .28 \end{bmatrix}$

(*Appendices continue*)

Appendix C

Parameter Values Used for Generating Graphs C and D in Figure 2

	Graph C	Graph D
$\boldsymbol{\Phi}_i$	$\begin{bmatrix} .7 & 0 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & 0 \\ .3 & .3 & .3 & .5 & .3 \\ 0 & 0 & 0 & 0 & .6 \end{bmatrix}$	$\begin{bmatrix} .6 & 0 & 0 & 0 & .2 \\ 0 & .7 & 0 & .2 & 0 \\ 0 & .25 & .6 & -.35 & 0 \\ .2 & .3 & -.35 & .6 & 0 \\ 0 & 0 & 0 & 0 & .7 \end{bmatrix}$
$\boldsymbol{\Sigma}_{\omega i}$	$\begin{bmatrix} .5 & .1 & .1 & .1 & .1 \\ .1 & .5 & .1 & .1 & .1 \\ .1 & .1 & .5 & .1 & .1 \\ .1 & 1 & .1 & .5 & .1 \\ .1 & .1 & .1 & .1 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & .0 & .0 & .0 & .0 \\ .0 & .5 & .0 & .0 & .0 \\ .0 & .0 & .5 & .0 & .0 \\ .0 & .0 & .0 & .5 & .0 \\ .0 & .0 & .0 & .0 & .5 \end{bmatrix}$
$\boldsymbol{\Sigma}_{\epsilon i}$	$\begin{bmatrix} .5 & .25 & .1 & .1 & .1 \\ .25 & .5 & .25 & .25 & .25 \\ .1 & .25 & .5 & .1 & .1 \\ .1 & .25 & .1 & .5 & .05 \\ .1 & .25 & .1 & .05 & .5 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1.4 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0 & 1.8 & 0 \\ 0 & 0 & 0 & 0 & .8 \end{bmatrix}$
rel_w	$\begin{bmatrix} .65 & -.15 & .00 & .10 & .00 \\ -.06 & .62 & -.08 & .04 & -.07 \\ .00 & -.11 & .58 & .06 & .01 \\ .06 & -.16 & .02 & .80 & .08 \\ -.01 & -.13 & .00 & .09 & .61 \end{bmatrix}$	$\begin{bmatrix} .47 & -.01 & .00 & .02 & .07 \\ -.02 & .56 & -.04 & .23 & .00 \\ .00 & -.03 & .54 & -.16 & .00 \\ .03 & .29 & -.24 & .58 & .01 \\ .06 & .00 & .00 & .00 & .54 \end{bmatrix}$
$\hat{\boldsymbol{\Phi}}_i$	$\begin{bmatrix} .46 & -.10 & .00 & .07 & .00 \\ -.03 & .31 & -.04 & .02 & -.03 \\ .00 & -.06 & .29 & .03 & .01 \\ .21 & -.01 & .16 & .49 & .21 \\ .00 & -.08 & .00 & .06 & .37 \end{bmatrix}$	$\begin{bmatrix} .29 & -.01 & .00 & .01 & .15 \\ .00 & .45 & -.07 & .27 & .00 \\ -.01 & .02 & .40 & -.25 & .00 \\ .11 & .35 & -.34 & .48 & .02 \\ .04 & -.00 & .00 & .00 & .38 \end{bmatrix}$

Received January 20, 2017
Revision received February 6, 2018
Accepted February 27, 2018 ■