Psychological Methods
2016, Vol. 21, No. 2, 206–221

© 2016 American Psychological Association
1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000062

How to Compare Cross-Lagged Associations in a Multilevel Autoregressive Model

Noémi K. Schuurman
Utrecht University

Emilio Ferrer
University of California Davis

Mieke de Boer-Sonnenschein
Eliagg Institute, Uithoorn, The Netherlands

Ellen L. Hamaker
Utrecht University

By modeling variables over time it is possible to investigate the Granger-causal cross-lagged associations between variables. By comparing the standardized cross-lagged coefficients, the relative strength of these associations can be evaluated in order to determine important driving forces in the dynamic system. The aim of this study was twofold: first, to illustrate the added value of a multilevel multivariate autoregressive modeling approach for investigating these associations over more traditional techniques; and second, to discuss how the coefficients of the multilevel autoregressive model should be standardized for comparing the strength of the cross-lagged associations. The hierarchical structure of multilevel multivariate autoregressive models complicates standardization, because subject-based statistics or group-based statistics can be used to standardize the coefficients, and each method may result in different conclusions. We argue that in order to make a meaningful comparison of the strength of the cross-lagged associations, the coefficients should be standardized within persons. We further illustrate the bivariate multilevel autoregressive model and the standardization of the coefficients, and we show that disregarding individual differences in dynamics can prove misleading, by means of an empirical example on experienced competence and exhaustion in persons diagnosed with burnout.

Keywords: multilevel autoregressive model, cross-lagged associations, intensive longitudinal data, dynamical modeling, time series

Supplemental materials: http://dx.doi.org/10.1037/met0000062.supp

Many questions in psychological research are concerned with the way two or more variables influence each other over time. Examples of such research questions are "How do concentration and job satisfaction influence each other?," "How does maternal stress influence a child's behavior, and vice versa?," and "How do anxiety and rumination influence each other?" Many of these questions cannot be investigated experimentally because of ethical limitations—and, as a result, researchers make use of correlational designs such as the cross-lagged panel design. In this approach, two or more variables are measured at two or more occasions, and the cross-lagged associations between the variables over time are examined while controlling for the effect that variables have on themselves (i.e., the autoregression; cf. Rogosa, 1980).

An important goal in many cross-lagged panel studies is to establish causal effects using cross-lagged regression coefficients, and then comparing these associations with respect to their strength (e.g., Christens, Peterson, & Speer, 2011; de Jonge et al., 2001; de Lange, Taris, Kompier, Houtman, & Bongers, 2004; Kinnunen, Feldt, Kinnunen, & Pulkkinen, 2008; Talbot et al., 2012). The strongest association is then judged to provide the most important causal influence that drives the system, which is also referred to as being "causally dominant" (cf. de Jonge et al., 2001; de Lange et al., 2004; Kinnunen et al., 2008). By taking multiple repeated measures and incorporating them in the cross-lagged model, two requisites for establishing causal relations are fulfilled, namely, establishing an association between the variables studied and taking into account the time order of the processes (e.g., the cause has to occur before the result). Such an association between variables, in which a variable x predicts future values of another variable y, is referred to as "Granger-causal": Variable x Granger-causes variable y (Granger, 1969).

Of course, establishing Granger-causal relations is not enough to infer a true causal relationship, or true "causal dominance," as that would require ruling out that any of these associations may be

This article was published Online First January 25, 2016.

Noémi K. Schuurman, Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University; Emilio Ferrer, Department of Psychology, University of California Davis; Mieke de Boer-Sonnenschein, Eliagg Institute, Uithoorn, The Netherlands; and Ellen L. Hamaker, Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University.

This study was supported by the Netherlands Organization for Scientific Research (NWO; VIDI Grant 452-10-007). We are grateful to Laura Bringmann for comments on the manuscript.

Correspondence concerning this article should be addressed to Noémi K. Schuurman, Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, P.O. Box 80140, 3508 TC, Utrecht, The Netherlands. E-mail: n.k.schuurman@uu.nl

spurious. However, comparing the relative strength of the Granger-causal cross-lagged associations can provide direction for studying cross-lagged associations in more depth. For instance, consider a treatment study in which rumination and stress have a reciprocal cross-lagged relation. Specifically, the association between rumination at a previous occasion and stress at a later occasion is much stronger than the association between stress at a previous occasion and rumination at a later occasion. In this situation, it may be most efficacious to focus most on the former association in further research, or in practice during therapy. The question in this and related scenarios is how such a comparison of the strength of the cross-lagged associations can best be made. The common approach is to standardize the cross-lagged regression coefficients—and then compare their absolute values (Bentler & Speckart, 1981).

In recent years, several alternatives for the cross-lagged panel design have gained popularity, including the experience sampling method (ESM), daily diary measurements, and ambulatory assessment. These methods result in more intensive longitudinal data (often with more than 30 repeated measurements per person), which are also more densely spaced in time (i.e., day-to-day, moment-to-moment, or even second-to-second), thus containing more detailed information about the process under investigation. Researchers, being aware of the richness of these data, are trying to find alternative ways to analyze them in order to extract as much information as possible. This has led to the implementation of autoregressive models and multilevel extensions of these models (Cohn & Tronick, 1989; Kuppens, Allen, & Sheeber, 2010; Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011; Madhyastha, Hamaker, & Gottman, 2011; Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001; Suls, Green, & Hillis, 1998).

The combination of intensive longitudinal designs and multilevel modeling has two important advantages over more traditional cross-lagged panel studies. First, handling the data in a multilevel model allows one to separate the within-person (WP) dynamics from stable between-person (BP) differences. This is essential for investigating the actual dynamics of a psychological (causal) process that operates at the WP level (Hamaker, Kuijper, & Grasman, 2015). Second, disregarding individual differences in dynamics and only investigating group effects can be misleading when these results are generalized to the Granger-causal processes that happen at a WP level (Borsboom, Mellenbergh, & van Heerden, 2003; Hamaker, 2012; Kievit et al., 2011; Molenaar, 2004; Nezlek & Gable, 2001). Multilevel modeling allows for group effects and investigating whether there are individual differences in the dynamics, for instance, in the reciprocal relations. For example, for some individuals, having higher levels of concentration may lead to more job satisfaction, whereas for others, having higher levels of concentration is mostly a result of their high levels of job satisfaction. Similarly, when investigating mother–child dyads, for some dyads, maternal stress may be the dominant force that affects the child's disruptive behavior, whereas for other dyads, the child's disruptive behavior is what triggers maternal stress.

For making a meaningful direct comparison of the strengths of the reciprocal associations, standardized coefficients are more suitable than unstandardized coefficients. The unstandardized coefficients vary in size depending on the variances of the variables, whereas standardized coefficients are reflective of the proportions of unique variance explained in an outcome variable by the predictors. However, there are different ways to standardize the coefficients, based on different variances for the variables (i.e., the total variance, the WP variance, and the BP variance), which may lead to different conclusions about which effect is the strongest. This issue is further complicated by the fact that one can standardize the fixed effects (i.e., the average parameters across individuals), but also the individual parameters (i.e., for each person separately).

The purpose of this study is therefore twofold: First, we illustrate the value of the multilevel model in studying Granger-causal cross-lagged relations, and the individual differences therein. Second, we examine how the cross-lagged parameters from multilevel bivariate autoregressive models can be standardized and discuss the substantive interpretation of these standardized parameters when the aim is to compare the relative strength of cross-lagged associations. We begin by introducing the multivariate multilevel autoregressive model. Next, we discuss the rationale of standardization in general and how to standardize the parameters of the multilevel multivariate autoregressive model. This is followed by an illustration of the model and the standardization procedure on an empirical data set. We end the article with a discussion in which we highlight our main conclusions.

The Multilevel Bivariate Autoregressive Model

In this section, we explicate how the multilevel multivariate autoregressive model is related to other cross-lagged models, respectively, the cross-lagged panel model and the autoregressive ($n = 1$) time series model. After this comparison, we discuss the specification of the multilevel model.

Relation to Other Cross-Lagged Models

The multilevel multivariate autoregressive model we consider throughout this work has strong links to the cross-lagged panel model, on the one hand, and the ($n = 1$) bivariate autoregressive time series model, on the other hand. Both cross-lagged panel models and multivariate autoregressive time series models are used to study Granger-causal processes of multiple variables and to establish which effect is causally dominant. As such, both models incorporate autoregressive coefficients, which represent the effect of a variable on itself at the next time point, and cross-lagged coefficients, which reflect the effects of variables on each other at the next time point. However, these models were developed for different types of data and, as a result, provide different perspectives.

Specifically, the cross-lagged panel model is fitted to panel data, which generally consist of a few repeated measures (two to five) taken from a large number of participants. The autoregressive effects of the cross-lagged panel model indicate how stable the individual differences in the scores are over time. The cross-lagged effects reflect the association between the individual differences of one variable with the individual differences of another variable at the next occasion. An advantage of the cross-lagged panel model is that it is fitted for a group of individuals at once, and in that sense, is easy to generalize to a larger population. On the other hand, these effects do not necessarily generalize to the dynamic process for any specific individual (Borsboom et al., 2003; Hamaker, 2012; Hamaker et al., 2015; Kievit et al., 2011; Molenaar, 2004)—first, because the cross-lagged model does not separate stable BP differences (differences in the intercepts) from the WP effects (cf. Hamaker et al., 2015), and second, because the panel model provides average group effects, and average effects do not necessarily apply to the individual effects the average was taken over. This is illustrated further in the empirical application.

On the other extreme, autoregressive time series models are fitted to one person who is repeatedly measured over time (e.g., 50 repeated measures or more; Hamilton, 1994; Madhyastha et al., 2011). The autoregressive effects in this model tell us how a specific individual's past measures influence his or her current measures. The cross-lagged effects tell us how past scores of one variable influence the current scores of another variable after controlling for all autoregressive effects. As such, it describes the intraindividual differences or dynamics for a specific person. A disadvantage of the time series approach is that because these models are fitted to one individual at a time, the results are hard to generalize to a larger population.

The multilevel autoregressive model that we consider here allows us to model the WP processes as in the $n = 1$ time series model simultaneously for multiple individuals, and to model group effects that allow us to generalize results to a larger population. Specifically, at the WP or first level, the time series model is specified to describe the dynamics of the process for each individual, whereas at the BP or second level, the individual differences in these dynamics are captured. As such, the multilevel autoregressive model provides a way to combine the best of two worlds. On the one hand, the model is an extension of the cross-lagged panel model, simply incorporating many more repeated measures, and allowing the intercepts or means and the regression coefficients to vary across persons. On the other hand, it can be seen as an extension of the $n = 1$ time series model, with the added assumption that the person-specific parameters come from a particular distribution; the characteristics of this distribution, such as its mean or variance, can then be used to say something about the average effects in the group of individuals. We discuss the specification of this model in more detail in the Model Specification section.

Model Specification

Let y_{1ti} and y_{2ti} represent the scores on variable y_1 and variable y_2 of person i at occasion t. Each score can be separated into two parts: (a) a trait part μ_{1i} and μ_{2i}, which remains stable over time and can be thought of as the individual's means or trait-score on the variables y_1 and y_2; and (b) a state part $\tilde{y}_{1ti}$ and $\tilde{y}_{2ti}$, which represents the individual's temporal deviations from the person's trait scores. In vector notation this can be expressed as

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

The temporal deviations $\tilde{y}_{1ti}$ and $\tilde{y}_{2ti}$ may depend on preceding deviations. For example, consider determination and self-confidence within individuals: If a person's determination or self-confidence is strong at a particular time point, this may also likely be the case at the following time point. Such relationships are modeled with autoregressive parameters ϕ_{1i}, ϕ_{2i}, which indicate how each variable y_1 and y_2 affects itself over time. A positive autoregression can be interpreted as the inertia—resistance to change—of the process (Kuppens et al., 2010; Suls et al., 1998): With a positive autoregressive effect, the current level of confidence will partly carry over to future levels of confidence, and as a result, when confidence is high at one occasion, it will only slowly revert back to baseline levels. If the autoregressive parameter is close to zero, this indicates that y does not depend much on its previous value so that it is hard to predict future values of confidence from past values of confidence, and that if confidence is high at one occasion, the process will return relatively quickly back to baseline levels. A negative autoregressive effect indicates that if y is high at one occasion, it is likely to be low at the next. Such an autoregressive association may be expected, for instance, for processes that concern daily intake, for instance, of the number of calories or the number of alcoholic beverages (e.g., Rovine & Walls, 2006).

Besides the autoregressive effects of determination and self-confidence on themselves, current high levels of determination may, for instance, also lead to subsequent high self-confidence, and, in turn, current high self-confidence may lead to elevated levels of determination. Such cross-lagged relationships can be investigated by adding cross-lagged regression parameters ϕ_{12i}, ϕ_{21i} to the model, which reflect the associations of variable y_1 and y_2 at time $t - 1$ with each other at time t for person i. The 2 × 2 matrix Φ_i contains the autoregression coefficients ϕ_{1i}, ϕ_{2i} for each variable on the diagonal, and the cross-lagged coefficients ϕ_{12i}, ϕ_{21i} on the off-diagonals for person i. In vector notation, this model can then be expressed as

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

$$\begin{bmatrix} \epsilon_{1ti} \\ \epsilon_{2ti} \end{bmatrix} \sim MvN\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right\}. \tag{3}$$

The innovations, ϵ_{ti}, reflect the effect of perturbations on the system by anything that is not explicitly measured and modeled for person i at time point t.[1] An elevation of determination as a result of reading an inspiring book is an example of what an innovation may (partly) represent. As can be seen from Equation 3, we assume that the innovations are normally distributed with means of zero, and covariance matrix Σ. Note that autoregressive models are stationary, a consequence of which is that the means and covariance structure of the outcome variables are fixed over time for each person. This results in certain restrictions on the matrix with regression parameters Φ_i, specifically that the eigenvalues of the matrix should lie within the unit circle (see Hamilton, 1994, p. 259).

The model as defined in Equations 1–3 forms level one of the multilevel model. The subject index i shows that the means, as well as the autoregressive and cross-lagged regressive parameters, are allowed to vary across persons. In the multilevel context, we assume such individual parameters come from a distribution, with a mean that is referred to as the fixed effect (denoted by $\boldsymbol{\gamma}$), and a person-specific part that is referred to as the random effect. We model this at level two as the vector $[\mu_{1i}, \mu_{2i}, \phi_{1i}, \phi_{12i}, \phi_{21i}, \phi_{2i}]'$

[1] Note that the innovations are not the same as measurement errors. For details, and on incorporating measurement error in $n = 1$ AR models, we refer to Schuurman, Houtveen, and Hamaker (2015). Multilevel AR modeling with measurement error is currently a work in progress.

(where $'$ indicates the transpose), which has a multivariate normal distribution with mean vector $[\gamma_{\mu 1}, \gamma_{\mu 2}, \gamma_{\phi 1}, \gamma_{\phi 12}, \gamma_{\phi 21}, \gamma_{\phi 2}]'$, and 6×6 covariance matrix Ψ. The fixed effects $\boldsymbol{\gamma}$ reflect the average individual autoregressive and cross-lagged effects, and the variances $\psi^2_{\mu 1}, \psi^2_{\mu 2}, \psi^2_{\phi 1}, \psi^2_{\phi 12}, \psi^2_{\phi 21}, \psi^2_{\phi 2}$ from the covariance matrix Ψ reflect the variation of the individual parameters around this mean. The variances of the person-specific means $\psi^2_{\mu 1}$ and $\psi^2_{\mu 2}$ are also referred to as the *BP variances* for variable y_1 and y_2 because they reflect the variance in the trait scores across persons. The covariances in matrix Ψ reflect the associations between the person-specific parameters. For instance, if persons with relatively high average confidence generally also have relatively high levels of determination compared with persons with lower levels of confidence, this would be reflected in a positive correlation between μ_{1i} and μ_{2i} (i.e., covariance element [1,2] in matrix Ψ). Finally, note that constraining all the effects to be fixed in the multilevel autoregressive model (i.e., constraining the elements of Ψ to zero, so that there are no random effects) leads to a cross-lagged panel model, whereas the model defined at level one is identical to an $n = 1$ autoregressive time series model.

Fitting the Model With Bayesian Techniques

We fit the multilevel bivariate autoregressive model and estimate the accompanying standardized coefficients (discussed in the following section) using Bayesian modeling techniques. There are several reasons for opting for a Bayesian approach here. First, in contrast to most multilevel software packages, software based on Bayesian analysis is very flexible with respect to model specification, and thus allows us to fit the complete bivariate model simultaneously. Second, it directly supplies the estimates for the fixed effects in the model, as well as the individual parameters, which are needed to standardize the results. Third, in Bayesian modeling, it is easy to calculate additional quantities, such as the standardized regression coefficients, and take into account the uncertainty about these new quantities, which is expressed in the posterior standard deviations and credible intervals for these quantities (i.e., the Bayesian equivalents of the standard error and confidence intervals in frequentist statistics; cf. Gelman, Carlin, Stern, & Rubin, 2003; Hoijtink, Klugkist, & Boelen, 2008). For an introduction to Bayesian statistics, we refer the interested reader to Gelman et al. (2003) and Hoijtink et al. (2008). We provide example R and WinBUGS code for simulating data based on the multilevel VAR model, fitting the model, and standardizing the parameters in the online supplemental materials (note that the model code can also be used within the software Openbugs and JAGS; Lunn, Spiegelhalter, Thomas, & Best, 2009; Lunn, Thomas, Best, & Spiegelhalter, 2000; Plummer, 2003). In Appendix B, we provide information on the prior specifications and convergence for the empirical application.

Standardizing Cross-Lagged Regression Coefficients in Order to Compare the Strength of Cross-Lagged Associations

In this section, we discuss the standardization of the regression coefficients in the multilevel bivariate autoregressive model in order to compare the relative strength of the cross-lagged effects. However, whereas some researchers will have no hesitation regarding the use of standardized coefficients for comparing the relative strength of associations, others may prefer to retain the original measurement scale, as they consider it to be more meaningful than a standardized scale. Because this is a key issue in the current discussion, we begin this section with an argument for the use of standardized parameters. We then discuss conceptual differences between multiple methods for the standardization of the parameters in the multilevel bivariate autoregressive model, and which method to use.

A Rationale for Standardized Cross-Lagged Parameters

An unstandardized regression parameter indicates the expected increase in measurement units in the outcome variable if the predictor were to increase by one measurement unit. Standardized regression parameters are parameters that would have been obtained if the variables had been standardized before the analysis, that is, if for each observation the mean was subtracted, and this centered score was divided by the standard deviation of the relevant variable. A standardized regression parameter indicates the expected increase in standard deviations in the outcome variable if the predictor were to increase by one standard deviation. A standardized parameter b^* can be calculated from the unstandardized parameter b, using the standard deviations of the predictor ω_x and the outcome variable ω_y, that is, $b^* = b\omega_x/\omega_y$.

Unstandardized parameters are generally considered unsuited for comparing relative strengths of associations because they are sensitive to differences in measurement units. When the sizes of these parameters are directly compared to infer which effect is the strongest, conclusions about the relative strength of the associations may change depending on what measurement unit was used for some of the variables. Of course, this is undesirable because the true underlying relative strength of the relationships does not change as a result of an arbitrary choice of measurement unit. Further note that even if two variables are measured on the same measurement scale, they may not have the same variances and therefore may not have equally large effects on the system, even if they have the same unstandardized parameters. Consider, for instance, the exchange in affect between a man and a woman in a relationship. If both individuals have the same cross-lagged effects on each other, but the woman is much more variable in her scores then the man is, then, in practice, she will produce more change in the dynamic system than he will. These differences in how likely a variable is to increase one unit are not taken into account when using unstandardized parameters.

In contrast, standardized coefficients are not sensitive to measurement units and take into account differences in the variances of the variables because they have standard deviations as units (see also Hunter & Hamilton, 2002; Luskin, 1991). As a result, the standardized parameters are reflective of the amount of unique explained variance in a dependent variable per predictor variable. The standardized coefficients are therefore often considered more suitable for comparing the relative strengths of associations than unstandardized regression coefficients.

It is important to note that when the predictor variables in a regression model are independent from each other, the squared standardized regression parameters represent the proportion of total variance in the outcome variable that is explained by each predictor variable (Cohen, Cohen, West, & Aiken, 2003). As a consequence, the predictor variable with the largest standardized parameter has, on occasion, been deemed the predictor variable that is "relatively most important" in the model. This interpretation has, however, been a point of controversy in the literature (e.g., see Blalock, 1967; Darlington, 1968; Greenland, Maclure, Schlesselman, Poole, & Morgenstern, 1991; King, 1986, 1991), because when the predictor variables are dependent, they partly explain the same variance in the outcome variable, and it is no longer possible to determine how much variance is accounted for by each separate predictor (cf. Cohen et al., 2003, pp. 64–79). When considering cross-lagged regression parameters, the predictor variables are almost always correlated, as are their residuals, such that—even if one could be sure that the lagged relations represent true causal mechanisms—the standardized parameters generally do *not* indicate which variable explains the most variance in the *model as a whole* or which variable is therefore "relatively most important" in the model as a whole. The standardized regression parameter is, however, a reflection of the proportion of *unique* explained variance, that is, the amount of variance in the outcome variable that is not shared with any of the other predictors.[2] As such, the standardized regression parameters indicate which predictor variable has the *strongest direct relationship* with an outcome variable, or has the most unique explained variance, regardless of the (in)dependence of the predictor variables.

WP, BP, and Grand Standardization in Multilevel Models

In order to establish the reciprocal cross-lagged effects and determine which variables have the strongest Granger-causal associations, we would like to compare the strength of each cross-lagged effect for each individual, but also for the group of individuals as a whole. For instance, in the context of a network of depression symptoms, we may want to know whether feelings of anxiety, anhedonia, or sleep problems are the strongest driving force within the network for a specific person (cf. Borsboom & Cramer, 2013; Bringmann et al., 2013). In addition, we would like to be able to determine, in general, across all individuals, which variable has the strongest cross-lagged effect. Therefore, we are interested in standardizing the individual cross-lagged parameters, but also in obtaining the standardized fixed effects from the multilevel bivariate autoregressive model.

Although standardization may be considered trivial in regression analysis, it is less straightforward in multilevel modeling. In fact, there are three ways to standardize the parameters from a two-level multilevel model: WP standardization, grand standardization, and BP standardization. For all three methods, the person-specific standardized cross-lagged coefficients ϕ^*_{jki} are calculated as the product of the person-specific unstandardized coefficient ϕ_{jki}, and the ratio of the standard deviations of the predictor variable y_k and the outcome variable y_j. Considering the standardization of the fixed effects, we believe that like the unstandardized fixed effects, the standardized fixed effects should reflect the average person-specific relations. That is, for each of the three methods, we determine the standardized fixed effects by taking the expectation with respect to the standardized person-specific parameters. However, WP, grand, and BP standardization are based on different standard deviations for the predictor variables and outcome variables, so that each method results in different standardized person-specific and fixed effects (Heck & Thomas, 2000). In the following, we will discuss the three methods in more detail. An overview of the different variances for each method and the equations for the person-specific parameters and fixed effects for each method are presented in Table 1.

WP standardization. WP standardization (also referred to as within-group standardization when the data consist of persons clustered in groups), is based on standardizing the parameters for each individual separately with their individual WP variances. Conceptually, the WP variance for a certain variable for a specific individual can be seen as the variance of that specific individual's repeated measures for that variable. That is, the WP variance for a specific variable for individual i is based solely on his or her person-specific parameters, as would be the case for an $n = 1$ autoregressive model. The WP variances ω^2_{1i} and ω^2_{2i} for variables y_1 and y_2 are the diagonal elements of the person-specific covariance matrix Ω_i. Based on the regression equation in Equation 2, this covariance matrix can be expressed as $\Omega_i = \Phi_i \Omega_i \Phi'_i + \Sigma$ (where Φ'_i is the transpose of matrix Φ_i). However, this not helpful in practice, because this equation includes Ω_i at both sides of the equation. To obtain an expression for the WP covariance matrix Ω_i in terms of Φ and Σ only, we can make use of the following expression instead:

$$\Omega_i = mat((I - \Phi_i \otimes \Phi_i)^{-1} vec(\Sigma)), \tag{4}$$

where $\otimes$ indicates the Kronecker product, function *vec*() transforms a matrix into a column vector, and *mat*() returns this vector back into a matrix (Kim & Nelson, 1999, p. 27).

As can be seen from Table 1, the WP standardized person-specific parameters equal the unstandardized person-specific parameters ϕ_{jki} multiplied by the ratio of the WP standard deviations ω_{ji} and ω_{ki}. The person-specific WP standardized parameters reflect the number of *person-specific standard deviations* that the dependent variable will increase, when the independent variable increases one *person-specific standard deviation*. Thus, given that the unstandardized cross-lagged parameters for a certain person are equal, the standardized parameter will be the largest for the predictor variable that varies the most *within that person over time*.

The WP standardized fixed effects are equal to the expectation of the person-specific parameters. The person-specific WP standardized parameters are a function of three dependent random variables (that vary across persons i), ϕ_{jk} (normally distributed), and ω_j and ω_k (both with a distribution of unknown form), so that the distribution of these parameters is not of a known form. Therefore, the fixed effects cannot be simplified further from $\underset{i}{E}\left[\phi_{jk}\frac{\omega_k}{\omega_j}\right]$, and should be calculated based on the person-specific standardized parameters. That is, $\underset{i}{E}\left[\phi_{jk}\frac{\omega_k}{\omega_j}\right] \neq \gamma_{\phi jk}\frac{\underset{i}{E}[\omega_k]}{\underset{i}{E}[\omega_j]}$, so that the WP standardized fixed effect *should not* be calculated using the

[2] When there are two predictors, x_1 and x_2, for the outcome variable y, the standardized regression parameter for the first predictor can be expressed in terms of correlation parameters using $b^*_1 = (r_{y1} - r_{12}r_{y2})/(1 - r^2_{12})$, where r_{y1} is the correlation between y and x_1; r_{12} is the correlation between x_1 and x_2; and r_{y2} is the correlation between y and x_2. The proportion of uniquely explained variance is equal to the squared semipartial correlation, which is expressed as $r_{y(1.2)} = (r_{y1} - r_{12}r_{y2})/\sqrt{(1 - r^2_{12})}$. Although this relationship is more complicated than taking the square of the standardized regression parameter, a larger (absolute) standardized regression parameter implies a larger proportion of unique explained variance.

Table 1
Equations for the Variances for WP, BP, and Grand Standardization for the Standardized Person-Specific Parameters and Fixed Effect Parameters

	WP	BP	Grand
Variance	ω_i^2	ψ_μ^2	$\underset{i}{E}[\omega^2] + \psi_\mu^2$
ϕ^*_{jki}	$\phi_{jki}\frac{\omega_{ki}}{\omega_{ji}}$	$\phi_{jki}\frac{\psi_{\mu k}}{\psi_{\mu j}}$	$\phi_{jki}\frac{\sqrt{\underset{i}{E}[\omega_k^2] + \psi_{\mu k}^2}}{\sqrt{\underset{i}{E}[\omega_j^2] + \psi_{\mu j}^2}}$
$\gamma^*_{\phi jk}$	$\underset{i}{E}\left[\phi_{jk}\frac{\omega_k}{\omega_j}\right]$	$\gamma_{\phi jk}\frac{\psi_{\mu k}}{\psi_{\mu j}}$	$\gamma_{\phi jk}\frac{\sqrt{\underset{i}{E}[\omega_k^2] + \psi_{\mu k}^2}}{\sqrt{\underset{i}{E}[\omega_j^2] + \psi_{\mu j}^2}}$

Note. The person-specific standardized parameters ϕ^*_{jki} are the product of the unstandardized parameters ϕ_{jki}, and the ratio of the standard deviation of the predictor variable k and the standard deviation of the outcome variable j. The standardized fixed effects $\gamma^*_{\phi jki}$ are calculated by taking the Expectation ($\underset{i}{E}[]$) over the standardized person-specific parameters for all persons i. The term ω_i^2 indicates the person-specific variance based on Equation 4. For variable j, this variance is referred to as ω_{ji}^2, and for variable k, as ω_{ki}^2. The term ψ_μ^2 indicates the variance of the person-specific means. For variable j, this becomes $\psi_{\mu j}^2$, and for variable k, this becomes $\psi_{\mu k}^2$. The term $\underset{i}{E}[\omega^2]$ indicates the expectation taken over all the person-specific variances ω_i^2. WP = within-person; BP = between-person.

unstandardized fixed effect $\gamma_{\phi jk}$ and the average WP standard deviations $\underset{i}{E}[\omega_j]$ and $\underset{i}{E}[\omega_k]$. The latter would disregard the dependencies between the random variables ϕ_{jk}, and ω_j and ω_k. It is unclear how different the results will be using $\gamma_{\phi jk}\frac{\underset{i}{E}[\omega_k]}{\underset{i}{E}[\omega_j]}$ rather than $\underset{i}{E}\left[\phi_{jk}\frac{\omega_k}{\omega_j}\right]$, because of the complicated nature of their dependencies and distribution, which will depend on many different parameters.

We estimate the standardized parameters as part of the Bayesian model fitting procedure. The person-specific standardized parameters can be estimated in a number of ways. One approach is to calculate the standardized coefficients directly in each Markov Chain Monte Carlo (MCMC) iteration, based on the estimated unstandardized regression coefficients using the equations in Table 1. Another, seemingly pragmatic, approach is to standardize the observed variables and fit the multilevel VAR model to these standardized data, resulting in standardizing regression coefficients. Note that in the case of WP standardization, these two methods for obtaining the standardized parameters rely on different assumptions about the distributions of the unstandardized and standardized parameters, and therefore may lead to different results. When the model described in the previous section is fitted to unstandardized data, the individual unstandardized parameters are assumed to be normally distributed. Then, the WP standardized individual regression parameters, which are a function of three dependent random variables, ϕ_{jk}, ω_k, and ω_j, will not be normally distributed. In contrast, if we fit the multilevel model to standardized data, we would assume that the standardized regression parameters, rather than the unstandardized regression parameters, are normally distributed. In this case, the unstandardized regression parameters have a distribution of unknown form. Hence, standardizing the data and standardizing the regression parameters are associated with different model assumptions, and thus may lead to different results. As such, when estimating the WP standardized and unstandardized coefficients, one needs to choose one of the two assumptions and calculate the coefficients that are not directly estimated (standardized or unstandardized) a posteriori using the equations in Table 1.

We opt here for the assumption that the unstandardized parameters are normally distributed and calculating the standardized coefficients afterward, for two reasons. First, this normality assumption is in line with conventions in multilevel research. Second, for standardizing either the data or the regression coefficients, estimates are needed for the person-specific means and the WP variances. When we standardize the coefficients a posteriori, we can easily use model-based parameter estimates for this purpose, rather than having to rely on sample means and variances. This is preferable because model-based estimates usually provide somewhat better estimates than sample statistics (given that the used model is correct), especially for smaller sample sizes, and, importantly, using model-based estimates allows us to take the uncertainty about these estimates into account.

Hence, to estimate the person-specific WP standardized parameters, we first calculate the WP variances based on Equation 4, within each iteration of the MCMC procedure. Subsequently, we calculate the person-specific standardized coefficients by means of the equation in Table 1 (i.e., the equation in the first column, second row), also within each iteration of the MCMC procedure. The standardized fixed effects are estimated by calculating the average person-specific standardized coefficient in each iteration of the MCMC procedure, because the distribution of the person-specific coefficients is of unknown form and thus cannot be taken analytically. In this way, a posterior distribution is obtained for each of the WP standard deviations, the standardized person-specific coefficients, and the standardized fixed effects, which can be used to derive point estimates, credible intervals, and posterior standard deviations for the standardized coefficients. The R code in the online supplemental materials includes example code for WP standardizing the cross-lagged coefficients.

Note that any other relevant statistics besides the standardized coefficients can be calculated in a similar way. For instance, we will make use of this in the empirical example, in which we determine the proportion of individuals for whom ϕ^*_{i12} is larger than ϕ^*_{i21} in each iteration of the MCMC procedure, resulting in a posterior distribution for the proportion of persons who have a larger directed association between Variable 1 at occasion t and Variable 2 at occasion t – 1, than between Variable 1 at occasion t – 1 and Variable 2 at occasion t.

Grand standardization. Grand standardization is based on the *grand or total variances*, which consist of the average WP variances (the average of all the person-specific variances), and the BP variances (the variances of the person-specific means). Conceptually, the grand variance is the variance taken over all the repeated measures for all individuals. The grand variances are the diagonal elements g_j^2 of the grand covariance matrix $\boldsymbol{G}$ (for which the derivation can be found in Appendix A),

$$\boldsymbol{G} = \underset{i}{E}[\Omega] + \psi_\mu, \tag{5}$$

where $\underset{i}{E}[\Omega]$ is the expected value taken over the person-specific covariance matrices Ω_i, and ψ_μ is the BP covariance matrix, that is, the covariance matrix of the person-specific means.

The person-specific parameters for grand standardization are simply the unstandardized regression parameters, each multiplied by a constant—the ratio of the grand standard deviations, as can be seen in Table 1. As a result, the grand standardized person-specific parameters will be normally distributed, just like the unstandardized parameters. Thus, the grand standardized fixed effects are equal to the product of the unstandardized fixed effect $\gamma_{\phi jk}$ and the ratio of the grand standard deviations (see Table 1). The grand standardized parameters reflect the number of grand standard deviations the outcome variable increases when the predictor variable increases one grand standard deviation. Thus, when the unstandardized parameters are equal, the predictor variable for which the combination of the BP variance and average WP variance is the largest will have the largest standardized parameter, and thus will be deemed to have the strongest Granger-causal effect.

The grand standardized parameters can be estimated by calculating the grand variances (which includes calculating the WP covariance matrix Ω_i for each person, and then calculating the average of the WP variances across all persons to estimate $\underset{i}{E}[\Omega]$) and standard deviations, and calculating the grand standardized person-specific coefficients and fixed effects by means of the equations in Table 1, in each iteration of the MCMC procedure. We have included example R code for grand standardization in the online supplemental materials.

BP standardization. BP standardization is based on the BP variance ψ^2_μ , the variability in the person-specific means μ_i across persons. In other words, BP standardization is based on the difference between the grand variance and the average WP variance. For BP standardization, the person-specific parameters are simply the unstandardized regression parameters, each multiplied by the ratio of the BP standard deviations. As a result, the BP standardization person-specific parameters will be normally distributed, and the fixed effects for BP standardization are equal to the product of the unstandardized fixed effect and the ratio of the BP standard deviations (see Table 1).

In BP standardization, the standardized parameters reflect the number of standard deviations for the person-specific means the dependent variable would increase, if the predictor variable increases one standard deviation of the person-specific means. This implies that when the unstandardized cross-lagged coefficients are equal, the BP standardized parameter will be the largest for the predictor variable for which *the person-specific means vary the most across persons*.

The BP variances are estimated as part of the multilevel VAR model, as discussed in the previous section, in which we presented the multilevel model. In order to estimate the BP standardized parameters, one would calculate the BP standard deviations in each iteration of the MCMC procedure based on these estimated BP variances, and then calculate the person-specific and fixed standardized coefficients in each iteration by means of the equations in Table 1. The supplementary materials include R code for BP standardizing the cross-lagged coefficients.

WP standardization, BP standardization, and grand standardization lead to different results. In general, WP, BP, and grand standardization will lead to different numerical results, and may lead to different conclusions about the relative strength of the Granger-causal effects. Differences between the grand, BP, and the WP standardized parameters will arise from differences between the respective variances used for standardization.

The first, most apparent difference between the WP variance and the grand and BP variance is that the latter two both include the BP variance—the variance of the random mean across persons for the variable in question—whereas the WP variance does not. As a result, differences between the grand standardized parameters and the WP standardized parameters will arise when the ratios of the BP variances, and the ratios of the WP variances, for the two variables are quite different. This may lead to different conclusions concerning the relative strength of the cross-lagged associations, both for the standardized random parameters and the standardized fixed effects.

We illustrate this point in Figure 1, in which simulated WP, grand, and BP standardized random and fixed parameters are plotted. A point above the plotted diagonal line indicates that the standardized coefficient ϕ^*_{12} is larger than the standardized coefficient ϕ^*_{21} , implying that Variable 2 is causally dominant; a point below the diagonal line indicates that ϕ^*_{21} is larger than ϕ^*_{12}, implying that Variable 1 is causally dominant. From both plots, it can be seen that grand, WP, and BP standardization do not give the same results: The plotted squares (grand standardization), circles (WP standardization), and triangles (BP standardization) do not match in location. In addition, in both cases, grand and BP standardization result in different conclusions from WP standardization: Grand standardization and BP stan-

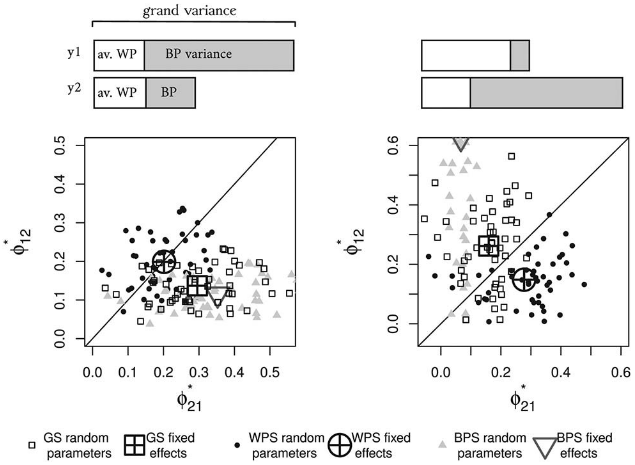

Figure 1. Plots of simulated individual cross-lagged parameters and accompanying fixed effects, with squares indicating grand standardized (GS) parameters, triangles indicating between-person standardized (BPS) parameters, and circles indicating within-person standardized (WPS) parameters. The area below the diagonal line implies that the association between y_1 with future y_2 is the strongest, and the area above the diagonal line implies that the association between y_2 with future y_1 is the strongest. When the ratios of the BP variances, grand variances for y_1 and y_2, and the WP variances for y_1 and y_2 are different, BP, grand, and WP standardization will result in different conclusions. These ratios for the simulated data are depicted above each of the two plots: The white rectangle indicates the average WP variance, and the gray rectangle indicates the BP variance; the total of these rectangles is equal to the grand variance. The fixed effects of the unstandardized parameters $\gamma_{\phi12}$ and $\gamma_{\phi21}$ were equal to .2 for both plots. The variances for ϕ_{i12} and ϕ_{i21} were equal to .005 for the first plot, and .01 for the second plot. av. WP = average within-person variance; BP variance = between-person variance.

dardization show that $\phi^*_{21} > \phi^*_{12}$ for most persons, whereas WP standardization shows that this is the case for some persons and that the reverse is true for other persons. Therefore, the grand and BP standardized fixed effects indicate that, on average, $\phi^*_{21} > \phi^*_{12}$, whereas the WP standardized fixed effects indicate that they are (approximately) equally large on average. The discrepancy between WP standardization and BP and grand standardization can be explained by the fact that the average WP variances for y_1 and y_2 are about equally large, whereas the BP variance for y_1 is larger than for y_2, resulting in a larger grand variance for y_1.

Similarly, in the second plot in Figure 1, the WP variances for y_2 are smaller than for y_1, whereas the grand variance for y_2 is larger than for y_1, and the unstandardized parameters are equal. As a result, BP and grand standardization actually give opposite conclusions from WP standardization, both for the individual parameters and the fixed effects: For grand and BP standardization, ϕ^*_{12} is larger than ϕ^*_{21}, whereas for WP standardization, it is the reverse. Finally, note that although grand and BP standardization both result in the conclusion that ϕ^*_{12} is larger than ϕ^*_{21}, the difference between ϕ^*_{12} and ϕ^*_{21} is much more extreme for BP standardization, because the difference between the BP variances for variables y_1 and y_2 is more extreme than the difference between the two grand variances.

A second difference between the WP variance, and the BP variance and grand variance, is that the latter two are fixed: They are the same for each person, whereas the WP variance varies across persons. Therefore, differences between the grand standardized and WP standardized parameters can arise when the person-specific variances deviate from the average person-specific variances, even if the BP variances were equal to zero. If the average WP variance for variable x is larger than the average WP variance for variable y, but for a specific individual, the WP variance for variable x is smaller than for variable y, this can result in opposite conclusions about the relative strengths of the cross-lagged associations based on the grand standardized and WP standardized parameters for that individual. Clearly, the method of standardization has the potential to strongly influence the results and, subsequently, the conclusions regarding which cross-lagged association is the strongest.

Why WP Standardization Should Be Preferred

Currently, there seems to be no consensus on the optimal standardization approach for comparing the relative strength of effects in the multilevel literature. For instance, Nezlek (2001) cautions against WP standardization, indicating that it seems more complicated and may result in different p values than those obtained for the unstandardized coefficients. Heck and Thomas (2000) state that all forms of standardization may be useful, depending on what variance one is interested in, but they do not specify why one variance may be more interesting than the others or which should be preferred in what situation. Notably, in later editions, this information on standardization has been removed. Many dedicated multilevel software packages, such as, HLM (Raudenbush et al., 2011), lme4 in R (Bates, Maechler, Bolker & Walker, 2015), and SPSS Mixed (IBM Corp, 2013), currently do not include the option to obtain standardized coefficients, perhaps because of the lack of consensus on how to standardize coefficients in multilevel models (see also Heck & Thomas, 2000, p. 134, for a software overview). An exception is the multilevel software STREAMS, for which the manual explicitly recommends and provides grand standardization (Gustafsson & Stahl, 2000, p. 118). Another exception is the structural equation modeling software Mplus (Muthén & Muthén, 1998 –2015), which also features multilevel modeling and standardizes "to the variance on within for within relationships," that is, Mplus seems to use WP standardization, although it is unclear exactly how this is achieved and what is the case for the fixed effects (Heck & Thomas, 2000; Muthén, 2008).[3]

The lack of consensus may partly be a result of different researchers having different modeling backgrounds: Some may take a bottom-up perspective based on an $n = 1$ time series modeling background, whereas others may take a top-down perspective based on a cross-lagged panel modeling background. Here, WP standardization can be considered to be in line with standardization in classical time series models: Given that, in this context, only one subject is modeled, the only way to standardize is by using the WP variances. In contrast, grand standardization is more in line with standardization as would be performed in cross-lagged panel models, in which the variances are naturally calculated across the scores of all persons, disregarding potential differences in variances between persons and the distinction between BP variance and WP variance (cf. Hamaker et al., 2015). Hence, for researchers who have a background in cross-lagged panel modeling and researchers who have a background in $n = 1$ modeling, different methods of standardization may seem more natural.

We argue that when standardization is used to compare the relative strength of different predictors, WP standardization should be the preferred approach. The reason for this is that we are interested in Granger-causal psychological processes, which happen within persons, at the level of the individual. It does not seem reasonable to conflate this WP variation with variation between persons, given that the person-specific Granger-causal processes are not concerned with differences in the means of these processes between individuals. Rather, we would like to obtain standardized coefficients with a similar interpretation as we would in a single-subject study.

To elaborate on this, consider a multilevel study on the effects of anxiety of mothers and that of their children on each other. Suppose that the person-specific mean levels for anxiety vary much more across different mothers than the person-specific mean levels vary across the children: That is interesting to consider in itself. However, if the interest is in determining how a mother's anxiety influences that of her child (and vice versa)—that is, the interest is in Granger-causal WP (or within-dyad) effects—such stable differences between persons are not directly relevant. Therefore, the cross-lagged regression parameters that reflect the personal Granger-causal processes that happen within persons (a dyad) should not be convolved. BP standardization (and, therefore, grand standardization) does convolve this information. Specifically, the BP standardized coefficient for the Granger-causal effect of the mother's anxiety on that of her child reflects the number of standard deviations for the average levels of anxiety across children in the population that the score of this child increases, when the child's mother's anxiety increases one standard deviation in the average levels of anxiety across mothers in the population. Yet

[3] This is based on Mplus forum responses by Muthén (2008).

there is no reason to suppose that the strength of the effect of a specific mother's anxiety on that of her child, or vice versa, over time is related in such a way with how the average level of anxiety differs across all mothers in the population, nor by how the average level of anxiety differs across all children in the population. Therefore, the standardized WP effects should not include the BP variance. This is the case for WP standardization, but not for grand or BP standardization.

Further note that the (unstandardized) cross-lagged effects reflect the increase in the dependent variable, given a unit increase in the predictor variable for a specific person. That is, the predictors explain the variation in the dependent variables that occurs within a specific person—not across different persons. As such, the standardized cross-lagged coefficients are only indicative of the proportion of uniquely explained variance for WP standardization, not for BP and grand standardization.

Finally, WP standardization takes into account that each individual may have a unique variance for each variable, whereas grand and BP standardization are based on standardizing with the same variance for each person and, as such, disregard this person-specific information. For these reasons, we prefer WP standardization over grand and BP standardization.

Empirical Illustration: Burnout Data

In this section, we begin by presenting an empirical data set concerned with moment-to-moment WP measurements of symptoms of burnout. After that, we apply the multilevel autoregressive model to the burnout data and present the unstandardized results. After that, we present and interpret the WP standardized results, and, finally, we compare these results with those obtained using BP and grand standardization.

Burnout Data: An ESM Study

Two core components of burnout are severe exhaustion and diminished experienced personal competence (Maslach & Jackson, 1981; Maslach, Jackson, & Leiter, 1996; World Health Organization, 2008). Our empirical application is concerned with the way these two components influence each other over time. Both exhaustion and competence were measured with multiple items, each scored on a 7-point Likert scale (Sonnenschein, Sorbi, van Doornen, & Maas, 2006; Sonnenschein, Sorbi, van Doornen, Schaufeli, & Maas, 2007). We obtained sum scores across the items "I feel tired," "I feel exhausted," "I feel dead tired," "I feel lethargic," "I feel energetic" (reversed), and "I feel fit" (reversed), to represent the current state of exhaustion, and sum scores across the items "I feel competent right now," "What I'm doing right now I can handle well," and "This activity is going well for me," to represent the current state of competence. Data were collected using experience sampling for a period of 2 weeks for 54 individuals with burnout. Each day the participants were alerted randomly throughout the day to fill out their questionnaire, and they filled in their diary right before sleep and after waking. For exhaustion, this resulted in an average of 80 repeated measures per person, and for competence, this resulted in an average of 40 repeated measures per person, as the latter was only measured during the day, but not in the morning after waking or in the evening before bedtime.[4]

Investigating whether and how experienced exhaustion and competence affect themselves and each other over time can provide important information for further research and the treatment of individuals with burnout. We are specifically interested in whether the association of competence at time $t - 1$ with exhaustion at time t is stronger or weaker than the association of exhaustion at time $t - 1$ with competence at time t. In more traditional longitudinal designs, such research questions are usually handled with cross-lagged panel modeling, using a few repeated measurements obtained from a large sample of individuals, which will result in a description of how BP differences of these variables are related over time. However, we want to understand the actual individual dynamics at play—whether competence and exhaustion affect each other, and if so, which one is the driving force in the perpetuation of burnout. Furthermore, we want to know whether and how this may differ across persons. The current rich data set allows us to model these WP processes and to investigate whether there are BP differences in these processes. For instance, it may be the case that for some individuals the association between current exhaustion and future competence is the strongest, whereas for others it is the reverse. Obtaining insights in such differences is desirable, as it would allow for a more person-tailored—and hopefully more effective—treatment of burnout.

Modeling Moment-to-Moment Exhaustion and Competence

To fit the model presented in Equations 1 to 3, we make use of WinBUGS (in combination with R, and the R packages r2winbugs and CODA), which is free software for conducting Bayesian analyses (Lunn et al., 2000; Plummer, Best, Cowles, & Vines, 2006; R Development Core Team, 2012; Sturtz, Ligges, & Gelman, 2005). The WinBUGS code we used for the modeling procedure for both applications can be found in the online supplemental materials, together with R code for simulating example data, fitting the model with WinBUGS, and standardizing the model using WP, BP, and grand standardization (note that WinBUGS model code can also be used within the software Openbugs and JAGS; Lunn et al., 2009; Plummer, 2003). Information on the prior specifications and convergence of the procedure can be found in Appendix B.

We present the unstandardized results for the multilevel bivariate autoregressive model discussed previously. Taking a bottom-up perspective, we will start by discussing the results for two of the 53 individuals from the burnout sample, whom we will refer to as "Arnold" and "Peter," in order to show how the

[4] Note that as a result of the general setup of ESM data collection, and because participants do not fill out the diaries during the night, the distance between measurements is not the same across all repeated measurements. Equidistant time intervals are an assumption of discrete time series models, including the one presented in this article. To correct for this, we added missing observations to the data set when the interval was particularly large (>5 hr, which occurred mostly at night when participants slept), which resulted in time intervals of, on average, 2.3 hr ($SD = 1.1$). After adding these missing values, the average rate of missing data was .55 ($SD = .17$). This should limit the effects of the nonequidistant time intervals. An alternative option is to use continuous time models, which do not require the assumption of equally spaced observations, assuming instead that the process changes continuously over time. However, current multilevel extensions of continuous time models have strong limitations: Either the lagged effects are fixed (Voelkle, Oud, Davidov, & Schmidt, 2012) or the random cross-lagged effects are assumed to be equal within a person (Oravecz, Tuerlinckx, & Vandekerckhove, 2009), which is clearly undesirable in the current context.

multilevel model can lead to different dynamics for individuals. We contrast the results for these individuals with the average results—the fixed effects. The estimated unstandardized model parameters for the burnout data can be found in Table 2.

In Figure 2, the observations for two individuals, Arnold and Peter, are plotted against time, where the black line represents scores on exhaustion and the gray line represents scores on competence. Breaks in these lines indicate a missing value at that measurement occasion. The dotted lines indicate Arnold and Peter's estimated mean scores on exhaustion and competence. The estimated individual regression parameters for Arnold and Peter are displayed in Figure 3. Both Arnold and Peter have large, positive autoregressive coefficients for exhaustion (i.e., .436, 95% credible interval [CI] [.228, .644], and .316, 95% CI [.102, .524], respectively), which implies that if they feel exhausted at one occasion, they are likely to feel similarly exhausted at the next occasion, and if they feel fit at one occasion, they are likely to feel fit the next occasion. As a result, when Arnold's or Peter's process of exhaustion is perturbed by a sudden late night, causing their exhaustion to increase, their exhaustion will only slowly return to baseline. This also holds for Arnold's feelings of competence (i.e., autoregression of .404, 95% CI [.124, .646]), implying that his feelings of high competence tend to last for some time, whereas his feelings of incompetence also tend to last for some time. In contrast, Peter's autoregressive parameter for competence is close to zero (i.e., .021, 95% CI [−.276, .269]), which implies that his current state of competence does not depend on his preceding state of competence. The differences between these two cases show the importance of allowing for individual variation in parameters: The inertia in these two components of burnout is not invariant across individuals.

When considering the cross-lagged relations, we see that, for Peter, the relation between past exhaustion with current competence is negative (i.e., −.289, 95% CI [−.530, −.078]), implying that higher levels of exhaustion lead to lower levels of competence. The relation of past competence with current exhaustion, however, is relatively close to zero (i.e., −.145, 95% CI [−.428, .159]), which indicates that there is little evidence that his feelings of competence predict his exhaustion. For Arnold, both cross-lagged coefficients are relatively close to zero (i.e., $\phi_{Cei} = -.179$, 95% CI [−.416, .086]; $\phi_{Eci} = -.099$, 95% CI [−.357, .155]), indicating that there is little evidence that his feelings of competence predict his exhaustion the next occasion, or vice versa. Again, this illustrates that there can be important individual differences in the dynamics of such processes.

Table 2
Unstandardized Parameter Estimates for the Multilevel Bivariate Autoregressive Model Studying the Association Between Exhaustion and Competence in Individuals Diagnosed With Burnout

Parameter	Median estimate [95% CI]
$\gamma_{\mu E}$	3.971 [3.787, 4.154]
$\gamma_{\mu C}$	4.919 [4.753, 5.085]
$\gamma_{\phi E}$	.427 [.367, .484]
$\gamma_{\phi C}$	.157 [.066, .248]
$\gamma_{\phi CE}$	−.091 [−.158, −.023]
$\gamma_{\phi EC}$	−.020 [−.110, .071]
$\psi^2_{\mu E}$	.413 [.280, .633]
$\psi^2_{\mu C}$	.336 [.227, .515]
$\psi^2_{\phi E}$	.024 [.013, .043]
$\psi^2_{\phi C}$	.047 [.024, .089]
$\psi^2_{\phi CE}$	.025 [.014, .048]
$\psi^2_{\phi EC}$	.066 [.035, .116]
σ^2_E	.787 [.749, .827]
σ^2_C	.742 [.697, .791]
σ_{CE}	−.324 [−.360, −.288]
ρ_{CE}	−.424 [−.460, −.385]

Note. CI = credible interval.

In addition to considering the particular dynamics of individuals, it is also of interest to consider the average group effects in order to be able to generalize conclusions to a broader population. For this purpose, the fixed effects representing the average parameters, and the variances and covariances of the random effects—representing the amount of BP differences in the individual parameters—can be used. The estimated fixed effects for the regression coefficients are presented in Figure 4. We found positive average autoregression coefficients for both exhaustion and competence (i.e., $\gamma_{\phi E} = .427$, 95% CI [.367, .484]; $\gamma_{\phi C} = .157$, 95% CI [.066, .248]), indicating that, averaged across individuals, feelings of exhaustion tend to carry over strongly to next observations, whereas feelings of competence only carry over a little. The fixed cross-lagged effect from exhaustion on competence was negative but small (i.e., $\gamma_{\phi CE} = -.091$, 95% CI [−.158, −.023]), and the fixed effect from competence on exhaustion was approximately zero (i.e., −.019, 95% CI [−.11, .071]). Thus, on average, there is a pattern of higher exhaustion being followed by lower competence (and thus lower exhaustion being followed by higher competence), but there is little evidence to suggest that competence predicts exhaustion at the next occasion.

This could imply that changes in feelings of competence in individuals with burnout are mostly the result of feeling more or less exhausted, and that in treatment, it would be most beneficial to focus on exhaustion, rather than on the feelings of competence. However, two points are worth considering here. First, there is considerable variation in these effects across persons, which is reflected in the variances of the individual parameters (i.e., $\psi_{\phi CE} = .025$, 95% CI [.014, .048]; $\psi_{\phi EC} = .066$, 95% CI [.035, .116]). In fact, the individual coefficients range from −.308 to .111 for ϕ_{CE}, and from −.571 to .532 for ϕ_{EC}. As such, merely inspecting the fixed effects here is misleading. This exemplifies a large pitfall of cross-lagged panel designs to evaluate Granger-causal processes: Cross-lagged panel designs evaluate average effects over a group of individuals, ignoring potentially substantial individual differences, as is the case in this empirical illustration.

Second, to compare these cross-lagged effects, we need to determine their relative strength by WP standardizing the coefficients and comparing the resulting standardized coefficients, rather than the unstandardized coefficients reported here. In the following section, we discuss the standardized results for the burnout data.

Standardized Results for the Burnout Data

In the following, we first discuss and interpret the results for the WP standardized coefficients. After that, we compare these results with those obtained for BP and grand standardization.

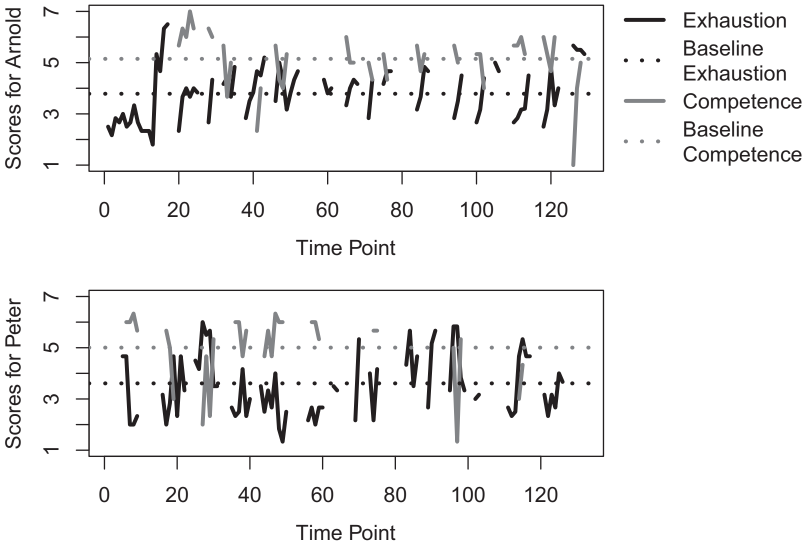

Figure 2. Time series for Arnold and Peter, representing their exhaustion (in black) and competence (in gray). Dotted lines represent the individuals' estimated mean scores μ_{Ci} and μ_{Ei}.

WP standardized results. The WP standardized cross-lagged coefficients for each participant and the corresponding fixed effects are displayed as circles in the left panel of Figure 5. For the unstandardized results, we find large individual differences in the unstandardized cross-lagged associations between exhaustion and competence. As can be seen from Figure 5, these individual differences are also reflected in the WP standardized coefficients. Specifically, the individual WP standardized coefficients range from −.265 to .103 for

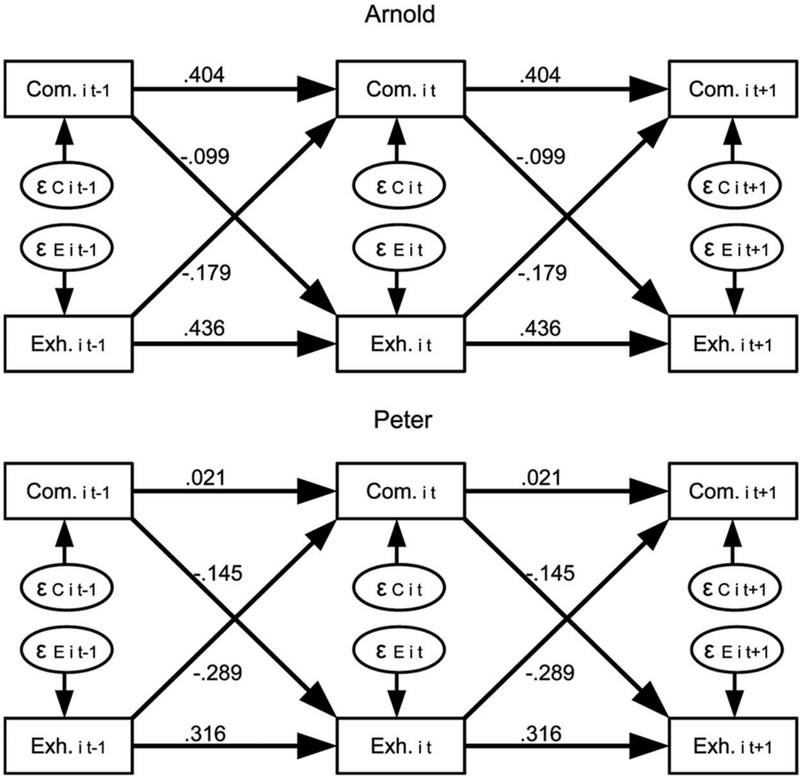

Figure 3. Estimated model parameters for the associations between Arnold and Peter's exhaustion and competence over time. Com. = competence; Exh. = exhaustion.

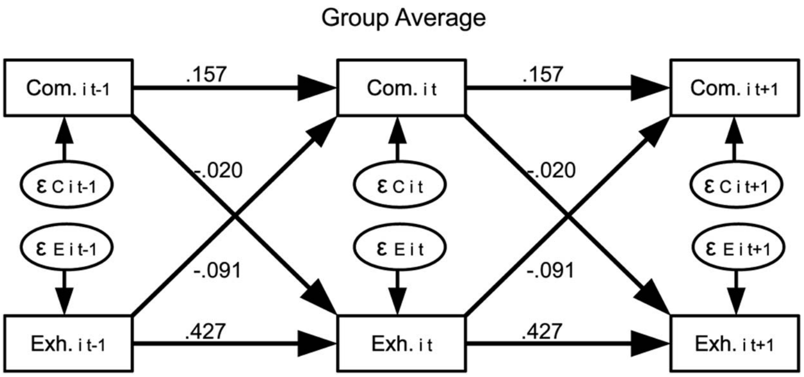

Figure 4. Estimated fixed effects for the multilevel autoregressive model studying the associations between exhaustion and competence for the group of individuals diagnosed with burnout. Com. = competence; Exh. = exhaustion.

ϕ^*_{CE}, and from −.471 to .54 for ϕ^*_{EC}. The fixed WP standardized effect of exhaustion on competence is equal to −.077 (95% CI [−.12, −.029]), and the fixed WP standardized effect for competence on exhaustion equal to −.004 (95% CI [−.065, .054]).

For some individuals, the effect of competence on exhaustion seems most likely to be positive; for others, the effect seems most likely to be negative. This has important implications for the use of the fixed effects: Given that some individuals have negative cross-lagged coefficients and others have positive cross-lagged coefficients, the fixed effects can give misleading results regarding which cross-lagged association is stronger on average, because the negative and positive coefficients cancel each other out. Therefore, we inspect the average absolute WP standardized cross-lagged coefficients (the absolute WP standardized fixed effects).

The absolute WP standardized cross-lagged coefficients for each individual and the fixed effects are displayed as circles in the right panel of Figure 5. It shows that for most persons in our sample, the effect of competence on exhaustion is stronger than that of exhaustion and competence. The absolute WP standardized fixed effect for the effect of exhaustion on competence is 0.121 (95% CI [.093, .154]), and the absolute fixed effect for the effect of competence on exhaustion is .197 (95% CI [.156, .235]). This indicates that the average effect of competence on exhaustion is *actually stronger* than the average effect of exhaustion on competence, whereas the nonabsolute fixed effects led to the opposite, misleading, conclusion.

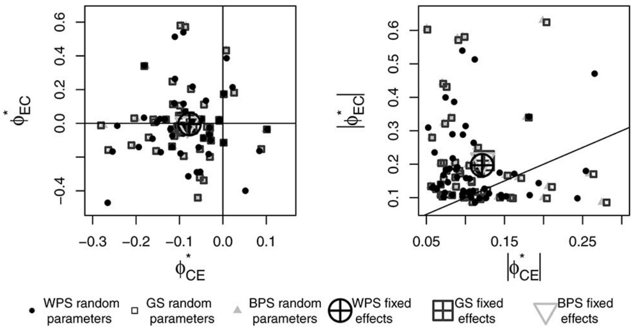

Figure 5. The left panel shows a plot of the point estimates of the within-person, between-person, and grand standardized random parameters and fixed effects for the cross-lagged associations between exhaustion and competence in individuals with burnout. The right panel shows the absolute values for the standardized random parameters and fixed effects. In the right panel, for individuals with estimated coefficients plotted below the diagonal line, the association of exhaustion with future competence is the strongest. For individuals with estimated coefficients above the diagonal line, the association of competence with future exhaustion is the strongest. WPS = within-person standardized; GS = grand standardized; BPS = between-person standardized.

Next to investigating which cross-lagged association is the largest on average, it is informative to investigate what proportion of the individuals has a larger (absolute) cross-lagged effect of competence at one occasion on exhaustion at the next occasion, and for what proportion the reverse is true. The estimated population proportion of individuals for whom the relationship between past competence and current exhaustion is larger than the relationship between past exhaustion and current competence is .66 (95% CI [.528, .774]).[5] This indicates that the cross-lagged effect of exhaustion on competence is not only weaker than the cross-lagged effect of competence on exhaustion on average across individuals, but this is also the case for the majority of the individuals.

In conclusion, there are large individual differences in the dynamics associated with burnout. We find that, for approximately 34% of persons diagnosed with burnout, the relation between feeling exhausted and subsequently feeling competent at the next occasion is stronger than the reverse. For most people, the relation between feeling competent and feeling exhausted at the next occasion is the strongest. Note, however, that for some persons, this relation is positive, whereas for others, it is negative. Perhaps, for some people, feeling good about themselves gives them a boost of energy, resulting in a negative relationship between competence and subsequent exhaustion, whereas for other people, feeling competent drives them to work harder, resulting in fatigue, and a positive relationship between competence and subsequent exhaustion.

[5] We estimate the population proportion of persons $|\phi^*_{i12}|$ is larger than $|\phi^*_{i21}|$ by calculating for each Gibbs sample for how many individuals $|\phi^*_{i12}| > |\phi^*_{i21}|$ and dividing this by the number of individuals, resulting in a posterior distribution for this proportion.

Comparison of the WP, BP, and grand standardized results. When we compare the coefficients that result from WP standardization with those that result from BP and grand standardization for the burnout data, we find that the results for the three methods are similar in some respects, but not identical. For some individuals, the standardized coefficients are markedly different for WP standardization compared with BP and grand standardization. However, for most individuals, the conclusions about which cross-lagged effect is the strongest are the same for the three methods. Note, however, that the conclusions about the strength of cross-lagged effects will not necessarily be the same for any other particular study, as demonstrated in Figure 1. Furthermore, the interpretations of the grand and BP standardized coefficients are not particularly sensible, so that we recommend against using them for comparing the strength of the cross-lagged effects in practice. In the following, we compare the WP, BP, and grand standardization results in more detail.

In the left panel of Figure 5, the standardized person-specific coefficients and fixed effects for each method are plotted together. As can be seen from this plot, the three standardization methods result in different values for the cross-lagged coefficients; the triangles, squares, and dots in the plot do not overlap perfectly. The differences between the three methods are the largest for coefficients that are farther away from zero (which is to be expected, because when the regression coefficients are near zero, the ratio of the variances will make a relatively small impact). The coefficients obtained with BP and grand standardization are generally very similar because the ratios of the grand standard deviations and BP standard deviations are very similar to each other. The differences between the coefficients obtained with BP and grand standardization and those obtained with WP standardization are larger, and can be quite large for some persons (see the bottom-most coefficients in the left panel of Figure 5). This indicates that for certain individuals, the ratio of the person-specific standard deviations is markedly different from the ratio of the grand standard deviations and the BP standard deviations. The fixed effects for WP, BP, and grand standardization are quite similar to each other, with for ϕ^*_{CE}, a fixed effect of −.083 (95% CI [−.128, −.035]) for grand standardization, −.082 (95% CI [−.138, .033]) for BP standardization, and −.077 (95% CI [−.12, −.029]) for WP standardization, and for ϕ^*_{EC}, −.021 (95% CI [−.083, .041]) for grand standardization, −.021 (95% CI [−.088, .042]) for BP standardization, and −.004 (95% CI [−.065, .054]) for WP standardization.

In the right panel of Figure 5, the absolute values of the standardized coefficients are plotted, where coefficients plotted below the diagonal line indicate that the association of exhaustion with future competence is the strongest, and coefficients plotted above this line indicate that the association of competence with future exhaustion is the strongest. As can be seen from this plot, for the majority of the individuals, and the fixed effects, the WP, BP, and grand standardized coefficients are on the same side of the diagonal line. This means that for most individuals and for the fixed effects, the conclusions about the strength of the cross-lagged associations are the same for grand, BP, and WP standardization. Specifically, the absolute fixed effects for ϕ^*_{CE} were 0.124 (95% CI [.093, .161]) for grand standardization, .122 (95% CI [.082, .179]) for BP standardization, and 0.121 (95% CI [.093, .154]) for WP standardization, and the absolute fixed effects for ϕ^*_{EC} were 0.211 (95% CI [.163, .257]) for grand standardization, 0.213 (95% CI [.147, .301]) for BP standardization, and .197 (95% CI [.156, .235]) for WP standardization. When we calculate the proportion of individuals for whom the effect of competence on future exhaustion is the strongest, we find a proportion of .623 (95% CI [.491, .775]) for grand standardization, .642 (95% CI [.434, .811]) for BP standardization, and .66 (95% CI [.528, .774]) for WP standardization.

Discussion

The aim of this study was twofold. First, we wanted to show the added value of the multilevel model in studying individual differences in Granger-causal cross-lagged relations. Second, we wanted to show how the cross-lagged parameters from multilevel bivariate autoregressive models should be standardized in order to compare the relative strength of these relations and study the individual differences therein. The ability to capture interindividual differences in WP processes is an important advantage of using multilevel time series modeling over techniques like cross-lagged panel modeling. Evaluating only average effects across persons can prove misleading because they do not necessarily apply to any specific individual. If we had focused only on the average (unstandardized or standardized) effects in the empirical example, we might have erroneously concluded that generally exhaustion has the strongest Granger-causal effect on competence, and that competence has no effect on exhaustion for persons diagnosed with burnout. However, the bottom-up multilevel approach allowed us to uncover large individual differences in the person-specific cross-lagged effects, with some persons having a positive association between past experienced competence and current exhaustion, and others having a negative association. Further, by standardizing the individual coefficients within each person, and then comparing the absolute WP standardized coefficients, we found that, actually, for most individuals, competence has the strongest effect on exhaustion rather than the reverse. We would not have established this if we had examined only fixed effects, as would have been the case in cross-lagged panel modeling. A next step in research could be to explain and predict the interindividual differences in the cross-lagged coefficients, and the interindividual differences in the relative strengths of these associations: Why is the effect positive for certain individuals and negative for others, and why is one association stronger than the other for certain individuals, whereas for others it is the opposite? These questions may be studied further by expanding the multilevel model by adding predictors for the random parameters.

We argued that in order to meaningfully compare the strength of cross-lagged associations and investigate potential interindividual differences herein, the estimated cross-lagged regression coefficients should be standardized within each person: firstly, because grand and BP standardization undesirably include the variance in means across persons in the process of standardization, whereas WP standardization does not; and secondly, because, WP standardization takes into account that each person may have unique standard deviations for the outcome and predictor variables, whereas the other methods of standardization—grand and BP standardization—do not. Although we focus here on the comparison of cross-lagged effects in a multilevel autoregressive model, we believe that the arguments to use WP standardization for comparing the relative strength of effects also generalizes to other multilevel models. Given that random effects are generally included to account for differences across subjects (be it persons, or groups, or classrooms, and so on), it makes sense to also account for these differences when comparing standardized coefficients by using WP standardization—even when the main interest is in the resulting fixed effects. The main appeal of the fixed effects is that they summarize the effects on the lower level—specifically, they reflect the average within-subject effects. As such, it is desirable that this interpretation of the fixed effects remains intact when comparing the strength of the fixed effects using the standardized fixed parameters. Given that the subject-specific parameters should be obtained by WP standardization, the standardized fixed effects should reflect the average WP standardized subject-specific parameters.

Finding out which direct effect is the strongest, and why, is valuable for providing direction in both further research, or in practice. Consider the effect of feeling competent on exhaustion, and vice versa, in the context of the treatment of burnout. For individuals for whom the effect of competence on exhaustion is the strongest for example, it may be most beneficial to focus on this relationship in treatment. This could be implemented by increasing the level of competence, and by altering the relationship between competence and exhaustion at the next occasion—for instance, by diminishing it if it is positive. Note that the focus here is not only on decreasing or increasing mean levels of variables but also on altering the harmful or beneficial relations between psychological variables, which may provide more resilience against negative events. This latter focus is central to the network perspective on psychological processes (e.g., Borsboom & Cramer, 2013; Bringmann et al., 2013; Schmittmann et al., 2013), a promising novel perspective in which psychological processes are conceptualized as networks of observed variables. The networks are represented in graphs, in which the reciprocal associations are displayed as arrows from the predictor variable to the dependent variable, and the strength of these relationships—inferred from the respective size of the cross-correlations or cross-lagged regression coefficients—is indicated in the graphs by the thickness of the arrows (Borsboom & Cramer, 2013; Bringmann et al., 2013; Schmittmann et al., 2013). In such a setting, comparing the relative strength of associations, and capturing individual differences herein, is one of the fundamental goals. How to compare the associations in a network in a meaningful way is an issue that has not received much attention thus far. We hope that the current article will contribute to this innovative area of research.

Of course, there are limitations to the use of standardized coefficients for comparing effects, especially when these are used to guide decisions concerning interventions in practice. For instance, standardization does not take into account how easily relevant associations or variables are manipulated in practice (by a clinician, for example) or how costly that would be. Further, when comparing cross-lagged coefficients, we are comparing the effects of predictors on two different dependent variables. The standardized coefficients may show which association is the strongest statistically; they do not take into account if changes in the dependent variables are equally important in practice. To illustrate, the standardized cross-lagged coefficients may indicate that the increase in standard deviations in stress associated with a standard deviation increase in depression is larger than the number of standard deviations increase in depression associated with a standard deviation increase in stress. However, a standard deviation increase in stress scores may be much less detrimental for the quality of life of a person than a standard deviation increase in depression scores. To complicate matters further, whether this is the case or not may also differ across persons.

Other aspects that were not considered in the current work are how to standardize models that include more than one lag, or how to standardize coefficients in nonstationary models. For a stationary model that includes multiple lags, one can simply calculate the standardized parameters based on the equations in Table 1, using the WP variances, and the raw coefficients for the relevant time lag. For nonstationary models, the standardization procedure becomes much more complex, given that the regression coefficients and variances may change over time. Another important question for future work is how to evaluate how change in one variable affects the system as a whole, considering multiple lags and variables, compared with other variables in the system, rather than comparing specific associations by comparing the standardizing coefficients directly, as was the focus here.

In summary, in this article, we showed how multilevel multivariate autoregressive models can be applied to psychological intensive longitudinal data, and that by standardizing the results within-person, the relative strengths of cross-lagged associations can be investigated. We believe that these techniques can provide an excellent basis for uncovering some of the hidden information in intensive longitudinal data, and we hope that these techniques will be applied more frequently to elucidate psychological processes.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Bentler, P. M., & Speckart, G. (1981). Attitudes “cause” behaviors: A structural equation analysis. *Journal of Personality and Social Psychology, 40,* 226–238.

Blalock, H. (1967). Causal inferences, closed populations, and measures of association. *The American Political Science Review, 61,* 130–136.

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.

Bringmann, L., 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.

Christens, B., Peterson, N., & Speer, P. (2011). Community participation and psychological empowerment: Testing reciprocal causality using a cross-lagged panel design and latent constructs. *Health Education & Behavior, 38,* 339–347.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

Darlington, R. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69,* 161–182.

de Jonge, J., Dormann, C., Janssen, P., Dollard, M., Landeweerd, J., & Nijhuis, F. (2001). Testing reciprocal relationships between job characteristics and psychological well-being: A cross-lagged structural equation model. *Journal of Occupational and Organizational Psychology, 74,* 29–46.

de Lange, A., Taris, T., Kompier, M., Houtman, I., & Bongers, P. (2004). The relationships between work characteristics and mental health: Examining normal, reversed and reciprocal relationships in a 4-wave study. *Work & Stress: An International Journal of Work, Health & Organisations, 18,* 149–166.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on an article by Browne and Draper). *Bayesian Analysis, 1,* 514–534.

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.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37,* 424–438.

Greenland, S., Maclure, M., Schlesselman, J., Poole, C., & Morgenstern, H. (1991). Standardized regression coefficients: A further critique and review of some alternatives. *Epidemiology, 2,* 387–392.

Gustafsson, J.-E., & Stahl, P. A. (2000). *Streams users guide, Version 2.5 for Windows*. Molndal, Sweden: MultivariateWare.

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

Hamaker, E. L., Kuijper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20,* 102–116.

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

Heck, R., & Thomas, S. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Erlbaum.

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

Hunter, J., & Hamilton, M. (2002). The advantages of using standardized scores in causal analysis. *Human Communication Research, 28,* 552–561.

IBM Corp. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Kievit, R., Romeijn, J., Waldorp, L., Wicherts, J., Scholte, H., & Borsboom, D. (2011). Mind the gap: A psychometric approach to the reduction problem. *Psychological Inquiry, 22,* 67–87.

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

King, G. (1986). How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science, 30,* 666–687.

King, G. (1991). "Truth" is stranger than prediction, more questionable than causal inference. *American Journal of Political Science, 35,* 1047–1053.

Kinnunen, M.-J., Feldt, T., Kinnunen, U., & Pulkkinen, L. (2008). Self-esteem: An antecedent or a consequence of social support and psychosomatic symptoms? Cross-lagged associations in adulthood. *Journal of Research in Personality, 42,* 333–347.

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

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

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine, 28,* 3049–3067.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337.

Luskin, R. (1991). Abusus non tollit usum: Standardized coefficients, correlations, and R2s. *American Journal of Political Science, 35,* 1032–1046.

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

Maslach, C., & Jackson, S. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior, 2,* 99–113.

Maslach, C., Jackson, S., & Leiter, M. (1996). *MBI: The Maslach Burnout Inventory: Manual*. Palo Alto, CA: Consulting Psychologists Press.

Moberly, N., & Watkins, E. (2008). Ruminative self–focus and negative affect: An experience sampling study. *Journal of Abnormal Psychology, 117,* 314–323.

Molenaar, P. (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.

Muthén, L. K. (2008, August 17). *Standardized solutions*. Retrieved from http://www.statmodel.com/discussion/messages/12/542.html?1380892370

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Authors.

Nezlek, J. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin, 27,* 771–785.

Nezlek, J., & Allen, M. (2006). Social support as a moderator of day-to-day relationships between daily negative events and daily psychological well-being. *European Journal of Personality, 20,* 53–68.

Nezlek, J., & Gable, S. (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., & Vandekerckhove, J. (2009). A hierarchical Ornstein–Uhlenbeck model for continuous repeated measurement data. *Psychometrika, 74,* 395–418.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Retrieved from https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6,* 7–11.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., Bryk, A. S., Cheong, A. S., Fai, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88,* 245–258.

Rovine, M., & Walls, T. (2006). A multilevel autoregressive model to describe interindividual differences in the stability of a process. In J. Schafer & T. Walls (Eds.), *Models for intensive longitudinal data* (pp. 124–147). New York, NY: Oxford.

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.

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (in press). A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*.

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in $n = 1$ psychological autoregressive modeling. *Frontiers in Psychology, 6,* 1038.

Sonnenschein, M., Sorbi, M., van Doornen, L., & Maas, C. (2006). Feasibility of an electronic diary in clinical burnout. *International Journal of Behavioral Medicine, 13,* 315–319.

Sonnenschein, M., Sorbi, M., van Doornen, L., Schaufeli, W., & Maas, C. (2007). Electronic diary evidence on energy erosion in clinical burnout. *Journal of Occupational Health Psychology, 12,* 402–413.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software, 12,* 1–16.

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

Talbot, L., Stone, S., Gruber, J., Hairston, I., Eidelman, P., & Harvey, A. (2012). A test of the bidirectional association between sleep and mood in bipolar disorder and insomnia. *Journal of Abnormal Psychology, 121,* 39–50.

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods, 17,* 176–192.

World Health Organization. (2008). *ICD-10: International Statistical Classification of Diseases and Related Health Problems* (10th rev. ed.). New York, NY: Author.

Appendix A

Derivation of the Grand Variance

For a vector of variables $\boldsymbol{X}$, the covariance matrix $\boldsymbol{\Theta}$ is derived as follows:

$$\boldsymbol{\Theta} = E[\boldsymbol{XX}'] - E[\boldsymbol{X}]E[\boldsymbol{X}]' \,, \tag{6}$$

where $E[]$ indicates the expectation, and symbol $'$ indicates the transpose. Then for a multilevel model with persons i and repeated measures t per person, the covariance matrix taken over the repeated measures t for all persons i—the grand covariance matrix $\boldsymbol{G}$—equals

$$\boldsymbol{G} = \underset{it}{E}[\boldsymbol{YY}'] - \underset{it}{E}[\boldsymbol{Y}]\underset{it}{E}[\boldsymbol{Y}]' \,. \tag{7}$$

And for person i in the multilevel var model,

$$\Omega_i = \underset{t}{E}[\boldsymbol{Y}_i\boldsymbol{Y}_i'] - \mu_i\mu_i' \,. \tag{8}$$

Then, for the multilevel model with i persons and t repeated measures per person, it follows that

$$\begin{aligned} \underset{i}{E}[\Omega] &= \underset{i}{E}[\underset{t}{E}[\boldsymbol{YY}'] - \mu\mu'] \\ &= \underset{it}{E}[\boldsymbol{YY}'] - \underset{i}{E}[\mu\mu'] \\ &= \underset{it}{E}[\boldsymbol{YY}'] - \left(\underset{i}{E}[\mu]\underset{i}{E}[\mu]' + \psi_\mu^2\right) \\ &= \underset{it}{E}[\boldsymbol{Y}]\underset{it}{E}[\boldsymbol{Y}]' + \boldsymbol{G} - \underset{i}{E}[\mu]\underset{i}{E}[\mu]' - \psi_\mu^2 \\ &= \boldsymbol{G} - \psi_\mu^2, \end{aligned} \tag{9}$$

such that,

$$\boldsymbol{G} = \underset{i}{E}[\Omega] + \psi_\mu^2 \tag{10}$$

Appendix B

Prior Specification and Convergence for the Empirical Illustration

The Bayesian analysis requires the specification of prior distributions for the individual parameters, the fixed effects, the innovation variances, and the variances and covariances of the random effects. We aimed to use uninformative prior specifications for all parameters. We specified normal distribution with means of zero and precision 10^{-9} for the fixed effects. For the innovation variances, we specified uniform(0,10) prior distributions and a uniform(-1,1) prior for the correlation between the innovations. It is notoriously difficult to specify uninformative priors for covariance matrices that are larger than 2 × 2, such as the covariance matrix Ψ for the random parameters. The conjugate prior for covariance matrices in a normal model is the Inverse-Wishart prior. Like the Inverse-Gamma prior, the Inverse-Wishart prior is relatively informative when the variance parameters are close to zero (Gelman, 2006; Schuurman, Grasman, & Hamaker, in press).

The solution that currently has been found to work the best is to specify the Inverse-Wishart prior based on prior estimates of these variances. We did this using maximum likelihood (ML) estimates as described in Schuurman et al. (in press). We checked the sensitivity of the result to this prior by also fitting a model with uniform priors specified for the variances, ignoring potential covariation between the random parameters. Both analyses gave very similar results and conclusions about the modeled processes. We report the results for the ML-based prior, given that with this prior, potential covariances between random parameters are taken in to account.

We evaluated the convergence of the fitted models by fitting three chains, each with 30,000 burn-in and 10,000 samples. Convergence was evaluated based on the mixing of the three chains, the Gelman-Rubin statistic (Gelman & Rubin, 1992) and autocorrelations. Based on the results, we concluded that 10,000 samples with 30,000 burn-in was sufficient for convergence. With and Intel Xeon 3.1 GHz CPU, it took approximately 24 hr to fit the model. We excluded one participant from the analyses because the regression coefficients for this participant did not converge, which was most likely a result of having only 24 observations for competence, which were also quite dispersed (note that the lack of convergence for this participant did not influence the group results). Throughout this article, we report the medians and 95% credible intervals of the posterior distributions for the parameters of interest.

Received June 6, 2014
Revision received July 28, 2015
Accepted August 5, 2015 ■