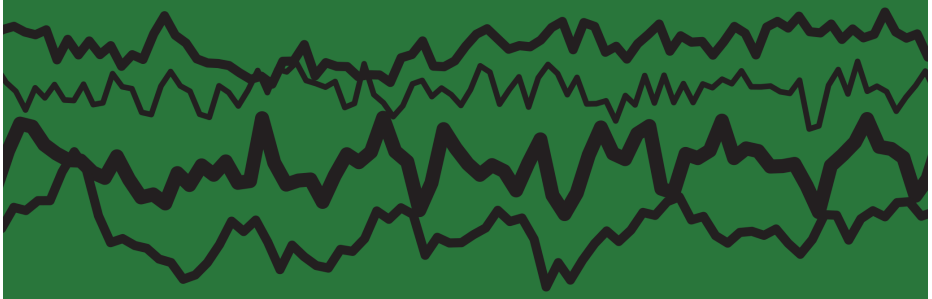
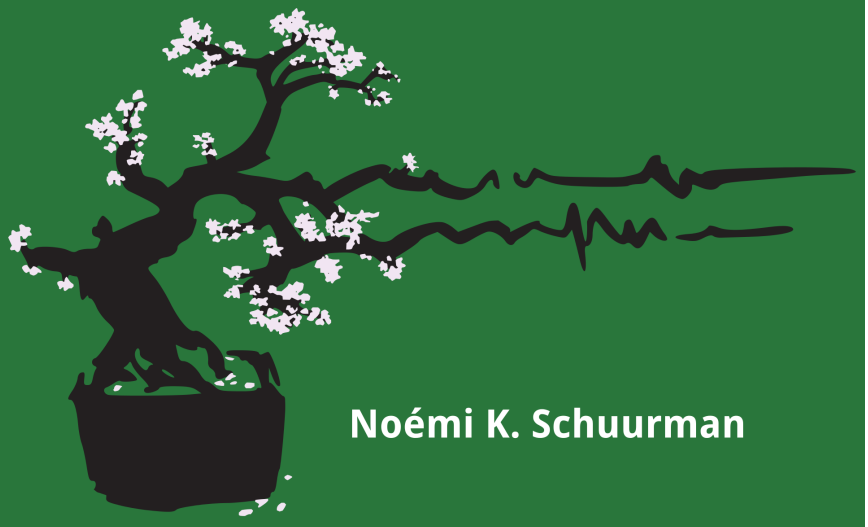
Multilevel Autoregressive Modeling in Psychology: *Snags and Solutions*





Noémi K. Schuurman

Multilevel Autoregressive Modeling in Psychology: Snags and Solutions

Noémi K. Schuurman

Schuurman, Noémi K.
Multilevel Autoregressive Modeling in Psychology: Snags and Solutions
Proefschrift Universiteit Utrecht, Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 978-90-393-6585-4
Druk: GVO drukkers & vormgeving
Cover design by Noémi K. Schuurman

Multilevel Autoregressive Modeling in Psychology: Snags and Solutions

Multilevel Autoregressieve Modellen in de Psychologie:
Problemen en Oplossingen
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op maandag 27 juni 2016
des middags te 12.45 uur

door

Noémi Katalin Schuurman

geboren op 5 juni 1988
te Amsterdam

Promotor:	Prof. dr. H.J.A. Hoijtink
Copromotor:	Dr. E.L. Hamaker

Beoordelingscommissie:

Prof. dr. D. Borsboom
Dr. E. Ceulemans
Prof. dr. P.G.M. van der Heijden
Prof. dr. P. de Jonge
Prof. dr. D.T.D. de Ridder

Contents

1 Introduction

Multilevel autoregressive models are statistical models that can be applied to intensive longitudinal data in order to model processes that occur within individuals over time, for many individuals at once. The aim of this dissertation is to further investigate, explicate, and if possible remedy certain difficulties in fitting and interpreting multilevel autoregressive models in the context of psychological science.

These two sentences however probably raise more questions than they answer. Some questions that may come to mind are: What exactly is intensive longitudinal data, and why should one care to collect such data? What kind of psychological processes are we talking about, and why should we care that they occur within individuals over time? Why model these processes for many individuals at the same time and how does that work? Where exactly do the 'autoregressive' part, and the 'multilevel' part come in and what do they both mean? And what are these difficulties that were mentioned and why should we care about them? In order to provide some context for the following chapters, I will try to provide brief answers to these questions. First, I will give examples of psychological processes, and discuss why it is important to measure and model such processes within individuals over time. This topic has gotten increasing attention in psychology, and the interested reader will find more elaborate and comprehensive works on the importance of studying the dynamics of individuals in the references of this chapter. Second, I will discuss why the autoregressive model is of interest for modeling psychological processes, and some limitations of this model. Third, I will briefly introduce the 'multilevel' extension of the autoregressive model. Finally, I will outline what is to come in the following chapters, and why the content of those chapters is relevant for people that want to use autoregressive models to study psychological processes.

1.1 Intensive Longitudinal Data for Studying Psychological Processes

Psychological processes are of interest in all areas of psychology, and all such processes occur at a within-subject level over time (where the subject may be a person, a dyad, or group of persons, and so on). Some examples of psychological processes are the regulation of affect throughout the day, the development of the mathematical

abilities of a child, the effects of therapy on someone diagnosed with a psychological disorder, the regulation of cortisol levels in relation to stress, the effects of a person's job motivation on his or her performance and vice versa, and social interactions between a parent and child.

Given that psychological processes happen within a subject over time, it makes sense that in order to study them, one needs to measure the variables for these processes over time. The more frequent these measurements are taken, and the longer the measurement period, the more the process will unfold itself to the researcher. If the goal is to generalize conclusions about these processes to a larger population, it is also desirable to take such repeated measures for many people sampled from that population. The type of data that is then obtained, consisting of many repeated measures for many individuals, is referred to as intensive longitudinal data (Walls & Schafer, 2005). By modeling such data over time using dynamic modeling techniques, we aim to capture the underlying psychological processes.

In psychological research practice, however, cross-sectional studies that are based on measurements of many individuals at one measurement occasion, or panel data based on a few repeated measurements (e.g., 2 to 5) of many individuals, are often used to make inferences about psychological processes. One reason for this may be that collecting intensive longitudinal data is relatively difficult, time intensive, and expensive compared to collecting one or a few measurements for many individuals. Another reason may be that the techniques that can be used to analyze them are still not well known.

However, the usage of cross-sectional or panel data for studying psychological processes is problematic, for mainly two reasons: Firstly, the way psychological variables are distributed within persons, is not necessarily the same as the way they are distributed across different persons (Adolf, Schuurman, Borkenau, Borsboom, & Dolan, 2014; Borsboom, Mellenbergh, & van Heerden, 2003; Hamaker, 2012; Kievit et al., 2011; Molenaar, 2004; Nezlek & Gable, 2001). Secondly, how a variable is distributed within persons, may be different from person to person (Adolf et al., 2014; Hamaker, 2012; Molenaar, 2004; Nesselroade, 2007). Many cross-sectional and panel study methods disregard these issues to varying degrees. I discuss both issues in more detail in the following.

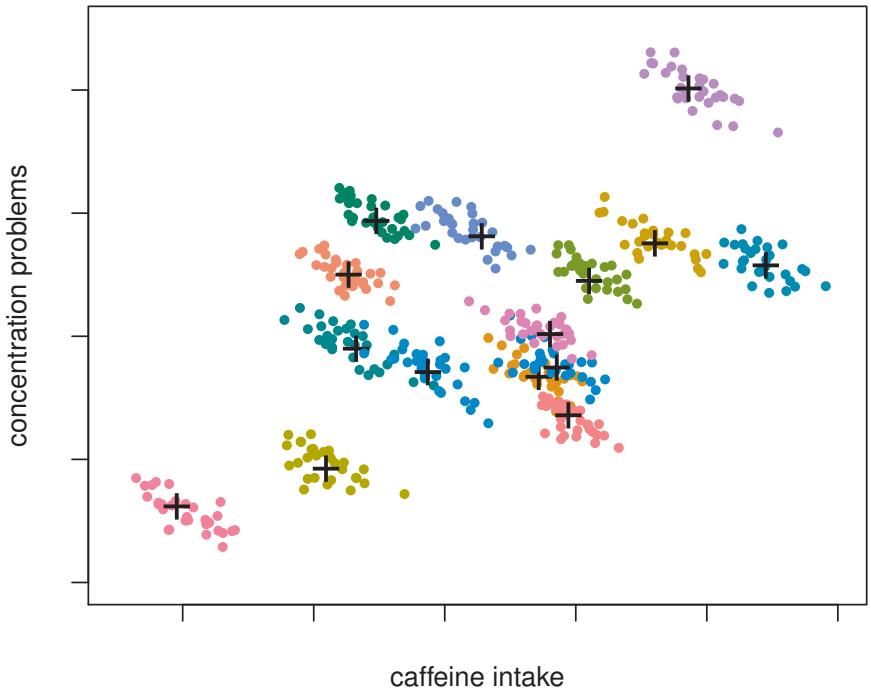


Figure 1.1: Scatterplot of simulated intensive longitudinal data on caffeine intake and concentration problems. The repeated measures for each person are indicated by a specific color. The average concentration problems and caffeine intake per person is indicated with the black plus signs. For each person there is a negative association between concentration problems and caffeine intake: the more caffeine intake, the less the concentration problems. Across persons however there is a positive association between concentration problems and caffeine intake: people with a higher average caffeine intake also tend to have higher average concentration problems.

Fluctuations Within Persons vs. Differences Between Persons

The fact that fluctuations in scores within persons over time are not necessarily distributed in the same way as differences in scores between persons is best illustrated with an example. Consider the relationship between caffeine intake and concentration problems. Say we find that persons who on average consume a lot of caffeine per week tend to have more trouble concentrating than persons who consume less caffeine per week. However, we also find that within each person over time caffeine actually improves concentration. An example of what the data would look like is presented in the scatter plot in Figure 1.1. In this Figure, there is a negative association between concentration problems and caffeine intake for each person (the data of each person is shown in a separate color). However, there is a positive association between the persons' average caffeine intake and their average concentration problems (indicated by the black plus signs). These results are not contradictory - they could be explained by the fact that people who experience concentration problems may start taking in caffeine to improve their concentration. On the other hand, people who do not have a lot of concentration problems may not find the need to drink as much caffeine, because they do not experience concentration problems. That is, the negative within-person effect of caffeine on concentration problems eventually results in a positive between-person association between caffeine and concentration problems.

In order to be able to generalize results about differences between persons to what happens on a within-person level, a very strict assumption needs to hold: That the distribution of a variable in a population of individuals, is the same as the distribution of a variable within each individual in that population (Adolf et al., 2014; Hamaker, 2012; Molenaar, 2004). That is, it requires ergodicity, which means that the moments for the variable in the population of individuals should be equal to those for each individual. This, for instance, implies that the mean score over time for a certain variable should be the same for each individual (and thus the same as the mean across the entire group of individuals). This implication alone seems very unrealistic for most, if not all, psychological variables that come to mind. For example, we generally would not expect that the average level of concentration problems across a period of time would the same for each person.

In practice, the observed variance in psychological data that consists of measures of multiple individuals will not, in all likeliness, consist of only variance that is the result of relatively stable differences between persons, or of only variance that is the

result of temporal fluctuations within each person. Rather, it will consist of some mix of both. For instance, it is likely that some people generally are better at concentrating than others; if we test the concentration ability of several persons, the stable differences in their ability to concentrate will result in stable differences between those persons' test scores. However, there will also be variation present in the test scores that is due to within-person fluctuations in their ability to concentrate: How well each person performs typically varies from occasion to occasion. Both types of variance will be present in our test scores.

In order to make meaningful inferences about psychological processes, these two sources of variance need to be distinguished from each other - if not, the results will represent a mix of the within-person and between-person effects, which may be neither informative about the psychological within-person process, nor about stable between-person differences (Cattell, 1967; Hamaker, 2012). This is a problem for correlational cross-sectional studies, but also for longitudinal studies that fail to separate stable between-person differences from within-person processes (Hamaker, Kuijper, & Grasman, 2015). Still, results from correlational, cross-sectional studies are on a fairly regular basis generalized to individuals in the psychological literature. Many (heatedly debated) examples of such generalizations can be found for instance in the context latent variable modeling in psychology (cf., Borsboom, 2015; Borsboom & Dolan, 2006), and in many longitudinal studies, most of which rely on panel data, stable between-person differences are not separated from within-person differences.

There are various ways to filter out stable between-person differences from within-person fluctuations. One option is to model only one subject at a time: If there is only one subject, there can be no between-subject differences (Cattell, Cattell, & Rhymer, 1947; Molenaar, 1985; Nesselroade, 2007). This is the approach taken when classical n=1 autoregressive models and other time series models are used to model intensive longitudinal data. Another option is to explicitly model the stable between person differences (Hamaker, Nesselroade, & Molenaar, 2007; Heck & Thomas, 2000; Hertzog & Nesselroade, 1987; Hox, 2010), for instance by including a random intercept or mean in the model. Examples of this approach can be found in growth curve modeling (cf., Hox, 2010, p.325), and in the random intercept cross-lagged panel model (Hamaker, Kuijper, & Grasman, 2015). It is also the approach taken in the multilevel autoregressive model(s) presented in this dissertation.

Within-Person Processes Differ Between Persons

Separating the variance that results from differences between persons from variance that is a result of within-person fluctuations is not enough to be able to make meaningful inferences about psychological processes. It is also important to take into account that these processes may differ to some extent from person to person. For example, for some persons how much external validation they receive may strongly affect how they perform their job, while for others it makes relatively little difference. In order to get an accurate picture of a psychological process, these differences between persons in the within-person processes should be accounted for.

Many (if not most) analysis techniques used in psychology however return results that concern average effects across subjects. This is the case for cross-sectional studies, but also for many longitudinal studies that are based on only a few repeated measurements. Average results are, however, not necessarily informative about specific individuals. For example, consider an experiment in which people are randomly assigned to a control group or an experimental group, in order to study the effectiveness of a certain therapy. The persons in the control group did not improve, while half of the persons in the experimental group improved, and the rest declined. Using standard analysis techniques we obtain an average result that indicates no evidence for a difference between the two groups. This average result of course does not reflect the effectiveness of the therapy for specific individuals, not even those that participated in the study. Although this example may be somewhat extreme, it should make clear that average results cannot simply be generalized to all, most, or even any specific individual.

By using intensive longitudinal data, it is possible to take into account that the within-person processes may differ from individual to individual. One option is to include relevant moderator and predictor variables that account for differences between person in the processes. However, this method is unlikely to account for all differences in the model parameters across subjects. Another option is to model the process for each person separately, potentially even specifically tailoring a model to each individual (Cattell et al., 1947; Madhyastha, Hamaker, & Gottman, 2011; Molenaar, 1985; Nesselroade, 2007; Snippe et al., 2015). A third option is to model all individuals at the same time, but to allow some or all model parameters to vary across individuals, which is the case in multilevel models, such as growth curve models for instance, and the multilevel autoregressive model(s) discussed in this dissertation (cf., Bringmann

et al., 2013; De Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2014; Kuppens, Allen, & Sheeber, 2010; Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011; Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001; Rovine & Walls, 2005; Wang, Hamaker, & Bergeman, 2012).

In sum, in order to study psychological processes that happen at a within-person level, we need to study these processes at the level of individuals. This can be done by obtaining many repeated measures per person, and using dynamic models to separate stable between-person differences from within-person processes, while taking into account that there may be between-person differences in those processes. The (multilevel) autoregressive model is one example of such a dynamic model. In the following section, I will start by discussing the n=1 autoregressive model, which is a building block for the multilevel autoregressive model that I will discuss after that.

1.2 Autoregressive Modeling

There are many approaches for modeling intensive longitudinal data, even though these approaches are not yet popularized in psychology. The list of potential models includes linear and non-linear ones, models for continuous dependent variables or discrete variables, and models with and without latent variables (Hamaker, Ceulemans, Grasman, & Tuerlinckx, 2015; Walls & Schafer, 2005). Which option to choose depends on the research question and data at hand. The autoregressive model (the AR model) may however be of particular interest for many psychological research, because of the interpretation of the model's regression parameters.

The n=1 AR Model

The n=1 AR model is a time series model, and is fitted for a single individual, for which many repeated measures are available. The approach of the AR model, and extensions of this model, is summarized well by a popular saying in folk psychology: "The best predictor of future behavior is past behavior". The basic univariate AR model is a linear model for a continuous outcome variable, with as a predictor the past observations of the outcome variable (hence *auto*regression). An AR model of 'order p' uses predictors for a number of p 'lags', where lags indicate the distance in time between two measurement occasions. For example, an AR(2) model for an outcome variable that is observed every day includes two predictors: The scores of

the previous day (lag 1) and the scores of two days ago (lag 2). Throughout this dissertation however, the focus will be on AR(1) models that only use the previous measurement occasions as a predictors for the current ones.

The autoregression parameters reflect the effect of each outcome variable at the previous occasion on itself at the current occasion. Say for example, that we measured the level of depression repeatedly for a certain individual. A positive autoregression coefficient for depression indicates that if this person's feelings of depression were high yesterday, they are likely to also be high today, and if they were low yesterday they are likely to also be low today. As a result of the autoregressive effect, the feelings of depression may linger above or below the person's baseline (average) level of depression across multiple measurement occasions, and will only slowly come back to that baseline level. This can be seen from the top row in Figure 1.2, in which a time series is plotted with a positive autoregressive effect. The stronger the autoregressive effect, the longer it will take for the feelings of depression to come back to baseline. On the other hand, if the autoregression parameter for depression is equal to zero for a certain person, this indicates that depression on the previous measurement does not influence the depression at the current occasion: Every day is a 'new day' for this person. As a result, their depression can vary freely around the baseline without lingering above or below it for a long amount of time, as can be seen from the time series plotted in the second row in Figure 1.2. This person therefore also can 'recover' relatively quickly after an increase in depression. A negative autoregressive effect indicates that if the observed score is high today, it is likely to be low the next day, as is illustrated in the third row of Figure 1.2. This may not seem very intuitive in the context of depression, but such negative autoregression effects may well be expected in other areas of psychology, especially those that concern intake. For instance, people that ate or drank a lot yesterday, may be likely to eat or drink less today, and more again tomorrow, and so on. The stronger this effect, the longer it will take for the process to come back to baseline levels after a relatively high (or low) score. In sum, the autoregression parameter is indicative of inertia: The stronger the autoregressive effect, the more 'resistant' the process is to change. Inertia is considered an important factor for many kinds of psychological processes, ranging from the regulation of affect (Kuppens et al., 2010; Suls, Green, & Hillis, 1998), to mood disorders (Koval, Kuppens, Allen, & Sheeber, 2012; Kuppens et al., 2010), and attention (Kirkham, Cruess, & Diamond, 2003).

AR models that include multiple outcome variables are referred to in the literature

as vector AR models (VAR models). In VAR models, the outcome variables are not only predicted by their own previous values, but also by the previous values of the other outcome variables, which is referred to as 'cross-lagged regression'. I provide an example of this in the bottom row of Figure 1.2, which shows a time series plot of two variables, one in blue and one in red. Both variables have a positive autoregression coefficient. The red variable has a cross-lagged effect on the blue variable, but the blue variable does not affect the red variable: This can be seen from the time series plot where a peak in the red process is often followed by a peak in the blue process one measurement occasion later, and a valley by a valley.

The cross-lagged parameters from VAR models can be used for investigating potential reciprocal effects between psychological variables. For instance, to investigate whether stress affects depression, whether depression affects stress, or if both affect each other, over a specific time interval. For many areas in psychology such reciprocal relationships are of interest for determining which associations or variables are 'causally dominant', or 'the strongest driving force' in the dynamic process, which may be useful for guiding interventions. Determining (the strength of) such effects is for instance central to network theories of psychology (Borsboom & Cramer, 2013; Schmittmann et al., 2013), and also has gotten a lot of attention in the context of cross-lagged panel modeling (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 classical AR model has a number of limitations. One limitation of the n=1 AR model is that the model is fitted for one individual at a time, which makes it difficult to generalize results to a population of individuals, which is usually the aim in psychological science. A solution to this problem is to extend the VAR model to a multilevel model, which is discussed in more detail in the following.

Multilevel AR Modeling

The classical AR model is fitted for one individual at a time. Because each individual may have his or her own model, the AR model takes into account that individuals may differ from each other in meaningful ways. However, it also disregards that the processes of individuals may be similar in certain ways. For instance, we may expect that overall there is a stronger positive autoregressive effect for day-to-day anxiety for adults diagnosed with generalized anxiety disorder, than for healthy adults. Or,

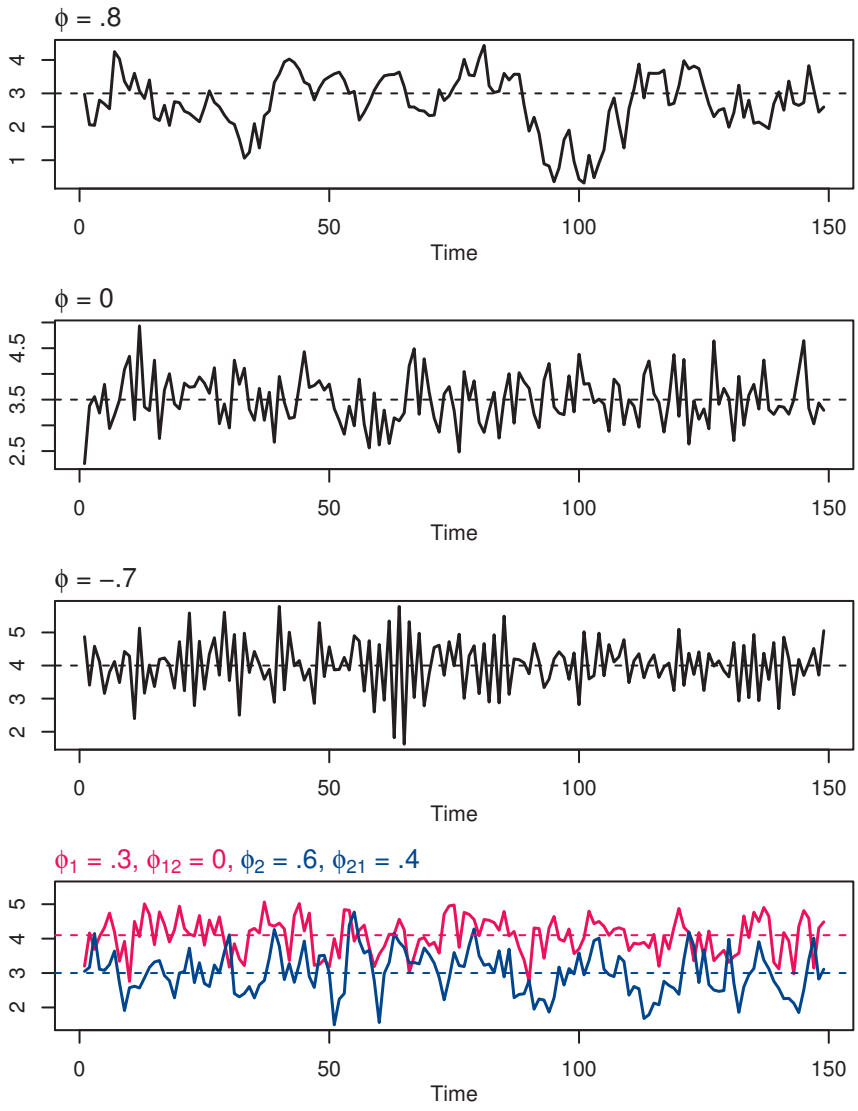


Figure 1.2: Time series plots of simulated autoregressive processes. The first panel shows an AR(1) process with a positive autoregression parameter equal to .8. The second panel shows an AR(1) process with the autoregression parameter equal to 0. The third panel shows an AR(1) process with a negative autoregression parameter equal to -.7. The last panel shows a VAR(1) process for two variables, with autoregression coefficients of .3 and .6 for the red and blue variable respectively, a cross-lagged effect of .4 of the red variable on the blue variable, and no cross-lagged effect of the blue variable on the red variable.

we may expect that for the large majority of individuals, the autoregressive effects for minute-to-minute affect are positive rather than negative. By fitting models for each individual separately this information is not taken into account, and it also makes it difficult to draw more general conclusions about a population of individuals.

Multilevel modeling allows for fitting the AR model at once for multiple individuals, while allowing the model parameters to be different across individuals. The model parameters for the different individuals are assumed to come from the same probability distribution, that is, it is assumed that they come from the same population. This assumption implies that the model parameters for one individual are informative for other individuals from the same population, and in this sense the model takes into account that people are to some extent similar to each other.

By evaluating the characteristics of the probability distribution for the individuals' model parameters, we can obtain information about the group of individuals, which we in turn may use to make inferences about the population. For instance, we can look at the variance of the distribution to see how much the model parameters of the individuals differ from the average parameter on average. We could use this information to infer what range of parameter values we may expect to see in the population. It is also possible to add predictors for the individual parameters to the multilevel model. For example, we may explain differences among the average levels of stress of individuals using their gender or occupation. In sum, with the multilevel model we can model the autoregressive processes per individual, while simultaneously modeling the differences in these processes across the individuals.

1.3 Outline of this Dissertation

This dissertation consists of four papers in which my co-authors and I investigate a number issues in fitting and interpreting multilevel AR models, in the context of psychological science. Throughout this dissertation, we fit the multilevel AR model using Bayesian modeling techniques. Although there are some frequentist techniques available for fitting multilevel AR models (Bringmann et al., 2013), there are practical reasons to opt for a Bayesian approach (next to any potential philosophical inclinations), which will be discussed throughout this dissertation. However, during preliminary simulations and modeling efforts of myself and other colleagues, we found unsettling biases in the estimates of the variances of the random autoregression and cross-lagged parameters, as a result of the prior specification for the covariance

matrix of the individual model parameters. In Chapter 2 the basis of this problem is described, and suggestions for the specification of the Inverse-Wishart prior distribution from the literature to remedy this problem are evaluated by means of a simulation study. Chapter 2 also contains an empirical example on the cross-lagged associations between worrying and positive affect, in which we find that for persons who worry relatively little, worrying seems to be beneficial to their positive affect, while for persons who worry a lot, it seems to be detrimental to their positive affect. This example provides an illustration of the potential of the multilevel VAR model.

Chapter 3 is about comparing the strength of cross-lagged effects for the multilevel VAR model, by making use of standardized cross-lagged regression coefficients. As mentioned previously in Section 1.2, the cross-lagged relationships between variables are often of interest to psychological researchers for determining which associations or variables are 'causally dominant', or 'the strongest driving force' in the dynamic process. Standardized cross-lagged regression coefficients can be used to determine which cross-lagged associations explain the most unique variance. However, standardization in multilevel models can be done in multiple ways. In Chapter 3, these different ways are discussed, and one of them is argued to be superior. Chapter 3 contains an empirical example on the reciprocal relationship between feeling competent and exhausted for persons diagnosed with burnout. This empirical example illustrates that average results can prove misleading, because they do not necessarily generalize to all individuals, as discussed previously in Section 1.1.

The last two chapters of this dissertation are concerned with the consequences of disregarding measurement errors, and how to account for them in the AR model. In Chapter 4 we discuss different ways of incorporating measurement errors in the classical n=1, univariate, AR model, and we compare these different options by means of a simulation study. This chapter includes an empirical application in which we model the mood of eight women, and we find that about 30% to 50% of the variance in the data (depending on the participant) is the result of measurement errors.

The results from Chapter 4 function as a stepping stone for the work described in Chapter 5, in which we discuss how to account for measurement errors in the multilevel VAR model, and how we can use that model to obtain estimates of the reliability of the repeated measurements for each person. Furthermore, the consequences of disregarding measurement error for the estimated cross-lagged and autoregression parameters in a multivariate model are discussed. Chapter 5 provides an empirical example of a bivariate multilevel VAR model, for which we relax the assumption that

the covariance matrix of the residuals is fixed across persons. In this example, we find that on average, the positive affect people feel about their romantic relationship affects their general positive affect the next day, but we find no evidence for the reverse.

In Chapter 6 the main findings of each chapter are summarized, and some limitations of the multilevel AR modeling approach are discussed, as well as directions for future research.

2 A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models

by N.K. Schuurman, R.P.P.P. Grasman, and E.L. Hamaker

Psychological processes occur within individuals: stress affecting a person's mood, a mother's self-esteem influencing her teenage daughter's self-esteem, an individual's job satisfaction affecting job performance, and so on. It is likely that many of these dynamical processes also differ across individuals (see for instance Adolf et al., 2014; Hamaker, 2012; Lodewyckx et al., 2011; Molenaar, 2004; Rovine & Walls, 2005; Wang et al., 2012). For instance, stressful situations may strongly affect the mood of one individual, while they have little effect on the mood of another individual. Multilevel autoregressive models are ideal for investigating these types of processes, because they allow for modeling how variables affect themselves and each other over time. Moreover, they allow for modeling these effects for each individual separately in the form of random parameters, and for the individuals on average, as a result of the inclusion of fixed effects.

Multilevel autoregressive models are complex models that can prove difficult to fit with software based on traditional maximum likelihood modeling, especially when considering multivariate or latent variable extensions, or models that include random residual variances. In contrast, with Bayesian modeling software, such as WinBUGS or OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009; Lunn, Thomas, Best, & Spiegelhalter, 2000), fitting these complex multilevel models is relatively trivial (see for instance Song & Ferrer, 2012; Wang et al., 2012; or Lodewyckx et al., 2011, for an implementation in R). Other benefits of Bayesian modeling are its flexibility in handling missing data, and that it directly provides the researcher with the estimated

This chapter is based on: Schuurman, N.K., Grasman, R.P.P.P., & Hamaker, E.L. (2016). A Comparison of Inverse-Wishart Prior Specifications for Covariance Matrices in Multilevel Autoregressive Models. *Multivariate Behavioral Research*.

Author contributions: Schuurman proposed the topic for the study, designed and performed the simulation study, analyzed, processed and interpreted the results and the empirical dataset, and wrote the paper. Grasman provided feedback on the study design and the written work. Hamaker gave extensive feedback on the design of the study, and the written work.

random parameters. To benefit from this flexibility of Bayesian modeling, it is necessary to specify prior distributions for the parameters that are to be estimated. The prior distributions may be specified based on a researcher's prior knowledge about the parameters in question, such as results from previous research. However, when there is little or no prior information available, or when the researcher wishes to take a more objective approach, it may be desirable to specify uninformative prior distributions, prior distributions that have a negligible influence on the estimated parameters.

In certain cases it can be difficult to specify uninformative prior distributions. One of these cases is specifying priors for variances or for covariance matrices when the variances are small (close to zero). Typical prior distributions chosen for variances and covariance matrices are Inverse-Gamma (IG) distributions and Inverse-Wishart (IW) distributions respectively (e.g., Gelman & Hill, 2007). These prior distributions, which are usually uninformative with certain hyperparameters, are quite informative when variances are small, resulting in a strong effect of the prior distribution on the parameter estimates (Gelman, 2006; see also Song & Ferrer, 2012, for an example). Gelman (2006) and Browne and Draper (2006) show that when a single variance is modeled, choosing a uniform distribution for the standard deviation or variance instead of the IG distribution results in parameter estimates that are negligibly affected by the prior distribution. However, the problem is harder to solve in the case of specifying a prior for a covariance matrix.

The issue is particularly relevant when considering multilevel models, because multilevel models are prone to having small variances in the covariance matrix of the random parameters. Small variances for the random parameters will result when the interindividual differences in the parameters are not very large, that is, the individuals have similar parameter estimates. However, it is important to note that the size of the variances also depends on the scale of the random parameters. For example, small variances for the random parameters may result as an artifact when the random parameters are restricted in range, which also restricts the size of their variance. For example, this may be the case when the random parameters are proportions, or probabilities. In the case of multilevel autoregressive modeling, the regression parameters are restricted in range as a result of the stationarity of the model (Hamilton, 1994). For instance, in a lag 1 autoregressive model (AR(1) model), where a variable is regressed only on itself at the preceding time point, the autoregression coefficient lies in the range from âĹŠ1 to 1, which necessarily results in a small variance for this coefficient across individuals. As a result, it is difficult to specify uninformative priors

for the covariance matrix of the random effects in multilevel autoregressive models (cf., Song & Ferrer, 2012).

Proper estimation of the covariance matrix of random parameters is essential for psychological research, in order to get an accurate impression of the magnitude of interindividual differences in the dynamics of individuals, and proper estimations of the covariances are necessary for getting an accurate impression of the associations among these interindividual differences. Therefore, the aim of this study is to compare the performance of several prior specifications for covariance matrices suggested in the literature, when one or more of the variances in the covariance matrix are close to zero. Specifically, we compare three IW prior specifications: a) a prior specification that is based on an identity matrix, and is often used as an uninformative prior in practice; b) a data-based prior that uses input from maximum likelihood estimations; and c) the default conjugate prior proposed by Kass and Natarajan (2006). Although we are especially interested in the Bayesian estimation of multivariate autoregressive models, in the simulation study we use univariate autoregressive models with one outcome variable and a time varying predictor variable for practical reasons (explained further in the following section). We illustrate a full multivariate model in an empirical example on repeated measures of positive affect and worrying for 129 participants.

The remainder of this article is organized as follows. We start by discussing the multilevel autoregressive model in more detail, followed by a section on the IW distribution and the three prior specifications for the covariance matrix of the random parameters. After that we present our simulation studies and their results, and we present an empirical application in which we compare the effects of the different prior specifications for a multivariate model. We end with our main conclusions and a discussion.

2.1 Multilevel Autoregressive Model

In autoregressive models variables are regressed on themselves and each other on a previous time point. In such a model, the autoregression coefficient reflects the influence the previous state of a variable has on its current state, and crossregression coefficients reflect the influence of the previous value of another variable has on the current state of this variable (Hamilton, 1994; Kim & Nelson, 1999). Multilevel extensions of these models allow for modeling these dynamic processes for multiple

persons, and to model the average intraindividual effects over the multiple subjects, which helps generalize the results to a larger population.

Although our main interest is in specifying uninformative priors for full multivariate multilevel autoregressive models with multiple outcome variables, for the simulation study we focus on a univariate multilevel autoregressive model with one outcome variable and a lagged predictor variable instead, for practical reasons: A bivariate multilevel autoregressive model contains six random effects (i.e., two autoregression parameters, two crossregression parameters, and two means), such that the covariance matrix of the random effects contains six variances and 15 covariances. Estimating such a model using WinBUGS is time intensive and computationally demanding, which would make a simulation study based on such a model challenging. Instead, we focus on a univariate multilevel autoregressive model with a lagged predictor, which contains only three random effects, such that the covariance matrix of the random effects contains three variances and three covariances. However, we emphasize that the model can be generalized to include more than one outcome variable. We illustrate such a multivariate model in the empirical example, for which we fit a bivariate model (this took approximately 24 hours with three chains, without parallel computing for the three chains). For a graphical representation of the univariate model with a lagged predictor, see Figure 2.1.

In the univariate multilevel AR(1) model with timevarying predictor, y_{tj} is a score on outcome variable y for person j at time point t. The scores y_{tj} are split in individual means μ_j, and a residual score z_{tj}. The autoregression and crossregression effects are modeled using residual scores z_{tj}: z_{tj} is regressed on z_{t-1j}, the residual score for outcome variable y_{t-1j} for person j at previous time point $t-1$, and on x_{t-1j}, the score on a time-varying predictor variable x for person j at time point $t-1$. Modeling the autoregression effects on z_{tj} rather than on y_{tj} directly allows us to estimate the means μ_j directly rather than the intercepts. The means represent the baseline score for an individual, which is more intuitive than the intercept, which represents the score of an individual when the predictor variables are zero. The autoregression coefficients ϕ_j represent the association between the outcome variable y at time t with itself at time $t-1$. The larger the absolute autoregression coefficient, the better future values of y can be predicted by the previous value of y. Positive autoregression coefficients are also interpreted as a measure of inertia — the larger the autoregression coefficient, the slower it will take for y to return to its baseline μ after a perturbation of the system (Suls et al., 1998). The crossregression coefficients β_j indicate how well

a past value of a predictor x predicts the future value of y. In multivariate models the crossregression coefficients can be used in investigate the reciprocality of the effects between multiple variables (Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001). Innovation e_{tj} represents anything that is not directly measured that may influence the system. These innovations are assumed to be normally distributed with a mean of zero and variance σ^2. In other words, at level 1 the multilevel model can be specified as:

$$\begin{aligned} y_{tj} &= \mu_j + z_{tj} \\ z_{tj} &= \phi_j z_{t-1j} + \beta_j x_{t-1j} + e_{tj} \end{aligned} \tag{2.1}$$

$$e_{tj} \sim N(0, \sigma^2). \tag{2.2}$$

In this model, three parameters are allowed to vary over individuals: μ_j, the mean for person j; ϕ_j, the autoregression coefficient for person j; and β_j, the crossregression coefficient for person j. We will refer to these individual parameters as random parameters, and assume that they are multivariate normally distributed, with means γ_μ, γ_ϕ, and γ_β, and 3×3 covariance matrix $\mathbf{\Psi}$. The means describe the average effects (i.e., fixed effects) for the group of individuals, and the covariance matrix describes the variations around these means for the group of individuals. Hence, at level 2 we have:

$$\begin{bmatrix} \mu_j \\ \phi_j \\ \beta_j \end{bmatrix} \sim MvN \left\{ \begin{bmatrix} \gamma_\mu \\ \gamma_\phi \\ \gamma_\beta \end{bmatrix}, \begin{bmatrix} \psi^2_\mu & & \\ \psi_{\mu\phi} & \psi^2_\phi & \\ \psi_{\mu\beta} & \psi_{\phi\beta} & \psi^2_\beta \end{bmatrix} \right\}. \tag{2.3}$$

We focus here on autoregressive processes that are stationary for each individual, meaning that the mean and variance of the outcome variable are stable over time for each individual. If the AR(1) process is stationary, the autoregression parameters ϕ_j will lie within a range of -1 to 1 (Hamilton, 1994).[1] When this holds for every indi-

[1]Note however because we assume that ϕ comes from a multivariate normal distribution, that technically autoregressive parameters outside of this range can occur. In our simulation study we chose the mean vectors and covariance matrices for the multivariate normal distribution so that parameters not in line with the stationary assumption are extremely unlikely. We chose these parameters to be in line with what we generally have encountered for autoregressive modeling in psychological practice: stationary processes with autoregressive parameters in a range of about 0 to .5. Note that encountering a non-stationary parameter value during estimation is not problematic for the estimation procedure, so that using a multivariate normal distribution rather than, for instance, a truncated multivariate normal distribution should not result in any technical (estimation) problems. In practice, encountering such a non-stationary parameter value would simply imply that the process is not stationary for that person. It may then be useful to consider different or extended models that models non-stationarity in an informative way (c.f., Hamilton, 1994).

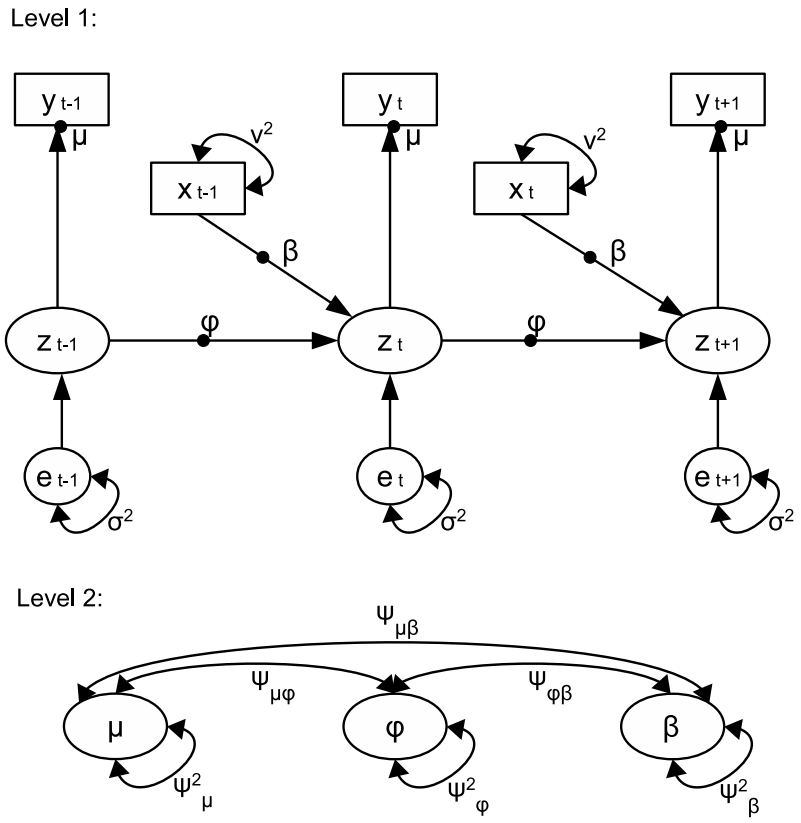


Figure 2.1: Multilevel AR(1) model with time-varying predictor. At level 1, the outcome variable y for individual j is regressed on itself at the previous time point $t-1$, and on a time varying predictor x at the previous time point $t-1$. The mean of y, μ, the autoregression coefficient ϕ, and the crossregression coefficient β, are allowed to vary over individuals j (indicated by the black dots). At level 2, the random coefficients are multivariate normally distributed, with the covariance structure as indicated in the figure.

vidual, the variance of the autoregression coefficients ψ^2_ϕ will be small. For instance, the mean of .4 and a variance of .04 for ϕ_j would result in a relatively large range of possible values for ϕ_j, namely a 95% interval of [.008, .792], whereas a variance of .01 would still result in a relatively large 95% interval of [.204, .596]. Similar ranges are found empirically, for instance by Wang et al. (2012): they studied daily measures of negative affect, and found a γ_ϕ of .15 and a variance ψ^2_ϕ of .04. The means μ_j and crossregression coefficients β_j are not restricted in range. Of course, the variances of μ_j and the random crossregression coefficients β_j may be small as well, due to the scale of the variables resulting in a small coefficient,[2] or simply due to minimal individual differences in these coefficients. Given that the standard priors for covariance matrices are very informative when variances are small, it will be difficult to specify the prior distribution for the covariance matrix of the random parameters $\mathbf{\Psi}$ such that it has a negligible influence on the results. In the next section we will go into more detail about the priors for covariance matrices, and why they are informative when variances are small.

2.2 Priors for the Covariance Matrix of the Random Parameters

For Bayesian estimation of the multilevel autoregressive model, prior distributions have to be specified for the random parameters (i.e., μ_j, ϕ_j, β_j for $j = 1, ..., n$), for the fixed effects (i.e., $\gamma_\mu, \gamma_\phi, \gamma_\beta$), for the innovation variance (i.e., σ^2), and for the covariance matrix of the random parameters (i.e., $\mathbf{\Psi}$). When an influence of the prior distributions on the results is undesirable, for instance when no relevant prior information is available, it is desirable to specify uninformative prior distributions that have a negligible influence on the end results. However, prior specifications that are uninformative in specific circumstances, may become informative under different circumstances. Our main interest here is in how to specify an uninformative prior distribution for $\mathbf{\Psi}$, so that the influence of the prior specification on the estimates

[2]Note that it is possible to increase the means μ_j and their variance, by transforming the relevant outcome variable (e.g., multiply the variable by ten). When the variance for the mean is increased so it is no longer close to zero, specifying the IW prior distribution for this coefficient will be relatively trivial. However, this is not possible for the autoregressive coefficient, because it is standardized by merit of the stationarity assumption that results in equal variances for y_t and y_{t-1} (Hamilton, 1994). While the transformation is possible for the crossregression coefficients, in a multivariate model increasing one crossregression coefficient results in decreasing the other crossregression coefficient — merely shifting the problem to another coefficient. Further, it may be difficult to determine in advance by how much to increase a coefficient, since its value is unknown a priori.

of the variances and covariances of the random effects is minimal, under the specific circumstance that the true sizes of some of these variances are small, as would be the case for the autoregression coefficients ϕ_j.

For this purpose we will first discuss the IW distribution, which is the conjugate prior for covariance matrices given normally distributed data, then we will go into more detail about the prior specification problem for covariance matrix $\mathbf{\Psi}$, and we will discuss three prior specifications for $\mathbf{\Psi}$ suggested in the literature.

The IW Prior Distribution

The prior distribution that is typically used for the covariance matrix of multivariate normally distributed variables, such as the covariance matrix $\mathbf{\Psi}$ for the random effects, is the IW distribution (Gelman, Carlin, Stern, & Rubin, 2003; Gelman & Hill, 2007). The IW distribution is a conjugate prior for the covariance matrix of multivariate normal distributed variables, which implies that when it is combined with the likelihood function, it will result in a posterior distribution that belongs to the same distributional family. Another important advantage of the IW distribution is that it ensures positive definiteness of the covariance matrix.

The IW distribution is specified with an $r \times r$ scale matrix $\mathbf{S}$, where r is equal to the number of random parameters, and with a number of degrees of freedom df, with the restriction that $df > r-1$. $\mathbf{S}$ is used to position the IW distribution in parameter space, and the df set the certainty about the prior information in the scale matrix. The larger the df, the higher the certainty about the information in $\mathbf{S}$, and the more informative is the distribution (Gelman, Carlin, Stern, & Rubin, 2003; Gelman & Hill, 2007). The least informative specification then results when $df = r$, which is the lowest possible number of df.

The means and covariance matrix of the IW distribution are a function of the elements s_{kl} on row k and column l from $\mathbf{S}$, r, and the df. That is, the density of the IW distribution is

$$\frac{\left|\boldsymbol{S}^{\frac{df}{2}}\right|}{2^{\frac{df\,r}{2}}\Gamma_r\left(\frac{df}{2}\right)}\left|\boldsymbol{X}\right|^{-\frac{df+r+1}{2}}e^{-\frac{1}{2}tr\left(\boldsymbol{S}\boldsymbol{X}^{-1}\right)} \tag{2.4}$$

where $tr\,()$ stands for the trace function, and Γ_r stands for the multivariate Gamma function. The mean of the IW distribution is

$$E[\boldsymbol{X}] = \frac{\mathbf{S}}{df - r - 1} \tag{2.5}$$

and the variance of each element of the IW distribution is

$$Var[x_{kl}] = \frac{(df - r + 1)s_{kl}^2 + (df - r - 1)s_{kk}s_{ll}}{(df - r)(df - r - 1)^2(df - r - 3)}). \tag{2.6}$$

The variances for the diagonal elements of the IW distribution simplify to

$$Var[x_{kk}] = \frac{2s_{kk}^2}{(df - r - 1)^2(df - r - 3)}. \tag{2.7}$$

It can be seen from Equations 2.6 and 2.7 that when the *df* increase, the denominator will increase more rapidly than the numerator, so that the variance will become smaller, which implies that the IW distribution will become more informative. It can also be seen that the size of the variance is partly determined by **S**: The smaller the elements of **S**, the smaller the variance of the IW distribution, and hence the more informative the prior will be. However, setting the scale to large values also influences the position of the IW distribution in parameter space, as can be seen from Equation 2.5. In other words, specifying a IW prior distribution requires balancing the size of **S** and the *df*.

A typically used relatively uninformative IW prior is a prior with small *df* and an identity matrix **S**. In many situations this prior specification will be uninformative enough for the the data to dominate the prior, so that the influence of the prior on the results will be minimal. However, when the variances are quite small, IW priors are informative, so that the estimates for the variances will be sensitive to the IW prior specification, resulting in over- or under-estimation of the variances depending on the specification of the prior distribution. The reason for this sensitivity when the variances are close to zero is that the IW distribution is bounded at zero for the variances: in consequence of this boundedness, slightly changing the central tendency of the distribution can have large effects on the weights placed on values close to zero.

We illustrate this in Figure 2.2, which shows eight plotted marginal densities for one of the diagonal elements of a bivariate IW distribution with varying *df* and $\boldsymbol{S}$. The four panels include two densities with the same diagonal $\boldsymbol{S}$, with respectively .001, .01, .1, or 1 as diagonal elements. For each panel, the density plotted in black has a larger *df* than the corresponding density plotted in gray. These plots further demonstrate that the IW distribution tends to place either a lot of weight on a specific

value close to zero (as in the upper panels), or place almost no weight close to zero (as in the lower panels). This shows that the IW- distribution is easily misspecified when variances are small. When the prior is specified too far from zero (e.g., IW prior with $\boldsymbol{S}$ as an identity matrix), this will result in an overestimation of the variances. However, specifying the central tendencies too close to zero will result in an underestimation of the variances, firstly because too much weight is shifted towards zero, and secondly because an element of the scale matrix set close to zero will also have a small variance (the density is more peaked). This is the case for instance for an IG distribution — which is basically a univariate simplification of the IW distribution — with a shape and scale close to zero (e.g., IG(10^{-5},10^{-5})). This IG distribution is often considered as an uninformative prior specification for a single variance, however it has been shown that this indeed results in an underestimation of the variance when this variance is small (Browne & Draper, 2006; Gelman, 2006). Although Gelman (2006) demonstrates that in the univariate case it is possible to use a Uniform or Inverse Half-Cauchy distribution instead of the conjugate IG distribution, giving good results, the solution to this problem is less simple for multivariate (IW prior) cases. In the following we will discuss three IW prior specifications that have been suggested in the literature.

Three Ways to Choose S for the IW Prior Distribution

In order to find the best way to specify the prior for $\boldsymbol{\Psi}$ when some of the variances are close to zero, we will evaluate the performance of three IW priors for $\boldsymbol{\Psi}$ that have been suggested in the literature, using a simulation study.[3] Note that for most Bayesian software, including the software WinBUGS that we use for the simulation study (Lunn et al., 2000), one actually specifies a Wishart distribution for the precisions, rather than the IW for the variances. The relation of the IW and the Wishart distribution is that if X (here, the precision matrix) is Wishart distributed with scale matrix V and degrees of freedom df, then variable X^{-1} (here, the covariance matrix)

[3]We considered the scaled Wishart described by Gelman and Hill (2007) as well, however, this specification resulted in traps in WinBUGS (e.g., the estimation procedure would crash). Further, we considered specifying the variances and covariances in a regression structure avoiding the use of the IW prior specification, specifying the model with univariate priors, and to transform the random parameters so that they have a larger variance, and specifying an IW prior for the covariance matrix of the transformed parameters. For this work however, we decided to focus on different specifications of IW specifications suggested in the literature. More information on the other specifications is available from the first author.

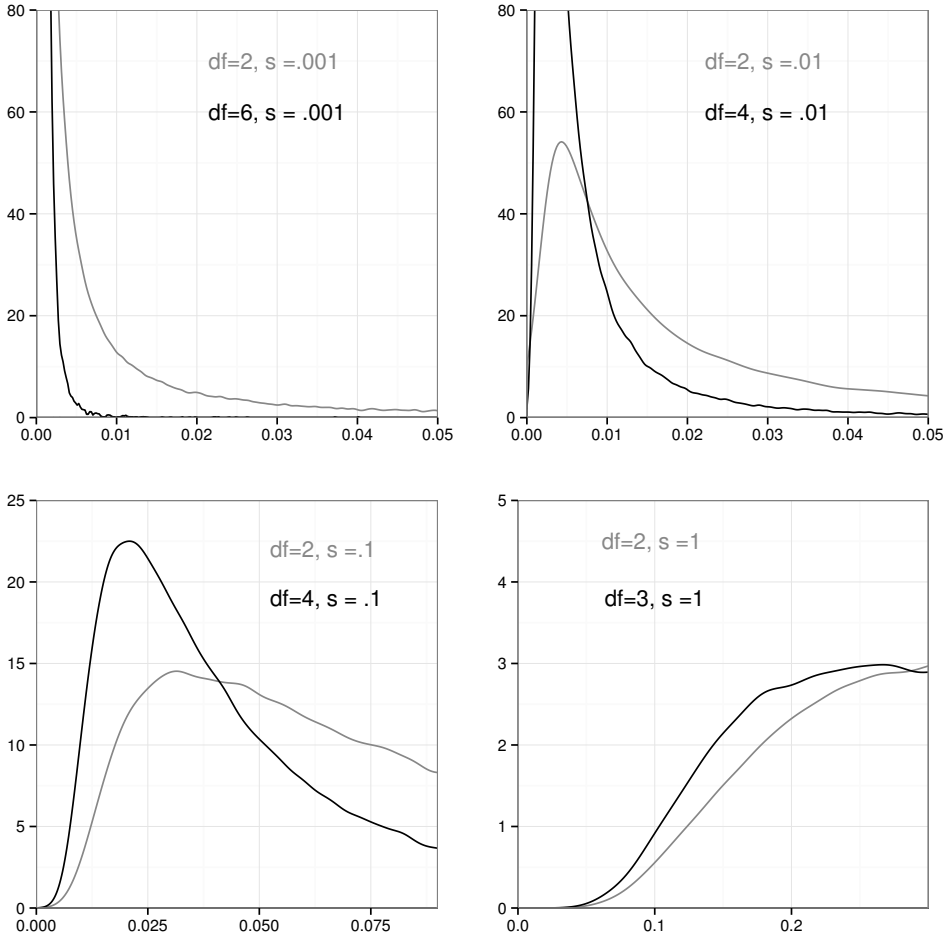


Figure 2.2: Eight Inverse-Wishart (IW) marginal probability densities, each specified with specific degrees of freedom *df* and scale matrix $\boldsymbol{S}$. The densities are based on samples from a bivariate IW distribution with a diagonal $\boldsymbol{S}$. All eight specifications are informative: In the area around zero, each specification gives either a lot of weight to a specific value around zero, or gives almost no weight around zero. As such, the IW prior distribution is easily misspecified if a IW distributed variable is close to zero.

is IW distributed with scale matrix V^{-1} and degrees of freedom df. Below, we discuss the prior specifications in terms of the IW distribution. For the corresponding Wishart specification, the scale matrix is simply inverted. For all three specifications the df are set equal to r (here $r = 3$), so that the priors are minimally informative (Gelman, Carlin, Stern, & Rubin, 2003; Gelman & Hill, 2007).

The first IW prior specification we will examine is the one that is commonly used as an uninformative prior specification, and which we will refer to as the Identity Matrix (ID) specification. In this specification the diagonal elements of scale matrix **S** are set to 1 and the off-diagonal elements are set to zero. We expect that this prior specification will prove to be quite informative in the current context where the variance of ϕ_j is small.

The second IW prior specification that will be examined is an IW prior specification in which the scale matrix is based on prior estimates of the variances of the random parameters. Using estimates of the variances as input for the IW prior specification ensures that the prior specification will be close to the data, and therefore should limit bias. However, this requires us to use the data twice; once for estimating the input for the prior, and again for the likelihood. When the data are used twice, the certainty about the estimated parameters is exaggerated (Kass & Steffey, 1989; see also discussions on the use of Empirical Bayes by Gill, p. 276–270 2008, and Lindley, 1969, p. 420–421). This can have statistical repercussions: because certainty about the point estimates is exaggerated, the standard deviations of the point estimates and their credible intervals will become too small (Kass & Steffey, 1989). How much the estimates will be influenced by using the data twice, will depend on how, and how much of the data is used. When the information in the used data is little, the effect will be negligible asymptotically (see for instance Berger & Pericchi, 1996; O'Hagan, 1995, on training data). Setting the df of the IW specification as small as possible minimizes the information value of the data-based prior, and therefore limits the effects of exaggerating the certainty about the point estimates. We will examine the effect of using the data twice as part of the simulation study, for instance by examining the coverage rates for the credible intervals of the estimated variances.

For the simulation study we will use a maximum likelihood procedure to obtain prior estimates of the variances, and we will refer to the prior specification based on these estimates as the maximum likelihood input (ML) specification. In this ML prior specification, we specify the Wishart prior distribution in WinBUGS so that the ML estimates of the variances of the random parameters are plugged into the Wishart

distribution scale matrix $\boldsymbol{S}^{-1}$ so that the mean of the Wishart distribution equals the estimated precisions (inverted variances). Note however that it is also possible to obtain estimates of the variances by other means — for instance by fitting a Bayesian model with uniform priors on the variances (ignoring any covariances), and base the IW or Wishart scale matrix on those estimates of the variances. Another option would be to fit an autoregressive model for each individual separately, provided that there are enough repeated measures per person to do this. Afterwards, the variances can be estimated by computing the variances of the estimated individual coefficients, which can then be used for the IW or Wishart prior specification. For our simulation study we opt for ML mainly because of its speed. We obtain the ML estimates by fitting the model in R (R Development Core Team, 2012) with the R-package lme4 (Bates, Maechler, & Bolker, 2012). In order to estimate the variance of the mean and not the variance of the intercept, the model in lme4 was fit on person-centered data. We used only the ML estimated variances as input for the prior specification, while setting the prior covariances to zero, because preliminary results showed that using estimates of the full covariance matrix decreased performance, probably because the ML-estimates for the covariances were not close to the true values.

The third IW specification we consider is the Default Conjugate (DC) prior proposed by Kass and Natarajan (2006). In the DC prior specification, the mean of the IW distribution is set to

$$\left(\frac{1}{n}\sum_{j=1}^{k}\boldsymbol{Z_j'W_jZ_j}\right)^{-1}, \tag{2.8}$$

where n is the number of participants, $\boldsymbol{Z_j}$ is the design matrix for person j, and $\boldsymbol{W_j}$ is the generalized linear model weight matrix for person j. The latter is based on (estimates of) model parameters. In the case of a normal model, $\boldsymbol{W_j}$ is a diagonal matrix with $1/\sigma^2$ on the diagonal (see Fahrmeir & Tutz, 1994). Given that we need an estimate of the residual variance σ^2 for the specification of the DC prior specification, we fit the multilevel model with maximum likelihood techniques in R-package lme4, and use ML-estimates of σ^2 as input for the generalized linear model weight matrix. Therefore, this specification is also data-based. However, the information in the data used will be little, so that the effect of using the data twice should be negligible, as is shown by Kass and Natarajan (2006).

The effect of the DC prior specification is that half of the prior weight on the random parameters is given to the common effects $(\gamma_\mu, \gamma_\phi, \gamma_\beta)$, and half of the weight

is given to estimates for each individual separately (μ, ϕ, β for each individual) as if a model was fit for each individual separately. This approach is directly related to shrinkage estimates (see Bryk & Raudenbush, 1992; Kass & Natarajan, 2006, but note that the weight on the random parameters is not necessarily one half for shrinkage estimates). In other words, the prior information is specified so that the prior weight is in between a parameter variance of zero (i.e., no individual differences) and the maximum parameter variance (i.e., maximum individual differences).

Kass and Natarajan (2006) compare the performance of the DC prior specification and ML specification for a Poisson model in a simulation study. In their study the DC prior outperformed the ML prior in terms of coverage rates, and squared and entropy loss. However, the model used was univariate with respect to the random parameters: only one parameter was random, so only one variance had to be modeled. Hence it remains unclear how the DC prior performs with regard to the estimation of the covariances between the random parameters. It also remains unclear how the DC prior performs when variances are close to zero. We will investigate these issues in the simulation study.

2.3 Simulation Study

Our simulation study consists of two parts: in the first part we examine the performance of the Wishart priors for different sizes of (small) variances in $\mathbf{\Psi}$, and in the second part we examine the performance of the IW prior specifications for different sample sizes and covariance structures when one or more variances in the covariance matrix are small. We compare three prior specifications for $\mathbf{\Psi}$ as discussed previously: the ID specification, the ML specification, and the DC specification (Kass & Natarajan, 2006). We will evaluate the performance of these three specifications against a specification that has the *df* set to 3, and the means of the IW distribution set to the true values. In practice this benchmark (BM) specification of course cannot be used, but we use it in the simulation study to determine optimal performance. For both parts the data are simulated according to the previously described model in open source software R (R Development Core Team, 2012). For both parts of the study the models are simulated 1,000 times (1,000 replications). In both parts of the simulation study γ_μ is set to 3, ψ^2_μ is set to 0.25, γ_ϕ is set to 0.3, γ_β is set to 0.35, and σ^2 is set to 1. The variance of the predictor variable x, ν^2, is set to 1.2.

For both parts we implemented and estimated all models in free Bayesian mod-

eling software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), in combination with the R-package R2WinBUGS (Sturtz, Ligges, & Gelman, 2005). We chose $Normal(0, 10^{-9})$ priors (specified in terms of precision instead of variance, which is required in WinBUGS) for the fixed effects γ_ϕ, γ_β, and γ_μ, and a $Uniform(0, 10)$ prior for σ^2, the residual variance at level 1. We evaluated the convergence of each model based on the visual inspection of the mixing of the three chains, and the Gelman-Rubin convergence diagnostic (Gelman & Rubin, 1992). We also evaluated the autocorrelations for the samples. Practically it was not possible to evaluate the convergence for each replication of each model in the simulation study (e.g., run each replication with three chains and visually inspect the convergence). Instead, we fitted and evaluated the convergence for one replication for each different condition in Part I and II of the study.

For all models convergence results were fairly similar. The three chains mixed well for all models and parameters. The Gelman-Rubin diagnostic was 1, or very close to 1, for the parameters in all models. Generally, the autocorrelations decreased exponentially to zero for the parameters σ^2, γ_μ, and ψ_μ, and the individual $\mu_j s$. For the remaining parameters, the autocorrelations were generally a bit slower to decrease, depending on the sample size and size of the variances of the random parameters. For most model specifications autocorrelations for these parameters diminished to zero after about 20 lags. Autocorrelations diminished to zero after about 40 to 60 lags when variance size is the smallest (.0025), and sample size is the smallest (25 persons and repeated measures). Based on the convergence results and the autocorrelations, we judged 40,000 iterations with 30,000 burn-in iterations as sufficient for convergence.

As point estimates for the parameters we used the means of the posterior distributions. Performance is evaluated using: a) coverage rates of the 95% (equal tailed) credibility intervals (CIs), which we would expect to be about .95 when the priors are uninformative; b) bias, which is computed by taking the average of the difference between the true value and the point estimate across all 1,000 replications; and c) the ratio of the average posterior standard deviation and the standard deviation of the posterior averages, which should be about 1 if the posterior standard deviations reflect the actual sampling variation.

Part I: The Effect of Small Variances in Ψ

In order to study the effect of the (small) size of the variances of the random parameters per prior for $\mathbf{\Psi}$, the variances ψ^2_ϕ and ψ^2_β were either set to .0025, .01, or .0225. These variances result in 95% intervals for the autoregression coefficients of respectively [.202, .398], [.104, .496], and [.006, .594]. These ranges are in line with autoregressive coefficients reported in the literature, which are usually small and positive (e.g, Moberly and Watkins (2008), Nezlek and Gable (2001), Suls et al. (1998), and Wang et al. (2012) report fixed autoregression effects between .08 and .3 approximately). The sample size is set to 50 individuals and 50 time points. All the correlations between the random parameters are set to .3. This results in a 4×3 (i.e., $priorspecification \times sizevariance$) simulation design. Below, we summarize the results for Part I of the simulation study. More detailed results for the simulation study are available as supplementary materials with the online paper (Schuurman, Grasman, & Hamaker, 2016) or at www.nkschuurman.com, and the simulated data are available upon request from the first author.

The results show that overall, the ML prior specification performs best. The bias of the ML specification is quite close to that of the BM specification. It can be seen from Figure 2.3 that even though coverage rates are lower than 95% for ψ^2_ϕ and ψ^2_β for this prior specification, it outperforms both the ID specification and DC specification. The coverage rates for ψ^2_ϕ and ψ^2_β are lower than .95 likely as a result of the double use of data: the data is used once in the prior and again in the likelihood, and as a result the information about the estimation is exaggerated, which in turn results in smaller credible intervals. The ID specification severely overestimates ψ^2_ϕ and ψ^2_β. The DC specification performs well only if the prior specification is close to the true values of ψ^2_ϕ and ψ^2_β. In this simulation study this was the case when the true variances were .01 or .0225, but not when they were .0025. Since in practice it is unknown if the DC prior is close to the true values of ψ^2_ϕ and ψ^2_β, it is an undependable prior to use when the aim is to use an uninformative prior for the covariance matrix, while some variances are close to zero. The ML specification on the other hand is by definition close to the information in the data and therefore performs relatively well.

Note that when a variance is further away from zero, all prior specifications perform reasonably well: The true value for the variance ψ^2_μ for random effect μ_j was .25, and it can be seen from Table 2.1 to 2.4 that all prior specifications perform well for this parameter. We discuss the results per prior specification in more detail below.

Table 2.1: Part I: Coverage rates for the 95% credible intervals, calculated over 1000 replications.

	$\psi^2_\beta = \psi^2_\phi = .0025$				$\psi^2_\beta = \psi^2_\phi = .01$				$\psi^2_\beta = \psi^2_\phi = .0225$			
	BM	ID	ML	DC	BM	ID	ML	DC	BM	ID	ML	DC
μ_j	.949	.951	.950	.948	.949	.952	.950	.949	.947	.949	.947	.946
β_j	.970	.977	.920	.990	.947	.972	.929	.959	.941	.964	.937	.939
ϕ_j	.971	.979	.893	.984	.948	.974	.928	.953	.941	.965	.937	.936
γ_μ	.937	.939	.935	.932	.967	.966	.965	.962	.948	.948	.946	.940
γ_β	.966	.999	.946	.981	.968	.999	.963	.972	.948	.991	.946	.948
γ_ϕ	.964	.999	.937	.979	.957	.999	.956	.962	.950	.992	.946	.947
ψ^2_μ	.948	.952	.930	.941	.937	.940	.917	.946	.965	.968	.945	.967
ψ^2_β	.990	.000	.730	.000	.984	.000	.889	.959	.967	.000	.895	.963
ψ^2_ϕ	.995	.000	.678	.003	.989	.000	.905	.974	.979	.000	.932	.965
$\psi_{\mu\beta}$	.991	1.000	.950	.988	.980	1.000	.968	.975	.978	.998	.973	.968
$\psi_{\mu\phi}$	.987	.999	.876	.832	.977	1.000	.953	.872	.975	.998	.964	.912
$\psi_{\phi\beta}$	1.000	1.000	.907	1.000	.985	1.000	.962	.986	.971	1.000	.958	.957
$\rho_{\mu\beta}$	.996	.848	.984	.951	.986	.940	.981	.975	.975	.959	.976	.966
$\rho_{\mu\phi}$	.991	.851	.978	.681	.980	.929	.984	.857	.976	.952	.970	.928
$\rho_{\phi\beta}$	.999	.676	.998	.987	.987	.849	.985	.980	.978	.921	.979	.971

*Note.*A coverage rate of .95 is considered optimal. The coverage rates are calculated for three different true values of variances ψ^2_ϕ and ψ^2_β. Coverage rates are shown for the benchmark (BM) prior specification, the identity matrix (ID) prior specification, the maximum likelihood (ML) input specification, and the default conjugate (DC) prior specification, and for the following parameters: the random effects μ_j, ϕ_j, and β_j, the fixed effects γ_{μ_j}, γ_{ϕ_j}, and γ_{β_j}, all elements from the covariance matrix Ψ for the random effects, and the correlations between the random effects $\rho_{\mu\phi}$, $\rho_{\mu\beta}$, and $\rho_{\phi\beta}$.

Table 2.2: Part I: Estimated bias for the estimated means of the posterior distributions, calculated over 1000 replications.

	$\psi^2_\beta = \psi^2_\phi = .0025$				$\psi^2_\beta = \psi^2_\phi = .01$				$\psi^2_\beta = \psi^2_\phi = .0225$			
True	BM	ID	ML	DC	BM	ID	ML	DC	BM	ID	ML	DC
$\mu_j =^a$	.001	-.001	.000	-.001	-.001	-.003	.000	-.001	-.001	-.004	-.001	-.002
$\beta_j =^a$	-.000	.000	-.001	-.000	-.000	-.000	-.000	-.000	-.000	-.000	-.000	-.000
$\phi_j =^a$	-.002	-.007	-.002	-.002	-.001	-.006	-.001	-.001	.000	-.004	.000	.001
$\gamma_\mu = 3$	-.000	-.002	-.001	-.002	-.003	-.007	-.004	-.005	.000	-.003	.000	-.001
$\gamma_\beta = 0.35$	-.001	-.000	-.001	-.000	-.000	-.000	-.000	-.000	-.000	-.000	-.000	-.000
$\gamma_\phi = 0.3$	-.002	-.007	-.002	-.003	-.000	-.005	-.000	-.000	.000	-.004	.000	.001
$\psi^2_\mu = 0.25$	.021	.023	.024	.009	.023	.026	.026	.012	.024	.027	.028	.014
$\psi^2_\beta =^b$	.002	.041	.002	.007	.002	.039	.002	.004	.002	.036	.002	.001
$\psi^2_\phi =^b$	.002	.041	.002	.006	.002	.038	.002	.003	.002	.034	.002	.000
$\psi^2_{\mu\beta} =^b$	-.000	-.001	-.003	-.001	.002	-.001	-.001	.001	.002	-.002	-.001	.000
$\psi^2_{\mu\phi} =^b$	-.001	-.001	-.002	-.012	.001	-.002	-.002	-.010	.001	-.003	-.002	-.008
$\psi^2_{\phi\beta} =^b$	-.000	-.000	-.000	-.000	-.001	-.001	-.001	-.001	.000	-.001	-.001	-.000
$\rho_{\mu\beta} = 0.3$	-.060	-.245	-.172	-.175	.002	-.181	-.061	-.052	.009	-.139	-.034	-.005
$\rho_{\mu\phi} = 0.3$	-.053	-.243	-.165	-.422	-.011	-.190	-.073	-.235	.001	-.148	-.042	-.118
$\rho_{\phi\beta} = 0.3$	-.091	-.287	-.249	-.263	-.111	-.254	-.125	-.138	.003	-.206	-.051	-.021

[a] The simulation values for these parameters vary over individuals and replications.

[b] The simulation values for these parameters vary for study I (see top row of table).

Note. A bias of zero is considered optimal. The estimated bias is calculated for three different true values of variances ψ^2_ϕ and ψ^2_β. The estimated bias is shown for the benchmark (BM) prior specification, the identity matrix (ID) prior specification, the maximum likelihood (ML) input specification, and the default conjugate (DC) prior specification per parameter.

Table 2.3: Part I: Bias relative to the true value of the parameter in percentages for the fixed effects, and variances and covariances for the random parameters, calculated over 1000 replications.

	$\psi^2_\beta = \psi^2_\phi = .0025$				$\psi^2_\beta = \psi^2_\phi = .01$				$\psi^2_\beta = \psi^2_\phi = .0225$			
True	BM	ID	ML	DC	BM	ID	ML	DC	BM	ID	ML	DC
$\gamma_\mu = 3$	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
$\gamma_\beta = 0.35$	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
$\gamma_\phi = 0.3$	-1%	-2%	-1%	-1%	0%	-2%	0%	0%	0%	-1%	0%	0%
$\psi^2_\mu = 0.25$	9%	9%	9%	3%	9%	10%	10%	5%	9%	11%	11%	5%
$\psi^2_\beta =^b$	67%	1657%	87%	295%	19%	391%	18%	43%	10%	160%	10%	7%
$\psi^2_\phi =^b$	67%	1638%	77%	236%	18%	383%	16%	29%	8%	153%	7%	1%
$\psi^2_{\mu\beta} =^b$	-3%	-18%	-37%	-13%	14%	-5%	-9%	4%	9%	-9%	-6%	1%
$\psi^2_{\mu\phi} =^b$	-9%	-13%	-31%	-156%	9%	-13%	-14%	-64%	6%	-15%	-9%	-35%
$\psi^2_{\phi\beta} =^b$	-13%	-21%	-58%	-46%	-24%	-22%	-30%	-23%	5%	-19%	-14%	-7%
$\rho_{\mu\beta} = 0.3$	-20%	-82%	-57%	-58%	1%	-60%	-20%	-17%	3%	-46%	-11%	-2%
$\rho_{\mu\phi} = 0.3$	-18%	-81%	-55%	-141%	-4%	-63%	-24%	-78%	0%	-49%	-14%	-39%
$\rho_{\phi\beta} = 0.3$	-30%	-96%	-83%	-88%	-37%	-85%	-42%	-46%	1%	-69%	-17%	-7%

[a] The simulation values for these parameters vary for study I (see top row of table).
Note. A relative bias of zero is considered optimal.

Table 2.4: Part I: Ratio of estimated average posterior standard deviations and calculated standard deviations of the estimated posterior means over 1000 replications.

	$\psi^2_\beta = \psi^2_\phi = .0025$				$\psi^2_\beta = \psi^2_\phi = .01$				$\psi^2_\beta = \psi^2_\phi = .0225$			
	BM	ID	ML	DC	BM	ID	ML	DC	BM	ID	ML	DC
γ_μ	.983	.992	.986	.964	1.088	1.100	1.091	1.069	1.028	1.042	1.035	1.014
γ_β	1.048	1.770	.820	1.189	1.083	1.625	1.074	1.129	1.031	1.385	1.030	1.022
γ_ϕ	1.087	1.847	1.037	1.196	1.062	1.627	1.057	1.080	1.038	1.404	1.028	1.012
ψ^2_μ	1.100	1.130	1.020	1.046	1.080	1.108	1.000	1.037	1.170	1.197	1.084	1.128
ψ^2_β	1.654	3.768	.720	1.946	1.350	2.879	.882	1.440	1.091	2.030	.877	1.067
ψ^2_ϕ	1.778	4.094	.714	1.934	1.394	3.226	.917	1.384	1.181	2.290	.959	1.095
$\psi_{\mu\beta}$	1.392	2.058	1.261	1.304	1.201	1.758	1.183	1.159	1.146	1.551	1.151	1.095
$\psi_{\mu\phi}$	1.284	1.976	1.137	1.125	1.181	1.788	1.162	1.066	1.152	1.612	1.158	1.050
$\psi_{\phi\beta}$	1.769	3.762	1.556	1.950	1.332	2.912	1.301	1.365	1.133	2.124	1.130	1.075
$\rho_{\mu\beta}$	1.456	2.029	1.553	1.296	1.200	1.764	1.256	1.153	1.124	1.551	1.153	1.070
$\rho_{\mu\phi}$	1.382	1.977	1.510	1.065	1.189	1.801	1.234	1.027	1.105	1.601	1.120	1.000
$\rho_{\phi\beta}$	2.071	3.716	2.411	2.035	1.388	2.889	1.453	1.407	1.143	2.136	1.157	1.080

Note. A ratio of 1 is considered optimal, with a ratio > 1 indicating an overestimation of the posterior standard deviations, and a ratio < 1 indicating an underestimation of the posterior standard deviations. The ratios are calculated for three different true values of variances ψ^2_ϕ and ψ^2_β. The ratios are shown for the benchmark (BM) prior specification, the identity matrix (ID) prior specification, the maximum likelihood (ML) input specification, and the default conjugate (DC) prior specification, and for the following parameters: the random effects μ_j, ϕ_j, and β_j, the fixed effects γ_{μ_j}, γ_{ϕ_j}, and γ_{β_j}, all elements from the covariance matrix Ψ for the random effects, and the correlations between the random effects $\rho_{\mu\phi}$, $\rho_{\mu\beta}$, and $\rho_{\phi\beta}$.

Performance ID specification

Overall, the ID specification performs poorly. From Figure 2.3 it can be seen that the coverage rates for ψ^2_ϕ and ψ^2_β are equal to zero regardless of the true sizes for ψ^2_ϕ and ψ^2_β, indicating that the true values were never within the credible interval. This is due to a large bias in these parameter estimates: The parameters ψ^2_ϕ and ψ^2_β are severely overestimated. The coverage rates for the remaining parameters on the other hand are equal or close to one (see Table 2.1), as a result of too conservative standard deviations for all parameters for the ID specification. This is illustrated by the the ratios of the average posterior standard deviations and calculated standard deviations of the posterior means, over 1,000 replications, reported in Table 2.4, which are much larger than one — indicating too large posterior standard deviations.

Performance ML specification

The coverage rates for the ML specification are similar to those for the BM specification, except for the estimates of the random effects ϕ_j, β_j, and their variances ψ^2_ϕ and ψ^2_β. The coverage rates for ψ^2_ϕ and ψ^2_β for the ML specification are low compared to the BM specification, ranging from approximately .70 when ψ^2_ϕ and ψ^2_β are .0025 to approximately .90 when ψ^2_ϕ and ψ^2_β are .01 or .0225 . These coverage rates are however considerably better than those for the ID specification. The coverage rates for the random parameters ϕ_j and β_j range from .893 to .936 (lowest rates for a true variance of .0025), which is relatively low compared to the other prior specifications, including the BM specification. These results are consistent with the relatively small posterior standard deviations for ψ^2_ϕ, ψ^2_β, ϕ_j, and β_j for the ML specification compared to the other prior specifications, as can be seen from the ratios of standard deviations in Table 2.4. The relatively small posterior standard deviations for these parameters are likely the consequence of the double use of data. Further, it can be seen from Tables 2.2 and 2.3 that the ML specification results in very little bias compared to the ID and DC specification, and that the amount of bias is actually similar to that of the BM specification.

Performance DC specification

The performance of the DC specification varies depending on the true values for the variances ψ^2_ϕ and ψ^2_β, as can be most clearly seen from a plot of the bias and

coverage rates in Figure 2.3. The DC specification performs well when ψ^2_ϕ and ψ^2_β are .01 and .0225: the coverage rates, bias, and ratios of the posterior standard deviations and standard errors for the DC specification are then close to those for the BM specification. However, when the variances are equal to .0025, performance strongly declines, with coverage rates for ψ^2_ϕ and ψ^2_β that are equal to zero. Closer inspection of the results indicates that this low coverage rate is due to an upward bias for these parameters. The ratio of the posterior standard deviations and standard errors also strongly increases when ψ^2_ϕ and ψ^2_β are .0025, indicating that the posterior standard deviations are overestimated. The upward bias for the parameters is so large however that it is not compensated by the relatively wide credible intervals. The DC prior has low coverage rates for the covariance and correlation between μ_j and ϕ_j. This is the result of a downward bias for this parameter, which seems due to the DC prior specification which sets $\rho_{\phi\mu}$ to approximately $-.90$ based on Equation 2.8. The coverage rates for the remaining parameters are high compared to the BM specification, due to relatively large posterior standard deviations for these parameters, as shown in Table 2.4.

The discrepancy in performance of the DC prior specification across the three values of ψ^2_ϕ and ψ^2_β probably results because the DC specification does not depend directly on ψ^2_ϕ and ψ^2_β, so that it does not change much in accordance with ψ^2_ϕ and ψ^2_β. Therefore, the input for the DC specification may be similar regardless of the true value for ψ^2_ϕ and ψ^2_β. When the information in the DC specification for ψ^2_ϕ and ψ^2_β is not close to the information from the data, it biases the estimates for ψ^2_ϕ and ψ^2_β. This can be seen most clearly from Figure 2.3, which shows that the bias increases when the true variance diverges from the DC prior specification: When ψ^2_ϕ and ψ^2_β are .0025 or .01 the bias is positive, and when they are .0025 the bias turns negative. Apparently the DC specification was close enough to ψ^2_ϕ and ψ^2_β when their true values were .01 and .0225, but not close enough when their true values were .0025.

Part II: The Effects of Sample Size and Covariance Structure

For Part II of our simulation study, we aim to study the effect of sample size and the sizes of the covariances or correlations on the parameter estimates for each prior specification for $\mathbf{\Psi}$. For this purpose we vary sample sizes between 25, 50, and 75 for both number of individuals and time points, and the correlations between the random parameters are either all set to 0 or all set to .3. The variances for both the

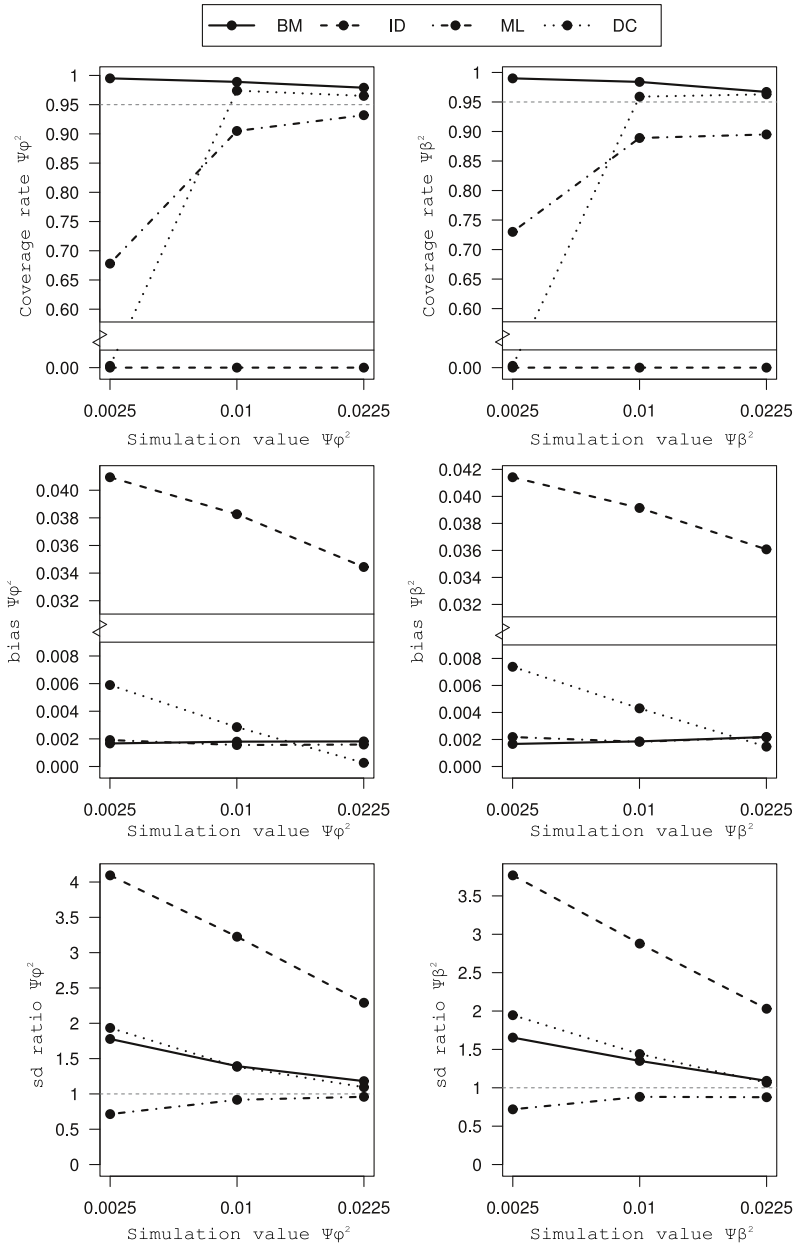


Figure 2.3: Part I coverage rates, estimated bias, and ratios of the average estimated posterior standard deviations and the standard deviations of the estimated posterior means, for ψ_ϕ and ψ_β calculated over 1000 replications. The coverage rates, bias, and ratios of standard deviations are shown for the benchmark (BM), identity matrix (ID), maximum likelihood input (ML), and default conjugate (DC) prior specification, for true values for ψ_ϕ and ψ_β of .0025, .01, and .0225.

crossregression and autoregression coefficients are set to .01 for this study, which is the medium value for the variances in Part I. This results in a $4 \times 3 \times 3 \times 2$ (i.e., $priorspecification \times timepoints \times individuals \times correlationmatrixspecification$) simulation design. The results for Part II of the simulation study are presented below.

Effects of sample size

In general, when sample size increased, parameter estimates improved as would be expected: The bias became smaller, the coverage rates became closer to .95, and the ratios of the posterior standard deviations and standard errors became closer to 1. Figure 2.4 contains graphs for the coverage rates, bias, and ratios of standard deviations for ψ^2_ϕ, and ψ^2_β for sample sizes of 25 time points and individuals, 50 time points and individuals, and 75 time points and individuals. The results for different combinations of time points and individuals, such as 25 time points and 50 individuals, were not included in Figure 2.4 to save space; these results, as well as the results for the other parameters are available as supplementary materials with the online paper (Schuurman, Grasman, & Hamaker, 2016) or at www.nkschuurman.com, and the simulated data are available upon request from the first author.

The estimates for μ_j, ϕ_j, and β_j improved when the number of time points increased, as would be expected for within-subject parameters. For the remaining parameters, estimates improved both when time points and number of individuals increased, as would be expected for between-subject parameters. Increasing the number of individuals seems most advantageous for these parameters. Noteworthy is that for all sample sizes and all prior specifications, including the BM specification, the credible intervals and posterior standard deviations for the correlations were quite large: For the BM specification the posterior standard deviations ranged from approximately .30 for the smallest sample size to .16 for the largest sample size. Although the accuracy of the estimates of the correlations increases as sample size increases, efficient estimates of the correlations clearly will require even larger sample sizes. We discuss the results per prior specification in more detail below.

The ID specification did not perform well regardless of sample size, as can be seen from Figure 2.4. The coverage rates for ψ^2_ϕ and ψ^j_β were equal to zero, regardless of number of time points or individuals. Although the bias in the parameter estimates decreased when sample sizes increased, it remained large, which was reflected in the

coverage rates. The ratios of standard deviations are larger than 1, and large compared to the other prior specifications, which indicates that the posterior standard deviations are relatively large across sample sizes, resulting in relatively large credible intervals.

For the ML specification, the coverage rates for ϕ_j and β_j were low when sample sizes were small, but they improved as sample size increased, from approximately .88 for 25 time points and individuals to .94 for 75 time points and individuals. The coverage rates for ψ^2_ϕ and ψ^2_β also improved when sample size increased, from approximately .73 for 25 time points and individuals to .90 for 75 time points and individuals (see Figure 2.4).

The performance of the DC prior specification also increased when sample size increased. However, Figure 2.4 shows that for small sample sizes the DC specification shows an especially sharp drop in coverage rates ψ^2_β, indicating that for this parameter the small sample was not enough to dominate the prior. In general, for all prior specifications and across both parts of the simulation study, the estimates for ψ^2_ϕ seem to be slightly less biased than those for ψ^2_β. In this case the true value for ψ^2_β seems to lie just outside the credible interval, whereas the true value for ψ^2_ϕ lies just within the credible interval, resulting in this sharp drop in coverage rates for ψ^2_β, but less so for ψ^2_ϕ. The estimates for the covariance $\psi_{\mu\phi}$ and correlation $\rho_{\mu\phi}$ improve strongly as sample size increases, with coverage rates ranging from approximately .55 to .92 for $\psi_{\mu\phi}$ and from .43 to .93 for $\rho_{\mu\phi}$, and bias ranging from approximately from $-.04$ to $-.003$ for $\psi_{\mu\phi}$, and from $-.7$ to $-.08$ for $\rho_{\mu\phi}$, for the smallest to the largest sample sizes respectively.

Effects of covariance structure

In general, performance did not differ much when the covariance structure was altered, except for the estimates of the covariances and correlations of the random parameters for the ID and ML prior specification. Note that the performance for the correlations and covariances will not necessarily be the same because the correlations are also affected by the estimates of the variances. However, for both the correlations and covariances estimates were better when the true values of the covariances were set to zero, which is not surprising since these prior specifications had covariances set to zero. When correlations of .3 were used to generate the data, the covariance and correlation estimates were downward biased for these specifications compared to the benchmark

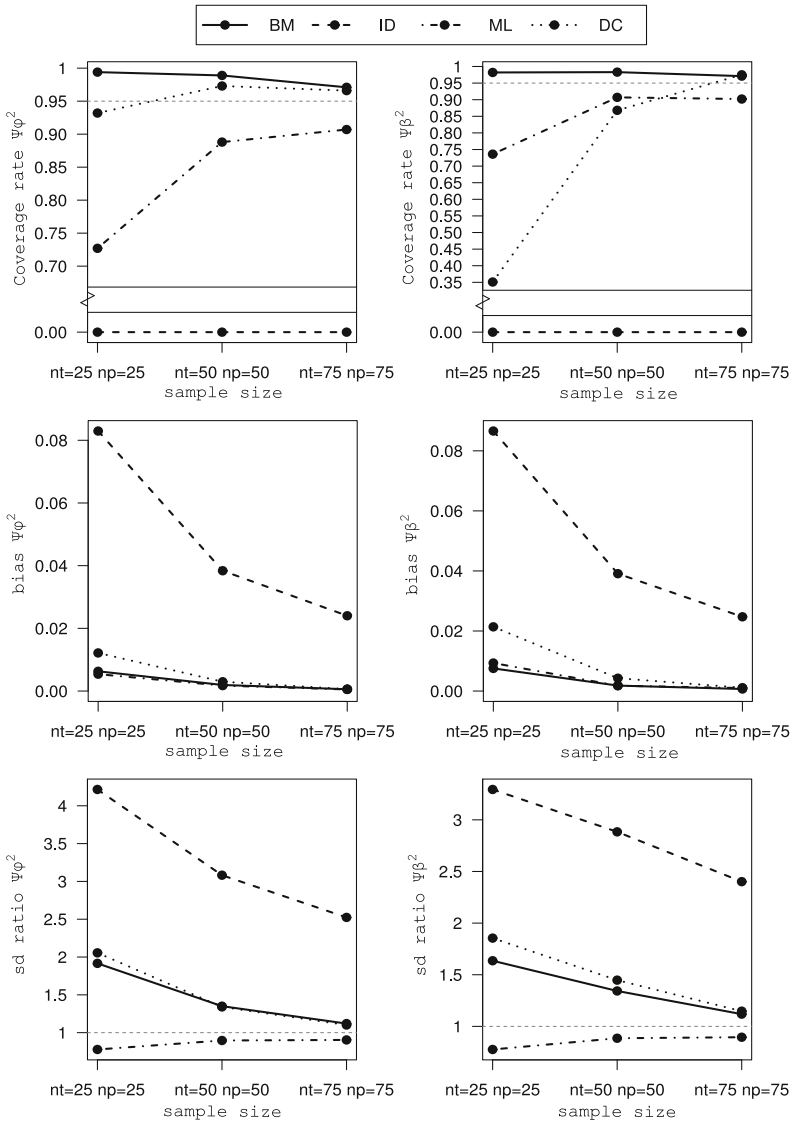


Figure 2.4: Part II coverage rates, estimated bias, and ratios of the average estimated posterior standard deviations and the standard deviations of the estimated posterior means, for ψ_ϕ and ψ_β calculated over 1000 replications for the models with correlations set to .3. The coverage rates, bias, and ratios of standard deviations are shown for the benchmark (BM), identity matrix (ID), maximum likelihood (ML), and default conjugate (DC) prior specification, for sample sizes of 25 time points and 25 individuals, 50 time points and 50 individuals, and 75 time points and 75 individuals.

specification. This relatively large bias compared to the benchmark specification was absent for the ID and ML specification when the true values of the correlations were equal to zero, and decreased when sample size increased so the data dominated the prior more. For the smallest to the largest sample sizes, for the BM specification this bias ranged from approximately $-.06$ to .004 for $\rho_{\mu\phi}$ and $\rho_{\mu\phi}$, and from $-.18$ to $-.04$ for $\rho_{\mu\phi}$. For the ID specification the bias ranged from approximately $-.24$ to $-.16$ for $\rho_{\mu\phi}$ and $\rho_{\mu\phi}$, and from $-.27$ to $-.23$ for $\rho_{\beta\phi}$. The bias for the ML specification was considerably less with the bias for $\rho_{\mu\phi}$ and $\rho_{\mu\phi}$ ranging from $-.19$ to $-.03$, and from $-.26$ to $-.04$ for $\rho_{\beta\phi}$. For the DC specification, the bias for the correlations was generally in between the bias for the ML and BM specification, except for $\rho_{\mu\phi}$, as was the case for this parameter for the DC specification for Part I of the simulation study. For all prior specifications, the coverage rates and ratios of standard deviations were not clearly affected by the different true correlation values. We briefly evaluated the performance of the BM and ML model for correlations equal to .7 rather than .3, with sample sizes of 25 occasions and persons, and 75 occasions and persons. As would be expected, the bias in the correlations for the ML specification was more severe when the correlations were equal to .7. For the rest, results were comparable to the condition for which the correlations were equal to .3.

Conclusion

Overall, the ML prior specification outperformed the other prior specifications. The ID specification, which is probably one of the most common choices in specifying uninformative priors for covariance matrices in practice, is not a good choice when variances may be small, because it results in severely overestimated variances even for relatively large sample sizes. The DC specification performs better than the ID specification, but gives inconsistent results. That is, it strongly influences the results when the DC prior information is not close enough to the information in the data. Given that there is no guarantee that the prior information from the DC will be close to the information in the data, the performance of the DC prior is unreliable when variances are small. The ML specification on the other hand, is directly based on maximum likelihood estimates of the variances from the data, which provide a good guess of the true value of the variances. As a result, the ML specification performs relatively well. The double usage of the data in the ML specification however does have consequences for the standard deviations and credible intervals of the variances:

these are too small. However, this effect diminishes when sample size increases.

A disadvantage of the ML specification is that when the models of interest become more complex, it may be difficult to fit these models with traditional ML procedures and software — in fact, this may be one of the reasons to opt for Bayesian estimation in the first place. For instance, multivariate multilevel modeling is often not available in frequentist software whereas it is relatively easy to fit with Bayesian software. Two other examples are multilevel multivariate autoregressive models that include latent variables, and models that include random residual variances — both may not be possible with frequentist software, whereas it is relatively trivial within Bayesian software. There are several ways to get estimates for the covariance matrix of the random parameters, or the variances in this matrix, when the models are too complex to fit with traditional techniques: Firstly, it may be possible to use simpler models that allow ML estimation to get preliminary estimates of the variances in question. For instance, if one aims to fit a multivariate multilevel autoregressive model, one option is to fit multiple univariate models with ML techniques in order to get preliminary estimates for the variances, and use these estimates in the prior specification. Secondly, an option is to fit state-space models per individual instead of one multilevel model, and calculate the variances based on the estimates of the individual parameters. Thirdly, one can fit the model with Bayesian estimation using uniform priors for the variances, while disregarding potential covariances, and use these estimates for the variances in the specification of the IW prior distribution. In the following section we will illustrate this last option, together with the ID, DC, and ML prior specification, using an empirical example.

2.4 Empirical Application on Positive Affect and Worrying

The data for this empirical illustration consist of ESM measurements (see Geschwind, Peeters, Drukker, van Os, & Wichers, 2011). Each participant was alerted randomly throughout each day to fill out the provided questionnaires, for six days, resulting in approximately 45 repeated measures per participant. Here we focus on baseline measures for 129 participants of positive affect (PA), measured with principal component scores for seven items (I feel 'happy', 'satisfied', 'strong', 'enthusiastic', 'curious', 'animated', and 'inspired') (for details, see Geschwind et al., 2011), and on baseline measures of worrying, measured with the item 'I am worrying'. All items were answered on a scale from 1 to 7 (with 1 being 'not at all' and 7 being 'very much

so'). Because an assumption for AR(1) models is that time intervals between measurement are about equal, we added observations and coded these observations as missing between measurements, when time intervals between random measurements were especially large (e.g., between the last observation of a day and the first observation of the following day), resulting in an average time lag of about 1.7 hours.

In the psychological literature worrying is considered to be both potentially productive and potentially destructive. That is, worrying is productive when it results in solving a (potential) problem, reducing negative affect that accompanied the problem. In that case, worrying is considered an adaptive emotion-regulating strategy (Ehring & Watkins, 2008; Nolen-Hoeksema, Wisco, & Lyubomirsky, 2008; Pyszcynski & Greenberg, 1987; Watkins, 2008). On the other hand, it may become destructive when the problem cannot be solved, and worrying becomes repetitive or compulsive in continuously trying to solve the problem, exacerbating negative emotions related to the problem. This repetitive worrying has been considered as a maladaptive strategy to regulate emotions, and has been related to affect, especially negative affects such as sadness and anxiety, to rumination, and to various depressive and anxiety disorders (Aldao, Nolen-Hoeksema, & Schweizer, 2010; Ehring & Watkins, 2008; Nolen-Hoeksema et al., 2008; Pyszcynski & Greenberg, 1987; Querstret & Cropley, 2013; Watkins, 2008). Within the current modeling framework, a positive autoregressive effect for worrying may serve as an indication of such repetitive or compulsive thinking — reflecting that a person tends to 'get stuck' in their worrying across multiple observations. Here, we will explore this autoregressive effect of worrying, and that of PA, and we will explore the reciprocal effects of worrying and PA on each other, by means of fitting a multilevel multivariate autoregressive model. Furthermore, we will investigate whether there are any associations between the individual autoregressive effects, cross-lagged effects, and individual means. Note that this is possible because we are using a model with a multilevel and multivariate structure.

Modeling Approach

Applications of multivariate multilevel autoregressive models are sparse (see Lodewyckx et al., 2011, for an exception). Univariate applications are more commonly seen in the psychological literature (e.g., Cohn & Tronick, 1989; De Haan-Rietdijk et al., 2014; Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001; Rovine & Walls, 2005; Suls et al., 1998). When researchers are interested in reciprocal lagged

effects between two or more variables, they typically estimate several univariate models instead (e.g., Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001). The reason for this may be that it is difficult to estimate multivariate multilevel models using traditional software. Here, a Bayesian approach is extremely valuable, because it can be easily extended to multivariate processes. An additional advantage of the Bayesian approach that is especially important for longitudinal designs is that it handles missing data well. We have some missing data for the measures on worrying and PA as a result of nonresponse, as well as the observations we added and coded as missing as noted previously. As such, the Bayesian approach will be quite helpful here.

In order to illustrate the effect of different prior specifications for the covariance matrix Ψ of the random parameters (two means, two cross-lagged parameters, and two autoregressive parameters per person), we fit the model with the ID prior, the DC prior, and the ML prior. Because fitting a multivariate multilevel autoregressive model currently is not an option in ML software, we fit two univariate models with package lme4 in R (Bates et al., 2012; R Development Core Team, 2012) in order to obtain estimates for the variances in Ψ to plug into the ML and DC prior. This may not be ideal, because 1) fitting two univariate models ignores any residual correlation between PA and worrying; and 2) In lme4 any missing observations are discarded from the analysis by means of listwise deletion, so that many observations are disregarded in this analysis: one missing value in the dependent variable, also means a missing value in the predictor at the preceding occasion, resulting in the list-wise deletion of two observations. Therefore, we also specify a second data-based prior, based on first fitting a bivariate Bayesian model with Uniform priors on the variances of the random parameters. Although this model ignores any correlation between the random parameters, it allows for a residual correlation between PA and worrying, and more importantly, it efficiently deals with the missing observations. We plug the estimates for the variances of the random parameters of this model into an Inverse-Wishart prior for a full multivariate model, and we will refer to this Inverse-Wishart prior specification as the Bayesian data based (BDB) prior specification.

In the first six panels of Figure 2.5 we provide plots of the resulting (marginal) IW prior distributions for the variances of the random parameters. For the variances of the means (two top most panels) the ML, BDB, and DC prior specifications are quite similar, with the exception of the DC prior specification for the mean of worrying which is more similar to the ID prior specification. For the variances of the

autoregressive and cross-lagged coefficients there are more dissimilarities between the prior specifications, as we would expect. The ML, DC, and BDB prior distributions all peak in the area close to zero for the variances of these regression coefficients. For these parameters, the prior distributions for the ML and BDB specification are most similar (except for $\beta_{Wo->PAj}$), but they do not overlap completely, especially in the area close to zero. As expected, the ID prior peaks quite far away from zero, and is most dissimilar to the other prior specifications. The final two panels of Figure 2.5 show plots for two of the fifteen correlations between the random parameters, specifically between $\beta_{Wo->PAj}$ and ϕ_{Woj}, and between ϕ_{Woj} and μ_{Woj}. For the correlation between $\beta_{Wo->PAj}$ and ϕ_{Woj} the prior covariance was set to zero for all prior specifications, resulting in a symmetric, saddle-shaped distribution. For the correlation between ϕ_{Woj} and μ_{Woj}, the prior covariance was set to zero for all specifications except the DC prior, for which the prior is shifted in favor of a negative correlation.

We fitted each model with three chains, with each 40,000 samples of which 20,000 were burn-in. We evaluated the convergence of each model through the visual inspection of the mixing of the three chains, the Gelman-Rubin convergence diagnostic (Gelman & Rubin, 1992), and autocorrelations. Based on these results we judged 40,000 iterations with 20,000 burn-in iterations as sufficient for convergence. Code for R and WinBUGS for simulating data and fitting the bivariate model, based on the ML prior specification and the BDB prior specification, is provided as supplementary materials with the online paper (Schuurman, Grasman, & Hamaker, 2016) or at www.nkschuurman.com.

Results

From Table 2.5 it can be seen that for most parameters, the estimates are quite similar across the different prior specifications. As would be expected, the largest differences are between the estimated variances of the random autoregressive and cross-lagged parameters (see the random effect for ϕ_{PA} and ϕ_{Wo}, and $\beta_{PA\rightarrow Wo}$ and $\beta_{Wo\rightarrow PA}$ in Table 2.5), and therefore, between the estimates of the individual random parameters. For the models with the ML based prior and the BDB prior we find very similar estimates for the variances. For the DC prior, we also find similar results, albeit a somewhat smaller point estimate for the variance of the cross-lagged effect of PA on worrying, compared to the models with ML and BDB priors. In the model with the ID prior specification, the variances are consistently estimated to be about twice as

Table 2.5: Parameter estimates for the multilevel bivariate autoregressive model on positive affect and worrying (Posterior means and 95% CI), for four different prior specifications for the covariance matrix of the random parameters.

Parameter	ID	DC	ML	BDB
Fixed Effects for:				
μ_{PA}	3.691 (3.551, 3.831)	3.687 (3.547, 3.828)	3.687 (3.545, 3.83)	3.689 (3.548, 3.83)
μ_{Wo}	2.937 (2.744, 3.13)	2.936 (2.741, 3.131)	2.940 (2.742, 3.137)	2.939 (2.743, 3.135)
ϕ_{PA}	.343 (.294, .390)	.354 (.312, .394)	.354 (.313, .395)	.356 (.316, .396)
ϕ_{Wo}	.267 (.217, .316)	.280 (.237, .323)	.277 (.235, .318)	.275 (.230, .318)
$\beta_{Wo\rightarrow PA}$	-.023 (-.058, .011)	-.020 (-.048, .007)	-.021 (-.047, .007)	-.020 (-.045, .006)
$\beta_{PA\rightarrow Wo}$	-.160 (-.223, -.096)	-.150 (-.205, -.094)	-.156(-.213, -.100)	-.157 (-.213, -.102)
Random Effects for:				
μ_{PA}	.590 (.456, .774)	.618 (.481, .810)	.612 (.474, .801)	.611 (.475, .800)
μ_{Wo}	1.126 (.869, 1.486)	1.132 (.874, 1.491)	1.187 (.918, 1.559)	1.186 (.915, 1.555)
ϕ_{PA}	.042 (.030, .060)	.024 (.015, .038)	023 (.015, .037)	.022 (.013, .035)
ϕ_{Wo}	.045 (.032, .064)	.024 (.013, .039)	.028 (.017, .044)	.026 (.015, .041)
$\beta_{Wo\rightarrow PA}$	.025 (.018, .034)	.009 (.006, .015)	.009 (.005, .014)	.006 (.003, .011)
$\beta_{PA\rightarrow Wo}$	.059 (.039, .090)	.017 (.007, .038)	.025 (.011, .049)	.028 (.013, .054)
Residuals:				
σ^2_{PA}	.815 (.785, .845)	.822 (.793, .853)	.822 (.792, .852)	.823 (.794, .854)
σ^2_{Wo}	1.918 (1.849, 1.990)	1.943 (1.873, 2.017)	1.935 (1.865, 2.008)	1.934 (1.865, 2.008)
ρ_{PAWo}	-.479 (-.498, -.458)	-.478 (-.498, -.458)	-.479 (-.499, -.459)	-.479 (-.499, -.459)

*Note.*Included are the average effects (fixed effects) and the variances (random effects) for the means of PA and worrying (μ_{PA}, μ_{Wo}), autoregressive effects of PA and worrying (ϕ_{PA}, ϕ_{Wo}), the cross–lagged effect of worrying on PA ($\beta_{Wo\rightarrow PA}$) and of PA on worrying ($\beta_{PA\rightarrow Wo}$). Further included are the residual variances (σ^2_{PA}, σ^2_{Wo}), and the residual correlation (ρ_{PAWo}).

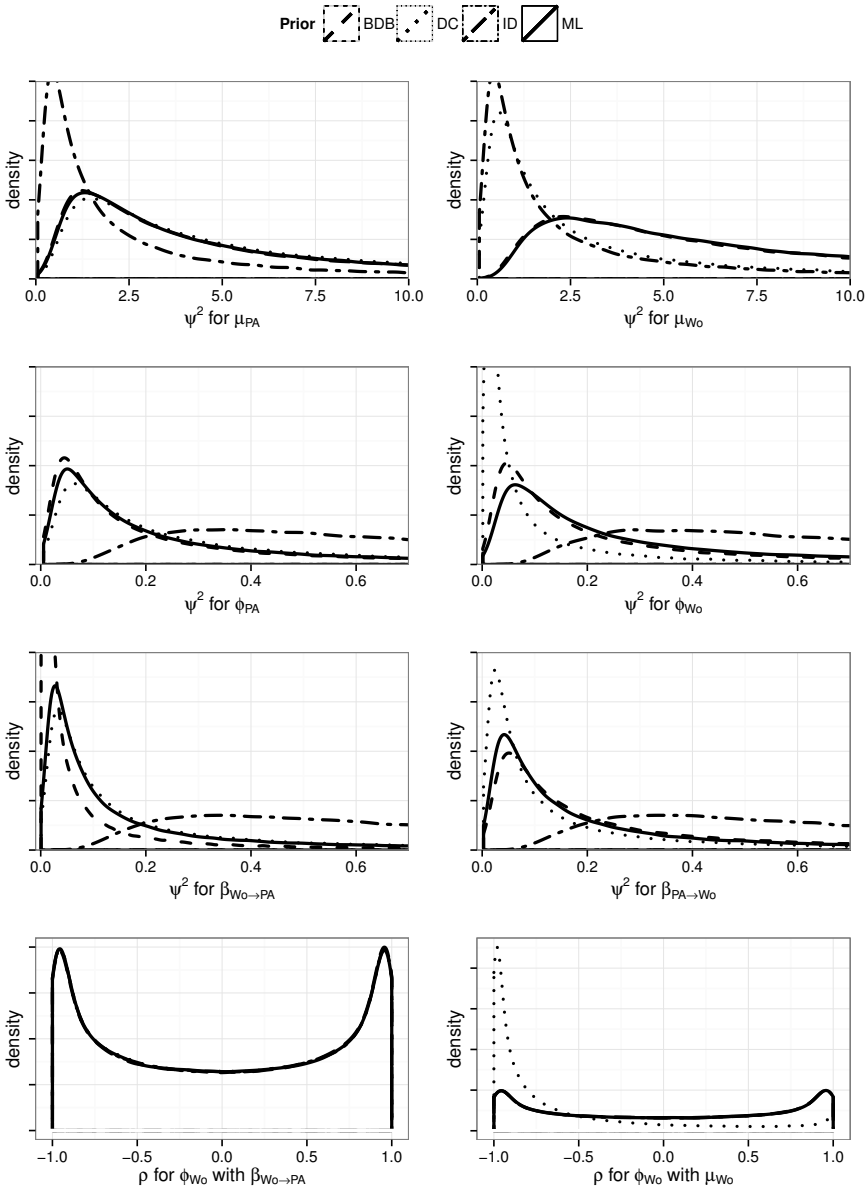


Figure 2.5: Plots of the (marginal) Inverse-Wishart prior distributions based on the maximum likelihood (ML), Bayesian data-based (BDB), default conjugate (DC), and identity matrix (ID) specification for the variances of the random parameters, and for the correlation between $\beta_{Wo->PAj}$ and ϕ_{Woj}, and $\beta_{Wo->PAj}$ and ϕ_{Woj}, and between ϕ_{Woj} and μ_{Woj}.

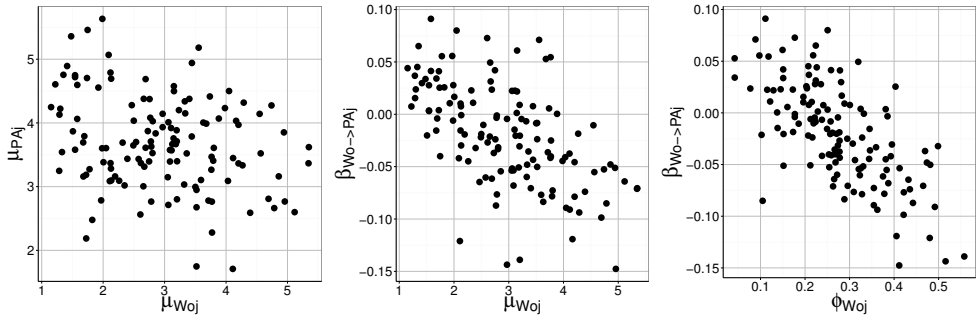


Figure 2.6: Scatter-plots of the point estimates of the random parameters based on the Bayesian data-based (BDB) prior specification, indicating negative correlations between the means for PA and worrying (μ_{PAj} and μ_{Woj}), the cross-lagged effects of worrying on PA and the means of worrying ($\beta_{Wo->PAj}$ and μ_{Woj}), and cross-lagged effects of worrying on PA and the autoregressive effects for worrying ($\beta_{Wo->PAj}$ and ϕ_{Woj}).

large compared to the estimates for the other prior specifications.

The fixed autoregressive effects are positive, which implies that on average, a participants' current PA is likely to be similar to their PA of the previous occasion, and a participants' current worrying is likely to be similar to their worrying of the previous occasion. Based on these point estimates for fixed effects, and the corresponding estimates of the variances based on BDB prior, we find an approximate 95% interval of .065 to .647 for the random autoregressive parameters of PA, and of -.041 to .591 for the autoregressive parameters of worrying. This indicates that the autoregressive coefficients are expected to be positive for most individuals. The average cross-lagged effect for the effect of worrying on PA is near zero, which implies that on average worrying on the preceding occasion does not affect PA at the current occasion. However, the variation around this average effect implies that for some persons the effect is actually positive, whereas for others it is negative: The point estimate of the fixed effect and of the corresponding variance imply a 95% interval of -.172 to .132 for the cross-lagged effects of worrying on PA. This may indicate that for some persons worrying is mostly a productive problem solving behavior, with successful problem solving leading to more positive affect, whereas for others worrying is ineffective, leading to less positive affect. The average cross-lagged effect of PA on worrying is negative, which implies that on average (across persons), higher PA on the

preceding occasion is likely to lead to less worrying at the current occasion, whereas diminished PA is likely to lead to more worrying. Based on the estimated fixed effect and corresponding variance, we find a 95% interval of -.485 to .171 for the random cross-lagged effects of PA on worrying, indicating that this effect is expected to be negative for most persons. This seems a logical result if worrying is problem-oriented: When there are problems to be solved, this may lead to lower PA, and to worrying in order to solve the problem, and vice versa.

For the correlations between the random parameters (not reported in Table 2.5 for reasons of space) we find that most correlations have quite wide credible intervals, with values ranging from strongly negative to strongly positive, so that we have too little information to draw conclusions about these correlations (similar to our findings in the simulation study). However, for three correlations we find credible intervals that include only negative values across the DC, ML, and DBD prior specifications, namely between the means for PA μ_{PAj} and worrying μ_{Woj} ($-.293$, 95% CI:$[-.453, -.115]$), between the mean of worrying μ_{Woj} and the cross-lagged effect of worrying on PA $\beta_{Wo\rightarrow PAj}$ ($-.360$, 95% CI:$[-.615, -.047]$), and between the autoregressive parameter for worrying ϕ_{Woj} and the cross-lagged effect of worrying on PA $\beta_{Wo\rightarrow PAj}$ ($-.551$, 95% CI:$[-.771, -.197]$; here we report the results based on the BDB prior, though results are similar across the other specifications). To gain more insight in the meaning of these correlations we made scatter plots of the individual parameters (see Figure 2.6), and we discuss each correlation in more detail below. First, the negative correlation between the mean of PA and of worrying (left panel of Figure 2.6), indicates that persons with higher average PA are likely to worry less on average compared to persons who generally have lower average PA. Second, the correlation between the cross-lagged effect of worrying on PA with the mean of worrying (middle panel of Figure 2.6) implies that individuals who worry a lot on average tend to have a negative cross-lagged effect of worrying on PA at the next occasion, whereas individuals who do not worry a lot on average tend to have a positive cross-lagged effect of worrying on PA. This may reflect the dual nature of worrying. For persons for whom worrying is effective in solving problems, worrying results in a higher positive affect because problems are being dealt with (i.e., a positive cross-lagged effect), and therefore may not need to worry as much (i.e., a low mean of worrying). In contrast, for persons for whom worrying is not effective, worrying may result in a lower PA (i.e., a negative cross-lagged effect) without the relief and accomplishment of solving the problem, and they may worry relatively a lot on average (i.e., a high mean for

worrying), because their problems are not going away. Third, the correlation between the cross-lagged effect of worrying on PA with the autoregressive effect of worrying (right panel of Figure 2.6) implies that persons who have high inertia in worrying (i.e., get stuck in worrying), tend to have a negative cross-lagged effect of worrying on PA, whereas persons that have little or no inertia in their worrying tend to have a positive cross-lagged effect of worrying on PA. This correlation also seems to illustrate the potential problem solving nature of worrying: When worrying results in solving the problem, worrying may result in a higher PA (i.e., a positive cross-lagged effect), and for persons for whom this is the case there may be little need to keep worrying (i.e., a relatively low inertia in worrying). In contrast, when worrying is ineffective, the futile worrying may result in a lower PA (i.e., a negative cross-lagged effect), and the persons for whom this is the case may continuously worry in order to keep trying to solve the problem, resulting in a relatively high positive inertia in worrying.

Finally, we note that there remains a strong negative association ($-.479$, 95% CI:$[-0.499, -0.459]$) between the residuals of PA and worrying. This residual correlation between PA and worrying after the lagged effects are taken into account, indicates that there is more to the relationship between PA and worrying. As such, it may be worthwhile to look at the relationship between PA and worrying at other time intervals than the interval of about 1.7 hours that was considered here, or to look for additional explanatory variables, for instance specific negative events, social interactions, stress, or psycho-physiological factors.

In sum, these results provide interesting considerations for future (confirmatory) research on the relationship between worrying and PA, and individual differences in this relationship. Based on the correlations between the random effects, we found that individuals who worry a lot on average, as well as individuals who get stuck in worrying, tend to have a negative cross-lagged effect of worrying on PA, indicating that for them worrying is a maladaptive coping strategy. In contrast, individuals who do not worry a lot on average, or who bounce back from worrying quickly, tend to have a positive cross-lagged effect from worrying to PA, indicating that for them worrying is a adequate tactic to solve current problems. Note that we were only able to find these results because we made use of a multivariate multilevel model, allowing for all random effects to be correlated. That is, had we used two separate multilevel models (i.e., for PA and worrying as dependent variables separately), we would not have obtained estimates of the correlations between these random effects. This illustrates the unique opportunities offered by the multivariate approach. Fur-

thermore, fitting several (data-based) priors helps evaluate the influence of specifying certain priors: The results for the ID specification considerably diverged from the results from the other prior specifications. The remaining three prior specifications however converged to approximately the same results, so that we feel that we can be reasonably confident about our results and conclusions based on these specifications.

2.5 Discussion

The multivariate multilevel autoregressive model is a valuable model for studying between-person differences in within-person processes. The Bayesian modeling framework provides a flexible environment for fitting this complex multilevel model. However, when some variances of the random parameters in the model are close to zero, the conjugate IW prior distribution for the covariance matrix of the random parameters becomes quite informative, unintentionally influencing the parameter estimates. In this study we evaluated the performance of three different IW prior specifications for the covariance matrix of the random parameters by means of a simulation study. In addition, by means of an empirical data set we demonstrated a sensitivity analysis for the IW prior specification, and illustrated the added value of the multivariate, multilevel modeling approach provided by the flexible Bayesian modeling environment.

The results from the simulation study indicate that the data-based ML prior specification for the covariance matrix of the random parameters performs the best compared to the ID specification and the DC specification. The ML specification performs well because it is based on estimates of the variances from the data. There are multiple ways to obtain estimates of the variances based on the data besides the ML procedure, that we discussed in the conclusion of the simulation study. The consequence of using the data twice is that the certainty about the parameter estimates is overestimated, resulting in too small posterior standard deviations. A solution to this problem may be to use a small part of the data for calibrating the prior specification (also referred to as training data, see Berger & Pericchi, 1996; O'Hagan, 1995), and using the remainder of the data for the model fitting procedure. Of course, this raises questions for future research on exactly how to do this. To cite two examples: Should you use part of the persons in your sample for calibrating the prior, or part of the repeated measures of each person? What sample size would provide good enough estimates for calibrating the prior?

An alternative specification that we considered, but was inoperative for the multilevel autoregressive model in WinBUGS (see footnote 3), is the scaled Wishart discussed in Gelman and Hill (2007). However, this may be a viable specification for other multilevel models. Other alternative IW prior specifications for $\mathbf{\Psi}$ that we did not investigate consist of specifying improper IW priors with *df* smaller than r, or to use a specification suggested in a recent study by Huang and Wand (2013), in which Half-Cauchy distributions for the standard deviations, and Uniform distributions for the correlations, are specified via an IW distribution. However, both WinBUGS and OpenBUGS require proper IW priors, and do not allow for setting priors within IW priors, so that these specifications are not available within this software. Still another option may be to transform the random parameters so that they have a larger variance, and specifying an IW prior for the covariance matrix of the transformed parameters. Finally, two potential specifications that circumvent the use of an IW distribution, are to specify the variances and covariances in a regression structure, or to specify uniform distributions for the variances of the random parameters, if disregarding the covariances does not affect the parameter estimates of interest. If the covariances between the random covariances are of primary interest, a possibility for the latter specification may be to correlate the random parameters a posteriori. Possible directions for future research are to compare the performance of the alternative specifications with the ML specification in a simulation study (in other software).

In conclusion, this study demonstrates that the IW prior specification for covariance matrices should not be taken lightly. When variances are small, the prior specification can have considerable consequences for the parameter estimates. In the multilevel autoregressive model, it is known in advance that some variances will be close to zero. We expect that our results will generalize to any multilevel model that has small variances in the covariance matrix of the random parameters, either as a result of the scale of the variables or parameters, or simply because there are only small individual differences in the parameters. Therefore, it seems imperative to include a prior specification sensitivity analysis for the covariance matrix of the random parameters in multilevel studies in psychology. Our empirical application provides an example of such an analysis, in which we compared the results for four different prior specifications: Three different (data-based) priors converged to approximately the same results, whereas the ID specification showed divergent results. Finally, we advise to include a data-based prior in such a prior sensitivity analysis. Although it may not be ideal to use the data twice in order to calibrate the prior, our simulation

study results indicate that a prior distribution based on estimates of the variances of the random parameters performs the best in this specific situation that some variances in the covariance matrix may be close to zero.

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

by N.K. Schuurman, E. Ferrer, M. de Boer-Sonnenschein, and E.L. Hamaker

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?," or "How do anxiety and rumination influence each other?." Many of these questions cannot be investigated experimentally due to 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 et al., 2011; de Jonge et al., 2001; de Lange et al., 2004; Kinnunen et al., 2008; Talbot et al., 2012). The strongest association is then judged to provide the most important causal influence that drives the system, also referred to as being 'causally dominant' (c.f., 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, where a variable x predicts future

This chapter is based on: Schuurman, N.K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E.L. (2016). How to Compare Cross-Lagged Associations in a Multilevel Autoregressive Model. *Psychological Methods*.

Author contributions: Schuurman designed the study, performed the analyses, processed and interpreted the results, and wrote the paper. Ferrer provided data for an empirical example that was later removed (which is now used in Chapter 5) and provided feedback on the written work. De Boer-Sonnenschein provided the data for the empirical example and provided feedback on the written work. Hamaker proposed the topic for the study (standardizing in the multilevel model), provided extensive feedback on the design of the study, and on the written work.

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 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 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 et al., 2010; Lodewyckx et al., 2011; Madhyastha et al., 2011; Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001; Suls et al., 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 dynamics from stable between-person differences. This is essential for investigating the actual dynamics of a psychological (causal) process that operates at the within-person 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 within-person level (Borsboom et al., 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, while 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, while 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 regression coefficients. The unstandardized coefficients vary in size depending on the variances of the variables, while 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 within-person variance, and the between-person 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: Firstly, we illustrate the value of the multilevel model in studying Granger-causal cross-lagged relations, and the individual differences therein. Secondly, 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 manuscript with a discussion in which we highlight our main conclusions.

3.1 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 modeling and multivariate autoregressive time series modeling 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 (2-5) 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, Kuijper, & Grasman, 2015; Kievit et al., 2011; Molenaar, 2004). Firstly, because the cross-lagged model does not separate stable between-person differences (differences in the intercepts) from the within-person effects (c.f., Hamaker, Kuijper, & Grasman, 2015). Secondly, 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 intra-individual 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 within-person processes as in the (n=1) time series model simultaneously for multiple individuals, and model group effects that allow us to generalize results to a larger population. Specifically, at the within-person or first level, the time series model is specified to describe the dynamics of the process for each individual, while at the between-person 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 below.

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: 1) 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 2) 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

$$\begin{bmatrix} y_{1ti} \\ y_{2ti} \end{bmatrix} = \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix} + \begin{bmatrix} \tilde{y}_{1ti} \\ \tilde{y}_{2ti} \end{bmatrix} \tag{3.1}$$

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, 2005).

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

$$\begin{bmatrix} \tilde{y}_{1ti} \\ \tilde{y}_{2ti} \end{bmatrix} = \begin{bmatrix} \phi_{1i} & \phi_{12,i} \\ \phi_{21,i} & \phi_{2i} \end{bmatrix} \begin{bmatrix} \tilde{y}_{1t-1i} \\ \tilde{y}_{2t-1i} \end{bmatrix} + \begin{bmatrix} \epsilon_{1ti} \\ \epsilon_{2ti} \end{bmatrix} \tag{3.2}$$

$$\begin{bmatrix} \epsilon_{1ti} \\ \epsilon_{2ti} \end{bmatrix} \sim MvN \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right\}. \tag{3.3}$$

The innovations, $\boldsymbol{\epsilon}_{ti}$, 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

[1] Note that the innovations are not the same as measurement errors. For details, and on incor-

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 3.3, we assume that the innovations are normally distributed with means of zero, and covariance matrix Σ. Note that autoregressive models as discussed above 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 above 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 $[\mu_{1i}, \mu_{2i}, \phi_{1i}, \phi_{12i}, \phi_{21i}, \phi_{2i}]'$ (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 *between-person 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 to 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.

porating 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.

Fitting the Model with Bayesian Techniques

We fit the multilevel bivariate autoregressive model, and estimate the accompanying standardized coefficients 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 et al., 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 (note that the model code can also be used within the software OpenBUGS and JAGS; Lunn et al., 2009, 2000; Plummer, 2003) for simulating data based on the multilevel VAR model, fitting the model, and standardizing the parameters in the supplementary materials published with the online paper (Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016) or at www.nkschuurman.com. In Appendix 3.A we provide information on the prior specifications and convergence for the empirical application.

3.2 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. Since 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, followed by which method to use.

A Rationale for Standardized Cross-Lagged Parameters

A 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 she is much more variable in her scores then he 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. [2]

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, page 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 therefore is '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.[3] 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.

[2]Further note,that when two predictor variables have the same probability distribution, a score larger than (for example) one standard-deviation is equally likely to occur in both variables. This is another reason that standardized parameters are deemed more comparable to each other than unstandardized parameters.

[3]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 $b_1^* = (r_{y1} - r_{12}r_{y2})/(1 - r_{12}^2)$, where r_{y1} is the correlation between y and x_1, r_{12} the correlation between x_1 and x_2, and r_{y2} the correlation between y and x_2. The proportion of uniquely explained variance is equal to the squared semi-partial correlation, which is expressed as $r_{y(1.2)} = (r_{y1} - r_{12}r_{y2})/\sqrt{(1 - r_{12}^2)}$. 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.

Within-Person, Between-Person, 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 (c.f., 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.

While 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: Within-person (WP) standardization, grand standardization, and between-person (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 3.1.

Within-person standardization

WP standardization (also referred to as within-group standardization when the data consists 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_{1i}^2 and ω_{2i}^2 for variables y_1 and y_2 are the diagonal elements of the person-specific covariance matrix $\boldsymbol{\Omega_i}$. Based on the regression equation in Equation 3.2, this covariance matrix can be expressed as $\boldsymbol{\Omega}_i = \boldsymbol{\Phi}_i\boldsymbol{\Omega}_i\boldsymbol{\Phi}_i' + \boldsymbol{\Sigma}$ (where $\boldsymbol{\Phi}_i'$ is the transpose of matrix $\boldsymbol{\Phi_i}$). However, this not helpful in practice, because this equation includes $\boldsymbol{\Omega}_i$ at both sides of the equation. To obtain an expression for the WP covariance matrix $\boldsymbol{\Omega_i}$ in terms of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ only, we can make use of the following expression instead

$$\boldsymbol{\Omega_i} = mat((\boldsymbol{I} - \boldsymbol{\Phi}_i \otimes \boldsymbol{\Phi}_i)^{-1} vec(\boldsymbol{\Sigma})), \tag{3.4}$$

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 3.1, 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), 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 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 3.1. 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 Section 3.1 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 described previously to standardized data, we would assume that the standardized regression parameters are normally distributed, rather than the unstandardized regression parameters. 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 3.1.

We opt here for the assumption that the unstandardized parameters are normally distributed and calculating the standardized coefficients afterwards, for two reasons. Firstly, this normality assumption is in line with conventions in multilevel research. Secondly, 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 3.4, within each iteration of the MCMC procedure. Subsequently, we calculate the person-specific standardized coefficients by means of the equation in Table 3.1 (i.e., the equation in the first column, the 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 for 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 supplemental materials includes example code for WP standardizing the cross-lagged coefficients (the supplemental materials can be found with the online paper (Schuurman, Ferrer, et al., 2016) or at www.nkschuurman.com).

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, where 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*, that 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 vari-

ance 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 3.B),

$$\boldsymbol{G} = \underset{i}{E}[\boldsymbol{\Omega}] + \boldsymbol{\psi_\mu}, \tag{3.5}$$

where $\underset{i}{E}[\boldsymbol{\Omega}]$ is the expected value taken over the person-specific covariance matrices $\boldsymbol{\Omega_i}$, and $\boldsymbol{\psi_\mu}$ 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 from Table 3.1. 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 3.1). 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 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 $\underset{i}{E}[\boldsymbol{\Omega}]$) and standard deviations, and calculating the grand standardized person-specific coefficients and fixed effects by means of the equations in Table 3.1, in each iteration of the MCMC procedure. We have included example R code for grand standardization in the supplementary materials.

Between-person standardization

BP standardization is based on the BP variance ψ_μ^2, 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 3.1).

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 Section 3.1. 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 3.1. 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 - while 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 3.1, 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 stan-

Table 3.1: Equations for the variances for WP, BP and grand standardization, for the standardized person-specific parameters ϕ^*_{jki} and fixed effect parameters $\gamma^*_{\phi jki}$ for outcome variable j and predictor variable k. The person-specific standardized parameters are the product of the unstandardized parameters, 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 are calculated by taking the expectation ($\underset{i}{E}[]$) over the standardized person-specific parameters for all persons i.

	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 term ω_i^2 indicates the person-specific variance based on Equation 3.4. 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.

dardized 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 standardization show that $\phi^*_{21} > \phi^*_{12}$ for most persons, while WP standardization shows that this is the case for some persons, while the reverse is true for other persons. Therefore, the grand and BP standardized fixed effects indicate that $\phi^*_{21} > \phi^*_{12}$ on average, while 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, while 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 3.1 the WP variances for y_2 are smaller than for y_1, while 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}, while for WP standardization it is the reverse. Finally, note that while 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, while 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 for

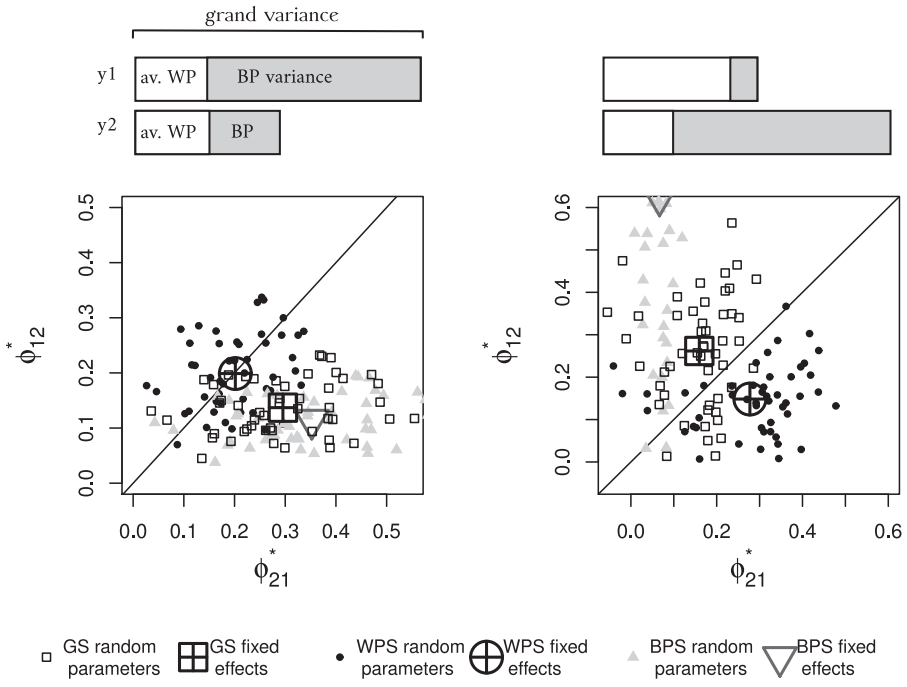


Figure 3.1: Plots of simulated individual cross-lagged parameters and accompanying fixed effects, with squares indicating grand standardized parameters, triangles indicating BP standardized parameters, and circles indicating WP standardized parameters. The area below the diagonal line implies that the association between y_1 with future y_2 is the strongest, while 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, while 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 $\gamma_{\phi 12}$ and $\gamma_{\phi 21}$ 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.

instance HLM, lme4 in R, and SPSS Mixed, 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, for a software overview, p. 134). 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 SEM modeling software Mplus, which also features multilevel modeling, and seems to standardize within-person ("to the variance on within for within relationships"), although it is unclear how this is achieved exactly, and if what is the case for the fixed effects (Heck & Thomas, 2000; Muthén, 2008).[4]

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 (c.f., Hamaker, Kuijper, & Grasman, 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 moth-

[4]This is based on Mplus forum responses by Muthén.

ers 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 also grand standardization) does convolve this information, with the following interpretation of the BP standardized coefficient for the Granger causal effect of the mother's anxiety on that of her child: the number of standard deviations for the average levels of anxiety across children in the population that the score of this child increases, when that child's mother's anxiety increases one standard deviation in the average levels of anxiety across mothers in the population. Yet, 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 (c.f., Section 3.2).

Finally, WP standardization takes into account that each individual may have a unique variance for each variable, while 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.

3.3 Empirical Application on 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 to those obtained using BP and grand standardization.

Burnout Data: An Experience Sampling Method Study

Two core components of burnout are severe exhaustion and diminished experienced personal competence (Maslach & Jackson, 1981; Maslach, Jackson, & Leiter, 1996; WorldHealthOrganization, 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 two 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, while 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.[5]

[5]Note that as a result of the general set-up 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 paper. To correct for this, we added missing observations to the data set when the interval was particularly large (> 5 hours, which occurred mostly at night when participants slept), which resulted in time intervals of on average 2.3 hours (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 non-equidistant 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),

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 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, while 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 3.1 to 3.3, we make use of WinBUGS (in combination with R, and the R–packages r2winbugs and CODA), which is free software for conducting Bayesian analysis (Lunn et al., 2000; Plummer, Best, Cowles, & Vines, 2006; R Development Core Team, 2012; Sturtz et al., 2005). The WinBUGS code we used for the modeling procedure for both applications can be found in the supplementary materials (available with the online paper (Schuurman, Ferrer, et al., 2016) or at www.nkschuurman.com), 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 3.A.

or the random cross-lagged effects are assumed to be equal within a person (Oravecz & Tuerlinckx, 2011), which is clearly undesirable in the current context.

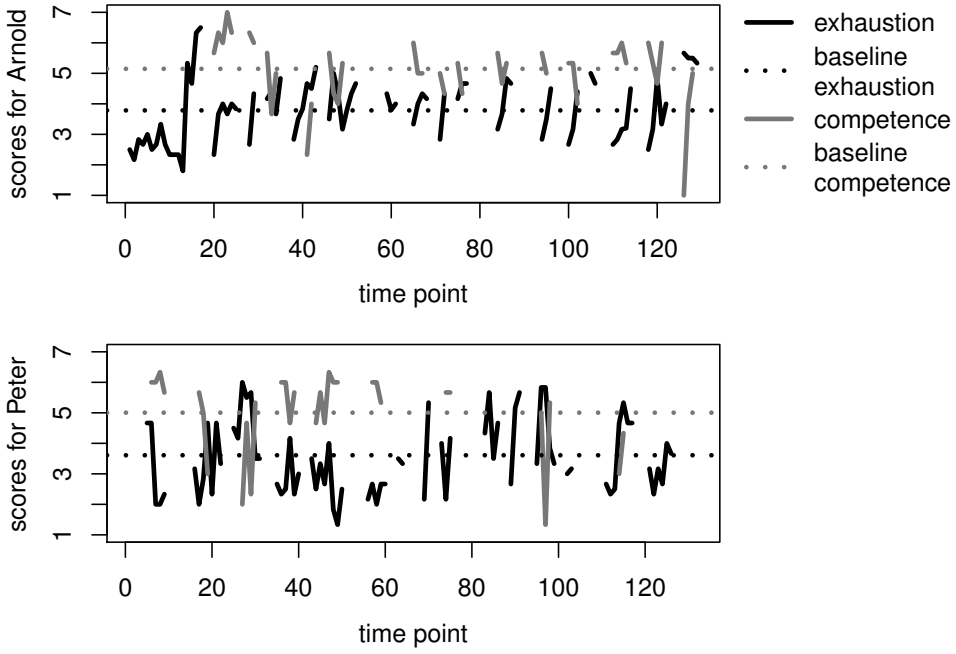


Figure 3.2: Time series for Arnold and Peter, representing their exhaustion (in black) and competence (in gray). Dotted lines represent the individuals' estimated mean scores μ_{Ci} and μ_{Ei}.

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 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 3.2.

In Figure 3.2 the observations for two individuals, whom we refer to as 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 indi-

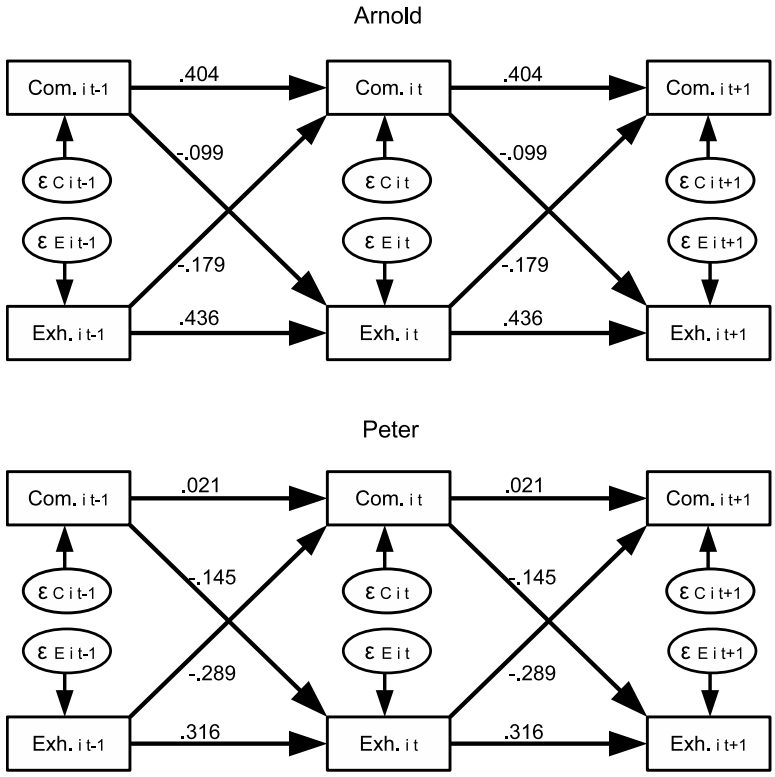


Figure 3.3: Estimated model parameters for the associations between Arnold and Peter's exhaustion and competence over time.

vidual regression parameters for Arnold and Peter are displayed in Figure 3.3. Both Arnold and Peter have large positive autoregressive coefficients for exhaustion (i.e., .436 with 95% CI [.228, .644], and .316 with 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 with 95% CI [.124, .646]), implying that his feelings of high competence tend to last for some time, while his feelings of incompe-

tence also tend to last for some time. In contrast, Peter's autoregressive parameter for competence is close to zero (i.e., .021 with 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., $-.289$ with 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., $-.145$ with 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.

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 3.4. 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, while 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 very 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

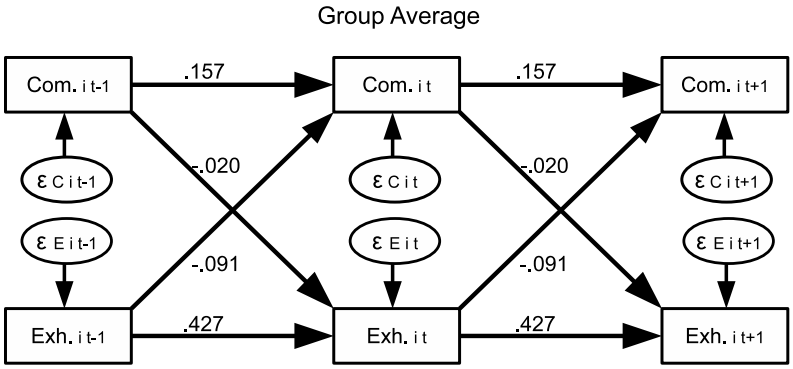


Figure 3.4: Estimated fixed effects for the multilevel autoregressive model studying the associations between exhaustion and competence for the group of individuals diagnosed with burnout.

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 $[.014, .048]$; and $\psi_{\phi EC} = .066$, 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 to those obtained for BP and grand standardization.

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 3.5. 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 3.5, these individual differences are also reflected in the WP standardized coefficients. Specifically, the individual WP standardized coefficients range from $-.265$ to .103 for 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 3.5. 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 - while the non-absolute fixed effects, led to the opposite, misleading, conclusion.

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 with a 95% CI of [.528, .774].[6] 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, while 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, while for other people feeling competent drives them to work harder, resulting in fatigue, and a positive relationship between competence and subsequent exhaustion.

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 to 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 Section 3.2. 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 3.5 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 co-

[6] 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 it by the number of individuals, resulting in a posterior distribution for this proportion.

Table 3.2: 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]

efficients; 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 3.5). 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 3.5, 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 of [.491, .775]) for grand standardization, .642 (95% CI of [.434, .811]) for BP standardization, and .66 (95% CI of [.528, .774]) for WP standardization.

3.4 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 inter-individual differences in within-person 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

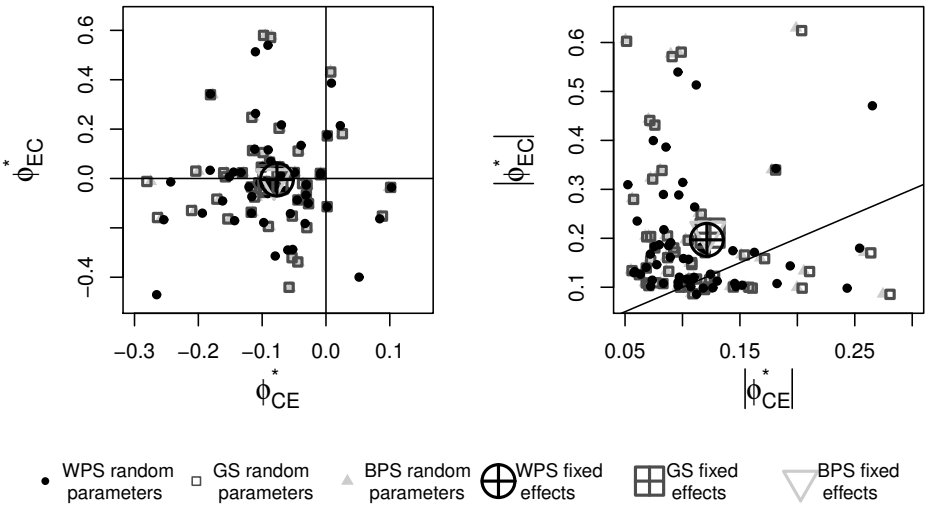


Figure 3.5: The left panel shows a plot of the point estimates of the WP, BP, and grand standardized random parameters and fixed effects for the cross-lagged associations between exhaustion and competence in individuals with burnout. The right panel shows the absolute values for the standardized random 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.

exhaustion for persons diagnosed with burn-out. 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 inter-individual differences in the cross-lagged coefficients, and the inter-individual 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, while 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 inter-individual 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, while WP standardization does not. Secondly, WP standardization takes into account that each person may have unique standard deviations for the outcome and predictor variables, while the other methods of standardization - grand and BP standardization - do not. While 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 - are 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 paper 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; it does 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 non-stationary models. For a stationary model that includes multiple lags, one can simply calculate the standardized parameters based on the equations in Table 3.1 using the WP variances, and the raw coefficients for the relevant time lag. For non-

stationary 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 to 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 paper we show 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.

Appendix 3.A Prior Specifications and Convergence for the Empirical Application

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, 2016).

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 ML estimates as described in Schuurman, Grasman, and Hamaker (2016). 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 3 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 hours 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 paper, we will report the medians, and 95% credible intervals of the posterior distributions for the parameters of interest.

Appendix 3.B Derivation of the Grand Variance

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

$$\boldsymbol{\Theta} = E\left[\boldsymbol{X}\boldsymbol{X'}\right] - E\left[\boldsymbol{X}\right]E\left[\boldsymbol{X}\right]', \tag{3.6}$$

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

$$\boldsymbol{G} = \underset{it}{E}\left[\boldsymbol{Y}\boldsymbol{Y'}\right] - \underset{it}{E}\left[\boldsymbol{Y}\right]\underset{it}{E}\left[\boldsymbol{Y}\right]'. \tag{3.7}$$

And, for person i in the multilevel VAR model

$$\boldsymbol{\Omega}_i = \underset{t}{E}\left[\boldsymbol{Y_i}\boldsymbol{Y_i'}\right] - \boldsymbol{\mu}_i\boldsymbol{\mu}_i'. \tag{3.8}$$

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

$$\begin{aligned}
\underset{i}{E}\left[\boldsymbol{\Omega}\right] &= \underset{i}{E}\left[\underset{t}{E}\left[\boldsymbol{Y}\boldsymbol{Y'}\right] - \boldsymbol{\mu}\boldsymbol{\mu}'\right] \\
&= \underset{it}{E}\left[\boldsymbol{Y}\boldsymbol{Y'}\right] - \underset{i}{E}\left[\boldsymbol{\mu}\boldsymbol{\mu}'\right] \\
&= \underset{it}{E}\left[\boldsymbol{Y}\boldsymbol{Y'}\right] - \left(\underset{i}{E}\left[\boldsymbol{\mu}\right]\underset{i}{E}\left[\boldsymbol{\mu}\right]' + \boldsymbol{\psi}_{\boldsymbol{\mu}}^{\boldsymbol{2}}\right) \\
&= \underset{it}{E}\left[\boldsymbol{Y}\right]\underset{it}{E}\left[\boldsymbol{Y}\right]' + \boldsymbol{G} - \underset{i}{E}\left[\boldsymbol{\mu}\right]\underset{i}{E}\left[\boldsymbol{\mu}\right]' - \boldsymbol{\psi}_{\boldsymbol{\mu}}^{\boldsymbol{2}} \\
&= \boldsymbol{G} - \boldsymbol{\psi}_{\boldsymbol{\mu}}^{\boldsymbol{2}},
\end{aligned} \tag{3.9}$$

such that,

$$\boldsymbol{G} = \underset{i}{E}\left[\boldsymbol{\Omega}\right] + \boldsymbol{\psi}_{\boldsymbol{\mu}}^{\boldsymbol{2}} \tag{3.10}$$

4 Incorporating Measurement Error in n=1 Psychological Autoregressive Modeling

by N.K. Schuurman, J.H. Houtveen, and E.L. Hamaker

The dynamic modeling of processes at the within-person level is becoming more and more popular in psychology. The reason for this seems to be the realization that inter-individual differences, in many cases, are not equal to intra-individual differences. Indeed, studies that compare interindividual differences and intraindividual differences usually do not harbor the same results, exemplifying that conclusions based on studies of group averages (including cross-sectional studies and panel data studies), cannot simply be generalized to individuals (Adolf et al., 2014; Borsboom et al., 2003; Ferrer, Steele, & Hsieh, 2012; Hamaker, 2012; Kievit et al., 2011; Madhyastha et al., 2011; Molenaar, 2004; Nezlek & Gable, 2001; Rovine & Walls, 2005; Wang et al., 2012).

The increased interest in analyses at the within-person level, and the increasing availability of technology for collecting these data, has resulted in an increase in psychological studies that collect intensive longitudinal data, consisting of many (say 25 or more) repeated measures from one or more individuals. A popular way to analyze these data currently is by autoregressive time series (AR) modeling, either by modeling the repeated measures for a single individual using classical n=1 AR models, or by using multilevel extensions of these models, with the repeated measures for each individual modeled at level 1, and individual differences modeled at level 2 (Cohn & Tronick, 1989; De Haan-Rietdijk et al., 2014; Kuppens et al., 2010; Lodewyckx et al., 2011; Madhyastha et al., 2011; Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001; Rovine & Walls, 2005; Suls et al., 1998; Wang et al., 2012).

This chapter is based on: Schuurman, N.K., Houtveen, J.H., & Hamaker, E.L. (2015). Incorporating Measurement Error in n=1 Psychological Autoregressive Modeling. *Frontiers in Psychology, 6*, 1038. doi:http://dx.doi.org/10.3389/fpsyg.2015.01038.

Author contributions: Schuurman designed and performed the study, analysed, processed and interpreted the results for the simulation study and the empirical example, and wrote the paper. Houtveen was part of the data collection in 2006 for the data used in the empirical example. Hamaker proposed the topic for the study (accounting for measurement error), collected the data for the empirical example in 2006, provided feedback on the design of the study, and on the written work.

In an AR model of order 1 (i.e., an AR(1) model), a variable is regressed on a lagged version of itself, such that the regression parameter reflects the association between this variable and itself at the previous measurement occasion (c.f., Chatfield, 2004; Hamilton, 1994). The reason for the popularity of this model may be the natural interpretation of the resulting AR parameter as inertia, that is, resistance to change (Suls et al., 1998). Resistance to change is a concept which is considered to be relevant to many psychological constructs and processes, including attention, mood and the development of mood disorders, and the revision of impressions and opinions (Geller & Pitz, 1968; Goodwin, 1971; Kirkham et al., 2003; Koval et al., 2012; Kuppens et al., 2010; Suls et al., 1998).

However, a problem with the regular AR(1) model is that it does not account for any measurement errors present in the data. Although AR models incorporate residuals, which are referred to as 'innovations' or 'dynamic errors', these residuals are to be distinguished from measurement error. Simply put, the distinction between dynamic errors and measurement errors is that dynamic errors carry over to next measurement occasions through the autoregressive relationship, while measurement errors are specific to one measurement occasion. Therefore, even though taking measurement errors into account is considered business as usual in many psychological studies of interindividual differences, it is largely neglected in AR modeling. Two exceptions are formed by Wagenmakers (2004) and Gilden (2001),[1] both of which concern studies on reaction time and accuracy in series of cognitive tasks. Gilden notes that there is evidence that some variance in reaction time is random (measurement) error as a result of key-pressing in computer tasks. Measurement error however is not limited to 'accidentally' pressing the wrong button or crossing the wrong answer, but is made up of the sum of all the influences of unobserved factors on the current observation, that do not carry-over to the next measurement occasion. Disregarding measurement error distorts the estimation of the effects of interest (Staudenmayer & Buonaccorsi, 2005). This is quite problematic, considering that in psychological studies it is often impossible to directly observe the variable of interest, and it therefore seems likely (and this seems generally accepted among psychological researchers) that psychological research in general is prone to having noisy data.

The aim of this study is therefore three-fold. First, we aim to emphasize the im-

[1] Other exceptions are of course dynamic factor models, and other latent variable models in which the measurement structure for multiple items is explicitly modeled. Here we focus on applications in which each construct is measured with one variable.

portance of considering measurement error in addition to dynamic error in intensive longitudinal studies, and illustrate the effects of disregarding it in the case of the n=1 autoregressive model. Second, we aim to compare two modeling strategies for incorporating measurement errors: 1) fitting an autoregressive model that includes a white noise term (AR+WN), and 2) fitting an autoregressive moving average (ARMA) model. These modeling strategies are the two most frequently suggested in the literature (e.g., in mathematical statistics, control engineering, and econometrics, c.f., Chanda, 1996; Chong, Liew, Zhang, & Wong, 2006; Costa & Alpuim, 2010; Deistler, 1986; Granger & Morris, 1976; Patriota, Sato, Blas, & G., 2010; Staudenmayer & Buonaccorsi, 2005; Swamy, Chang, Mehta, & Tavlas, 2003). Third, our aim is to compare the performance of these models for a frequentist and a Bayesian estimation procedure. Specifically, for the frequentist procedure we will focus on a Maximum Likelihood (ML) procedure based on the state-space modeling framework, which is a convenient modeling framework for psychological longitudinal modeling, as it readily deals with missing data, and is easily extended to multivariate settings, or to include latent variables (Harvey, 1989). The Bayesian alternative shares these qualities, and has the additional advantage that the performance of the estimation procedure is not dependent on large samples (Dunson, 2001; Lee & Wagenmakers, 2005), while the performance of the frequentist ML procedure depends on asymptotic approximations, and in general requires large samples. This is convenient for the modeling of intensive longitudinal data, given that large amounts of repeated measures are often difficult to obtain in psychological studies. By means of a simulation study we will evaluate the parameter recovery performance of the Bayesian procedure for the ARMA(1,1) and the AR+WN model, and compare it to the ML procedure.

This paper is organized as follows. We start by introducing the AR(1) model, ARMA(1,1) model, and the AR(1)+WN model, and discussing their connections. After that, we present the methods for the simulation study, followed by the results. We present an empirical application concerning the daily mood of eight women, in order to further illustrate the consequences of disregarding measurement error in practice, and we end with a discussion.

4.1 Models

In this section we present the AR(1) model, and explain the difference between the dynamic errors that are incorporated in the AR(1) model, and measurement errors.

After that we will introduce models that incorporate measurement errors, namely the autoregressive model with an added white noise term (AR(1)+WN model), and the autoregressive moving average (ARMA) model.

The AR(1) Model

In order to fit an AR model, a large number of repeated measures is taken from one individual. Each observation, or score, y_t in the AR model consists of a stable trait part - the mean of the process denoted as μ, and a state part $\tilde{y}_t$ that reflects the divergence from that mean at each occasion. In an AR model of order 1, the state of the individual at a specific occasion $\tilde{y}_t$ depends on the previous state $\tilde{y}_{t-1}$, and this dependency is modeled with the AR parameter ϕ. Specifically, the AR(1) model can be specified as

$$\begin{aligned} y_t &= \mu + \tilde{y}_t \\ \tilde{y}_t &= \phi \tilde{y}_{t-1} + \epsilon_t \end{aligned} \tag{4.1}$$

$$\epsilon_t \sim N\left(0, \sigma_\epsilon^2\right). \tag{4.2}$$

For a graphical representation of the model, see Figure 4.1A. A positive value for ϕ indicates that the score at the current occasion will be similar to that at the previous occasion - and the higher the positive value for ϕ, the more similar the scores will be. Therefore, a positive AR parameter reflects the inertia, or resistance to change, of a process (Suls et al., 1998). A positive AR parameter could be expected for many psychological processes, such as that of mood, attitudes, and (symptoms of) psychological disorders. A negative ϕ indicates that if an individual has a high score at one occasion, the score at the next occasion is likely to be low, and vice versa. A negative AR parameter may be expected for instance in processes that concern intake, such as drinking alcoholic beverages: If an individual drinks a lot at one occasion, that person may be more likely to cut back on alcohol the next occasion, and the following occasion drink a lot again, and so on (Rovine & Walls, 2005). An AR parameter close to zero indicates that a score on the previous occasion does not predict the score on the next occasion. Throughout this paper we consider stationary models, which implies that the mean and variance of y are stable over time, and ϕ lies in the range from -1 to 1 (Hamilton, 1994). The innovations ϵ_t reflect that component of

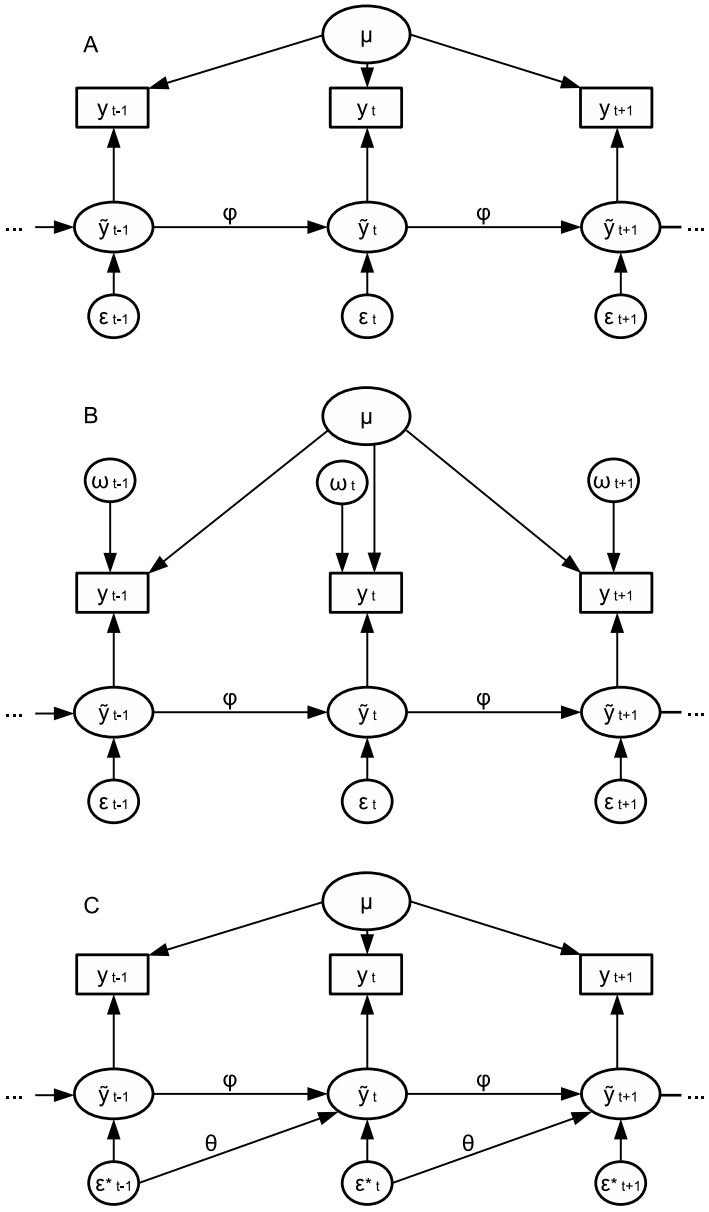


Figure 4.1: A) Graphical representation of an AR(1) model; B) Graphical representation of an AR(1)+WN model; C) Graphical representation of an ARMA(1,1) model.

each state score $\tilde{y}_t$ that is unpredictable from the previous observation. The innovations ϵ_t are assumed to be normally distributed with a mean of zero and variance σ^2_ϵ.

Dynamic Errors vs. Measurement Errors

The innovations ϵ_t perturb the system and change its course over time. Each innovations is the result of all unobserved events that impact the variable of interest at the current measurement occasion, of which the impact is carried over through the AR effect to the next few measurement occasions. Take for example hourly measurements of concentration: Unobserved events such as eating a healthy breakfast, a good night sleep the previous night, or a pleasant commute, may impact concentration in the morning, resulting in a heightened concentrating at that measurement occasion. This heightened concentration may then linger for the next few measurement occasions as a result of an AR effect. In other words, the innovations ϵ_t are "passed along" to future time points via ϕ, as can be seen from Figure 4.1A, and this is why they are also referred to as "dynamic errors."

Measurement errors, on the other hand, do not carry over to next measurement occasions, and their effects are therefore restricted to a single time point. Classical examples of measurement error, which are moment-specific, are making an error while filling in a questionnaire, or accidentally pressing a (wrong) button during an experiment (e.g., Gilden, 2001). However, any unobserved effect of which the influence is not carried over to the next measurement occasion may also be considered as measurement error, rather than dynamic error. The only distinguishing characteristic of measurement errors and dynamic errors is that the latter's influence lingers for multiple measurement occasions. Therefore, in practice, what unobserved effects will end up as measurement error, and what effects will end up as dynamic error, will depend largely on the measurement design of the study, such as on the frequency of the repeated measures that are taken. For example, some unobserved effects may carry-over from minute to minute (e.g., having a snack, listening to a song), but not from day to day - if measurements are then taken every minute, these unobserved effects will end up in the dynamic error term, but if measurements are taken daily, such effects will end up in the measurement error term. As such, the more infrequent measurements are taken, the more measurement errors one can expect to be present in the data, relative to the dynamic errors.

In psychological research, measurement is complicated and likely to be noisy. As such, the contribution of measurement error variance to the total variance of the measured process may be considerable. Ignoring this contribution will result in biased parameter estimates. Staudenmayer and Buonaccorsi (2005) have shown that in the case of an AR(1) model, ϕ will be biased toward zero. Specifically, the estimated AR coefficient $\hat{\phi}$ will be equal to $(1 - \lambda) * \phi$, where ϕ is the true AR parameter and λ is the proportion of measurement error variance to the total variance. Hence, in order to prevent the measurement error from biasing estimates of ϕ, it is necessary to take measurement error into account in the modeling procedure. This approach has two advantages: First, it leads to less biased estimates of ϕ, and second, it allows us to investigate to what extent the measurements are determined by measurement error.

Incorporating Measurement Error: The AR(1)+WN Model

A relatively simple way to incorporate measurement error in dynamic modeling is to add a noise term to the model, typically white noise, to represent the measurement error. White noise is a series of random variables that are identically and independently distributed (Chatfield, 2004). For the AR model with measurement error (AR(1)+WN), the white noise ω_t is simply added to each observation y_t (see Figure 4.1B). We assume that this white noise is normally distributed with a mean of zero and variance σ_ω^2. This results in the following model specification for the AR(1)+WN model

$$\begin{aligned} y_t &= \mu + \tilde{y}_t + \omega_t \\ \tilde{y}_t &= \phi \tilde{y}_{t-1} + \epsilon_t \end{aligned} \tag{4.3}$$

$$\epsilon_t \sim N\left(0, \sigma_\epsilon^2\right) \tag{4.4}$$

$$\omega_t \sim N\left(0, \sigma_\omega^2\right). \tag{4.5}$$

Important to note is that when ϕ is equal to zero, the measurement error and dynamic error will no longer be discernible from each other, because they are only discernible from each other from the merit that the innovations are passed to future time points through ϕ, while the measurement errors are not. In that case, the AR(1)+WN model is no longer identified, which is problematic for estimating the model parameters. Further note that when ϕ is nonzero, the higher $|\phi|$, the easier it will be to

discern measurement error from the innovations, and as such the model will be easier to identify empirically, and likely easier to estimate. Hence, in this sense the (empirical) identification of the AR(1)+WN model may be seen as dimensional rather than dichotomous, ranging from unidentified when ϕ is zero, to maximally empirically identified when $|\phi|$ is one.

Incorporating Measurement Error: The ARMA(1,1) Model

Another way to incorporate measurement error into an AR(1) model that is relatively frequently suggested in the literature on dynamic modeling with measurement error, is to use an autoregressive moving average (ARMA) model (see for instance: Chanda, 1996; Chong et al., 2006; Costa & Alpuim, 2010; Deistler, 1986; Granger & Morris, 1976; Patriota et al., 2010; Staudenmayer & Buonaccorsi, 2005; Swamy et al., 2003; Wagenmakers et al., 2004). Granger and Morris (1976) have shown that the AR(p)+WN model is equivalent to an ARMA(p,p) model, where p stands for the number of lags included in the model. As a result, an ARMA(1,1) model can be used as an indirect way to fit an AR(1) model and take measurement error into account (Granger & Morris, 1976; Staudenmayer & Buonaccorsi, 2005; Wagenmakers et al., 2004). One advantage of fitting an ARMA(1,1) model rather than fitting an AR(1)+WN model directly, is that it can be estimated with a wide range of estimation procedures, and a wide range of software, including for instance SPSS. A second important advantage is that the ARMA(1,1) is identified when the value of ϕ is equal to zero, so that in practice it may be easier to estimate than the AR(1)+WN model.

An ARMA(1,1) process consists of an AR(1) process, and a moving average process of order 1 (MA(1)). In an MA(1) process, the current state $\tilde{y}_t$ depends not only on the innovation, ϵ_t^*, but also on the previous innovation ϵ_{t-1}^*, through moving average parameters θ.[2] For example, consider the daily introverted behavior for a specific person. On a certain day, the person has a shameful experience, resulting in a strong boost (e.g., an innovation or perturbation) in introverted behavior. The next day, this person may display lingering heightened introverted behavior from the previous day as a result of an AR effect, but there may also be a delayed response to the perturbation from yesterday, for instance because the person remembers the events of the previous day. The strength of the delayed response depends on the size

[2] We add the * to ϵ, to distinguish the innovations for the ARMA(1,1,) model from the innovations of the AR(1)+WN model.

of θ. The ARMA(1,1) model, which is depicted in Figure 4.1C, can be specified as:

$$\begin{aligned} y_t &= \mu + \tilde{y}_t \\ \tilde{y}_t &= \phi\tilde{y}_{t-1} + \theta\epsilon^*_{t-1} + \epsilon^*_t \end{aligned} \tag{4.6}$$

$$\epsilon^*_t \sim N\left(0, \sigma^{2*}_\epsilon\right). \tag{4.7}$$

The ARMA(1,1) model is characterized by four parameters, that is, the mean μ, AR parameter ϕ, moving average parameter θ, and innovation variance σ^{2*}_ϵ. The model is stationary when ϕ lies between -1 and 1, and is invertible if θ lies between -1 and 1 (Chatfield, 2004; Hamilton, 1994).

If the true underlying model is an AR(1)+WN model, the ϕ and μ parameter in an ARMA(1,1) will be equal to those of the AR(1)+WN model. Granger and Morris (1976) have shown that the innovation variance σ^2_ϵ and measurement error variance σ^2_ω can be calculated from the estimated θ, ϕ, and σ^{2*}_ϵ as follows (see also Staudenmayer & Buonaccorsi, 2005),

$$\sigma^2_\omega = (-\phi)^{-1}\theta\sigma^{2*}_\epsilon, \tag{4.8}$$

$$\sigma^2_\epsilon = (1+\theta^2)\sigma^{2*}_\epsilon - (1+\phi^2)\sigma^2_\omega. \tag{4.9}$$

It is important to note that while the AR(1)+WN models is equivalent to an ARMA(1,1) model, an ARMA(1,1) models is not necessarily equivalent to an AR(1)+WN model. That is, it is only possible to transform the ARMA(1,1) parameters to AR(1)+WN model parameters under these restrictions in line with an underlying AR(1)+WN model (Granger & Morris, 1976; Staudenmayer & Buonaccorsi, 2005):

$$\frac{1}{1+\phi^2} > \frac{\theta}{1+\theta^2}(-\phi^{-1}) \geq 0 \tag{4.10}$$

4.2 Simulation Study Methods

We present a simulation study in which we simulate data according to an AR process with added measurement error. We fit an AR(1) model to the data in order to illustrate the effects of ignoring any present measurement error, and compare the performance of the AR(1) model to the AR(1)+WN, and ARMA(1,1) model, which

both account for measurement error. Furthermore, we will compare the performance of the Bayesian and frequentist estimation of these models.

Frequentist Estimation

For the frequentist estimation of the AR(1) model and the ARMA(1,1) model a relatively wide range of procedures and software is available. Potential estimation procedures for fitting the AR(1)+WN model include specially modified Yule-Walker equations, and modified Least Squares estimation procedures (Chanda, 1996; Dedecker, Samson, & Taupin, 2011; Staudenmayer & Buonaccorsi, 2005). However, we opt to use the (linear, Gaussian) state-space model, for which the Kalman Filter (Harvey, 1989; Kim & Nelson, 1999) is used to estimate the latent states, while Maximum Likelihood is used to estimate the model parameters (c.f., Staudenmayer & Buonaccorsi, 2005, for this approach, but with the measurement error variance considered as known). This is an especially convenient modeling framework for psychological longitudinal modeling, as it readily deals with missing data, and is easily extended to multivariate settings, or to include latent variables (c.f., Hamilton, 1994; Harvey, 1989; Kim & Nelson, 1999).

In the state-space model representation, a vector of observed variables is linked to a vector of latent variables – also referred to as 'state variables' – in the *measurement equation*, and the dynamic process of the latent variables is described through a first-order difference equation in the *state equation* (Hamilton, 1994; Harvey, 1989; Kim & Nelson, 1999). That is, the measurement equation is

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{d} + \boldsymbol{F}\tilde{\boldsymbol{y}}_t + \boldsymbol{\omega}_t \\ \boldsymbol{\omega}_t &\sim MvN\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\omega\right), \end{aligned} \tag{4.11}$$

where $\boldsymbol{y}_t$ is an $m \times 1$ vector of observed outcome variables, $\tilde{\boldsymbol{y}}_t$ is an $r \times 1$ vector of latent variables, $\boldsymbol{d}$ is an $m \times 1$ vector with intercepts for the observed variables, $\boldsymbol{F}$ is an $m \times r$ matrix of factor loadings, and $\boldsymbol{\omega}_t$ is an $m \times 1$ vector of residuals that are assumed to be multivariate normally distributed with zero means and $m \times m$ covariance matrix $\boldsymbol{\Sigma}_\omega$. The state equation (also referred to as the transition equation) is specified as

$$\begin{aligned} \tilde{\boldsymbol{y}}_t &= \boldsymbol{c} + \boldsymbol{A}\tilde{\boldsymbol{y}}_{t-1} + \boldsymbol{\epsilon}_t \\ \boldsymbol{\epsilon}_t &\sim MvN\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon\right), \end{aligned} \tag{4.12}$$

where $\boldsymbol{c}$ is an $r \times 1$ vector of intercepts for the latent variables, $\boldsymbol{A}$ is an $r \times r$ matrix of structural coefficients, and $\boldsymbol{\epsilon}_t$ is an $r \times 1$ vector of residuals, which are assumed to be multivariate normally distributed with zero means and $r \times r$ covariance matrix $\boldsymbol{\Sigma}_\epsilon$.

The previously discussed AR(1) and AR(1)+WN model are both already specified in terms of a state-space representation in Equations 4.1 through 4.5 (simplified where possible). For the state-space model specification for the ARMA(1,1) model vector $\boldsymbol{d}$ is μ, $\boldsymbol{F}$ is $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$, $\tilde{\boldsymbol{y}}_t$ is $\begin{bmatrix} \tilde{y1}_t & \tilde{y2}_t \end{bmatrix}^T$, $\boldsymbol{\Sigma}_\omega$ is a zero matrix, $\boldsymbol{c}$ is a zero vector, $\boldsymbol{A}$ is 2×2 matrix $\begin{bmatrix} \phi & 0 \\ 1 & 0 \end{bmatrix}$, and 2×2 matrix $\boldsymbol{\Sigma}_\epsilon$ is equal to $\boldsymbol{H^T H}$ with $\boldsymbol{H}$ equal to $\begin{bmatrix} \sigma_{1\epsilon^*} & \theta\sigma_{1\epsilon^*} \end{bmatrix}$, where superscript T indicates the transpose.

To fit the frequentist state-space models we use R, with R packages FKF (Kalman Filter; Luethi, Erb, & Otziger, 2010) combined with R base package optim (for maximum likelihood optimization; R Development Core Team, 2012). Within optim we used optimization method l-bfgs-b, with lower bounds and upper bounds for ϕ and θ of -1 and 1, -Inf and Inf for μ, and 0 and Inf for σ^2_ϵ, σ^2_ω, and σ^{2*}_ϵ.

Bayesian Estimation

Bayesian modeling shares a lot of conveniences with the frequentist state-space modeling framework: For instance, like frequentist state-space modeling procedures, Bayesian modeling can deal conveniently with missing data, is flexible in modeling multivariate processes, and in including latent variables in the model. Particular to Bayesian modeling is the relative ease in extending models to a hierarchical or multilevel setting (e.g., De Haan-Rietdijk et al., 2014; Lodewyckx et al., 2011). Another advantage may be the possibility to include prior information in the analysis, based, for instance, on expert knowledge or results from previous research (e.g., Rietbergen, Groenwold, Hoijtink, Moons, & Klugkist, 2014; Rietbergen, Klugkist, Janssen, Moons, & Hoijtink, 2011). Finally, the Bayesian estimation procedures are not dependent on large sample asymptotics like the frequentist procedures, and may therefore perform better for smaller samples (Dunson, 2001; Lee & Wagenmakers, 2005). Because currently there is no literature on the Bayesian estimation performance for the AR(1)+WN model, we will compare the performance of the Bayesian AR(1), ARMA(1,1), and AR(1)+WN model with the frequentist modeling equivalents in a simulation study.

In Bayesian estimation the information in the data, provided through the likelihood, is combined with a prior distribution using Bayes' rule (c.f., Gelman et al., 2003; Hoijtink et al., 2008). The prior distribution is specified such that it contains

prior information the researcher would like to include in the analysis. Here we prefer to specify uninformative prior distributions that contain minimal prior information, such that their influence is minimal. Specifically, we use the following prior specifications across the three models: A $uniform(0, 20)$ prior on σ^2_ω, σ^2_ϵ, and σ^{2*}_ϵ, a $uniform(-1, 1)$ prior on ϕ and θ, and a $normal(0, .001)$ prior for μ (specified with precision rather than variance). When the prior distribution and the likelihood are combined using Bayes' rule, this results in the posterior probability distribution or density of the estimated parameters. Summary statistics based on this distribution can then be used to summarize the information on the estimated parameters, for instance, the mean or median may be used to obtain a point estimate for an estimated parameter, and the posterior standard deviation can be used to describe the uncertainty around that point estimate.

Although it is possible to obtain the posterior distribution analytically for some simple models, the Bayesian estimation of more complex models is usually done with Markov Chain Monte Carlo algorithms, such as Gibb's sampling, which relies on consecutively samples from the conditional distributions of the parameters (rather than directly from their joint distribution, c.f., Casella & George, 1992); when the procedure has converged, one effectively samples from the (joint) posterior distribution. These samples can then be used as an approximation of the underlying posterior distribution, which in turn can be used to obtain point estimates for the parameters. A particularly desirable feature of MCMC procedures is that, based on the samples of the estimated parameters, it is also possible to calculate new statistics and obtain their posterior distribution. For instance, based on the estimated parameters θ, ϕ, and σ^{2*}_ϵ for the ARMA(1,1) model, we will calculate the innovation variance σ^2_ϵ and measurement error variance σ^2_ω in each sample, such that we obtain posterior distributions for these parameters. In our simulations we use the free open source software JAGS (Plummer, 2003) which employs a Gibb's sampling algorithm, in combination with the R package Rjags (Plummer, Stukalov, & Plummer, 2014).

Simulation Conditions

Throughout the simulation study, we simulated 1000 data sets per condition according to the AR(1)+WN model specified in Equations 4.3 to 4.2 using R (R Development Core Team, 2012). For all conditions, the mean of the model is fixed to 2. The study consists of three parts. First, we examine the effect of *the proportion of measurement*

error variance to the total variance, on parameter recovery. The total variance for the AR(1)+WN is the sum of the variance for an AR(1) model and the measurement error variance: $\sigma^2_{total} = \sigma^2_\epsilon/(1-\phi^2)+\sigma^2_\omega$ (c.f., Harvey, 1989; Kim & Nelson, 1999). To vary the proportion of σ^2_ω to the total variance, ϕ and σ^2_ϵ are both fixed to .5 in this study while the measurement error variance is varied. Specifically, the measurement error variance takes on the values 0, .1, .2, .3, .5, .7, 1, 2, 4, and 12, which results approximately in the following proportions of measurement error variance to the total variance: 0, .13, .23, .31, .43, .51, .6, .75, .86, and .95.

Second, we examine the effect of *the size of* ϕ on parameter recovery. We vary ϕ over the values -.75, -.5, -.25, 0, .25, .5, and .75. The proportion of measurement error variance to the total variance of the AR(1)+WN process is fixed to .3 here, through varying the innovation variances σ^2_ϵ by approximately 1.2, 1.1, .9, .5, .9, 1.1, and 1.2 respectively.

Third, we examine the effects of *sample size*. In part 1 and 2 of the study we use a sample size 100 repeated measures. We based this number roughly on what one may expect for research in psychology: Typically, what we see in time series applications in psychology is a range of about 60 to 120 repeated measures per person (e.g., see Adolf et al., 2014; Ferrer et al., 2012; Madhyastha et al., 2011; Nezlek & Gable, 2001; Rovine & Walls, 2005; Wang et al., 2012). However, in preliminary analyses we found difficulties in estimating the model with a small sample size, especially for the frequentist estimation procedure, that pointed to empirical underidentification (we elaborate on this in the next section). Therefore, we varied sample size by 100, 200, and 500. For this part of the study σ^2_ϵ, σ^2_ω, and ϕ were fixed to .5, implying a proportion of measurement error variance to the total variance of .43.

We judge the performance of each model based on: a) its bias in the estimates; b) the absolute error in the estimates; and c) coverage rates for the 95% confidence or credible intervals. It is not clear whether Bayesian 95% credible intervals should have exactly 95% coverage rates, however, with uninformative priors we would expect this to be the case. Moreover, we consider it informative to see how often the true value lies within the credible interval across multiple samples (e.g., if this occurs very rarely this seems problematic for making inferences).

For the coverage rates of the variances estimated with the frequentist ML procedure, we calculate the confidence intervals based on a χ^2 distribution with degrees of freedom *df* as follows: $CI(\frac{(df)s^2}{\chi^2_{1-\alpha/2}}, \frac{(df)s^2}{\chi^2_{\alpha/2}})$, where n is the sample size, and s^2 is the

estimated variance ($df = n - 2$ for σ_ϵ^2 in the AR(1) model, and $df = n - 3$ for σ_ϵ^2 and σ_ω^2 in the ARMA(1,1) and the AR(1)+WN model).

Expectations

For part 1, we expect that all models will decrease in performance (i.e., more bias and absolute error, lower coverage rates) as the proportion of measurement error variance increases, because an increase in random noise should make it harder to distinguish an (autoregressive) effect. Furthermore, we expect that the decrease in performance will be larger for the AR(1) model than for the ARMA(1,1) and AR(1)+WN model. Specifically, based on Staudenmayer and Buonaccorsi (2005), we expect a bias in the estimates of ϕ in the AR(1) model of approximately 0, -.07, -.12, -.16, -.21, -.26, -.30, -.38, -.43, and -.47, given that the proportions of measurement error variance are 0, .13, .23, .31, .43, .51, .6, .75, .86, and .95.

For part 2, we expect that the AR(1)+WN and ARMA(1,1) models will improve in performance as the value of $|\phi|$ increases, given that σ_ω^2 and σ_ϵ^2 should be more easily distinguished from each other as $|\phi|$ approaches 1. We are specifically interested in the performance of the AR(1)+WN model compared to the ARMA(1,1) model when $|\phi|$ is relatively small. Given that the ARMA(1,1) model is identified regardless of the value of ϕ, we expect the ARMA(1,1) model may converge better, and therefore to perform better when ϕ is relatively close to zero than the AR(1)+WN model, which is no longer identified when ϕ is equal to zero.

For part three, we expect that performance will improve as sample size increases for the ARMA(1,1) model and the AR(1)+WN model, both in the frequentist and Bayesian estimation procedure. Finally, we expect that the Bayesian procedure will perform better than the frequentist state-space procedures for smaller sample sizes, given that both modeling procedures have similar benefits, but the Bayesian estimation procedure is not dependent on large sample asymptotics (Dunson, 2001; Lee & Wagenmakers, 2005).

4.3 Simulation Study Results

In this section we present the results of the simulation study. As was mentioned before, for a sample size of 100 we found some convergence issues especially for the

frequentist ML procedure. Given that convergence is an important precondition for obtaining reasonable parameter estimates, we start by discussing the convergence of the Bayesian models and frequentist models across the different parts of the simulation study. After that, we discuss the parameter recovery performance for each condition specific for each of the three parts of the simulation study. We end with a summarizing conclusion.

Convergence of the Bayesian Procedures

For the Bayesian procedures we obtained three chains of 40,000 samples each for each replication, half of which was discarded as burn-in. We judged convergence based on the multivariate Gelman-Rubin statistic and autocorrelations for all replications, and we inspected the mixing of the three chains visually a number of replications (c.f., Brooks & Gelman, 1998; Gelman & Rubin, 1992). For the AR(1) model the chains mixed well, the Gelman Rubin statistic was generally equal to one, and the autocorrelations for the parameters decreased exponentially across all conditions.

For the ARMA(1,1) the chains generally mixed well, and the Gelman Rubin statistic was equal to one across all conditions.[3] The autocorrelations for the parameters decreased slower than for the AR(1) model, and decreased most slowly when the proportion of measurement error variance was higher than 50% or $|\phi|$ was zero.

For the AR(1)+WN model, overall the chains mixed well and the Gelman Rubin Statistic was equal to one for most replications. For approximately 1% to 2% of the data sets the Gelman Rubin statistic was larger than 1.1, indicating possible non-convergence, with the exception of the condition where $\phi = .75$, for which it was 8%. Closer inspection indicated that these problems usually originated and were limited to μ. The percentage of non-convergence is larger for the condition $\phi = .75$, most likely because when ϕ is strong and positive it is most difficult to estimate μ because observations may tend to linger longer above or below the mean. The autocorrelations for the AR(1)+WN model are higher overall, and slower to decrease than those for the AR(1) and ARMA(1,1) model across all conditions. More measurement error and a closer $|\phi|$ to zero, was associated with more slowly decreasing autocorrelations.

[3]By visually inspecting the chains for μ in the ARMA(1,1) model, we found some extreme values for some of the Gibb's samples (visible as large 'spikes' in the chains). To limit these extreme values we adjusted the normal prior for μ to have a smaller variance (10), however this did not resolve the issue completely. As a result, the posterior standard deviation for μ was very large, however, the effects on the point estimates and credible intervals seem limited when we compare these results for μ to those of the other models.

Convergence of the (Frequentist) ML State-Space Modeling Procedures

For the ML procedure we encountered three types of problems: 1) negative standard errors for the estimated parameters , 2) optim failing to initialize (more rarely), and 3) Heywood cases (negative variances) for the measurement error variance or the innovation variance. The first and second type of problem could usually be resolved by providing alternative starting values and rerunning the model. For a small percentage of data sets, five sets of starting values still did not resolve these issues (for the number of data sets per condition, see Table 4.2 in Appendix 4.A). These data sets are excluded from the parameter recovery results. When sample size was increased to 200 or 500 repeated measurements, these problems were no longer encountered.

The third type of problem - Heywood cases - was much more prevalent, and could generally not be resolved by providing different starting values. For the AR(1)+WN model, for 10% to 55% of the replications σ^2_ω, or more rarely σ^2_ϵ, were estimated at the lower bound of zero. For the ARMA(1,1) model, we similarly detected Heywood cases for σ^2_ω and σ^2_ϵ (note that σ^2_ω and σ^2_ϵ are calculated a posteriori based on the estimated ϕ, θ and σ^{2*}_ϵ by means of Equation 4.8 and 4.8). In the case that for the AR(1)+WN model σ^2_ω or σ^2_ϵ were estimated at the lower bound, usually a Heywood case would also observed for the ARMA(1,1) model for that replication. The proportions of Heywood cases for σ^2_ω and σ^2_ϵ across all conditions are reported in Table 4.2 in the Appendix 4.A.

The number of Heywood cases increased when: 1) ϕ got closer to zero, such that it is harder to discern measurement errors from innovations 2) when there was very little measurement error, such that σ^2_ω was already close to zero, and 3) There was a lot of measurement error, such that all parameter estimates were uncertain (large standard errors). This indicates issues of empirical identification, and as such we expected these issues to decrease as sample size increases.

The Heywood cases for σ^2_ϵ and σ^2_ω decreased as sample size increased - however, the issues were not resolved completely: For n=200 almost 30% of the data sets still returned a Heywood case, and for n=500 almost 13% still returned a Heywood case. Given that for smaller sample sizes (e.g., less than 500), which are much more common in psychological studies, the proportion of replications with Heywood cases was quite large for many conditions, this seems quite problematic. In practice, encountering such a result may lead a researcher to erroneously conclude that there most likely is

no considerable measurement error variance, so that a regular AR(1) model should suffice.

In the following sections, where we discuss the parameter recovery results, the data sets with Heywood cases for σ^2_ω or σ^2_ϵ are included in the results, because to exclude so many data sets would make a fair comparison to the Bayesian procedure (for which no data sets need to be excluded) problematic. However, the results with these data sets excluded for the ML AR(1)+WN model and ARMA(1,1) model are presented and discussed in Appendix 4.A. Finally note that, in contrast to our expectations, in the ML procedure the ARMA(1,1) model does not seem to converge more easily than the AR(1)+WN model. In general it seems that in order to properly estimate and distinguish the measurement error variance from the innovation variance using ML, quite large sample sizes are required.

Parameter Recovery for Different Proportions of Measurement Error

In general, as the proportion of measurement error increases, the estimated parameters become increasingly more biased, the absolute errors become larger, and coverage rates become lower, as expected. In Figure 4.2 we provide plots of the 95% coverage, absolute errors, and bias for each model, condition, and parameter. As can be seen from this figure, overall, the Bayesian AR(1)+WN model outperforms the other procedures in terms of coverage rates and absolute errors, and for the variance parameters also in terms of bias. The ML state-space AR(1)+WN model performs second-best overall, and performs the best for ϕ in terms of bias. The Bayesian and frequentist AR(1) and ARMA(1,1) models perform relatively poorly in all respects. However, the ARMA(1,1) models result in better coverage rates for ϕ than the AR(1) models, so that an ARMA(1,1) model is still preferred over a simple AR(1) model. Below, we will discuss the results in more detail, per parameter.

For μ, all models perform similarly well in terms of bias and absolute error, as can be seen from the top-left panel of Figure 4.2. In terms of coverage rates, the Bayesian AR(1) and AR(1)+WN model outperform the other models for μ, most pronouncedly when the proportion of measurement error is high.

For ϕ, the models that perform the best in terms of bias are the ML AR(1)+WN model, followed by the Bayesian AR(1)+WN model (see the top-right panel in Figure 4.2). The bias for ϕ in both AR(1) models is in line with our expectations, increasing from approximately zero to -.5 as measurement error increases. As can

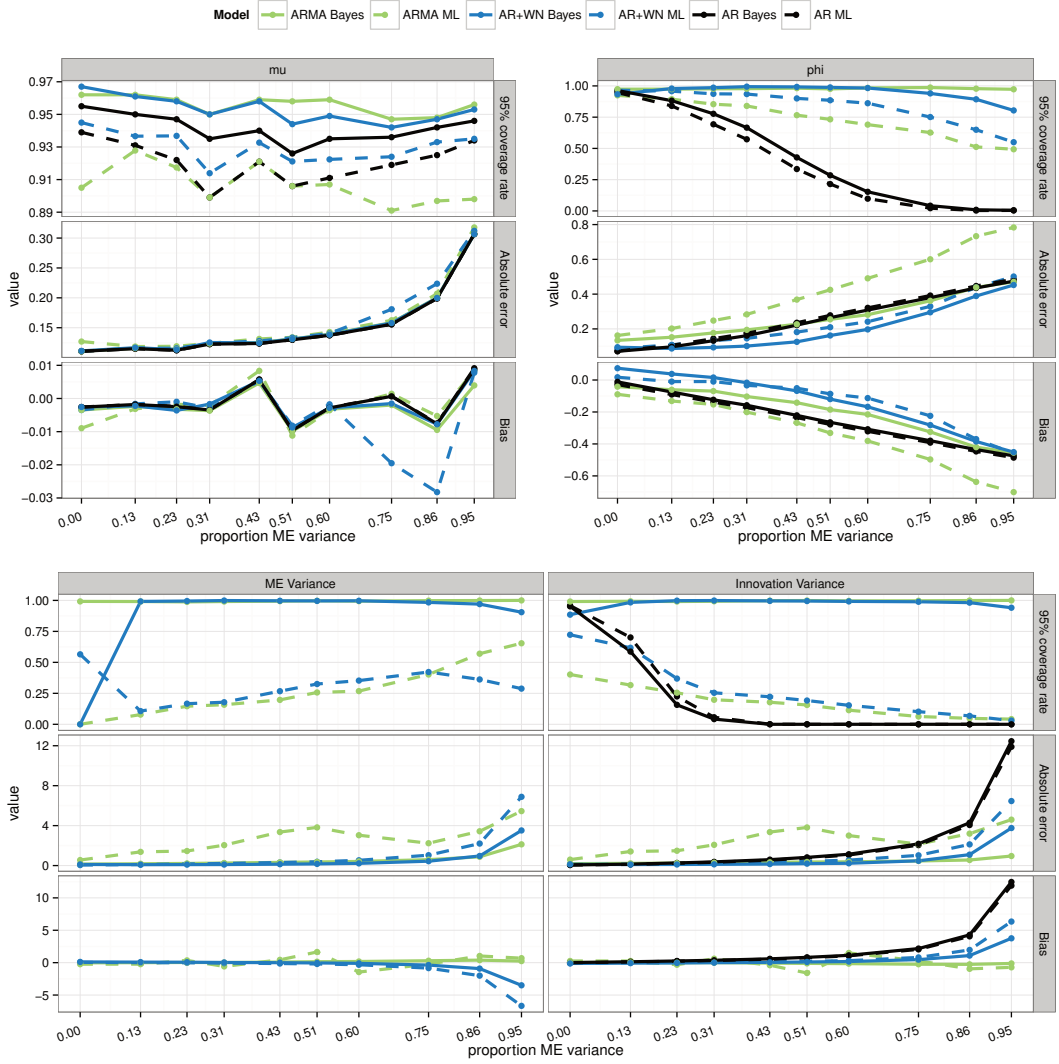


Figure 4.2: Coverage rates, absolute errors, and bias for the parameter estimates for the frequentist and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across different proportions of measurement error variance to the total variance.

be seen from the top-right panel of Figure 4.2, in terms of absolute error for ϕ, the Bayesian AR(1)+WN model performs the best, followed by the ML AR(1)+WN model. The top-right panel of Figure 4.2 shows that the coverage rates for ϕ based on the 95% CI's for the Bayesian ARMA(1,1) model are consistently higher than those for the other models, however, this is a result of having wider credible intervals, rather than a result of more precise estimates for ϕ. The coverage rates for the Bayesian AR(1)+WN model are most stable across the different proportions of measurement error variance. The coverage rates for this Bayesian model are generally higher than .95,[4] only dropping below .95 when 75% or more of the total variance is measurement error variance. In comparison, the ML AR(1)+WN model starts with a coverage rate of approximately .95 for ϕ when measurement error is absent, and the coverage decreases as measurement error increases (with a lowest coverage of .55 when 95% of the variance is due to measurement error). The ML ARMA(1,1) model and the Bayesian and ML AR(1) models perform the worst, as can be seen from Figure 4.2. Note that for the AR(1) models, the coverage rates for ϕ are already below 90% when the proportion of measurement error variance is as little as .13.

In the bottom panel of Figure 4.2 the results for σ^2_ω and σ^2_ϵ are displayed. When the proportion of error variance is larger than about .3, the Bayesian AR(1)+WN model starts to outperform the ML AR(1)+WN model in terms of bias for σ^2_ω and σ^2_ϵ. Further, it can be seen from Figure 4.2 that for the AR(1)+WN models, when the proportion of measurement error is small, the measurement error variance is slightly overestimated, while when the proportion of measurement error is large, the measurement error variance is underestimated. The coverage rates are the highest for the Bayesian AR(1)+WN and ARMA(1,1) model. Note that for the ARMA(1,1) model σ^2_ω and σ^2_ϵ are calculated based on the estimated ARMA(1,1) parameters. For the Bayesian model this was done in each Gibbs sample by means of Equations 4.8 and 4.9, resulting in a posterior distribution for σ^2_ω and σ^2_ϵ. However, depending on the specific values of the ARMA(1,1) parameters in each Gibbs sample, σ^2_ω and σ^2_ϵ may become quite large or even negative. As a result, the posterior standard deviations and credible intervals for σ^2_ω and σ^2_ϵ in the Bayesian ARMA(1,1) model can be quite large, including negative and large positive values. The confidence intervals for

[4]While it may seem undesirable that the Bayesian model has 'too high' coverage rates, indicating too large credible intervals or exaggerated uncertainty about the estimated parameters, it is important to note that compared to the ML model, the Bayesian estimates actually have smaller posterior standard deviations than the ML standard errors (except for parameter μ, for which it is the reverse).

the variances parameters in frequentist procedures are consistently too narrow, which results in low coverage rates, as can be seen from the bottom panel of Figure 4.2. As such, for the two variances, the Bayesian AR(1)+WN model performs best in terms of coverage rates, followed by the Bayesian ARMA(1,1) model (which has higher coverage rates, but much wider intervals), and the ML AR(1)+WN model. The same pattern holds for the absolute errors as can be seen in Figure 4.2.

Parameter Recovery for Different Values of ϕ

For this part of the study, the value of ϕ was varied from -.75, to -.5, -.25, 0, .25, .5 and .75. As can be seen from the top-left panel of Figure 4.3, for μ all the models perform very similarly in terms of bias, absolute errors, and coverage rates. The absolute errors and bias increase as ϕ becomes larger, because when ϕ is strong and positive, observations may tend to linger longer above or below the mean than when ϕ is weak or negative, making it harder to estimate μ.

As can be seen from the top-right and bottom panel of Figure 4.3, the results for ϕ and the variance parameters are symmetric for negative and positive values of ϕ (or mirrored in the case of bias). As such, we will discuss these results in terms of $|\phi|$. For the parameters ϕ, σ^2_ϵ and σ^2_ω, performance increases as $|\phi|$ increases, except the AR(1) models, for which it is the opposite. Overall, the Bayesian AR(1)+WN performs best, followed by respectively the ML AR(1)+WN model, the Bayesian ARMA(1,1) model, and the ML ARMA(1,1) model. The performance of the latter three models decreases considerably more as $|\phi|$ decreases than that of the Bayesian AR(1)+WN model, as can be seen from Figure 4.3.[5] For the two variances, the ML AR(1)+WN model outperforms the Bayesian model in terms of bias. Finally, we find that when $|\phi|$ is relatively close to one, the measurement error variance is underestimated, however, when $|\phi|$ is relatively small, the measurement error variance was actually overestimated, as can be seen from the bottom panel of Figure 4.3.

[5]The diverging patterns in the bias and absolute errors for the ML ARMA(1,1) model is a result of the Heywood cases discussed in section 4.3; when the Heywood cases are removed the pattern is similar to the patterns of the other models, as can be seen in Figure 4.31 to 4.32 in Appendix 4.A

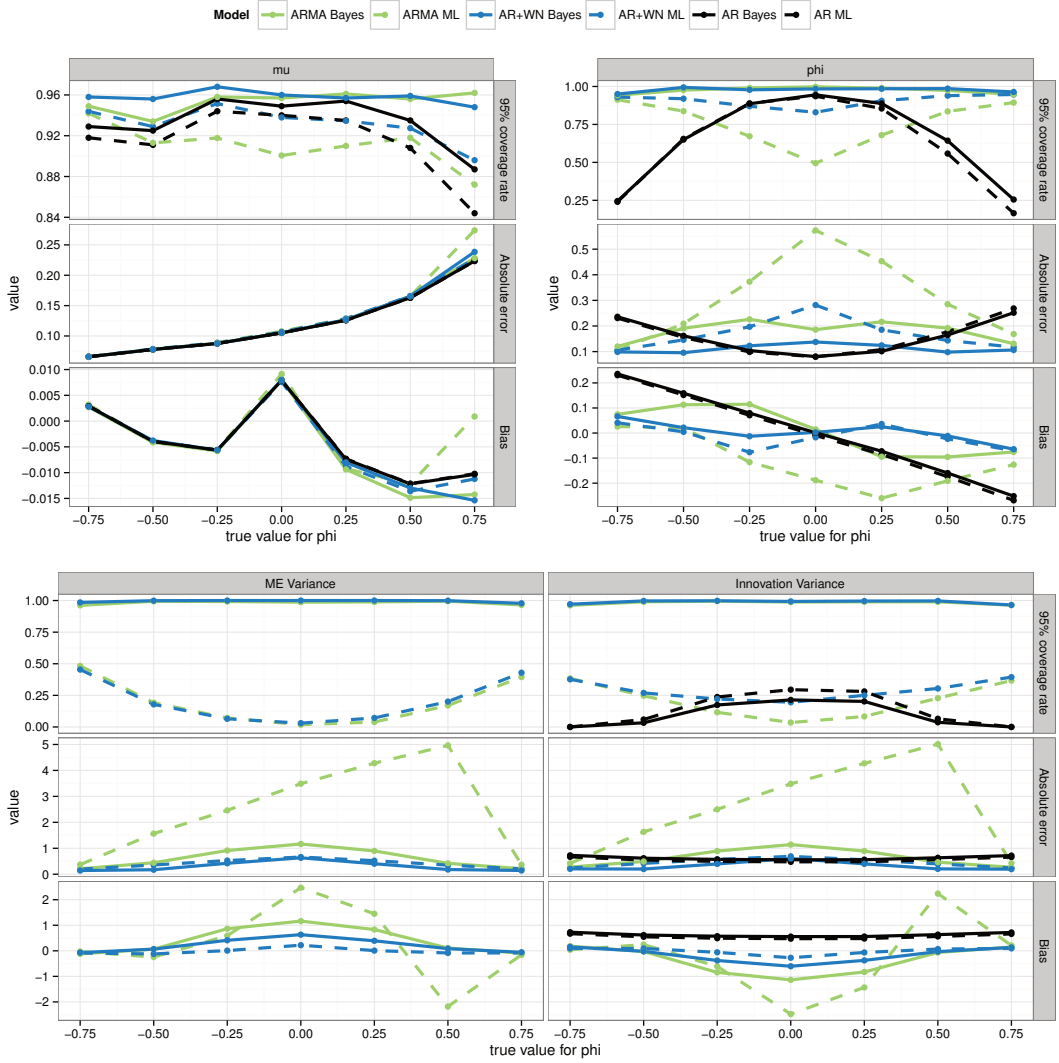


Figure 4.3: Coverage rates, absolute errors, and bias for the parameter estimates for the frequentist and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across different values for ϕ.

Parameter Recovery for Different Sample Sizes

For this part of the simulation study, the sample size was varied from 100 to 200 and 500. As shown in Figure 4.4, as sample size increases, parameter recovery improves: Bias and absolute errors decrease, while coverage rates become closer to .95. We Further, the ML AR(1)+WN results become more similar to those of the Bayesian AR(1)+WN model as sample size increases, although the Bayesian model still outperforms the ML model in terms of absolute error and coverage: The Bayesian procedure results in higher coverage rates, but less wide intervals, that is, in more precise estimates than the ML procedure for ϕ. Note that the performance of the ML and Bayesian ARMA(1,1) models only near the performance of the AR(1)+WN models as sample size has increased to 500 observations.[5]

Conclusion

Overall, the Bayesian AR(1)+WN model performs better than the other five procedures we considered. We expected that the ARMA(1,1) models may outperform the AR(1)+WN models in parameter recovery, because we expected this model to have less trouble with identification and convergence. Interestingly, although the Bayesian ARMA(1,1) model seems to converge more easily than the Bayesian AR(1)+WN model, the AR(1)+WN model still outperforms the ARMA(1,1) model in terms of parameter recovery, even when ϕ is close or equal to zero. The ML AR(1)+WN model and ARMA(1,1) models are both unstable for small sample sizes (n=100), frequently resulting in Heywood cases for the innovation and measurement error variances. However, the ML AR(1)+WN model still performs relatively well for estimating ϕ compared to the AR(1) models. For a smaller sample size of 100 observations the Bayesian procedure outperforms the frequentist ML procedure. When sample sizes are larger, the discrepancies between the Bayesian and frequentist AR(1)+WN model decrease, although the confidence intervals for the variance parameters in the frequentist procedures are consistently too narrow. As expected, the AR(1) models severely underestimate $|\phi|$, which is reflected in large bias and absolute errors, and low coverage rates. Finally, we note that although the AR(1)+WN models perform considerably better than the AR(1) models, some bias in ϕ still remains, because the innovations and measurement errors cannot be perfectly discerned from each other. Generally, the more measurement error and the lower $|\phi|$, the more the estimate of $|\phi|$ will be biased, even when measurement error is taken into account by the model.

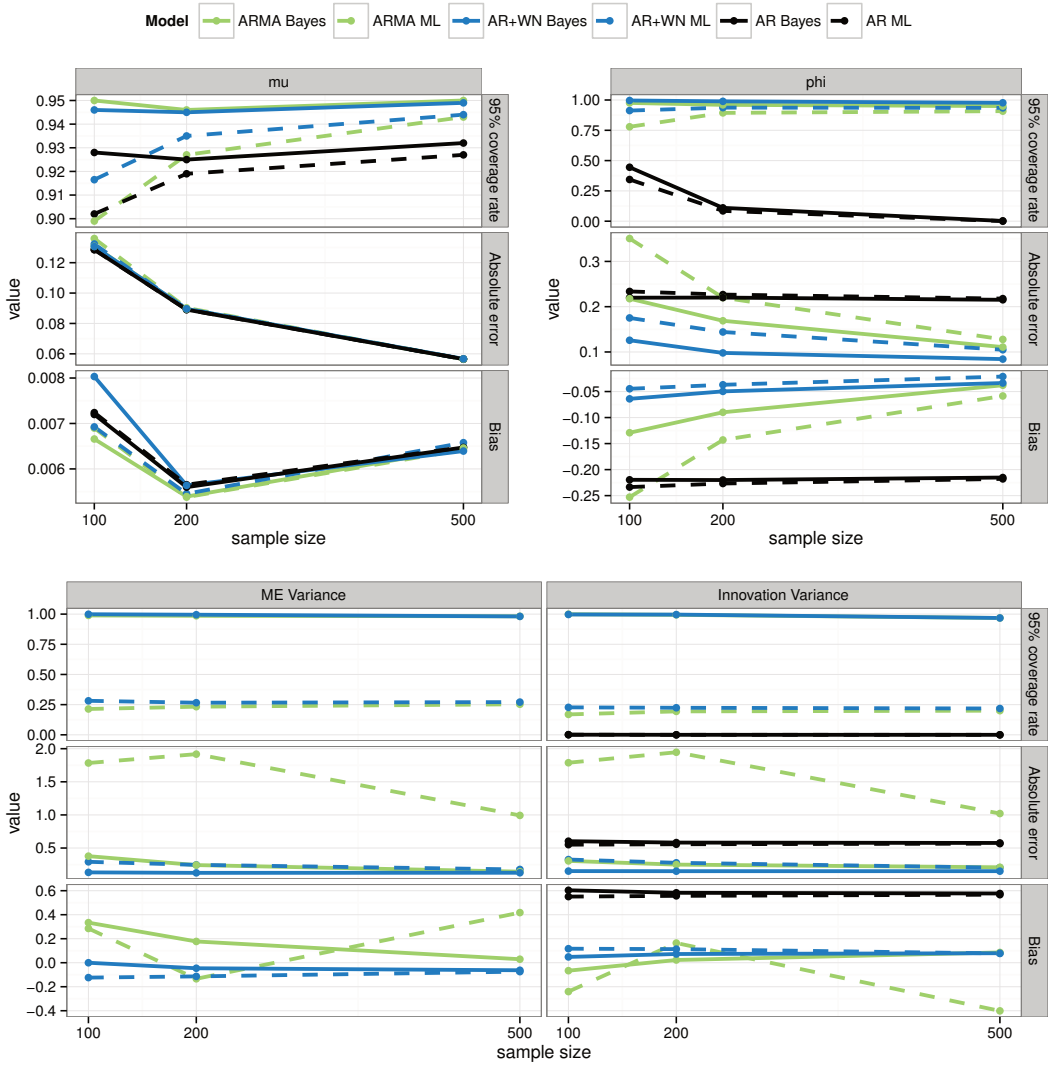


Figure 4.4: Coverage rates, absolute errors, and bias for the parameter estimates for the frequentist and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across sample sizes.

4.4 Empirical Application on Mood Data

To further illustrate the AR(1), ARMA(1,1), and AR(1)+WN model discussed above, we make use of time series data that was collected from female first year social science students at Utrecht University in 2007. Eleven women kept a daily electronic diary for approximately three months (across participants the minimum was 90 observations, the maximum 107 observations), in which they filled out how they felt that day on a scale from 1 to 100 - 1 meaning worst ever, and 100 meaning best ever. Three of the eleven women were excluded from the current study because of non-compliance, issues with the electronic devices, and one woman had very little variation in her scores. For the remaining women the average number of missing observations was approximately nine. Values for these missing observations will be automatically imputed as part of the estimation procedure, based on the specified model.

We are interested in finding out to what extent current mood influences mood the following day. As such, we are interested in fitting an AR(1) model, and specifically in the AR effect reflected in parameter ϕ. However, the mood of each person is not likely to be perfectly measured. For instance, it is possible that participants accidentally tapped the wrong score when using the electronic diary stylus to fill in the questionnaire. Furthermore, the participants evaluate their mood for the day on average, such that momentary influences around the time of filling out the diary may have colored their evaluation of the whole day (i.e., a form of retrospective bias). In fact, anything that is not explicitly measured and modeled, and of which the influence does not carry-over to the next day, can be considered measurement error. As such, it seems likely that there is at least some measurement error present in the data. Therefore, we fit the AR(1)+WN model to take this measurement error into account, and for illustrative purposes compare it to an ARMA(1,1) model, and an AR(1) model (which disregards measurement error). We make use of a Bayesian modeling procedure, given that the results from our simulation study indicate that the parameter recovery performance of the Bayesian procedure is better and more stable for this number of repeated measures. The data and code used for fitting the models are available as supplementary materials with the online publication (Schuurman et al., 2015) or at www.nkschuurman.com. The priors we use for the models are aimed to be uninformative, specifically: A $uniform(0, 500)$ prior distribution for all variance parameters, a $uniform(-1, 1)$ prior distribution for ϕ and θ, and a $normal(0, .001)$ prior distribution for μ (specified with a precision rather than a variance).

We evaluated the convergence of the AR(1), ARMA(1,1), and AR(1)+WN model by visually inspecting the mixing of the three chains, the Gelman Rubin statistic, and the autocorrelations. For the AR(1) and AR(1)+WN model the chains mixed well, the Gelman Rubin statistic was approximately equal to one, and the autocorrelations for the parameters decreased within 50 to 100 lags across all participants. For the ARMA(1,1) model this was the case, except for participants 3 and 8.[6] We included the ARMA(1,1) estimates for these participants in Table 4.1, but these should be interpreted with caution.

The parameter estimates of the mean μ, AR parameter ϕ, innovation variance σ^2_ϵ, measurement error variance σ^2_ω, and moving average parameter θ for each person are presented in Table 4.1. For most of the eight individuals, the baseline mood is estimated to be around 60-70, which indicates that on average they are in moderately good spirits. Further, we see that across models and persons, the AR parameters are either estimated to be positive, or nearly zero. Participant 8 has an AR effect near zero in both the AR(1) model and the AR(1)+WN model, so that for her, everyday seems to be a 'new day': How she felt the previous day does not predict her overall mood today. On the other hand, for participants 2, 4, 5, and 6, the credible intervals for ϕ include only positive values across models: how they feel today depends in part on how they felt yesterday. For the remaining individuals, 1, 3, and 7, the point estimates for ϕ are also positive, however, the credible intervals including negative and positive values for ϕ.

When we compare the results for the AR(1) model and the AR(1)+WN model, we find that for all participants except participant 8, the AR parameter is estimated to be higher in the AR(1)+WN model: Because the AR(1) model does not take measurement error into account, the AR parameter is estimated to be lower than for the AR(1)+WN model. The extent to which the estimate for ϕ differs across the AR(1) and AR(1)+WN model, differs from person to person. The larger the estimated measurement error variance relative to the total variance, the larger the difference between the estimated ϕ in the AR(1) and AR(1)+WN model. For instance, for participants 4 and 6 their estimates of ϕ in the AR(1) model are quite similar to each other (i.e., .21 and .27), but because the measurement error variance for participant

[6] For participants 3 and 8 we found that the estimates for ϕ and θ in the ARMA(1,1) model were very dispersed, varying across the entire range of -1 to 1, switching from negative to positive values. A density plot of their samples revealed a bimodal distribution for ϕ and θ (with one peak around negative values, and one for positive values): This seems to be some form of label switching, which is indicative of (empirical) under–identification of the ARMA(1,1) model for these two participants.

4 is estimated to be much larger than that for participant 6 (i.e., 70 versus 10), her ϕ in the AR(1)+WN model ϕ is also estimated to be larger (i.e., .69 vs. .33).

Note that the ARMA(1,1) and AR(1)+WN model should not necessarily give the same results: Although the AR(1)+WN model is equivalent to the ARMA(1,1) model, the reverse is not the case. In other words, it is possible that the ARMA(1,1) model captures a different pattern of variation in the data than the AR(1)+WN model, giving different results. However, when we compare the results for the ARMA(1,1) and AR(1)+WN model, we do find fairly similar results for most of the participants (with exception of participants 3 and 8, who had convergence issues for the ARMA(1,1) model), especially for participants 2 and 5. However, a clearly notable difference is that the ARMA(1,1) model has less precise estimates than the AR(1)+WN model, as can be seen from the relatively wide credible intervals for the ϕ parameters in Table 4.1.

Finally, we note that when we calculate the estimated proportion of measurement error variance relative to the total variance based on the AR(1)+WN model for each participant, we find a range of .34 to .50 (i.e., .36, .47, .48, .50, .46, .42, .46, and .34 respectively). This implies that across these eight women, between one third to half of the observed variance is estimated to be due to measurement error.

4.5 Discussion

In this paper we demonstrate that it is important to take measurement error into account in AR modeling. We illustrated the consequences of disregarding measurement error present in the data both in a simulation study, and an empirical example based on a replicated time series design. Further, we compared the parameter recovery performance for the Bayesian and frequentist AR(1)+WN and ARMA(1,1) models that account for measurement error. Ignoring measurement error present in the data is known to result in biased estimates toward zero of the AR effects in AR(1) models, with the extent of the bias depending on the proportion of measurement error variance and the size of ϕ (Staudenmayer & Buonaccorsi, 2005). Our simulations also demonstrated this bias, and showed large absolute errors and importantly, very poor coverage rates for the AR effect when measurement error is disregarded, regardless of sample size. For research in psychology, for which it is very difficult or perhaps impossible to measure error-free, it seems imperative to consider this potentially large source of variance in our (AR) time series models. In our empirical application for

Table 4.1: Parameter estimates for the AR(1), ARMA(1,1), and AR+WN model for the mood of eight women, estimated with Bayesian software.

Pp	Model	μ (95% CI)	ϕ (95% CI)	σ^2_ϵ (95% CI)	σ^2_ω (95% CI)	σ^{2*}_ϵ (95% CI)	θ (95% CI)
1	AR1	75 (72, 79)	.08 (-.17, .32)	166 (122, 235)	-	-	-
	ARMA	76 (72, 81)	.53 (-.32, .90)	21.34 (-91, 180)	125 (-6, 278)	160 (117, 227)	-.41 (-.81, .29)
	ARWN	76 (72, 79)	.39 (-.23, .77)	42 (3, 160)	112 (16, 193)	-	-
2	AR1	63 (59, 68)	.36 (.13, .57)	188 (141, 256)	-	-	-
	ARMA	63 (58, 69)	.48 (-.21, .97)	103 (-740, 1087)	69 (-870, 960)	189 (142, 257)	-.13 (-.64, .49)
	ARWN	63 (58, 68)	.52 (.15, .84)	101 (20, 208)	77 (7, 184)	-	-
3	AR1	63 (61, 66)	.21 (0,.42)	108 (81, 148)	-	-	-
	ARMA	64 (61, 66)	.02 (-.72,.81)	-1 (-288, 251)	109 (-134, 418)	105 (79, 144)	.19 (-.64,.95)
	ARWN	64 (61, 67)	.40 (-.01, .82)	38 (4, 112)	64 (6,118)	-	-
4	AR1	56 (53, 58)	.21 (.01, .42)	103 (78, 141)	-	-	-
	ARMA	54 (40, 59)	.85 (.35,.99)	7(1,47)	75 (44,112)	95 (71, 130)	-.68 (-.87,-.14)
	ARWN	55 (49, 59)	.69 (.07, .97)	19 (2, 88)	70 (17, 111)	-	-
5	AR1	69 (64, 75)	.48 (.28, .67)	174 (131, 239)	-	-	-
	ARMA	69 (62, 77)	.67 (.20, .92)	86 (24, 348)	61 (-139, 143)	173 (130, 237)	-.26 (-.58,.24)
	ARWN	69 (62, 77)	.67 (.37, .91)	90 (27, 190)	66 (6,140)	-	-
6	AR1	73 (71, 74)	.27 (.07, .46)	31 (24, 42)	-	-	-
	ARMA	73 (71, 74)	.18 (-.43,.66)	22 (-305, 349)	8 (-314, 339)	31 (24, 42)	.09 (-.45,.61)
	ARWN	73 (71, 74)	.33 (.01, .62)	21 (4, 35)	10 (.51, 30)	-	-
7	AR1	71 (69, 73)	.08 (-.13, .28)	105 (79, 144)	-	-	-
	ARMA	71 (65, 75)	.48 (-.77,.99)	7 (-132, 175)	87 (-63, 248)	104 (78, 142)	-.36 (-.90, .77)
	ARWN	71 (68, 74)	.26 (-.57, .92)	23 (1, 101)	76 (8, 123)	-	-
8	AR1	73 (71, 74)	.03 (-.18, .24)	59 (44, 80)	-	-	-
	ARMA	73 (71, 74)	-.22 (-.81, .84)	-5 (-131, 102)	67 (-41, 197)	57 (43, 78)	.31 (-.98, .95)
	ARWN	73 (71, 74)	-.03 (-.65, .51)	16 (.35, 61)	42 (2, 70)	-	-

Note. Note that the negative values for in the credible interval for σ^2_ϵ and σ^2_ω for the ARMA(1,1) models result, because they are calculated a posterior based on the samples for ϕ, θ, and σ^{2*}_ϵ based on Equations 4.8 and 4.9: It is possible that for certain combinations of these parameters that σ^2_ϵ and σ^2_ω become negative. For participant 3 and 8 the ARMA(1,1) model did not converge properly, so that these results should be interpreted with caution.

instance, between one third to half of the variance in the data is estimated to be due to measurement error.

Comparing the parameter recovery for the models that incorporate measurement error - the Bayesian and ML ARMA(1,1) model and AR(1)+WN model - revealed that the Bayesian AR(1)+WN model performed best in terms of parameter recovery. It proved relatively tricky to properly estimate the ML ARMA(1,1) and AR(1)+WN model, even for larger sample sizes of 500 repeated measures: These models are prone to Heywood cases in the measurement error variance and to a lesser extent in the innovation variance. This was especially common (up to 55% of the replications) when AR effect was closer to zero, or the amount of measurement error was large. In practice, hitting such a lower bound for the measurement error variance may erroneously suggest to researchers that the model is overly complex, and that there is no notable measurement error present in the data, which is problematic.

Note that while 100 observations may be small for estimation purposes, it is quite a large number of repeated measures to collect in practice. In psychological research using intensive longitudinal data, we usually see no more than about 120 observations per person (to illustrate, 120 observations would arise from about 4 months of daily measurements, or for more intense two weeks regime, measuring someone 9 times a day). Fortunately, the Bayesian AR(1)+WN model provides a good option even for such small sample sizes. Still, the models that incorporate measurement error need more observations to give as precise estimates as the basic AR(1) model, which has relatively small credible/confidence intervals (although this is precision around a wrong estimate when there actually is measurement error present in the data). Therefore, it seems good practice to take potential measurement error into account in the design of the study, thus collecting more repeated measures in order to compensate for any potential measurement error that has to be filtered out later. Expectedly, and as is shown in the simulation study, this becomes especially important when the proportion of measurement error variance is relatively large, or when the AR effects are (expected to be) relatively small. One option to improve the estimates may be to use (weakly) informative prior specifications based on previous research, or expert knowledge. However, prior information on the model parameters may currently prove difficult to obtain, given that studies that estimate measurement error or take it into account are very rare, and that the model parameters differ from person to person, and from variable to variable. Another option could be to extend the AR+WN model to a multilevel model, assuming a common distribution for the parameters of multiple

individuals, and allowing the model parameters to vary across persons. By making use of this hierarchical structure that can take similarities between persons into account, a relatively low number of time points may be compensated for to some extent by a large number of participants, which may be easier to obtain (for examples of the multilevel AR(1) model, see De Haan-Rietdijk et al., 2014; Lodewyckx et al., 2011; Rovine & Walls, 2005).

The reader may wonder how one may determine if there is, or isn't, measurement error present in the data. One way to do this is to use information criteria to compare the AR(1) model with the ARMA(1,1) or AR(1)+WN model. Although a thorough study of model selection is beyond the scope of the current paper, we provide some preliminary evaluations of the model selection performance of the AIC, BIC, and DIC, in Appendix 4.B. We find that these criteria frequently incorrectly selects the simpler AR(1) model over the (true) AR(1)+WN model and ARMA(1,1) model, so that these criteria seem inappropriate for selecting between the AR(1) and the ARMA(1,1) model or the AR(1)+WN model in this context. Selecting between an AR(1)+WN model and an ARMA(1,1) model will also be problematic using standard information criteria, because the AR(1)+WN model may be considered a restricted (simpler) version of the ARMA(1,1) model (see Equation 4.8), while they have the same number of parameters, and thus the same penalty for complexity for many fit criteria. In that sense, when they have equal fit, the AR(1)+WN model may be preferred because it is the simpler model, but if this is not the case, it becomes more complicated to choose between the two. Directions for future research therefore are to establish information criteria for selecting between the AR(1)+WN model and the AR(1) and ARMA(1,1) model, perhaps using information criteria or Bayes factors developed for restricted parameters (c.f., Dudley & Haughton, 1997; Klugkist & Hoijtink, 2007; Kuiper, Hoijtink, & Silvapulle, 2012). Although model selection using information criteria may prove complicated, it is important to note that the estimates for ϕ in the AR(1)+WN models seem to be reasonably accurate, even when there is no measurement error present in the data. Combined with the intuition that most psychological measurements will contain at least some measurement error, fitting the model that incorporates measurement error seems a relatively 'safe bet.'

Another interesting topic for future work is how measurement error affects estimates of the effects variables have on each other over time, that is, the cross-lagged effects. This may be especially relevant for individual network models of psychological processes (Schmittmann et al., 2013). For example, in a network model for an

individual diagnosed with a depressive disorder, the depression symptoms constitute the nodes in the network, and the AR and cross-lagged effects between the symptoms constitute the connections in this network (Borsboom & Cramer, 2013; Bringmann et al., 2013). It would be interesting to investigate to what extent measurement error in each variable affects the resulting network.

Finally, while incorporating measurement error into time series models is likely to decrease distortions as a result of ignoring measurement error to the parameter estimates, we emphasize that it is not a cure-all. Even in the models that incorporate measurement errors, the AR parameters may be slightly under- or over-estimated, because measurement error variance and innovation variance are not completely discernible from each other. The more measurement error present in the data, the more difficult it will be to pick up any effects. Therefore, there is still a strong argument for preventing measurement errors in the first place. One option to potentially improve the measurements is to use multiple indicators to measure the relevant construct. However, in a intensive longitudinal data setting, using multiple items for each variable would strongly increase the burden on the participant, who would have to repeatedly fill out all these questions. What remains are classical ways of preventing measurement error: Improving the respective measurement instruments, the circumstances under which participants are measured, and explicitly measuring and modeling potential sources of measurement error. Still, any remaining measurement error that could not be prevented, should be taken into account in the respective model. That is, prevention is better than cure - but a cure is better than ignoring the issue.

Appendix 4.A Heywood Cases

In Table 4.2 we provide the proportions of data sets for which the ML AR(1)+WN and ARMA(1,1) procedure failed, and the proportion of data sets for which σ_ω and σ_ϵ were estimated at the lower bound, or to be negative (a Heywood case). In the main text we present results were the data sets for which the procedure failed (Prop failed in Table 4.2) are excluded for the ML AR(1)+WN and ARMA(1,1) model (not for the remaining models), but data sets with Heywood cases are included. In Figure 4.A.1, 4.A.2 and 4.A.3 we provide results with both the data sets for which the procedure failed, and the data sets with Heywood cases are excluded for the ML AR(1)+WN and ARMA(1,1) model (not for the remaining models). As can be seen from these figures, the results for the frequentist AR(1)+WN and ARMA(1,1) are more similar to the results of the Bayesian procedures. The Bayesian AR(1)+WN model overall outperforms the remaining Bayesian and frequentist models.

Table 4.2: Proportion of data sets for which the state space AR+WN and ARMA models would not initialize or had negative standard errors (Prop failed), had Heywood cases or hit the lower bound in the estimates of σ_ϵ (Prop Heywood or lower bound σ_ϵ) and σ_ω (Prop Heywood or lower bound σ_ω), across different proportions of measurement error, different values for ϕ, and different sample sizes.

σ_ϵ^2:	0	.1	.2	.3	.5	.7	1	2	4	12
AR+WN										
Prop failed	.041	.012	.009	.011	.023	.030	.060	.084	.183	.256
Prop lower bound σ_ω	.544	.456	.361	.370	.324	.288	.296	.358	.399	.355
Prop lower bound σ_ϵ	.002	.003	.006	.005	.020	.024	.038	.065	.075	.087
ARMA										
Prop Failed	.179	.030	.013	.005	.020	.040	.054	.088	.106	.112
Prop Heywood σ_ω	.511	.462	.307	.328	.209	.145	.129	.078	.044	.042
Prop Heywood σ_ϵ	.024	.068	.136	.147	.256	.298	.338	.445	.499	0.537

ϕ:	-0.75	-0.5	-0.25	0	0.25	0.5	0.75
AR+WN							
Prop failed	.002	.024	.045	.122	.066	.011	.004
Prop lower bound σ_ω	.089	.362	.479	.430	.471	.321	.104
Prop lower bound σ_ϵ	0	.002	.035	.115	.051	.003	.001
ARMA							
Prop failed	.002	.008	.062	.104	.042	.006	.001
Prop Heywood σ_ω	.070	.229	.140	.049	.164	.263	.131
Prop Heywood σ_ϵ	.001	.066	.266	.544	.457	.166	.017

N:	100	200	500
AR+WN			
Prop failed	.031	0	0
Prop lower bound σ_ω	.293	.219	.102
Prop lower bound σ_ϵ	.020	0	0
ARMA			
Prop failed	.016	.005	0
Prop Heywood σ_ω	.218	.218	.112
Prop Heywood σ_ϵ	.228	.084	.016

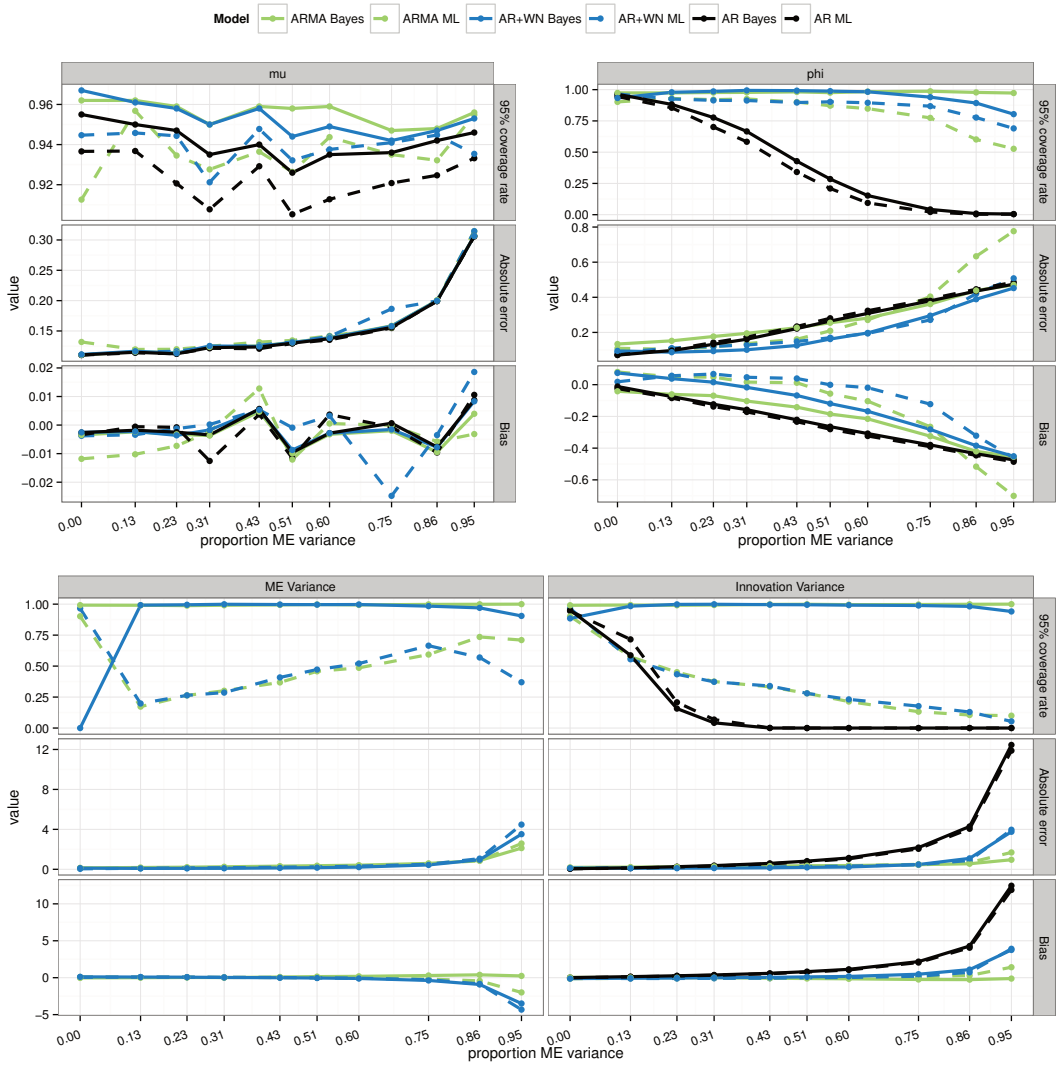


Figure 4.A.1: Coverage rates, bias, and absolute errors for the parameter estimates for the frequentist State-space and Bayesian, AR(1), ARMA(1,1), and AR(1)+WN models across different proportions of measurement error variance to the total variance. Data sets with Heywood cases for the frequentist ARMA(1,1) and AR(1)+WN models are excluded here.

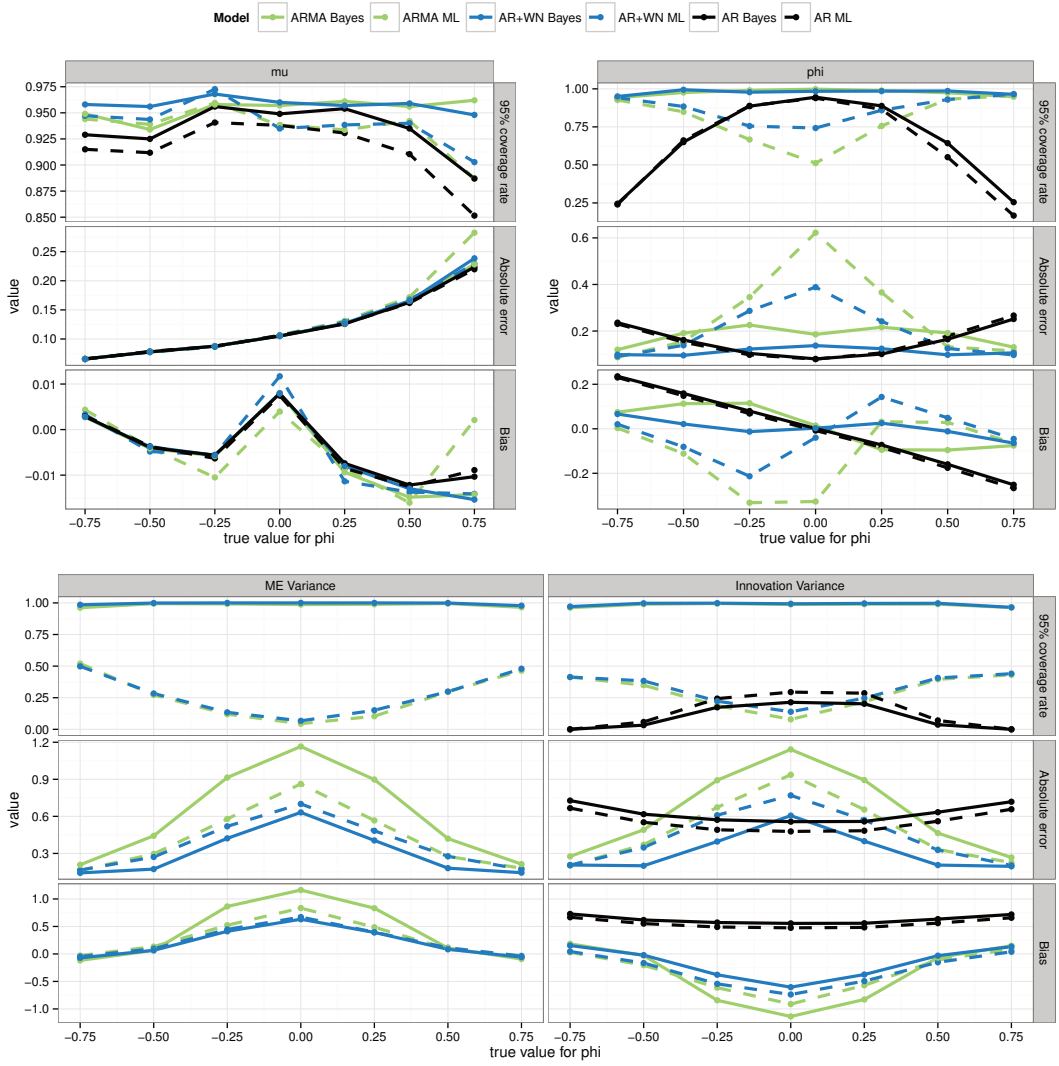


Figure 4.A.2: Coverage rates, bias, and absolute errors of the parameter estimates for the frequentist ML State-space and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across different values for ϕ. Data sets with Heywood cases for the frequentist ARMA(1,1) and AR(1)+WN models are excluded here.

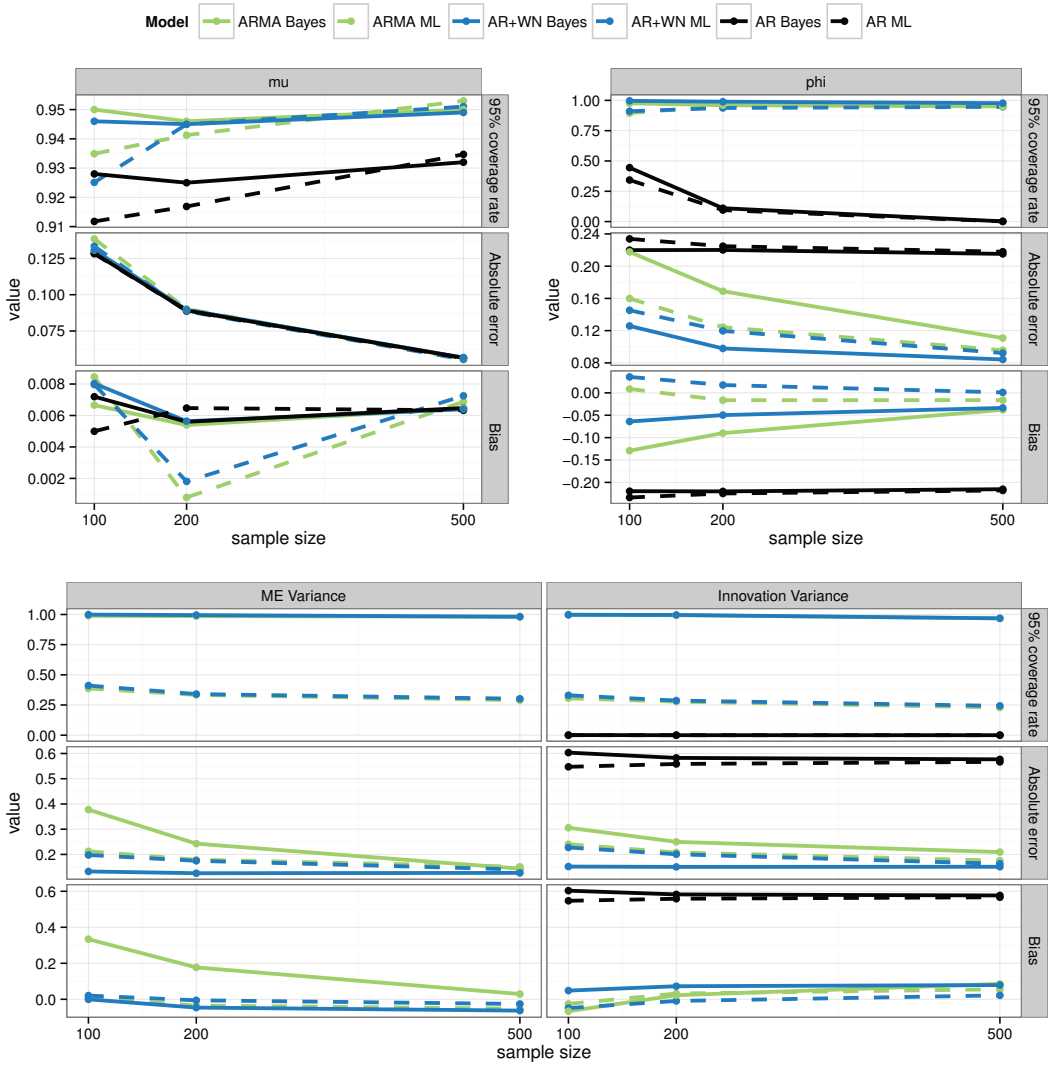


Figure 4.A.3: Coverage rates, bias and absolute errors of the parameter estimates for the frequentist ML and Bayesian AR(1), ARMA(1,1), and AR(1)+WN models across differentsample sizes. Data sets with Heywood cases for the frequentist ARMA(1,1) and AR(1)+WN models are excluded here.

Appendix 4.B Information Criteria Results

While model selection is beyond the scope of this work, we provide some preliminary evaluations here of the model selection performance of the AIC and BIC for the frequentist estimation procedures, and the DIC for the Bayesian estimation procedures. In Figure 4.B the average information criteria, as well as the proportion of the 1000 replications for each of the information criteria that the ARMA model was selected over the AR(1) model, and the AR(1)+WN model was selected over the AR(1) model are presented. Based on the AIC, BIC, and DIC, the AR(1) model is selected in favor of the AR(1)+WN and ARMA model for the large majority of replications, even while the latter are the true models. Although the rate the right model is selected improves as sample size increase, for 500 observations the percentage of data sets for which the AR(1)+WN model is correctly selected is still only 50% for the AIC, 40% for the BIC, and 32% for the DIC. As such, the AIC, BIC and DIC are not appropriate for selecting between an AR(1) model and an AR(1)+WN model. The reason for this may be that the measurement error variance and innovation variance are not completely distinct from each other - this depends on the value of ϕ, the higher $|\phi|$ the better the can be distinguished from each other. This is supported by the results presented in the middle panels of Figure 4.B, which show that as $|\phi|$ increases the proportion of correctly selected models increases.

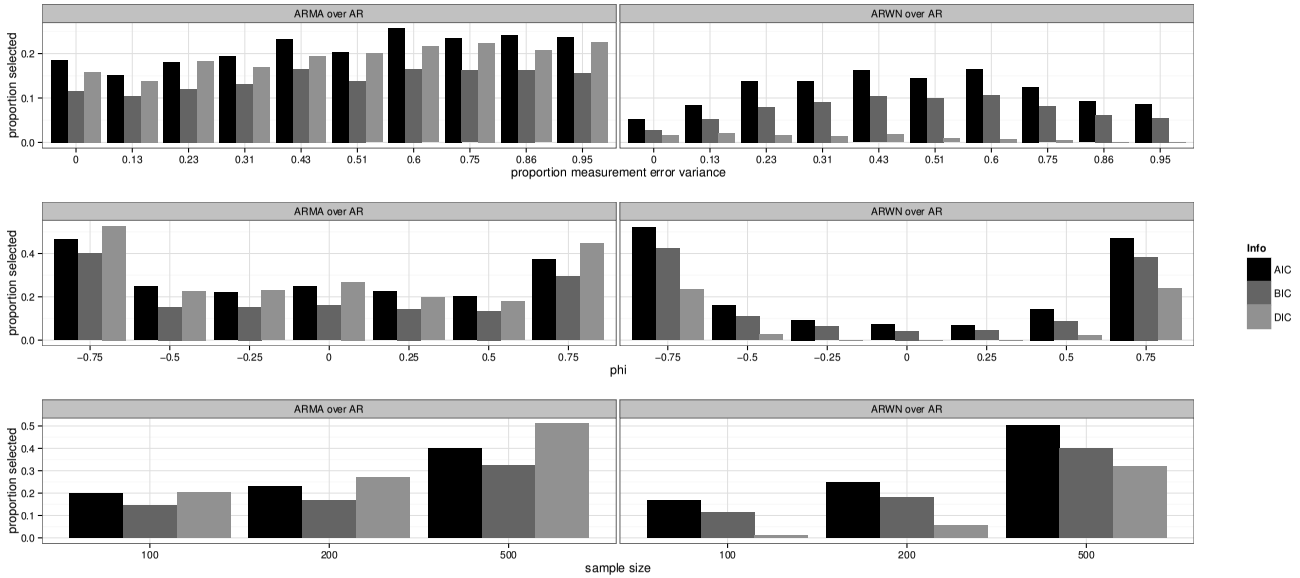


Figure 4.B: Plots of the proportion the ARMA(1,1) (left panels) and the AR(1)+WN (right panels) that are selected over the AR(1) model per simulation condition, based on the AIC and BIC for the frequentist procedures, and the DIC for the Bayesian procedures.

5 Measurement Error and Person-Specific Reliabilities in Multilevel Autoregressive Modeling

by N.K. Schuurman, J.H. Houtveen, and E.L. Hamaker

In psychology there is an increased attention for modeling within-person, dynamical processes, using intensive longitudinal data. Intensive longitudinal data consist of many repeated measures, say twenty or more, typically for multiple individuals. These kinds of data are becoming readily available to psychological researchers due to the development of personal devices such as smart-phones. As a result, psychological scientists are reaching for new modeling techniques that get the most out of these rich, complex data.

A promising approach for analyzing intensive longitudinal data is multilevel vector autoregressive (VAR) modeling. These multilevel models are based on classical VAR models, which are fitted for single subjects (e.g., a person, dyad, country, stock prices and so on) for which many repeated measures were taken. In VAR modeling, multiple variables are regressed on themselves and each other at a previous measurement occasion. This makes it possible to investigate how current values of a variable affect future values of that variable - the autoregressive effect. The autoregressive effect reflects the resistance to change, or inertia, of the psychological process (Suls et al., 1998). Given that the concept of inertia is of interest for many psychological processes (Goodwin, 1971), such as those of affect (Kuppens et al., 2010; Suls et al., 1998), attention (Kirkham et al., 2003), mood disorders (Koval et al., 2012; Kuppens et al., 2010), job performance (Kunze, Boehm, & Bruch, 2013), shopping behavior (van Putten, Zeelenberg, & van Dijk, 2013), and the revision of impressions, opinions and attitudes (Geller & Pitz, 1968; Goodwin, 1971), to be able to estimate the autoregressive effect is an attractive quality of VAR modeling. In addition, VAR modeling makes it possible to investigate potential reciprocal effects between different variables over time, for example: Does stress affect future feelings of depression, do feelings of depression affect future stress, or are both the case?

Author contributions: Schuurman designed the study, performed the analyses, processed and interpreted the results, and wrote the paper. Hamaker proposed the topic for the study (accounting for measurement error, reliability), and provided feedback on the written work.

By extending these models to a multilevel model, it is possible to fit these models for multiple individuals simultaneously, taking into account that people may be similar to some extent, and making it much easier to generalize results to a larger population than for classical n=1 VAR models. At the same time, the multilevel model allows for the model parameters to vary across individuals, so that differences between individuals are taken into account. Currently, both n=1 time series modeling, and multilevel autoregressive modeling are finding more and more applications in psychology, especially in the area of affect regulation and dyadic interactions (Bringmann et al., 2013; Cohn & Tronick, 1989; De Haan-Rietdijk et al., 2014; Kuppens et al., 2010; Lodewyckx et al., 2011; Madhyastha et al., 2011; Moberly & Watkins, 2008; Nezlek & Allen, 2006; Nezlek & Gable, 2001; Rovine & Walls, 2005; Snippe et al., 2015; Suls et al., 1998; van der Krieke et al., 2015; Wang et al., 2012).

In most multilevel autoregressive modeling applications in psychology it is implicitly assumed that the observed variables are free of measurement errors. However, it is unlikely that measurements of psychological constructs will be perfectly reliable, because most of these constructs are not directly observable and measuring them is complex. In line with this notion, many cross-sectional psychological studies take the reliability of measurements into account by measuring a single construct with multiple exchangeable items, and modeling this measurement structure with latent variable models, such as factor- or IRT-models (Ferrer et al., 2012; Lodewyckx et al., 2011; Oravecz & Tuerlinckx, 2011). However, it is relatively complex to fit these models, especially in the multilevel context, and they require a lot of data to fit properly. Furthermore, single-item measures or single variables often play a central role in longitudinal studies (Lucas & Donnellan, 2012); Using multiple items for each construct of interest severely increases the burden on the participants, that may have to fill out these items on a daily or even hourly basis. Moreover, latent variable models may be considered inappropriate theoretically, because the items cannot be considered exchangeable, for instance, when each item is considered to play a unique role within a network of items (cf., Borsboom & Cramer, 2013; Borsboom et al., 2003; Schmittmann et al., 2013). Regardless of these reasons however, ignoring the reliability of measurements is problematic, because it leads to substantially biased regression parameters. For example, it has been shown that for n=1 univariate AR(1) models the regression parameters will be biased towards zero (Schuurman et al., 2015; Staudenmayer & Buonaccorsi, 2005).

Therefore, we introduce a multilevel VAR model in the current paper that takes

measurement errors into account. In this model, we allow the means, regression parameters, variances, and covariances to vary across persons, by making use of Bayesian modeling techniques. As a result, it is possible to estimate the reliability of repeated measures for each individual, which allows us to look at the concept of reliability in a new light. While it has occasionally been acknowledged in psychological studies that the reliability of measurements may be different for each person (e.g., as early as 1968 by Lord and Novick), it is generally not accounted for in psychological studies. By incorporating measurement error into the multilevel VAR model, we can estimate and take into account the reliability for the within-person measurements of each individual, in addition to the reliability of the measurements with regard to between-person differences.

In the remainder, we will first introduce the concept of reliability as defined in classical test theory, and how reliability is usually evaluated in longitudinal research. After that, we introduce the extended multilevel autoregressive model that takes measurement error into account, we discuss the accompanying reliability estimates, and we discuss the consequences of disregarding measurement error in the data. We illustrate the model, person-specific reliabilities estimates, and consequences of disregarding measurement error using an empirical example on the effects of men and women's general positive affect, and their positive affect specifically concerning their romantic relationship. Finally, we end with a discussion in which we consider the concept of reliability and measurement error further in light of our findings.

5.1 Measurement Errors and Reliability

Reliability concerns the consistency of measurements. That is, in the hypothetical situation that we would replicate our measurements of a certain quality of interest while the quality of interest has not changed, perfectly reliable measurements would give the same result for each replication. In contrast, measurements that are unreliable can result in different scores for each replication. The unreliable part of a score is due to *random measurement errors*,[1] while the reliable part is what is consistent across replications, and includes the true value of the actual quality of interest and any consistent errors in the measurements (e.g., consistently measuring a person as 2 pounds heavier than he or she really is). As such, reliable measurements are not necessarily valid, but obtaining reliable measurements is a precondition for obtaining

[1]Throughout the text we will refer to 'random measurement errors' as measurement errors.

valid measurements.

Although we are primarily interested in reliability in the context of the autoregressive modeling of within-person differences, the roots of reliability lie in cross-sectional studies of between-person differences. Therefore, in this section we will start by discussing the definition of reliability from classical test theory, in which reliability was first defined, and reliability estimates in the context of cross-sectional studies. After that, we will discuss how reliability is currently handled within the context of longitudinal (autoregressive modeling) studies of within-person differences.

Reliability in Classical Test Theory

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

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

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

Although the true score and measurement errors in classical test theory are defined on the level of a specific individual, reliability is defined for the measurements of a *specific population of individuals* (cf., Lord & Novick, 1968; Mellenbergh, 1996). The focus lies on the distribution of the observed and true scores across all individuals

in the population. The expected value of the observed scores across the individuals in the population is equal to the expected value of the true scores of these individuals, that is, $\underset{i}{E}[y] = \underset{i}{E}[\theta]$. The variance of the observed scores $V(y)$ is the sum of the variance of the true scores τ^2 and the measurement error variance σ^2_ω, that is, $V(y) = \tau^2 + \sigma^2_\omega$. The reliability $rel(y)$ of the set of measurements is then defined as the proportion of variance in the observed scores that is due to the variance in the true scores, $rel(y) = \tau^2/V(y) = 1 - \sigma^2_\omega/V(y)$. As such, the maximum reliability is equal to 1, indicating that variable y is measured without error in the population, and the minimum reliability is 0, indicating that the measurements consist of only measurement error in the population.

In practice, the true score(s) for any individual are of course unknown, such that in order to be able to take the reliability of our measurements into account, the true scores (or inversely, the measurement errors) have to be estimated from the data. In the following, we fill first discuss briefly how this is done in the context of cross-sectional studies. After that, we will discuss how reliability is currently handled within the context of longitudinal (autoregressive modeling) studies of within-person differences.

Reliability for Cross-sectional Studies

In psychological cross-sectional studies, which focus on interindividual differences, one usually encounters one of two approaches for dealing with measurement errors in the data. In the first approach, the reliability of the measurements is estimated a priori - before doing any other analyses on the measurements. The estimated reliability is then judged to be either sufficient or insufficient for further usage of the measurements to answer a particular research question. The second approach is to account for the reliability of the measurements during the statistical analyses, by explicitly modeling the measurement errors. This second approach is generally preferable, because in this way the results of the analyses are corrected for the measurement errors and an estimate of the reliability of the measurements is obtained. However, the main idea behind both approaches is the same: Find out what part of the observed scores remains constant across replications (i.e., what part is due to the true scores), and what part fluctuates randomly across replications (i.e., what part is due to measurement errors).

In order to achieve this, it is necessary to obtain replicate measurements for the

construct of interest, specifically, replications of the same kind as described in the thought experiment cited by Lord and Novick (1968, p. 29; citing Lazarsfeld, 1959). Obtaining such replications is not an easy feat, and the different reliability measures available are in general based on different ideas on how to obtain these replications (Cronbach, 1947). The most well known methods for estimating the reliability of measurements are parallel-test reliability methods, internal consistency methods, and test-retest reliability methods. *Parallel-test reliability* is based on the construction and administration of two or more 'parallel tests' to each individual, that is, tests that are constructed to be equivalent (cf., Borsboom, 2003; Cronbach, 1947, 1990; Lord & Novick, 1968). Parallel test reliability seems to be mostly used for a priori estimates of reliability, rather than for modeling reliability. *Internal consistency reliability* is used for composite scores, and circumvents the construction of parallel tests by treating the components from which the scores are composed as the replications of the construct under scrutiny (cf., Borsboom, 2003; Cronbach, 1947, 1990; Lord & Novick, 1968). For instance, for a test-score that consist of multiple items, each item may be considered a replicate measurement of the construct, such that the correlations between these items may be used as an indicator of reliability. A well-known estimator of reliability in this category is Cronbach's alpha. It is also the idea behind most measurement models that are used to account for measurement error during analysis, such as item response theory and factor models: In these models, common variation between the items is explained by a latent variable (i.e., the true score variable), and what is uncommon ends up in the residual terms for the items (i.e., the measurement error; Mellenbergh, 1996). Finally, *test-retest reliability* is based on the repeated administration of the same test, that is, each individual fills out the same test multiple times (cf., Borsboom, 2003; Cronbach, 1947; Lord & Novick, 1968). The test-retest reliability is then equal to the squared correlation between the measurements obtained at two occasions.

Reliability for Longitudinal Studies

One way to obtain reliability estimates for studies that focus on intraindividual differences, consists of simply generalizing reliability estimates obtained for inter-individual differences to the within-person context: For certain questionnaires used in longitudinal studies, reliability estimates may be available based on results from cross-sectional studies. It is important to note however that it has been widely established in psychol-

ogy that results based on between-person differences do not automatically generalize to within-person differences, such that reliability estimates from cross-sectional studies cannot be simply generalized to (intensive) longitudinal measurements (Adolf et al., 2014; Borsboom et al., 2003; Hamaker, 2012; Kievit et al., 2011; Molenaar, 2004; Nezlek & Gable, 2001). Furthermore, we may not only wish to obtain an estimate of the reliability of our repeated measures, but also to correct our parameter estimates for this reliability. Therefore, we need to estimate and account for the reliability of our repeated measures within our longitudinal model.

Test-retest reliability may seem a natural way to account for reliability in the context of longitudinal studies, given that in such studies repeated measures are available by definition, and that the variables of interest often have only a single indicator, thus precluding the use of parallel tests or internal consistency methods. Classical test-rest reliability however is only appropriate when the true score remains stable across replications - any within-person variation across time is considered to be a result of measurement error - and therefore is not a valid option for (intensive) longitudinal data (Hertzog & Nesselroade, 1987).

More specifically, for longitudinal studies it is insufficient to only separate the variance of the observed scores into variance due to between-person differences and variance due to within-person differences. The reason for this is that the within-person variance will consist of variation in true scores over time, and variation in measurement errors. As such, it is necessary to establish not only what part of the variance in the longitudinal measurements is due to systematic between-person differences, but also what part of the within-person differences is due to systematic within-person dynamic processes and what part of the within-person differences is due to (within-person) measurement error.

The additional step of separating the measurement error variance from within-person variance due to a systematic dynamic process, has gotten attention in the literature on autoregressive modeling in the context of *panel data*. Panel data consist of a few repeated measures (say between 2 to 5 measurement occasions) for many participants, and are usually analyzed by means of structural equation modeling (SEM) techniques. In certain panel data models, measurement error is separated from systematic within-person differences that are the result of an autoregressive process by using multiple indicators in a factor structure (e.g., Edmondson et al., 2013). The Trait-State-Error (TSE) model suggested by Kenny and Zautra (1995), also accounts for the reliability of single indicator variables, and can be seen as an extension of

the Quasi-Simplex model (Jöreskog, 1970). In the TSE model, systematic between-person differences ('traits'), are separated from systematic within-person differences that are the result of the autoregressive process ('states'), and measurement errors.

A downside of the TSE model is - as is the case for many panel data models - that the model ignores potential individual differences in the dynamic processes: It seems quite unlikely that these dynamic processes are the same for each person (Kenny & Zautra, 1995; Molenaar, 2004). Furthermore, it also seems quite unlikely that the measurement error variances are the same for each person (cf., Schuurman et al., 2015). In fact, Lord and Novick (1968, p. 32) already mentioned this in their discussion of classical test theory, stating

> ...Thus we allow the possibilities that some persons' responses are inherently more consistent than those of others, and that we are able to measure some persons' responses more accurately than others.

An alternative to the TSE model, which does take into account that the processes for individuals may differ from each other, is n=1 (autoregressive) time series modeling, in which models are fitted for each person separately. The advantage of the this approach is that the model can be tailored to each person, so that individual differences in dynamics are taken into account, and that there is no between-person variance to filter out. Although the time series models used in psychological practice generally do not take measurement error into account, it is possible to do so (cf., Schuurman et al., 2015). Downsides of the n=1 approach are however that you need relatively many repeated measures per person to fit these models, and that it is hard to generalize the results for specific individual to a larger population.

Fortunately, it is possible to extend the n=1 models, including those that take measurement error into account, to a multilevel setting (or similarly, extend a multilevel VAR model so it incorporates measurement error, or extend the TSE model to incorporate random effects for all parameters of the within-person process). The multilevel approach allows us to fit the model for all individuals at once, and evaluate to what extent the within-person process differ across persons, making it easier to generalize results to the population of individuals. Furthermore, by allowing the measurement error variance and the systematic dynamic process to be different for each person, we take into account that we can measure some persons' responses more accurately than others as mentioned by Lord and Novick, such that we can obtain

estimates of the reliability of the measurements for each individual. We introduce the extended multilevel VAR(1) model in the following section.

5.2 Accounting for Measurement Errors in the Multilevel VAR(1) Model

In the following, we will introduce the interpretation and specification of the extended multilevel VAR(1) model, which we refer to as the VAR(1)+ White Noise (VAR+WN) model. The multilevel VAR(1)+WN model consists of two levels. At level 1, the within-person process for each individual is specified, and at level 2, the between-person differences in this process across individuals is specified. We will start by discussing the model at level 1. We will focus here on the specification of a bivariate model. This model can easily be extended to models that include more dependent variables and predictors, if the innovations and measurement error (co)variances are fixed to be the same across persons, or if the covariances are disregarded while the variances are allowed to vary across persons. When the (co)variances are allowed to vary across persons, this model is theoretically still easily extended to include more variables, but currently it is not possible, to our knowledge, to fit these models in the available software (this is discussed further at the end of this section).

Level 1 of the VAR(1)+WN Model

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

$$\begin{bmatrix} y_{1ti} \\ y_{2ti} \end{bmatrix} = \begin{bmatrix} \mu_{1i} \\ \mu_{2i} \end{bmatrix} + \begin{bmatrix} \tilde{y}_{1ti} \\ \tilde{y}_{2ti} \end{bmatrix} + \begin{bmatrix} \omega_{1ti} \\ \omega_{2ti} \end{bmatrix} \tag{5.1}$$

$$\begin{bmatrix} \omega_{1ti} \\ \omega_{2ti} \end{bmatrix} \sim MvN \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\omega 11i} & \\ \sigma_{\omega 12i} & \sigma^2_{\omega 22i} \end{bmatrix} \right\}. \tag{5.2}$$

The vector $\boldsymbol{\mu}_i$ contains the person-specific means μ_{1i} and μ_{2i} for each variable, for individual i. These means are stable across the repeated measurements for each individual, and therefore reflect a baseline, or 'trait' part, for each persons' scores.

For example, some persons are on average more extroverted than others. The differences between the trait scores μ_{1i} and μ_{2i} across persons, reflect *systematic, trait-like, between-person differences.*

The vectors $\tilde{\boldsymbol{y}}_{ti}$ and $\boldsymbol{\omega}_{ti}$ together reflect the within-person fluctuations around the person-specific trait scores: While some persons are on average more extroverted than others, a specific person i may display more or less extroverted behavior across different occasions t. The terms ω_{1ti} and ω_{2ti} capture the *measurement errors* for person i at occasion t, and are assumed to be serially uncorrelated, and multivariate normally distributed with means equal to zero, and 2×2 covariance matrix $\boldsymbol{\Sigma}_{\omega i}$. Variables with such a distribution are often called (Gaussian) 'white noise' in the time series literature, hence the model name VAR+WN.

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

$$\begin{bmatrix} \tilde{y}_{1ti} \\ \tilde{y}_{2ti} \end{bmatrix} = \begin{bmatrix} \phi_{11i} & \phi_{12,i} \\ \phi_{21,i} & \phi_{22i} \end{bmatrix} \begin{bmatrix} \tilde{y}_{1t-1i} \\ \tilde{y}_{2t-1i} \end{bmatrix} + \begin{bmatrix} \epsilon_{1ti} \\ \epsilon_{2ti} \end{bmatrix} \tag{5.3}$$

$$\begin{bmatrix} \epsilon_{1ti} \\ \epsilon_{2ti} \end{bmatrix} \sim MvN \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{11i} & \\ \sigma_{12i} & \sigma^2_{22i} \end{bmatrix} \right\}. \tag{5.4}$$

That is, the variables $\tilde{y}_{1ti}$ and $\tilde{y}_{2ti}$ depend on themselves and each other at the previous measurement occasion, such that they constitute a VAR(1) process. The regression coefficients of the VAR(1) process are gathered in 2×2 matrix Φ_i. The relationship between the variables and themselves at the previous measurement occasion for person i is reflected in the autoregressive coefficients ϕ_{11i} for variable $\tilde{y}_{1i}$, and ϕ_{22i} for variable $\tilde{y}_{2i}$. Positive autoregressive coefficients indicate that the score of the current measurement occasion will be similar to that of the previous measurement occasion - the larger the autoregressive coefficient the more similar the scores will be. As such, autoregressive parameters reflect the resistance to change in a process, which is also referred to as inertia (Kuppens et al., 2010; Suls et al., 1998). For instance, when a person's feelings of depression spike because of (say) an unpleasant encounter with an ex, and this person has a positive autoregressive effect for depressive feelings, the heightened feelings of depression will linger above baseline for a long time. On the other hand, an autoregressive coefficient of zero or near zero indicates that the

previous value of a variable does not, or hardly carries over to the next occasion. In other words, if a person with an autoregressive effect of zero for depressive feelings experiences a negative event, the effect of this event is specific to that occasion, and does not carry-over to future occasions.[2]

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

The residuals for the transition equation, ϵ_{1ti} and ϵ_{2ti}, reflect perturbations of the dynamic process for person i at occasion t, and are referred to as innovations. As can be seen from Equation 5.4, we assume here that these innovations are normally distributed with means of zero, and covariance matrix $\mathbf{\Sigma}_\epsilon$.

Innovations versus measurement errors

The innovations ϵ_{1ti} and ϵ_{2ti} and the terms that capture the measurement errors ω_{1ti} and ω_{2ti} are substantially different: The innovations ϵ_{1ti} and ϵ_{2ti} represent any unmeasured effects on the observed variables that are carried over from one measurement occasion to the next through the autoregressive and cross-lagged effects. This is visible from Figure 5.1, where the effect of ϵ_{t-1i} is passed along through $\tilde{y}_{t-1i}$, and to $\tilde{y}_{ti}$ via the autoregressive effect ϕ_i. Because the innovations affect the system across multiple occasions, the innovations are sometimes also referred to as dynamic errors. Measurement errors, on the other hand, are specific to one occasion. Consider, for instance, the classical examples of a measurement error, where someone accidentally checks the wrong answer on a questionnaire, or presses the wrong button during a computer task. The effects of these errors do not carry over to the next measurement occasion, but are specific to that moment. Such occasion-specific effects are not captured by the innovations, but should be modeled separately. By including ω_{1ti}

[2] A negative autoregression coefficient indicates that a relatively high score at the previous measurement occasion is usually followed by a relatively low score at the current measurement occasion, and vice versa. Negative autoregressive effects are relatively rare in psychological research, but can be expected for processes that concern intake, such as smoking, drinking, and eating behaviors(e.g., Rovine & Walls, 2005). For example, a negative autoregressive effect may be expected for the number of calories that are consumed by persons that turn successively to the restriction of food, and binging, during diets.

and ω_{2ti} in the measurement equation, the measurement errors are separated from the autoregressive processes for $\tilde{y}_{ti}$, as can be seen from Figure 5.1, so that they can be distinguished from dynamic errors ϵ_{1ti} and ϵ_{2ti}. In traditional (multilevel) VAR models the innovations are incorporated in the model, whereas the terms ω_{1ti} and ω_{2ti} are not, and as such potential measurement errors are disregarded.

Some additional remarks about the terms ω_{1ti} and ω_{2ti} are in place in the current context: Although these terms will capture measurement errors present in the data, they may also capture other within-person fluctuations that are specific to occasion t. In fact, they will capture *anything* that affects the the variables at one occasion, of which the effect is dissipated before the next measurement occasion. For example, if someone fills out an hourly questionnaire on mood while eating a tasty snack, this may influence that person's mood at that occasion, but this effect on mood may have dissipated before the next measurement occasion an hour later. Hence, this effect would end up in the terms ω_{ti}, even though it does not reflect an actual error of measurement but a true, occasion-specific, fluctuation in mood. Hence, while we refer to the terms ω_{ti} as "measurement errors", they actually represent a mix of occasion-specific fluctuations of the true score and measurement errors. We will return to this issue in the discussion.

A second issue that we need to point out is that the innovations and the occasion-specific fluctuations are distinguishable *only* by merit of the autoregressive effect. Therefore, if the autoregressive effect is equal to zero, the measurement errors and innovations cannot be distinguished from each other, and as a result, the model will no longer be identified (Schuurman et al., 2015).[3]

Level 2 of the VAR(1)+WN Model

At level 2 of the multilevel model the individual differences in the dynamic processes of the individuals are modeled. It seems natural that the means, autoregressive and cross-lagged regressive coefficients differ from person to person, and in most multilevel VAR applications, this is accounted for (e.g., Bringmann et al., 2013; De Haan-Rietdijk et al., 2014; Jongerling, Laurenceau, & Hamaker, 2015; Lodewyckx et al., 2011; Schuurman, Grasman, & Hamaker, 2016). As such, we allow the means, autoregressive, and cross-lagged regressive coefficients to vary across persons. We assume

[3]Note however that Schuurman et al. (2015) found for n=1 models, that even if the true autoregressive effect was zero, a Bayesian AR+WN model still provided reasonable estimates of the model parameters.

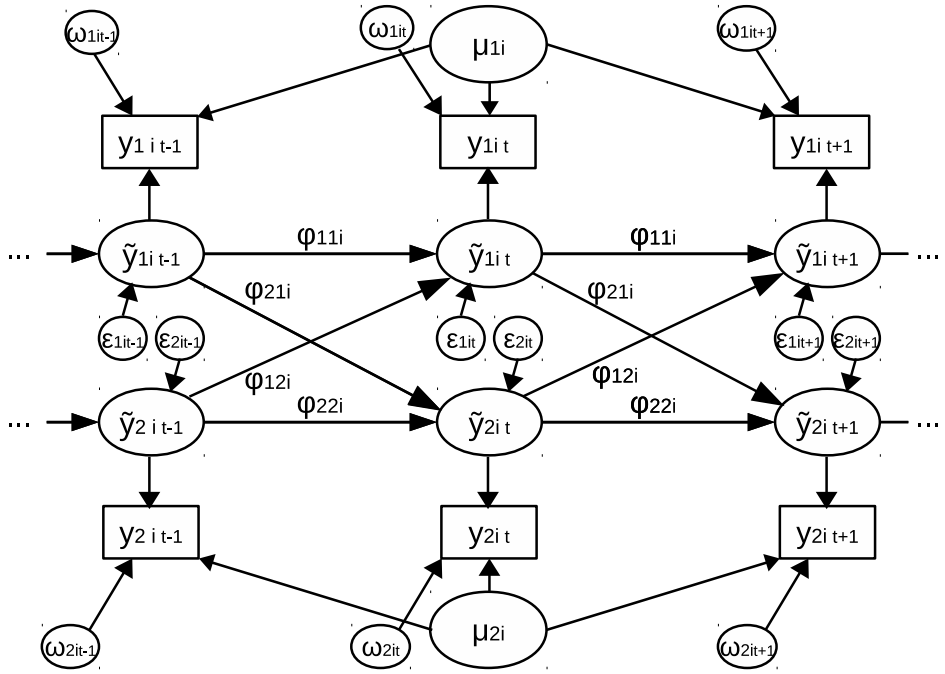


Figure 5.1: A graphical representation of the multilevel VAR+WN model, a VAR model which takes measurement errors into account: They are captured in the terms ω_i.

that each individual's parameter comes from a common population, with a common probability distribution. Characteristics from this distribution, such as its mean and variance, can be used to make inferences about the between-person differences in the within-person dynamics of the individuals. Specifically, we assume that the means μ_i and the regression parameters ϕ_{11i}, ϕ_{22i}, ϕ_{12i}, and ϕ_{21i} are multivariate normally distributed, with means $\gamma_{\mu 1}$, $\gamma_{\mu 2}$, $\gamma_{\phi 11}$, $\gamma_{\phi 12}$, $\gamma_{\phi 21}$, and $\gamma_{\phi 22}$, and a 6 by 6 covariance matrix Ψ. The means γ, also referred to as fixed effects, reflect population averages for the individual means (the trait scores), autoregressive, and cross-lagged effects. The personal deviations from the fixed effects are also referred to as random effects, and their variances are included in Ψ. The covariances in Ψ reflect the associations between the person-specific parameters across persons.

In addition to the trait scores and regression effects, it is also important to con-

sider that the variability of the measurement errors and innovations may differ across individuals. Variance parameters are usually considered fixed across persons in the multilevel literature, and also in multilevel time series applications in psychology (cf., Jongerling et al., 2015). This seems more practically than theoretically motivated, as in many multilevel modeling software including a random variance parameter is not possible. However, it is sensible to assume that the variance parameters to differ across persons in many cases (cf., Lord & Novick, 1968). The innovation variance and the measurement error variance may indicate sensitivity to external events (that either carry over from one measurement occasion to the next, or are measurement occasion specific). Some persons may be more sensitive to external events than others: One person's concentration may be highly impacted by their surroundings, and as such their level of concentration fluctuates a lot due to a variable environment, while another person may remain relatively steady in their concentration regardless of fluctuations in their surroundings. The first individual would then have a relatively large measurement error and/or innovation variance for concentration, indicating that external events have a relatively strong impact, while the second individual would have a relatively small variance, indicating that external events have a relatively weak impact. Further, each individual may experience different external events, that may have more or less impact on each variable that is measured. As such, it seems likely that the variances of the innovations and measurement errors, and the covariances or correlations between the innovations or measurement errors, vary in size across persons. These individual differences should be taken into account in the multilevel model.

In a bivariate model, it is reasonably straightforward to allow for individual differences in the covariances matrices of the innovations and measurement error, by specifying separate hierarchical distributions for the variances, the correlations between the innovations, and the correlations between the measurement errors. Specifically, at the second level of the VAR+WN multilevel model, we assume that the variances and correlations are truncated (univariate) normally distributed, that is, the variances truncated at zero, the correlations truncated at -1 and 1. Each variance and correlation term has a mean, respectively $\gamma_{\sigma^2_{\epsilon 11}}$, $\gamma_{\sigma^2_{\epsilon 22}}$, $\gamma_{\sigma_{\epsilon 12}}$, $\gamma_{\sigma^2_{\omega 11}}$, $\gamma_{\sigma^2_{\omega 22}}$, and $\gamma_{\sigma_{\omega 12}}$, and a variance, respectively $\psi^2_{\sigma^2_{\epsilon 11}}$, $\psi^2_{\sigma^2_{\epsilon 22}}$, $\psi^2_{\rho_{\epsilon 12}}$, $\psi^2_{\sigma^2_{\omega 11}}$, $\psi^2_{\sigma^2_{\omega 22}}$, and $\psi^2_{\rho_{\omega 12}}$. Here, the means indicate the average measurement error variances, average dynamic error variances, the average correlation between the measurement errors, and the average correlation between the dynamic errors across persons. The variances ψ indicate the average

divergence from the means γ across persons.[4]

For covariance matrices that are larger than 2×2, it becomes considerably more complex to allow for random covariance matrices, because it is more difficult to ensure that positive definite matrices are sampled in the Bayesian estimation procedure. An option would be to (for instance) sample the person-specific covariance matrices from an Inverse-Wishart distribution, with (hyper)prior distributions set on the scale matrix of this distribution, or to use specific hierarchical sampling schemes that ensure that sampled covariance matrices for each individual are positive definite (see for instance Tokuda, Goodrich, Van Mechelen, Gelman, & Tuerlinckx, n.d.). These options are however not available in current software to our knowledge.

Fitting the VAR(1)+WN Model

We make use of Bayesian modeling to fit the VAR+WN model. A main reason to opt for a Bayesian approach here, is the flexibility to fit complex multilevel models, allowing us to fit a full bivariate multilevel VAR+WN model including random means, regression parameters, and covariance matrices. Especially the option of modeling random variances and covariances is often unavailable in traditional software. Another important reason to opt for a Bayesian approach is that it is relatively easy to estimate new quantities based on the estimated model parameters, and obtain information about the uncertainty around these estimated quantities in the form of credible intervals and posterior standard deviations (which may be considered the Bayesian equivalents of confidence intervals and standard errors). This quality of the Bayesian approach will prove especially useful for obtaining reliability estimates based on the estimated model parameters. For an introduction to Bayesian statistics, we refer the reader to Hoijtink et al. (2008) and Gelman et al. (2003).

For fitting the model we make use of (free) Bayesian Markov Chain Monte Carlo sampling software WinBUGS, in combination with R and R-package R2winbugs. We provide details on the Bayesian model fitting procedure, including the WinBUGS model code in Appendix 5.A and 5.B.

[4]Note that we do not allow these variance and correlations to correlate with each other, as this would require the specification a high-dimensional truncated multivariate normal distribution. This would likely require a specifically developed procedure, as such distributions are not typically available in currently available software to our knowledge.

5.3 Reliability Estimates Obtained from the Multilevel VAR(1)+WN Model

The model parameters of the VAR(1)+WN model are corrected for the reliability of the data, by modeling the measurement errors using the terms ω_i. This correction is implicit in the model – the VAR(1)+WN model does not directly estimate the reliability of the data. However, it is desirable to obtain an estimate of the reliability of the data, because this gives us an impression of the composition of the data: What proportion of the data is due to between-person variance, and what proportion is due to within-person variation over time, and of the within-person variation, how much is the result of the dynamic process, and how much is due to measurement error? This information can in turn be used to determine the extent of the bias in the estimated parameters if we would not account for the reliability of our measurements, as will be discussed in more detail in Section 5.4. Below, we will first discuss the composition of the variance of the multilevel VAR+WN model. After that, we discuss how to derive an estimate for the between-person reliability, and within-person reliabilities for each person.

The reliability for a specific variable can be calculated as the proportion of true score variance to the total variance for that variable, or equivalently, as 1 minus the proportion of measurement error variance to the total variance. The total variance V(y) for each variable in the VAR(1)+WN model, taken over all participants' repeated measures, can be decomposed into three parts: The between-person variance or trait score variance ψ_μ^2 (i.e., the variance of the person-specific means), the expected value of the person-specific variances for the VAR process $\underset{i}{E}\left[\tau^2\right]$, and the expected value of the person-specific measurement error variances γ_ω. Hence, we have

$$V(y) = \psi_\mu^2 + \underset{i}{E}\left[\tau^2\right] + \gamma_\omega. \tag{5.5}$$

The person-specific variance τ_i^2 in the term $\underset{i}{E}\left[\tau^2\right]$ in this equation, is equal to the diagonal element for variable y of the person-specific covariance matrix $\boldsymbol{T}_i$ for person i. This person-specific covariance matrix is equal to

$$\boldsymbol{T}_i = mat((I - \Phi_i \otimes \Phi_i)^{-1} vec(\Sigma_{\epsilon i})), \tag{5.6}$$

where I is an identity matrix, $\otimes$ indicates the Kronecker product, function $vec()$ transforms a matrix into a column vector, and $mat()$ transforms a vector into a matrix (cf., Kim & Nelson, 1999, p 27).

Based on the variance decomposition in Equation 5.5, we can calculate various reliability estimates for our measurements of variable y. For example, we can determine an overall reliability of our measurements y of both within-person or between-person differences, by calculating $rel(y) = (\psi^2_\mu + \underset{i}{E}\left[\tau^2\right]/V(y)$. However, most studies have a main interest in either within-person differences, or between-person differences. For example, if we want to use the observed score to order people on their baseline level of agreeableness, verbal skills, or stress for a certain time period, priority lies in being able to establish between-person differences well. On the other hand, if the goal is to evaluate or predict fluctuations in people's agreeableness, verbal skills, stress, and so on, being able to reliably establish between-person differences is of a secondary importance compared to being able to establish within-person differences. Therefore, we argue it is more insightful to calculate separate reliabilities for each, than muddling the two in one reliability estimate. In the following, we will discuss a reliability estimate for between-person differences based on Equation 5.5, and after that we will discuss (person-specific) estimates for the reliability of within-person differences. In Appendix 5.A we provide further details on how to estimate these reliabilities as part of the Bayesian model fitting procedure.

Reliability for Systematic Between-Person Differences

The systematic between-person variance in intensive longitudinal data is captured in ψ^2_μ, the variance of the person-specific means. Therefore, we can obtain an estimate of the between-person reliability of variable y using

$$rel_b(y) = \frac{\psi^2_\mu}{V(y)}. \tag{5.7}$$

The reliability for establishing between-person differences may be quite small when a lot of within-person variance is captured in the measurements, either as a result of a systematic autoregressive process captured by $\underset{i}{E}\left[\tau^2\right]$ in $V(y)$, or of measurement error captured by γ_ω in $V(y)$. That is, in order to obtain reliable measurements of differences between persons, for instance for rank ordering people by their ability, the measurements should capture as little within-person fluctuations over time as possible, as is well known from classical test theory.

Reliabilities for Within-Person Fluctuations (per person)

When the main interest of a study is in establishing within-person differences, this is usually because researchers want to be able to infer something about individuals. For example, we may want to know whether the mood of a certain person will benefit more from increasing his/her amount of exercise or from increasing or decreasing his/her amount social interactions. Another example could be that we may want to predict if someone is likely to binge-eat the next day, based on his/her caloric intake the previous day. Because such (autoregressive) psychological processes may differ substantially from individual to individual, we need to know the reliability of the measurements *for each individual*. The VAR+WN takes these individual differences into account, and as a result we can obtain estimates of each individual's reliability, based on the estimates of the person-specific variances for each individual. For the VAR+WN model, the total variance of variable y_i for a specific person i equals

$$v(y_i) = \tau_i^2 + \sigma_{\omega i}^2. \tag{5.8}$$

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

$$rel_w(y_i) = \frac{\tau_i^2}{v(y_i)}. \tag{5.9}$$

From this equation, it can be seen that differences between the reliabilities of different individuals can arise when their autoregressive or cross-lagged associations differ, because then the terms τ_i^2 differ; when the variability of their innovations differs, because then the terms τ_i^2 differ; or when the variability of their measurement errors differ, because then the terms $\sigma_{\omega i}^2$ differ.

Note that these person-specific reliabilities can only be calculated when there is enough data available for each person to fit the multilevel VAR+WN model (or an n=1 model). However, we may also want to know what kind of reliability we can expect for other individuals from the population, but whom we have not observed before. To do this, we can use the information on the reliabilities for each individual we measured. That is, we can estimate the expected value of the reliability across all individuals to obtain an estimate for an individual we have not observed. For example, we can estimate the average person-specific reliability based on our sample of individuals, and we can estimate the variance of the person-specific reliabilities to get an impression

of the range of person-specific reliabilities in the population of individuals under study. Furthermore, if level 2 (person level) predictors are available, such as gender or personality traits, these may also be used to obtain more accurate predictions for the person-specific reliabilities (e.g., we can use these level 2 predictors for the person-specific parameters, and based on the predicted person-specific parameters, determine the associated person-specific reliability).

5.4 Consequences of Disregarding Measurement Error in VAR Modeling

Given that the model parameters and the reliability of each variable can be different for each person, the consequences of disregarding measurement errors in the data can also be different for each person. In the following, we will therefore discuss the effects of disregarding occasion-specific fluctuations for the person-specific parameter estimates of individuals. Note however that because the fixed effects in the multilevel model equal the average person-specific parameters, the effect of disregarding measurement error on the fixed effects depends entirely on the effects of disregarding measurement error for each person-specific parameter. Hence, the bias in the fixed effects can be determined by taking the expectation of the biases for each person-specific parameter.

For a univariate AR(1) model, it is known that disregarding measurement error in the data results in AR parameter estimates that are pulled towards zero, so that $|\phi_i|$ will be underestimated (Schuurman et al., 2015; Staudenmayer & Buonaccorsi, 2005). How much $|\phi_i|$ will be underestimated in such a univariate model depends directly on the person-specific reliability $rel_w(y_i)$ that was defined in Equation 5.9, that is:

$$\hat{\phi}_i = rel_w(y_i)\phi_i \tag{5.10}$$

, where $\hat{\phi}_i$ is the expected estimated AR parameter and ϕ_i is the true AR parameter, such that the bias in $\hat{\phi}_i$ is equal to 1 minus the person-specific reliability.

For a multivariate VAR(1) model, the effects of disregarding measurement error are more complicated. The bias in the estimated matrix of autoregression and cross-lagged effects in $\hat{\Phi}_i$ when measurement errors are disregarded depends on the person-specific reliability matrix $\boldsymbol{rel_w}(\boldsymbol{y}_i)$ (p. 108-109, Buonaccorsi, 2010; Gleser, 1992), which is equal to

$$\boldsymbol{rel_w}(\boldsymbol{y}_i) = \boldsymbol{T}_i \left(\boldsymbol{\Sigma}_{\omega i} + \boldsymbol{T}_i\right)^{-1} \tag{5.11}$$

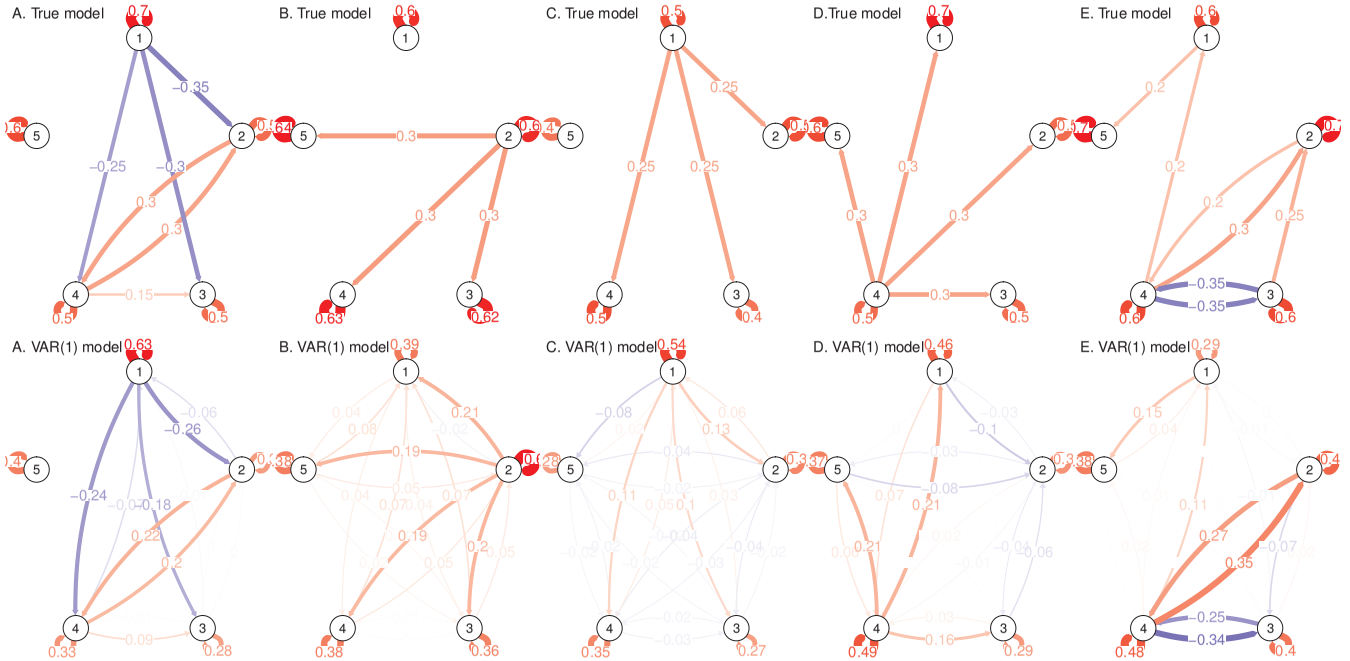


Figure 5.2: Five sets of graphs that provide examples of the potential effects of disregarding measurement error (using a regular VAR(1) model) on the regression coefficients. In these graphs, the circles (nodes) represent the measured variables, and the arrows between the nodes (edges) represent the regression relationships between these variables. Edges that point from one variable to the same variable represent autoregressive regression coefficients, and edges that point from one variable to another variable represent cross-lagged regression coefficients. Red arrows indicate positive regression coefficients, and blue arrows indicate negative regression coefficients. The larger the absolute variable of a regression coefficient, the thicker the arrow. Graphs on the top row (A1,B1,C1,D1,E1) represent the data-generating model, while the accompanying graphs on the bottom row (A2, B2, C2, D2, E2) represent the parameter estimates that would be obtained disregarding measurement error (and disregarding other occasion-specific fluctuations in the variables).

where I is an identity matrix of the same dimension as that of the covariance matrix of the measurement errors $\boldsymbol{\Sigma}_{\omega i}$. Each element in the reliability matrix is a rather complex function of the covariances and variances of the true scores and the measurement errors. For instance, in the bivariate case this results in

$$\boldsymbol{rel_w}(\boldsymbol{y}_i) = \begin{bmatrix} \frac{\tau_{11i}^2\left(\tau_{22i}^2+\sigma_{\omega 22i}^2\right)-\tau_{12i}(\tau_{12i}+\sigma_{\omega 12i})}{\left(\tau_{11i}^2+\sigma_{\omega 11i}^2\right)\left(\tau_{22i}^2+\sigma_{\omega 22i}^2\right)-(\tau_{12i}+\sigma_{\omega 12i})^2} & \frac{\tau_{12i}\left(\tau_{11i}^2+\sigma_{\omega 11i}^2\right)-\tau_{11i}^2(\tau_{12i}+\sigma_{\omega 12i})}{\left(\tau_{11i}^2+\sigma_{\omega 11i}^2\right)\left(\tau_{22i}^2+\sigma_{\omega 22i}^2\right)-(\tau_{12i}+\sigma_{\omega 12i})^2} \\ \frac{\tau_{12i}\left(\tau_{22i}^2+\sigma_{\omega 22i}^2\right)-\tau_{22i}^2(\tau_{12i}+\sigma_{\omega 12i})}{\left(\tau_{11i}^2+\sigma_{\omega 11i}^2\right)\left(\tau_{22i}^2+\sigma_{\omega 22i}^2\right)-(\tau_{12i}+\sigma_{\omega 12i})^2} & \frac{\tau_{22i}^2\left(\tau_{11i}^2+\sigma_{\omega 11i}^2\right)-\tau_{12i}(\tau_{12i}+\sigma_{\omega 12i})}{\left(\tau_{11i}^2+\sigma_{\omega 11i}^2\right)\left(\tau_{22i}^2+\sigma_{\omega 22i}^2\right)-(\tau_{12i}+\sigma_{\omega 12i})^2} \end{bmatrix}. \tag{5.12}$$

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

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

$$\hat{\Phi}_i = \Phi_i \boldsymbol{rel_w}(\boldsymbol{y}_i), \tag{5.13}$$

which results in the following for the bivariate case,

$$\hat{\Phi}_i = \begin{bmatrix} \phi_{11}r_{11} + \phi_{12}r_{21} & \phi_{11}r_{21} + \phi_{12}r_{22} \\ \phi_{21}r_{11} + \phi_{22}r_{21} & \phi_{21}r_{12} + \phi_{22}r_{22} \end{bmatrix}, \tag{5.14}$$

where r_{pq} indicates an element of the reliability matrix. It is important to note from Equations 5.11 to 5.14, that the bias in each regression parameter partly depends on the other regression parameters: For instance, the bias in ϕ_{11} depends on r_{11}, but it also depends on the product of ϕ_{12} and r_{12}, such that the larger ϕ_{12} and r_{12}, the stronger the bias in ϕ_{11} may become. As a result, the more variables that are included in the model, the more complicated and severe the bias can become, because the more variables are included, the more biasing terms will be included in each element of $\hat{\Phi}_i$ (e.g., in a 3×3 model the bias in phi_{11} will not only depend on r_{11} and the product ϕ_{12} and r_{12}, but also on the product of ϕ_{13} and r_{31}). Furthermore, note that even if one variable is measured without error, the autoregressive parameter for that variable (for instance) may still be biased as a result of measurement errors in other variables in the model. One may further observe from Equations 5.11 to 5.14 that the impact of disregarding measurement error on estimates will depend roughly on two aspects: The person-specific reliability for each variable, and the correlations

between the measurement errors of the different variables in the model. More specifically, the lower the reliabilities of the variables, the more severe the bias in the VAR parameters. The stronger the (either positive or negative) correlations between the measurement errors, the more easily spurious effects will arise for the VAR model.

In Figure 5.2 we present five examples of the effects of disregarding measurement error on the estimated regression parameters in a VAR model. Figure 5.2 consists of five graphs each showing a VAR process for five variables. In the graphs, the circles represent the variables, and the arrows between the circles represent the regression relationships between the variables. Arrows that point from one variable to the same variable represent autoregressive regression parameters, and arrows that point from one variable to another variable represent cross-lagged regression parameters. Red arrows indicate positive regression parameters, and blue arrows indicate negative regression parameters. The closer a regression parameter is to zero, the thinner the arrow. The top row of Figure 5.2 shows five network graphs of the true data-generating VAR(1)+WN models, and the bottom row shows the corresponding graphs for the results of a regular VAR(1) model that disregards measurement error. Note that these graphs are based on Equation 5.13 - there was no data simulated - such that sampling error is not an issue in these graphs. In the five examples in Figure 5.2 the reliabilities for the variables range in between .85 to approximately .5 (for the generating values for all the relevant parameters, see Appendix 5.C). These reliabilities are similar to those found in an n=1 empirical example by Schuurman et al. (2015) about the daily mood of eight women, where it was found that depending on the participant, approximately 30% to 50% of the variance was estimated to be due to measurement errors. They are also similar to the results of the empirical example of the current work, which are presented in the following section. These percentages may seem high, but this is not necessarily surprising. We will discuss this issue further in the discussion.

Disregarding measurement error can have various effects on the estimated regression coefficients: The regression coefficients in the VAR model may be underestimated, or they may be overestimated, to the extent that effects may 'disappear' or even switch signs, and spurious effects may arise. Graphs A1 and A2 of Figure 5.2 show an example where the autoregressive and cross-lagged effects are underestimated, similar to what would happen in a univariate AR(1) model: The AR and cross-lagged parameters are pushed towards zero, so that the edges in the VAR-based graph A2 are thinner than those in graph A1 of the true process. The extent of

the underestimation is different for each variable, depending on the reliability of that variable.[5] Graphs B1 and B2 and C1 and C2 illustrate that spurious effects may arise as a result of disregarded measurement error. Graphs B1 and B2 show a strong positive spurious relationship between variable 2 and 1, where the spurious effect is actually the strongest effect in the model. Graphs C1 and C2 show a negative spurious relationship between variable 1 and 5. Furthermore, note that across all sets of graphs, many small spurious cross-lagged effects arise. Graphs D1 and D2 illustrate how an effect may become underestimated to the point that it 'disappears' or even changes signs (i.e., the relationship between variable 4 and 2 changes signs from .3 in graph D1 to -.04 in graph D2). Graphs E1 and E2 illustrate that associations may be overestimated (i.e., for the relationships between variable 4 and 2), and that many of the effects of disregarding measurement error may occur together. Together, these examples show that disregarding measurement errors in the VAR model can seriously distort the estimated regression parameters.

5.5 Empirical Application on Dyadic Affect Data

In this empirical application we focus on affect measurements from a daily diary study including 70 heterosexual couples from a local community (Ferrer et al., 2012; Ferrer & Widaman, 2008). We focus on two measures of positive affect: a) daily relationship positive affect (RelPA), that is, the PA each person experienced specifically about their romantic relationship that day; and b) general positive affect (GenPA), the PA each person experienced generally that day. RelPA was measured with nine 5-point Likert scale items (1 indicated very slightly or not at all, and 5 indicated extremely), for which the participants indicated to what extent they felt the following ways about their relationship that day: "emotionally intimate", "trusted", "committed", "physically intimate", "free", "loved", "happy", "loving", and "socially supported". GenPA was measured with the PANAS (Watson, Clark, & Tellegen, 1988). The man and woman from each couple reported on their RelPA and GenPA at the end of each day, for approximately 90 days. For each person, daily average scores were calculated for both types of PA.

Here, we will investigate how GenPA and RelPA influence each other within a

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

person, separately for the men and women. Specifically, we want to know if a) the temporal evaluation of one's relationship spreads to other areas in daily life; b) general affective tone colors the evaluation of one's relationship; or c) both are the case. The multilevel VAR(1)+WN approach allows us to study this question by establishing associations between GenPA and RelPA, and identifying individual differences in these associations.

We fit two bivariate models (as specified in Section 5.2) for the RelPA and GenPA of the men and women respectively, making use of the Bayesian software WinBUGS (Lunn et al., 2000), in combination with R and the R-package R2winbugs (R Development Core Team, 2012; Sturtz et al., 2005). We provide details on the Bayesian model fitting procedure and the estimation of the reliabilities in Appendix 5.A and 5.B. In the following, we present the results of the VAR(1)+WN model. Below we first discuss the results for the estimated mean scores and lagged effects. To illustrate the effect of disregarding measurement error, we compare these results to those of the regular multilevel VAR model. After that we discuss the estimated innovation (co)variances, measurement error (co)variances, and the estimated reliabilities for the VAR+WN model.

Results for the Dynamics of General and Relationship Positive Affect

The results for daily RelPA and GenPA were quite similar for men and women, as can be seen from the point estimates of the fixed effects and the variances of the random effects, and their 95% credible intervals (CI) presented in respectively Table 5.1 and 5.2. We discuss the results for the various model parameters below.

Autoregressive effects

The estimated means and variances for the autoregressive effects for men and women (see Tables 5.1 and 5.2), indicate that for most individuals the autoregressive effects of General and RelPA are expected to be positive. For RelPA we would expect that the autoregressive coefficients for 95% of the men range from about .2 to .9, and for women from about .3 to .8. For GenPA we find ranges from about .1 to .8 for the men, and from about .2 to .8 for the women. This indicates that for the most if not all of the couples there is inertia present in the regulation of both types of positive affect. In other words, if the general positive affect or the positive affect about the relationship of an individual is perturbed - for instance resulting in a relatively high

Table 5.1: Parameter estimates for the bivariate multilevel VAR(1)+WN and VAR model for men, modeling the relationship between daily relationship and general positive affect.

Men	Fixed Effects		Random Effects	
Parameter	VAR+WN [95% CI]	VAR [95% CI]	VAR+WN [95% CI]	VAR [95% CI]
μ_g	2.9 [2.7, 3.1]	2.9 [2.7, 3.1]	.44 [.31, .64]	.45 [.32, .65]
μ_r	3.5 [3.4, 3.7]	3.5 [3.4, 3.7]	.45 [.32, .65]	.45 [.32, .64]
ϕ_g	.45 [.39, .51]	.29 [.24, .35]	.04 [.03, .06]	.03 [.02, .05]
ϕ_r	.55 [.49, .60]	.37 [.32, .42]	.03 [.02, .06]	.04 [.02, .05]
ϕ_{rg}	.04 [.00, .08]	.03 [-.01, .07]	.02 [.01, .03]	.01 [.01, .02]
ϕ_{gr}	.09 [.04, .14]	.08 [.04, .11]	.02 [.01, .03]	.01 [.01, .02]
$\sigma^2_{\epsilon g}$	.16 [.13, .19]	.27 [.23, .31]	.01 [.00, .02]	.02 [.01, .03]
$\sigma^2_{\epsilon s}$	.12 [.10, .15]	.22 [.18, .25]	.01 [.00, .02]	.02 [.01, .03]
$\rho_{\epsilon gs}$	.37 [.29, .44]	.39 [.34, .43]	.05 [.03, .08]	.03 [.02, .04]
$\sigma^2_{\omega g}$	.13 [.10, .16]		.01 [.00, .02]	
$\sigma^2_{\omega s}$	.10 [.08, .12]		.01 [.00, .01]	
$\rho_{\omega gs}$	.35 [.25, .43]		.06 [.03, .11]	
$rel_w(g)$	.65 [.60, .69]		.03 [.02, .04]	
$rel_w(r)$	.67 [.63, .71]		.03 [.02, .04]	

Note. Parameter estimates for the fixed effects (group means) and the variances of the random effects (group variances) are presented for the person-specific means (μ_g, μ_r), autoregression effects (ϕ_g, ϕ_r), cross-lagged effects (ϕ_{gr}, ϕ_{rg}), the innovation variances ($\sigma^2_{\epsilon g}, \sigma^2_{\epsilon s}$) and correlation ($\rho_{\epsilon gs}$), the measurement error variances ($\sigma^2_{\omega g}, \sigma^2_{\omega s}$) and correlation ($\rho_{\omega gs}$), and person-specific reliabilities ($rel_w(g), rel_w(r)$) of each model.

PA - the positive affect will linger some time above the average level of PA. However, the same holds when someone experiences a relatively low PA: In this case PA will linger for some time below baseline levels due to the autoregressive effect.

Cross-lagged effects

The estimates of the average cross-lagged effect of GenPA at the previous day on current RelPA are .04 (95% CI$[0, .08]$) for men and .03 (95% CI$[-.01, .06]$) for women, which indicates there is no evidence that GenPA colors the RelPA the following day on average for either men or women. For women, there is very little variance around the average (estimated at approximately 0; 95% CI$[0, .01]$), while across men the variance for this cross-lagged effect of GenPA on RelPA is estimated to be a bit larger (at .02; 95% CI$[.01, .03]$). There is evidence that RelPA positively influences GenPA the next day for most of the couples, although the effect may be small: The

Table 5.2: Parameter estimates for the bivariate multilevel VAR(1)+WN and VAR model for women, modeling the relationship between daily relationship and general positive affect.

Women	Fixed Effects		Random Effects	
Parameter	VAR+WN [95% CI]	VAR[95% CI]	VAR+WN [95% CI]	VAR [95% CI]
μ_g	2.8 [2.7, 3.0]	2.8 [2.7, 3.0]	.38 [.27, .54]	.37 [.27, .53]
μ_r	3.6 [3.4, 3.7]	3.6 [3.4, 3.7]	.57 [.41, .82]	.41 [.57, .81]
ϕ_g	.48 [.44, .53]	.30 [.27, .34]	.02 [.01, .03]	.01 [.01, .02]
ϕ_r	.55 [.50, .60]	.38 [.33, .42]	.02 [.01, .03]	.03 [.02, .04]
ϕ_{rg}	.03 [-.01, .06]	.03 [.00, .05]	.00 [.00, .01]	.00 [.00, .01]
ϕ_{gr}	.06 [.02, .11]	.08 [.04, .12]	.01 [.01, .03]	.01 [.01, .02]
$\sigma^2_{\epsilon g}$	.18 [.15, .22]	.31 [.27, .35]	.01 [.01, .02]	.03 [.02, .04]
$\sigma^2_{\epsilon s}$	.14 [.11, .17]	.23 [.19, .27]	.01 [.01, .02]	.02 [.02, .04]
$\rho_{\epsilon gs}$	.33 [.25, .41]	.33 [.28, .37]	.05 [.03, .09]	.03 [.02, .04]
$\sigma^2_{\omega g}$	.15 [.12, .18]		.01 [.00, .03]	
$\sigma^2_{\omega s}$	.10 [.08, .13]		.00 [.00, .02]	
$\rho_{\omega gs}$	.27 [.17, .36]		.07 [.04, .12]	
$rel_w(g)$	.63 [.59, .67]		.02 [.02, .03]	
$rel_w(r)$	.67 [.63, .70]		.02 [.02, .03]	

Note. Parameter estimates for the fixed effects (group means) and the variances of the random effects (group variances) are presented for the person-specific means (μ_g, μ_r), autoregression effects (ϕ_g, ϕ_r), cross-lagged effects (ϕ_{gr}, ϕ_{rg}), the innovation variances ($\sigma^2_{\epsilon g}, \sigma^2_{\epsilon s}$) and correlation ($\rho_{\epsilon gs}$), the measurement error variances ($\sigma^2_{\omega g}, \sigma^2_{\omega s}$) and correlation ($\rho_{\omega gs}$), and person-specific reliabilities ($rel_w(g), rel_w(r)$) of each model.

estimated average cross-lagged effect was .09 (95% CI[.04, .14]) across men and .06 (95% CI[.02, .11]) across women, with variances of respectively .02 95% CI[.01, .03], and .01 (95% CI[.01, .03]). Based on the estimated means and the variances around these means we would expect that the cross-lagged effects of RelPA on GenPA for 95% of the men would lie in between approximately -.2 and .4, and for the women in between -.1 and .3.

VAR+WN model results vs. the VAR model results

Tables 5.1 and 5.2 include results for the VAR model next to the results for the VAR+WN model. When we compare the estimated autoregressive effects of the VAR+WN model to those of the VAR model, we find that the inertia is estimated to be much weaker for the latter, with average effects of about .29 (95% CI[.24, .35]) and .3 (95% CI[.27, .34]) for the GenPA of men and women respectively, and fixed

effects of .37 (95% CI[.32, .42]) and .38 (95% CI[.33, .42]) for their RelPA. In fact, the credible intervals for the fixed autoregressive effects for the VAR and VAR+WN model do not even overlap, such that inferences about the strength of the inertia are markedly different for the two models. The estimated average cross-lagged effects on the other hand are very similar for the two models. That is, in this example, we do not reach different conclusions for the cross-lagged effects in this application. However, note that this may not be the case for other empirical data sets, as discussed in Section 5.4 and illustrated in Figure 5.2.

Trait scores

Based on the estimated means and variances of the traits scores (see Tables 5.1 and 5.2), we find that individuals on average feel moderately positive to quite positive about their relationship, although there is considerable variance in this across the couples. Across the men and women, 95% would be expected to have a trait score between approximately 2, indicating they on average feel slightly positive, and 5, indicating they feel extremely positive about their relationship. The average experienced GenPA is estimated to be a bit lower on average (see Tables 5.1 and 5.2), and based on the estimated means and variances of the trait scores across men and women, we would expect 95% to have a trait score between approximately 1.6 (very slight GenPA) and 4 (much GenPA).

Correlations between the random parameters

When we inspect the estimated correlations between the trait scores and the regression parameters for the VAR+WN model, we find that the traits scores for GenPA and RelPA are positively correlated (.6 95% CI[.41, .73] for men, and .5 95% CI[.3, .67] for women). For the correlations between the remaining random parameters the credible intervals for the correlations are quite wide (intervals including negative and positive values), as a result of a limited number of participants, so that the results for these correlations are not very informative.

Innovations and Measurement Errors

Finally, when we inspect the variances and correlations of the innovations and the measurement errors, we find that the estimated average variances all lie within a range of .1 to .2, and the variances around these average variances are almost all estimated

at approximately .01 (see Tables 5.1 and 5.2). The average correlation between the innovations of general and RelPA is .37 for men (95% CI[.29, .44]), and .33 for women (95% CI[.25, .41]). This indicates that there is a considerable part of the concurrent association between RelPA and GenPA that cannot be explained by the experienced PA at the previous occasion, that seems to be due to unobserved influences of which the effects are passed along across multiple measurement occasions. The average correlation between the measurement errors of general and RelPA is .35 for men (95% CI[.25, .43]), and .27 for women (95% CI[.17, .36]). This indicates that there is also a considerable part of the concurrent association between RelPA and GenPA that seems to be due to unobserved, occasion-specific effects.

Results for the Reliabilities for Relationship and General Positive Affect

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

Between-person reliability

We estimated the *between-person reliability*, that is, the proportion of variance that is due to stable differences across persons, for RelPA and GenPA based on Equation 5.7 (cf., Appendix 5.A, for details on how we calculated the reliabilities as part of the Bayesian modeling procedure). We found that for GenPA about half of the variance in the observed scores across all persons and repeated measures is estimated to be due to systematic differences between persons, while the other half is due to differences within persons (for men rel_b = .53, 95% CI[.31, .74], for women rel_b = .48, 95% CI[.39, .57]). For RelPA, a bit more than half of the total variance is due to systematic differences between persons (for men rel_b= .58, 95% CI[.32, .81], for women rel_b = .63, 95% CI[.54, .72]).

Within-person reliabilities

We can estimate *person-specific reliabilities* for relationship and GenPA using Equation 5.9, based on the estimated person-specific regression parameters, the person-specific covariance matrices of the innovations, and the person-specific covariance matrices for the measurement errors (cf., Appendix 5.A, for details on the estimation of the reliabilities as part of the Bayesian modeling procedure). In the previous subsection we found considerable variation across persons in the regression parameters,

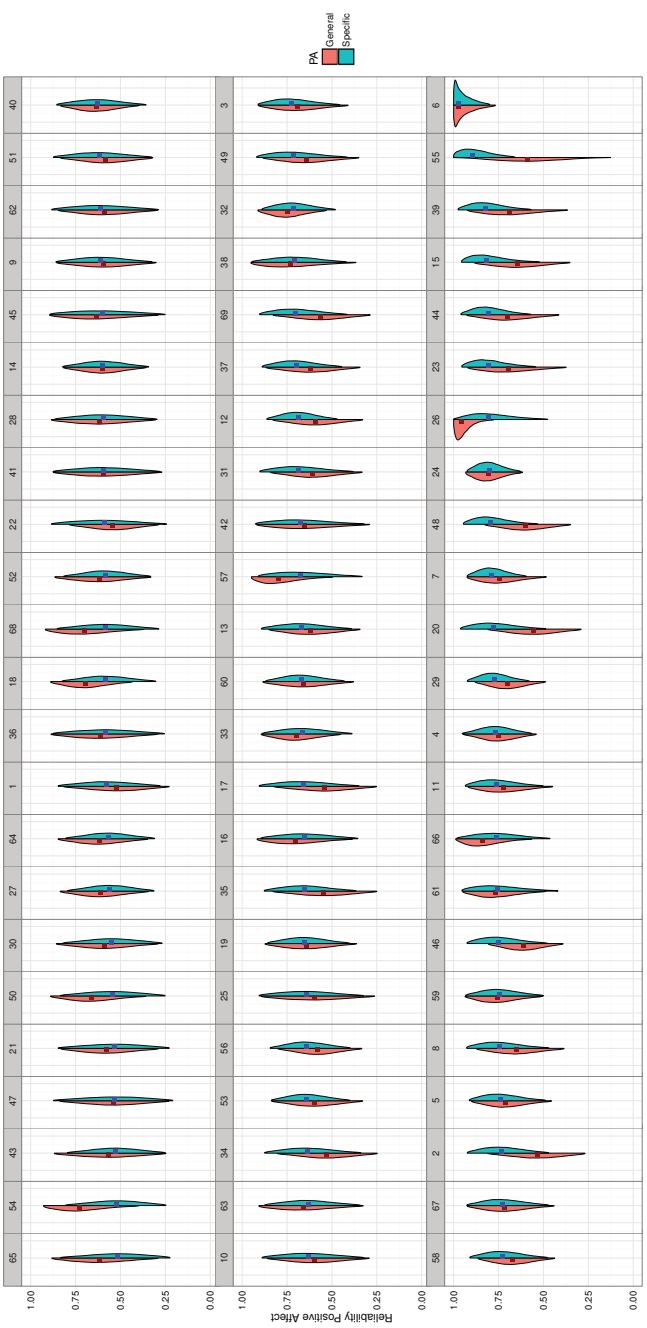


Figure 5.3: Plots of the posterior distributions for the reliability of relationship positive affect (blue), and general positive affect (red) for each man (the couple number for each man is presented above each set of posterior distributions). The dot in each distribution represents the median for that distribution. The tails of the posteriors are trimmed at their 95% credible intervals. The posterior are ordered based on the median reliabilities for relationship positive affect, from low to high (top left to bottom right).

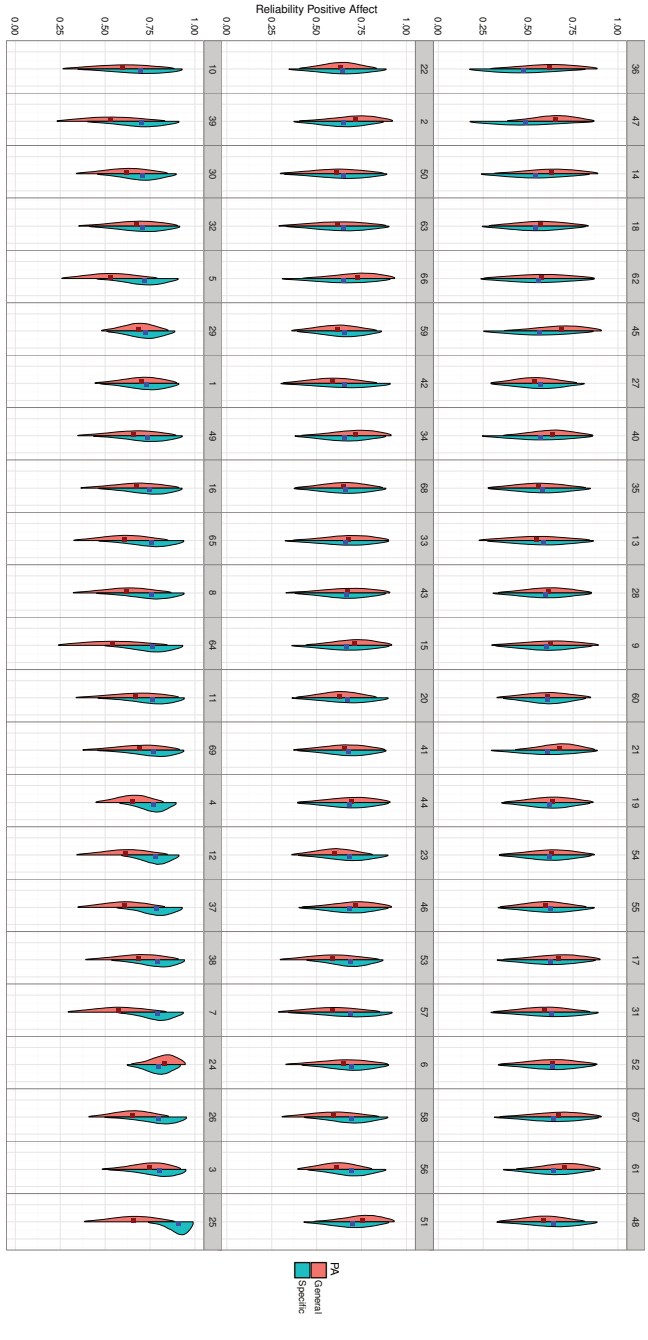


Figure 5.4: Plots of the posterior distributions for the reliability of relationship positive affect (blue), and general positive affect (red) for each woman (the couple number for each woman is presented above each set of posterior distributions). The dot in each distribution represents the median for that distribution. The tails of the posteriors are trimmed at their 95% credible intervals. The posterior are ordered based on the median reliabilities for relationship positive affect, from low to high (top left to bottom right).

but also in the variances - and especially the correlations - for the innovations and measurement errors. As a result, the reliabilities for general and RelPA will also differ from person to person.

Figures 5.3 and 5.4 are for men and women respectively, and contain plots of the posterior distributions for each individual's person-specific reliability. The red (left) distributions are the posterior distributions for GenPA, and the blue (right) distributions are the posterior distributions for RelPA. The tails for each distribution are trimmed at their respective 95% CI, and the dots in each distribution represent the median reliability. In both figures the posterior distributions for the individuals are ordered on the estimated median reliability for RelPA (i.e., with the person with the lowest reliability on the top-right, and the highest on the bottom-left). The lowest and highest estimated reliabilities for RelPA were approximately .52 and .98 for the men, and .47 and .91 for the women. For GenPA the lowest and highest estimated reliability were .52 and .97 for the men, and .53 and .82 for the women. However, it can also be seen from Figures 5.3 and 5.4 that there is a fair amount of uncertainty about the person-specific reliabilities, that is, they have wide CIs. Still, there is evidence that the reliabilities of the PA measurements are likely to be lower than .8 for many individuals (i.e., most of the posterior distribution's mass lies below a reliability of .8). In other words, a considerable part of the variation in the observations for most individuals is due to measurement error. This is also reflected in the average person-specific reliabilities (see also Tables 5.1 and 5.2), of .65 for the GenPA of men and .63 for women (95% CI[.60, .68], 95% CI[.59, .67]), and of .67 for the RelPA of both men and women (95% CI[.63, .71], 95% CI[.63, .70]).

Finally, we investigated whether there is an association across persons between the between the reliabilities for relationship and GenPA. Scatter plots of the point estimates of the person-specific reliability for men and women are shown in Figure 5.5. We found no convincing evidence for a positive relationship for men, with an estimated correlation of .26 (95% CI[−.03, .46]), or for women, with a correlation of .09 (95% CI[−.15, .32]). This may indicate that reliability may not be a personal common trait, as it seems that a high reliability for one type of affect (i.e., GenPA), does not necessarily indicate a high reliability for another type of affect (i.e., RelPA).

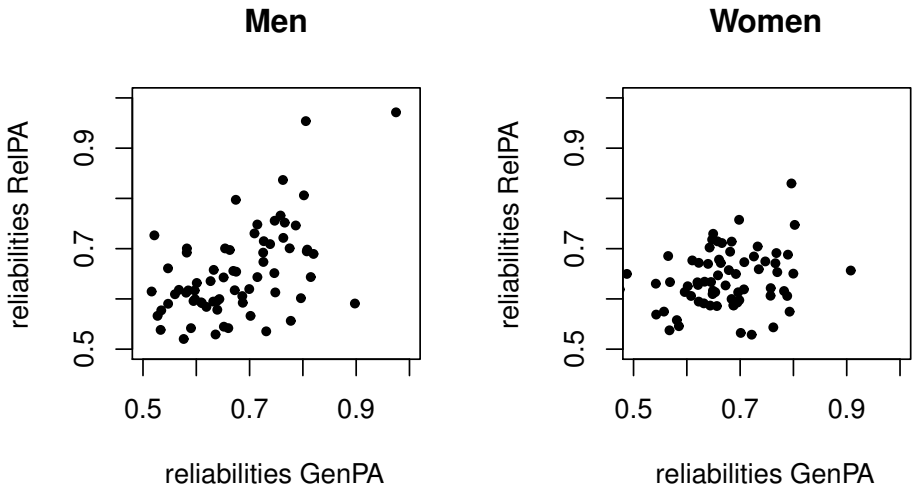


Figure 5.5: Scatter plots of the point estimates (medians of the posterior distributions) of the person-specific reliabilities for general positive affect (GenPA) en relationship positive affect (RelPA), for men and women.

5.6 Discussion

Intensive longitudinal data generally are a mix of between person variation, within-person variation due to dynamic processes, and occasion-specific, random within-person variation, including measurement errors. The VAR+WN model separates these three sources of variation, while also taking into account that there may be differences in the sources of within-person variation across persons. As a result, we can evaluate the reliabilities of a measurement instrument for between-person differences in a specific population of individuals, but also person-specific reliabilities for within-person fluctuations.

As noted in Section 5.1, reliability is defined as specific to a certain population and the measurement instrument (Mellenbergh, 1996). That is, reliability estimates for one population (e.g., men) cannot simply be generalized to another (e.g., women). In the case of reliability for within-person psychological processes, each person may have a unique psychological process, and as such may be considered a single subpopulation

(within a larger population of individuals). For example, in the empirical example we saw that people differ in their levels of inertia for both their general positive affect and their positive affect concerning their relationship, and for some persons their feelings about their relationship influence their general positive affect, while for other this is not the case. Furthermore, one can imagine that some people experience more, or are more easily affected by, external events than other people, or some persons may take more care in filling out self-report measures than others. From this perspective, it seems not very informative to state one reliability estimate for all individuals, disregarding that there may be considerable variation in the psychological processes of these individuals, and hence in their reliabilities. By taking a multilevel VAR+WN modeling approach, we can take this into account, and we can evaluate the average group reliability, and the variability around this average reliability amongst the group of individuals.

It is important to note again however, that the measurement error terms in the multilevel VAR model do not only contain variance that is due to measurement errors - they also contain any 'true' occasion-specific fluctuations in the construct of interest (as was discussed in Section 5.2). As such, reliability estimates based on this measurement-error variance can in practice at most provide an estimate of the *lower bound* of the reliability of the observed scores (as is the case for all other reliability estimates in psychology discussed; see also Borsboom, 2003; Guttman, 1945; Ten Berge & Sočan, 2004). One way to separate true occasion specific fluctuations from measurement errors further, would be to also make use of internal consistency reliability measures in the VAR+WN model, by including multiple indicators to reflect one psychological construct (as is done in dynamic factor modeling); if an occasion-specific fluctuation occurs in all indicators at the same time, this may be indicative of a true occasion specific fluctuation rather than a measurement error (see Edmondson et al., 2013, for an example of this in the context of panel modeling). However, this would require that the indicators function to some extent as parallel tests, *for every single individual*, which may be difficult to achieve. Further, in some cases using multiple indicators for each construct may severely increase the burden on the participants.

Regardless however of whether the 'unreliable part' of the data is a result of real occasion-specific fluctuations in the construct, or of errors in measurement, it it important to take this type of variation in the data into account. Disregarding this type of variation, as is the case in regular (multilevel) VAR models, results in severe bias in the estimated autoregression and cross-lagged parameters. As we have

shown, depending on the reliabilities of the variables and the correlations among the occasion-specific fluctuations, the regression parameters may be under- or over-estimated, may switch signs, and spurious cross-lagged relationships may emerge. The lower the person-specific reliability of a specific variable, the more severe these biasing effects will be. An important question is therefore what kind of reliabilities we may expect in practice. Empirical examples that provide estimates of person-specific reliabilities are still very rare. In an n=1 example by Schuurman et al. (2015) about the daily mood of eight women, their reliabilities ranged from about .5 to .7. In the empirical application of the current work on general and relationship PA we found similar results, with average reliabilities of .65 across persons.

These reliabilities may seem quite low, however, this result is less surprising considering that ϵ_{ti} includes any true occasion-specific fluctuations in the variable of interest in addition to measurement errors. Such fluctuations may occur as a result of a wide range of both internal and external influences, including for example the weather, hormone levels, getting a phone call or email, eating a snack, hearing a certain song, and so on. Whether these effects are dissipated before the next measurement will depend to a large extent on the frequency of measurements: If one measures once every second, these effects most likely carry over to the next measurement occasion through the autoregressive effect and will be part of the innovations, but if one measures once every hour, or once a day, or even once a week, such effects may not carry over to the next measurement occasion and become part of the error term ϵ_{ti} instead. As a result, depending on the construct of interest and the frequency of measurements, the proportion of variance due to occasion-specific fluctuations, including measurement errors, may be considerable. As a consequence, the bias in the estimated VAR parameter may also be considerable when these fluctuations (measurement errors or not) are disregarded.

In this context, it is also important to consider if, and how, this bias can be decreased by preventing measurement errors, next to accounting for the measurement errors in the model. To some extent, the classical measurement errors, such as making a mistake filling out a questionnaire, may be circumvented by designing better measurement instruments or improving the (control over the) measurement conditions. This is considerably more complex for 'true' occasion-specific fluctuations in the psychological construct, because these fluctuations are truly part of the psychological process of interest, and as such should be reflected in valid measurements of that process. It may be possible, however, to decrease the proportion of

such 'true' occasion-specific fluctuations by taking more frequent measurements, so that these fluctuations essentially become innovations rather than measurement errors. However, this is not always practically possible, for instance because it increases the burden on participants, and it seems unlikely that these fluctuations can be completely avoided.

Therefore, it is imperative to always account for measurement errors in the VAR models. Furthermore, it is important to take into account that the proportion of variance that is due to measurement errors in our repeated measurements may differ from person to person. The Bayesian multilevel VAR+WN model presented here provides a relatively flexible environment for accomplishing these two goals.

Appendix 5.A Fitting the Bivariate VAR+WN Model Using Bayesian Software

Here we will provide some details about the Bayesian model fitting procedure we used for the empirical example, after which we discuss how we obtained the reliability estimates. For fitting the Bayesian model, prior distributions need to be specified for the fixed effects and the (co)variances of the random effects for all the random parameters in the model, that is, for the person-specific means, the person-specific cross-lagged and autoregression parameters, the person-specific innovation-covariance matrices, and the person-specific measurement error covariance matrices.

For the empirical example, we specified the following prior distributions for the model parameters: γ_{μ_g}, γ_{μ_r},γ_{ϕ_g}, γ_{ϕ_r}, $\gamma_{\phi_{rg}}$, $\gamma_{\phi_{gr}}$ were all $\sim N(0, 1e-09)$ (specified with a precision, rather than a variance); $\gamma_{\sigma^2_{\epsilon r}}$, $\gamma_{\sigma^2_{\epsilon g}}$, $\gamma_{\sigma^2_{\omega r}}$, $\gamma_{\sigma^2_{\omega g}}$ were all $\sim U(0, 2)$; $\gamma_{\rho_{\epsilon rg}}$, $\gamma_{\rho_{\omega rg}}$ were both $\sim U(-1, 1)$; $\psi^2_{\sigma^2_{\epsilon r}}$, $\psi^2_{\sigma^2_{\epsilon g}}$, $\psi^2_{\sigma^2_{\omega r}}$, $\psi^2_{\sigma^2_{\omega g}}$ were all $\sim U(0, 2)$; $\psi^2_{\rho_{\epsilon rg}}$, $\psi^2_{\rho_{\omega rg}}$ were both $\sim U(0, 1)$. For the precision matrix of the random means and regression effects we specify a Wishart distribution $\sim W(df, S)$ with $df = 6$ (equal to the dimension of the covariance matrix of the random parameters), and Scale matrix S based on prior estimates of the variances of the random parameters. The reason for using a data-based prior distribution here, is that the Wishart prior can be very informative when variances are close to zero, which would be expected for the variances of the autoregressive and cross-lagged effects (because these parameters are restricted in range in a stationary model). In the study by Schuurman, Grasman, and Hamaker (2016), such a data-based prior was found to work the best in this situation, although this will result in slightly too small credible intervals for the estimated variances. In the current study we obtain prior estimates of the variances of the random means and regression parameters by fitting the model as described in Section 5.2, with the prior distributions as discussed previously, except with uniform priors for the variances of the random means and regression parameters (disregarding any covariance between these parameters). The estimated variances for the random means and regression parameters are plugged into the Wishart distribution such that the mean of the Wishart is equal to the estimated precisions of the random parameters, while the covariances between the random parameters are set to zero.

We fit the model using the MCMC procedure provided by WinBUGS, with 3 chains of each 50000 iterations, with 30000 iterations burn-in. We evaluate the convergence of the model by (visually) inspecting the mixing of the three chains, and by

inspecting the Gelman-Rubin statistics for the parameters (Gelman & Rubin, 1992). Based on this we judged 50000 iterations with 30000 iterations burn-in as sufficient for convergence. Throughout the paper, we report the medians of the posterior distributions as point estimates, and the equal-tailed credibility intervals reported by WinBUGS as the credible intervals.

Obtaining Reliability Estimates from the Bayesian VAR+WN Model

Obtaining the reliability estimates while fitting the VAR+WN model using Bayesian techniques is fairly straightforward. For the between-person reliability of a specific variable, we make use of Equation 5.5 to 5.7: First, in each iteration of the MCMC procedure, we calculate the person-specific covariance matrix $\boldsymbol{T_i}$ for each person i using Equation 5.6, based on the estimated person-specific regression parameters and the estimated person-specific innovations covariance matrices. From the relevant diagonal element of each person-specific covariance matrix, we obtain the person-specific variances τ_i^2 we need to estimate $\underset{i}{E}\left[\tau^2\right]$. To obtain an estimate of $\underset{i}{E}\left[\tau^2\right]$ - which we need to calculate the total variance $V(y)$ - we calculate the average person-specific variance across persons in each iteration. We then calculate the total variance using Equation 5.5 in each iteration, based on the estimate we obtained for $\underset{i}{E}\left[\tau^2\right]$, the estimated variance of the person-specific means ψ_μ^2, and the fixed effect for the measurement error variance γ_ω. Finally, in each iteration we calculate the between-person reliability making use of Equation 5.7, based on the estimate we obtained for the total variance $V(y)$ and the the estimated variance of the person-specific means ψ_μ^2. This results in a sample of the between-person reliability for each iteration of the MCMC procedure, which together form a posterior distribution. Based on this posterior distribution we determine a point estimate for the between-person reliability (we use the median of the posterior distribution) and the credible interval for the between-person reliability.

For the person-specific reliabilities for a specific variable we make use of Equation 5.6, 5.8, and 5.9: First, in each iteration of the MCMC procedure, we calculate the person-specific covariance matrix $\boldsymbol{T_i}$ for each person i using Equation 5.6, based on the estimated person-specific regression parameters and the estimated person-specific innovations covariance matrices. From the relevant diagonal element of this covariance matrix, we obtain the variance τ_i^2 for our variable of interest, which we need to estimate the person-specific total variance $v(y_i)$. Based on this estimate and the esti-

mated measurement error variance $\sigma^2_{\omega i}$, we then calculate the total variance $v(y_i)$ for each person using Equation 5.8 in each iteration of the MCMC procedure. Finally, in each iteration we calculate the person-specific reliability making use of Equation 5.9 (or equivalently, we could calculate $1-\sigma^2_{\omega i}/v(y_i)$), based on the estimate we obtained for the total variance $v(y_i)$ and the the estimated person-specific variance τ^2_i. This results in an estimate of the person-specific reliability for each iteration of the MCMC procedure for each person, which together result in a posterior distribution of the person-specific reliability for each person (see Figures 5.3 and 5.4). Based on these posterior distributions we can determine a point estimate for each person's reliability and a credible interval for the between-person reliability. Finally, to obtain an estimate of the average person-specific reliability across persons, and the variance of the person-specific reliabilities across persons, we simply calculate the mean and variance across the person-specific reliabilities in each iteration of the MCMC procedure.

Appendix 5.B WinBUGS Model Code

```
model{
#####model timepoint 2:nt ######
for (i in 1:np)
{
        for (t in (startcom[i]+1):endcom[i])
        {
        ## this reads: for (t in (rownumber of second
        ## observation for participant i):
        ## (rownumber of last observation for participant i)).
        ## These rownumbers startcom and endcom are passed to winbugs as data.

                y[t,1:2]~dmnorm(muy[t,1:2],Epre[i,1:2,1:2])
                ## y (the data) are multivariate normal distributed
                ## with mean muy and measurement error
                ## precision matrix Epre, both vary across participants

                muy[t,1:2] ~ dmnorm(muytilde[t,1:2], Ipre[i,1:2,1:2])
                ## muy has a multivariate normal distribution
                ## with means muytilde and innovation precision matrix Ipre,
                ## both vary across participants

                muytilde[t,1] <-b[i,5] + b[i,1]*zytilde[t-1,1]
                + b[i,2]*zytilde[t-1,2]
                ##muytilde has an autoregressive process with
                ##mean b[i,5], autoregression par b[i,1],
                ## and crossregression par b[i,2]
                muytilde[t,2] <-b[i,6] + b[i,3]*zytilde[t-1,2]
```

```
                + b[i,4]*zytilde[t-1,1]
                ##muytilde has an autoregressive process with
                ##mean b[i,6], autoregression par b[i,3],
                ##and crossregression par b[i,4]

                zytilde[t,1] <- muy[t,1] - b[i,5]
                ##zytilde essentially contains centered muy scores
                zytilde[t,2] <- muy[t,2] - b[i,6]
                ##zytilde contains centered muy scores

        }
}

####timepoint 1#######
##distributions for the first observations of the participants###
for (i in 1:np)
{
        muy[startcom[i],1] <- y[startcom[i],1]  - z[i,1]
        muy[startcom[i],2] <- y[startcom[i],2]  - z[i,2]
        zytilde[startcom[i],1] <- muy[startcom[i],1] - b[i,5]
        zytilde[startcom[i],2] <- muy[startcom[i],2] - b[i,6]

        z[i,1:2]~dmnorm(zm[i,1:2],Epre[i,1:2,1:2])
        zm[i,1]<-0
        zm[i,2]<-0
}

####priors###########

for (i in 1:np)
{
        Epre[i,1:2,1:2] <- inverse(Evar[i,1:2,1:2])
        Ecor[i]~dnorm(Ecormu,Ecorpre) I(-1,1)
        Evar[i,1,1] ~dnorm(Evarmu1,Evarpre1)I(0,)
        Evar[i,2,2] ~dnorm(Evarmu2,Evarpre2)I(0,)
        Evar[i,1,2]<-Ecor[i]*sqrt(Evar[i,1,1])*sqrt(Evar[i,2,2])
        Evar[i,2,1]<-Evar[i,1,2]
}

Ecormu ~dunif(-1,1)
Evarmu1 ~dunif(0,2)
Evarmu2 ~dunif(0,2)
Icormu~dunif(-1,1)
Ivarmu1 ~dunif(0,2)
Ivarmu2 ~dunif(0,2)

Evarpre1 <- 1/Evarvar1
Evarpre2 <- 1/Evarvar2
Evarvar1 ~dunif(0,2)
Evarvar2 ~dunif(0,2)
Ecorpre <- 1/Ecorvar
Ecorvar ~dunif(0,1)
```

```
Ivarpre1 <- 1/Ivarvar1
Ivarpre2 <- 1/Ivarvar2
Ivarvar1 ~dunif(0,2)
Ivarvar2 ~dunif(0,2)
Icorpre <- 1/Icorvar
Icorvar ~dunif(0,1)

for (i in 1:np)
{
        Ipre[i,1:2,1:2] <- inverse(Ivar[i,1:2,1:2])
        Icor[i]~dnorm(Icormu,Icorpre)I(-1,1)
        Ivar[i,1,1] ~dnorm(Ivarmu1,Ivarpre1)I(0,)
        Ivar[i,2,2] ~dnorm(Ivarmu2,Ivarpre2)I(0,)
        Ivar[i,1,2]<-Icor[i]*sqrt(Ivar[i,1,1])*sqrt(Ivar[i,2,2])
        Ivar[i,2,1]<-Ivar[i,1,2]
}

for (j in 1:np)
{
        b[j,1:6]~dmnorm(bmu[1:6],bpre[1:6,1:6])
                ## random means and regression parameters
                ## are mvnormal distributed
}

bmu[1]~dnorm(0,.000000001)
bmu[2]~dnorm(0,.000000001)
bmu[3]~dnorm(0,.000000001)
bmu[4]~dnorm(0,.000000001)
bmu[5]~dnorm(0,.000000001)
bmu[6]~dnorm(0,.000000001)

bpre[1:6,1:6]~dwish(W[,],df)
##W is input provided as data, based on
## on prior estimates of the variances of the
##random means and regression parameters. See Appendix A.
df<- 6

bcov[1:6,1:6] <- inverse(bpre[1:6,1:6])


for (d in 1:6)
{
        for (g in 1:6){
                bcor[d,g] <- bcov[d,g] / ( sqrt(bcov[d,d]) * sqrt(bcov[g,g]) )
        }
}

} # end model
```

Appendix 5.C Parameter Values for Generating Figure 5.2

Table 5.C.1: Parameter values used for generating Graphs A and B in Figure 5.2.

	Graph A	Graph B
$\boldsymbol{\Phi_i}$	$\begin{bmatrix} .7 & -.35 & -.3 & -.25 & 0 \\ 0 & .5 & 0 & .3 & 0 \\ 0 & 0 & .5 & 0 & 0 \\ 0 & .3 & .15 & .5 & 0 \\ 0 & 0 & 0 & 0 & .6 \end{bmatrix}$	$\begin{bmatrix} .6 & .0 & .0 & .0 & .0 \\ .0 & .61 & .3 & .3 & .3 \\ .0 & .0 & .62 & .0 & .0 \\ .0 & .0 & .0 & .63 & .0 \\ .0 & .0 & .0 & .0 & .64 \end{bmatrix}$
$\boldsymbol{\Sigma_{\epsilon i}}$	$\begin{bmatrix} .5 & .0 & .0 & .0 & .0 \\ .0 & .5 & .0 & .0 & .0 \\ .0 & .0 & .5 & .0 & .0 \\ .0 & .0 & .0 & .5 & .0 \\ .0 & .0 & .0 & .0 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & -.15 & -.1 & -.1 & -.1 \\ -.15 & .5 & .2 & .2 & .21 \\ -.1 & .2 & .5 & .2 & .2 \\ -.1 & .2 & .2 & .5 & .15 \\ -.1 & .2 & .2 & .15 & .5 \end{bmatrix}$
$\boldsymbol{\Sigma_{\omega i}}$	$\begin{bmatrix} .5 & 0 & 0 & 0 & 0 \\ 0 & .6 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & 0 \\ 0 & 0 & 0 & .5 & 0 \\ 0 & 0 & 0 & 0 & .4 \end{bmatrix}$	$\begin{bmatrix} .5 & -.28 & -.25 & -.25 & -.25 \\ -.28 & .5 & .15 & .15 & .17 \\ -.25 & .15 & .5 & .15 & .1 \\ -.25 & .15 & .15 & .5 & .1 \\ -.25 & .17 & .1 & .1 & .5 \end{bmatrix}$
$\boldsymbol{rel_w}$	$\begin{bmatrix} .83 & -.07 & -.02 & -.08 & .00 \\ -.08 & .55 & -.01 & .09 & .00 \\ -.02 & -.01 & .57 & .01 & .00 \\ -.08 & .07 & .01 & .60 & .00 \\ .00 & .00 & .00 & .00 & .66 \end{bmatrix}$	$\begin{bmatrix} .65 & -.04 & .06 & .07 & .07 \\ .17 & .87 & .02 & .02 & .01 \\ .12 & .08 & .59 & .01 & .03 \\ .11 & .09 & .01 & .60 & .00 \\ .12 & .08 & .04 & .00 & .60 \end{bmatrix}$
$\hat{\Phi}_i$	$\begin{bmatrix} .63 & -.26 & -.18 & -.24 & .0 \\ -.06 & .30 & .00 & .22 & .0 \\ -.01 & .00 & .28 & .01 & .0 \\ -.07 & .20 & .09 & .33 & .0 \\ .00 & .00 & .00 & .00 & .4 \end{bmatrix}$	$\begin{bmatrix} .39 & -.02 & .04 & .04 & .04 \\ .21 & .60 & .20 & .19 & .19 \\ .07 & .05 & .36 & .01 & .02 \\ .07 & .05 & .00 & .38 & .00 \\ .08 & .05 & .02 & .00 & .38 \end{bmatrix}$

Table 5.C.2: Parameter values used for generating Graphs C, D and E in Figure 5.2.

	Graph C	Graph D	Graph E
$\boldsymbol{\Phi_i}$	$\begin{bmatrix} .5 & .25 & .25 & .25 & .0 \\ .0 & .5 & .0 & .0 & .0 \\ .0 & .0 & .4 & .0 & .0 \\ .0 & .0 & .0 & .5 & .0 \\ .0 & .0 & .0 & .0 & .4 \end{bmatrix}$	$\begin{bmatrix} .7 & 0 & 0 & 0 & 0 \\ 0 & .5 & 0 & 0 & 0 \\ 0 & 0 & .5 & 0 & 0 \\ .3 & .3 & .3 & .5 & .3 \\ 0 & 0 & 0 & 0 & .6 \end{bmatrix}$	$\begin{bmatrix} .6 & 0 & 0 & 0 & .2 \\ 0 & .7 & 0 & .2 & 0 \\ 0 & .25 & .6 & -.35 & 0 \\ .2 & .3 & -.35 & .6 & 0 \\ 0 & 0 & 0 & 0 & .7 \end{bmatrix}$
$\boldsymbol{\Sigma_{\epsilon i}}$	$\begin{bmatrix} .5 & .25 & .25 & .25 & .25 \\ .25 & .5 & .1 & .1 & .1 \\ .25 & .1 & .5 & .05 & .05 \\ .25 & .1 & .05 & .5 & .05 \\ .25 & .1 & .05 & .05 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & .1 & .1 & .1 & .1 \\ .1 & .5 & .1 & .1 & .1 \\ .1 & .1 & .5 & .1 & .1 \\ .1 & 1 & .1 & .5 & .1 \\ .1 & .1 & .1 & .1 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & .0 & .0 & .0 & .0 \\ .0 & .5 & .0 & .0 & .0 \\ .0 & .0 & .5 & .0 & .0 \\ .0 & .0 & .0 & .5 & .0 \\ .0 & .0 & .0 & .0 & .5 \end{bmatrix}$
$\boldsymbol{\Sigma_{\omega i}}$	$\begin{bmatrix} .5 & .25 & .25 & .25 & .25 \\ .25 & .5 & .1 & .1 & .1 \\ .25 & .1 & .5 & .05 & .05 \\ .25 & .1 & .05 & .5 & .05 \\ .25 & .1 & .05 & .05 & .5 \end{bmatrix}$	$\begin{bmatrix} .5 & .25 & .1 & .1 & .1 \\ .25 & .5 & .25 & .25 & .25 \\ .1 & .25 & .5 & .1 & .1 \\ .1 & .25 & .1 & .5 & .05 \\ .1 & .25 & .1 & .05 & .5 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1.4 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0 & 1.8 & 0 \\ 0 & 0 & 0 & 0 & .8 \end{bmatrix}$
$\boldsymbol{rel_w}$	$\begin{bmatrix} .93 & -.04 & -.08 & -.06 & -.08 \\ .11 & .71 & -.07 & -.07 & -.08 \\ .08 & -.05 & .68 & -.05 & -.04 \\ .09 & -.05 & -.06 & .69 & -.05 \\ .06 & -.06 & -.04 & -.04 & .70 \end{bmatrix}$	$\begin{bmatrix} .65 & -.15 & .00 & .10 & .00 \\ -.06 & .62 & -.08 & .04 & -.07 \\ .00 & -.11 & .58 & .06 & .01 \\ .06 & -.16 & .02 & .80 & .08 \\ -.01 & -.13 & .00 & .09 & .61 \end{bmatrix}$	$\begin{bmatrix} .47 & -.01 & .00 & .02 & .07 \\ -.02 & .56 & -.04 & .23 & .00 \\ .00 & -.03 & .54 & -.16 & .00 \\ .03 & .29 & -.24 & .58 & .01 \\ .06 & .00 & .00 & .00 & .54 \end{bmatrix}$
$\tilde{\Phi}_i$	$\begin{bmatrix} .54 & .13 & .10 & .11 & -.08 \\ .06 & .35 & -.04 & -.04 & -.04 \\ .03 & -.02 & .27 & -.02 & -.02 \\ .05 & -.03 & -.03 & .35 & -.02 \\ .02 & -.02 & -.02 & -.02 & .28 \end{bmatrix}$	$\begin{bmatrix} .46 & -.10 & .00 & .07 & .00 \\ -.03 & .31 & -.04 & .02 & -.03 \\ .00 & -.06 & .29 & .03 & .01 \\ .21 & -.01 & .16 & .49 & .21 \\ .00 & -.08 & .00 & .06 & .37 \end{bmatrix}$	$\begin{bmatrix} .29 & -.01 & .54 & .01 & .15 \\ .00 & .45 & -.07 & .27 & .00 \\ -.01 & .02 & .40 & -.25 & .00 \\ .11 & .35 & -.34 & .48 & .02 \\ .04 & .00 & .00 & .00 & .38 \end{bmatrix}$

6 Summary and Discussion

The aim for this dissertation was to further investigate, explicate, and if possible remedy certain difficulties in fitting and interpreting multilevel autoregressive models in the context of psychological science. In Chapter 2 we investigated a specific difficulty in estimating the model parameters using Bayesian techniques. In Chapter 3 we discussed how to standardize the multilevel model such that we can make meaningful comparisons of the strength of the cross-lagged effects in a VAR model. In Chapters 4 and 5 we investigated the consequences of ignoring measurement errors in the data for the estimation of n=1 and multilevel autoregressive model parameters, and we showed how to account for measurement errors. Below, I will give a brief summary of the main findings for each chapter. After that, I will discuss some limitations of the multilevel autoregressive modeling approach, as well as some directions for future research.

Chapter 2

In Chapter 2 we discussed a difficulty in specifying an Inverse-Wishart prior distribution for the covariance matrix of the random parameters, for fitting the multilevel autoregressive model in a Bayesian framework. Specifically, the issue is that the Inverse–Wishart distribution tends to be informative when variances are close to zero. This is problematic for the multilevel autoregressive model, because the random autoregressive parameters in a stationary model are restricted in range, and as a result will have a small variance across persons. We therefore compared the performance of three Inverse-Wishart prior specifications suggested in the literature, when one or more variances for the random effects in the multilevel autoregressive model are small, by means of a simulation study. Our results indicated that a data-based prior specification — a prior specification that uses plug-in estimates of the variances — performs the best out of the three specification we compared, even though by using a data-based prior specification the certainty we had about our estimates was overestimated. We recommended that for any multilevel model, especially autoregressive models, a sensitivity analysis is performed for the prior specifications for covariance matrices of random parameters. We illustrated such an analysis in Chapter 2 for an

empirical application on repeated measures data on worrying and positive affect, in which we find that for persons who worry a lot, worrying seems to be detrimental to their positive affect, while for persons who worry relatively little, it seems to be beneficial to their positive affect.

Chapter 3

Multivariate autoregressive (VAR) models can be used to investigate Granger-causal cross-lagged associations between variables. The aim of the study presented in Chapter 3 was two-fold: Firstly, to illustrate the added value of a multilevel multivariate autoregressive modeling approach for investigating cross-lagged associations over more traditional techniques. Secondly, to discuss how to directly compare the strength of the cross-lagged effect in the multilevel autoregressive model. In order to directly compare the strength of the cross-lagged effects by comparing the size of the cross-lagged regression coefficients, these regression coefficients first need to be standardized. However, in the multilevel model subject-based statistics or group-based statistics can be used to standardize the coefficients, and each method may result in different conclusions. In Chapter 3 we argued that in order to make a meaningful comparison of the strength of the cross-lagged associations, the coefficients should be standardized within persons. The chapter contains an empirical application on experienced competence and exhaustion in persons diagnosed with burnout, in which we illustrated the multilevel VAR model and the standardization of its coefficients, and that disregarding individual differences in dynamics can prove misleading.

Chapter 4

It is safe to assume that psychological data contain measurement errors. However, this is not taken into account in the vast majority of autoregressive modeling applications in psychology. This is problematic, because failing to account for measurement errors leads to biased estimates of the autoregressive parameters. In Chapter 4 we discussed two n=1 autoregressive models that account for measurement errors: an autoregressive model that includes a white noise term (AR+WN model), and the autoregressive moving average (ARMA) model. We compared the performance of these models, and the regular autoregressive model that does not account for measurement errors in a simulation study. Furthermore, we investigated whether these models perform better using a frequentist approach, or a Bayesian approach. Our results

indicated that overall, the AR+WN model recovers the model parameters the best, and that psychological research would benefit from a Bayesian approach in fitting this model, especially for smaller sample sizes. We illustrated the effect of disregarding measurement error in an AR(1) model with an empirical application on mood data in women. In this example we found that, depending on the participant, approximately 30-50% of the total variance was due to measurement error and other measurement occasion-specific fluctuations. Disregarding this type of variance in the data resulted in a substantial underestimation of the autoregressive parameters.

Chapter 5

In Chapter 5 we presented a multilevel VAR+WN model that accounts for measurement errors — or equivalently, for the reliability of our measurements — in a multilevel, multivariate context. The VAR+WN model presented in this chapter is an extension of the model presented in Chapter 4. In the model the means, regression parameters, innovations variances and covariances, and measurement error variances and covariances are allowed the vary across persons. As a result, the reliabilities of the measurements also vary across persons. Next to accounting for measurements errors within the model, we showed how the multilevel VAR model can be used to obtain estimates for the reliability of the within-person measurements of each individual, in addition to the reliability of the measurements with regard to between-person differences.

Because the reliabilities for a specific variable may differ from person to person, the consequences of failing to account for measurement errors for the VAR model may also differ from person to person. In Chapter 5 we discussed these, potentially severe, consequences: The cross-lagged effects may be underestimated or overestimated, true cross-lagged effects may 'disappear', and spurious cross-lagged may arise in the VAR model estimates. In the empirical example on general and relationship positive affect in couples, we found average person-specific reliabilities across persons of approximately .65 for both general and relationship positive affect. That is, like in Chapter 4, we found that the proportion of variance in the data that is due to fluctuations specific to the measurement occasion is quite large, which exemplifies the need of accounting for this type of variance in our autoregressive models.

Limitations and Future Directions

Although the collection and dynamic modeling of intensive longitudinal data is steadily gaining in popularity in psychology, the methodology for both is still very much in the development phase. In this dissertation important steps forward are made in the modeling of intensive longitudinal data with (multilevel) autoregressive models. However, many challenges and limitations that have not been the focus of this dissertation still remain, and are important to consider. In the following, I will briefly discuss some of these challenges, as well as some suggestions for future research.

One fairly obvious limitation that the reader has encountered throughout this dissertation is the stationarity assumption of the autoregressive model: The mean and variance of the psychological process should remain the same across time for each person. This means that, as is, the autoregressive model is not suited for modeling processes that involve some kind of trend, as would be expected to be the case for many developmental processes. It is also not suited for processes in which the autoregression parameters or residual (co)variances change across time, either continuously or abruptly. However, it is possible to extend the autoregressive model to incorporate trends (Hamilton, 1994, , p. 454), abrupt changes in the dynamics (e.g., De Haan-Rietdijk et al., 2014; Kim & Nelson, 1999), or continuous changes in the dynamics (Harvey, 1989, p. 341). Challenges for future research will be to extend such models to a multilevel context, and to figure out when to use which technique in practice (e.g., how to distinguish and choose between a positive autoregressive process and a non-stationary process that includes a trend, for relatively short time series).

A perhaps less obvious, but important, limitation of the autoregressive model is that it requires that the time intervals between the measurement occasions are equal for all occasions, that is, measurements should be taken each week at the same time, or every hour, every minute and so on. Obtaining such measurements is not always possible, and sometimes it is even purposely avoided. For instance, in Experience Sampling Method (ESM) studies, participants are randomly prompted to report what they are doing or how they feeling at the time. This is done randomly so that the feelings or behavior of the participants will not change as a result of the participants' anticipation of the next prompt. Unequally spaced observations are problematic for VAR models, because in these models time is treated as if it is discrete. As a result, the regression parameters indicate the effect of one variable on another across a specific time interval (e.g., across weeks, or hours, or minutes, seconds, and so on). Thus,

if the observations are not equally spaced across time, the estimate of the effect for a specific interval will not be correct. An ad hoc way to deal with unequally spaced observations to some extent that the reader has encountered in Chapter 2 and 3, is to add missing values between the observations to make the time intervals between measurements (more) equal. A more sophisticated solution to this issue would be to treat time as continuous, rather than discrete.[1] However, the current multilevel extensions of continuous time models still have strong limitations: either the random cross-lagged effects are assumed to be equal within a person (Oravecz & Tuerlinckx, 2011), or the lagged effects are assumed to be the same across persons (Voelkle et al., 2012). Many developments in this area are expected however in the near future.

Although the multilevel autoregressive model presented throughout this dissertation provides a flexible framework for modeling individual differences in psychological processes, the multilevel extension brings in some additional assumptions. For example, it rests on the assumption that the means and regression parameters are multivariate normally distributed. This assumption may not be tenable: the random parameters may have completely different distributions, either known or unknown. Future directions here could be to evaluate the robustness of the model against violations of the model assumptions, to use non-parametric multilevel techniques to fit the model, or to use a-posteriori techniques to group participants (see for example, Gates & Molenaar, 2012).

Finally, many limitations and challenges in fitting and extending multilevel autoregressive models and dynamic models in general, lie in the available technology for collecting (enough) intensive longitudinal data, the capabilities of software, and computational power. For example, all of these may need to be improved to be able to successfully fit large multivariate multilevel autoregressive models. Fortunately, the technology for collecting intensive longitudinal data (e.g., smartphones and smartwatches), as well as the available software for fitting dynamic models are rapidly being developed. Therefore, it seems only a matter of time until the dynamic modeling of intensive longitudinal data will become one of the main focal areas in psychological science.

[1] It should be noted that theoretically, by adding an infinite amount of missing data between all observations, such that the time intervals become infinitesimally small, the autoregressive effects of the discrete model will approach those for the continuous time model.

References

Adolf, J., Schuurman, N. K., Borkenau, P., Borsboom, D., & Dolan, C. V. (2014). Measurement invariance within and between subjects: A distinct problem in testing the equivalence of intra- and inter-individual model structures. *Frontiers in Psychology*.

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review*, *30*(2), 217–237. doi: http://dx.doi.org/10.1016/j.cpr.2009.11.004

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 0.999999-0)

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

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *J. R. Stat. Soc.*, *91*, 109–122.

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

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

Borsboom, D. (2015). What is causal about individual differences?: A comment on weinberger. *Theory & Psychology*, 0959354315587784.

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., & Dolan, C. V. (2006). Why g is not an adaptation: a comment on kanazawa (2004). *Psychological Review*, *113*(2), 433 - 437.

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. doi: 10.1371/jour-

nal.pone.0060188

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *8*, 434–455.

Browne, W. J., & Draper, D. (2006). A comparison of bayesian and likelihood–based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473–514.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

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

Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, *46*, 167–174.

Cattell, R. B. (1967). The structuring of change by p-technique and incremental r-technique. In C. W. Harris (Ed.), *Problems in measuring change.* Madison, WI: The University of Wisconsin Press.

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

Chanda, K. C. (1996). Asymptotic properties of estimators for autoregressive models with errors in variables. *The Annals of Statistics*, *24*(1), 423–430.

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

Chong, T. T., Liew, V., Zhang, Y., & Wong, C. (2006). Estimation of the autoregressive order in the presence of measurement errors. *Economics Bulletin*, *3*, 1–10.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/-correlation analysis for the behavioral sciences (3d ed.).* New Jersey: Lawrence Erlbaum Associates.

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

Costa, M., & Alpuim, T. (2010). Parameter estimation of state space models for univariate observations. *Journal of Statistical Planning and Inference*, *140*(7),

1889–1902.

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

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

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

Dedecker, J., Samson, A., & Taupin, M. (2011). Estimation in autoregressive model with measurement error. *ESAIM: Probability and Statistics*.

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

Deistler, M. (1986). Linear dynamic errors-in-variables models. *Journal of Applied Probability*, 23–39.

de Jonge, J., Dormann, C., Janssen, P. P. M., Dollard, M. F., Landeweerd, J. A., & Nijhuis, F. J. N. (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. H., Taris, T. W., Kompier, M. A. J., Houtman, I. L. D., & 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 & amp; Organisations*, *18*, 149–166.

Dudley, R. M., & Haughton, D. (1997). Information criteria for multiple data sets and restricted parameters. *Statistica Sinica*, *7*(2), 265–284.

Dunson, D. B. (2001). Commentary: practical advantages of bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*(12), 1222–1226.

Edmondson, D., Shaffer, J., Chaplin, W., Burg, M., Stone, A., & Schwartz, J. (2013). Trait anxiety and trait anger measured by ecological momentary assessment and their correspondence with traditional trait questionnaires. *Journal of Research in Personality*, *47*(6), 843 - 852. doi: http://dx.doi.org/10.1016/j.jrp.2013.08.005

Ehring, T., & Watkins, E. R. (2008). Repetitive negative thinking as a transdiagnostic process. *International Journal of Cognitive Therapy*, *1*(3), 192–205. doi: http://dx.doi.org/10.1680/ijct.2008.1.3.192

Fahrmeir, L., & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models*. New York: Springer-Verlag.

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

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

Gates, K. M., & Molenaar, P. C. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310 - 319. doi: http://dx.doi.org/10.1016/j.neuroimage.2012.06.026

Geller, E. S., & Pitz, G. F. (1968). Confidence and decision speed in the revision of opinion. *Organizational Behavior and Human Performance*, *3*, 190–201.

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., & Hill, L. S. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

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

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: a randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*, 618–628. doi: http://dx.doi.org/10.1037/a0024595

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, *108*, 33–56.

Gill, J. (2008). *Bayesian methods: A social and behavioral sciences approach (2nd ed.)*. Boca Raton, FL: Chapman & Hall/CRC.

Gleser, L. J. (1992). The importance of assessing measurement reliability in multivariate regression. *Journal of the American Statistical Association*, *87*(419), 696-707.

Goodwin, W. (1971). Resistance to change. *American Behavioral Scientist*, *14*, 745–766.

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

Granger, C. W. J., & Morris, M. J. (1976). Time series modelling and interpretation. *Journal of the Royal Statistical Society. Series A (General)*, *139*, 246–257.

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 userâĂŹs guide. version 2.5 for windows*. Molndal, Sweden: MultivariateWare.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255 - 282.

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

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State-of-the-art and future challenges. *Emotion Review*, *7*(4), 316–322. doi: http://dx.doi.org/10.1177/1754073915590619

Hamaker, E. L., Kuijper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. doi: http://dx.doi.org/10.1037/a0038889

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. (2007). The integrated traitâĂŞstate model. *Journal of Research in Personality*, *41*(2), 295 - 315. doi: http://dx.doi.org/10.1016/j.jrp.2006.04.003

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

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

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

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development*, *58*(1), 93-109.

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

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*, 439–452.

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

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

Jöreskog, K. (1970). Estimation and testing of simplex models. *ETS Research Bulletin Series*, *2*, i - 45.

Kass, R. E., & Natarajan, R. (2006). A default conjugate prior for variance components in generalized linear mixed models (comment on an article by browne and draper). *Bayesian Analysis*, *1*, 535–542.

Kass, R. E., & Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *JASA*, *84*, 717–726.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of consulting and clinical psychology*, *63*(1), 52. doi: http://dx.doi.org/10.1037/0022-006X.63.1.52

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

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

Kirkham, N. Z., Cruess, L., & Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Developmental Science*, *5*, 449–476.

Klugkist, I., & Hoijtink, H. (2007). The bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379.

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition and Emotion*, *26*(8), 1412-1427. doi: 10.1080/02699931.2012.667392

Kuiper, R., Hoijtink, H., & Silvapulle, M. (2012). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of statistical planning and inference*, *142*(8), 2454–2463.

Kunze, F., Boehm, S., & Bruch, H. (2013). Age, resistance to change, and job performance. *Journal of Managerial Psychology*, *28*(7/8), 741-760. doi: 10.1108/JMP-06-2013-0194

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

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

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: comment on trafimow (2003). *Psychological Review*, *112*(3), 662–668.

Lindley, D. V. (1969). Discussion of *Compound decisions and Empirical Bayes*, j.b. copas. *J. R. Stat. Soc., Series B*, *31*, 397–425.

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

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

Lucas, R., & Donnellan, M. (2012). Estimating the reliability of single-item life satisfaction measures: Results from four national panel studies. *Social Indicators Research*, *105*(3), 323–331. doi: 10.1007/s11205-011-9783-z

Luethi, D., Erb, P., & Otziger, S. (2010). Fkf: Fast kalman filter [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=FKF` (R

package version 0.1.1)

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. J. (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. L., & Gottman, J. (2011). Investigating spousal influence using moment-to-moment affect data from marital conflict. *J. Fam. Psychol.*, *25*, 292–300.

Maslach, C., & Jackson, S. (1981). The measurement of experienced burnout. *J. Organiz. Behav.*, *2*, 99–113.

Maslach, C., Jackson, S., & Leiter, M. (1996). *Mbi: The maslach burnout inventory: Manual*. Palo Alto: Consulting Psychologists Press.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response theory models. *Psychological Methods*, *1*(3), 293 - 299.

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self–focus and negative affect: an experience sampling study. *J. Abnorm. Psychol.*, *117*, 314–323.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202.

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

Muthén, L. K. (2008). *Standardized solutions.* Retrieved 2015-17-08, from `http://www.statmodel.com/discussion/messages/12/542.html?1380892370`

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

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Pers Soc Psychol Bull*, *27*, 771–785.

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

Nezlek, J. B., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day–to–day psychological adjustment. *Pers. Soc. Psychol. Bull.*, *27*, 1692–1704.

Nolen-Hoeksema, S., Wisco, B., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, *3*(5), 400–424. doi: 10.1111/j.1745-6924.2008.00088.x

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *J. R. Stat. Soc., Series B*, *57*, 99–138.

Oravecz, Z., & Tuerlinckx, F. (2011). The linear mixed model and the hierarchical ornstein–uhlenbeck model: Some equivalences and differences. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 134–160. doi: http://dx.doi.org/10.1348/000711010x498621

Patriota, A. G., Sato, J. R., Blas, A., & G., B. (2010). Vector autoregressive models with measurement errors for testing granger causality. *Statistical Methodology*, *7*(4), 478–497.

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*(1), 7–11. Retrieved from `http://CRAN.R-project.org/doc/Rnews/`

Plummer, M., Stukalov, A., & Plummer, M. M. (2014). Package âĂŸrjagsâĂŹ. *update*, *16*, 1.

Pyszcynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychological Bulletin*, *102*(1), 122 - 138. doi: http://dx.doi.org/10.1007/978-1-4613-9191-3 _6

Querstret, D., & Cropley, M. (2013). Assessing treatments used to reduce rumination and/or worry: A systematic review. *Clinical Psychology Review*, *33*(8), 996 - 1009. doi: http://dx.doi.org/10.1016/j.cpr.2013.08.004

R Development Core Team. (2012). R: A language and environment for statisti-

cal computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/ (ISBN 3-900051-07-0)

Rietbergen, C., Groenwold, R. H. H., Hoijtink, H. J. A., Moons, K. G. M., & Klugkist, I. (2014). Expert elicitation of study weights for bayesian analysis and meta-analysis. *Journal of Mixed Methods Research*, 1558689814553850.

Rietbergen, C., Klugkist, I., Janssen, K. J. M., Moons, K. G. M., & Hoijtink, H. J. A. (2011). Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary clinical trials*, *32*(6), 848–855.

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

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

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

Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods*. doi: http://dx.doi.org/10.1037/met0000062

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

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

Snippe, E., Bos, E. H., van der Ploeg, K. M., Sanderman, R., Fleer, J., & Schroevers, M. J. (2015, 10). Time-series analysis of daily changes in mindfulness, repetitive thinking, and depressive symptoms during mindfulness-based treatment. *Mindfulness*, *6*(5), 1053–1062. doi: 10.1007/s12671-014-0354-7

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

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.

Staudenmayer, J., & Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association*, *100*(471), 841–852.

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.

Swamy, P. A. V. B., Chang, I., Mehta, J. S., & Tavlas, G. S. (2003). Correcting for omitted-variable and measurement-error bias in autoregressive model estimation with panel data. *Computational Economics*, *22*(2-3), 225–253.

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

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625. doi: 10.1007/BF02289858

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., & Tuerlinckx, F. (n.d.). Visualizing distributions of covariance matrices. *Unpublished Manuscript*. Retrieved from `http://www.stat.columbia.edu/ gelman/research/unpublished/Visualization.pdf`

van der Krieke, L., Emerencia, A. C., Bos, E. H., Rosmalen, J., Riese, H., Aiello, M., ... de Jonge, P. (2015). Ecological momentary assessments and automated time series analysis to promote tailored health care: A proof-of-principle study. *JMIR research protocols*, *4*(3), e100. doi: 10.2196/resprot.4000

van Putten, M., Zeelenberg, M., & van Dijk, E. (2013). How consumers deal with missed discounts: Transaction decoupling, action orientation and inaction inertia. *Journal of Economic Psychology*, *38*, 104 – 110. doi: http://dx.doi.org/10.1016/j.joep.2012.09.008

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An sem approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological methods*, *17*(2), 176. doi: http://dx.doi.org/10.1037/a0029251

Wagenmakers, E., Farrell, S., & Racliff, R. (2004). Estimation and interpretation of 1/fÎś noise in human cognition. *Psychonomic Bulletin & Review*, *11*, 579–615.

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

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

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin*, *134*(2), 163 - 206. doi: http://dx.doi.org/10.1037/0033-2909.134.2.163

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, *54*, 1063–1070.

WorldHealthOrganization. (2008). *Icd-10: International statistical classification of diseases and related health problems (10th rev. ed.).* New York, NY: Author.

List of Figures

List of Tables

Nederlandse Samenvatting

Multilevel autoregressieve modellen zijn statistische modellen die gebruikt kunnen worden voor het analyseren van intensieve longitudinale data, dat wil zeggen, data die bestaan uit tijdreeksen voor meerdere personen. De insteek van autoregressieve modellen wordt goed geduid met de uitspraak "Gedrag uit het recente verleden is de beste voorspeller van toekomstig gedrag". In een klassiek autoregressief model worden huidige scores op de afhankelijke variabele namelijk geregresseerd op voorgaande scores op diezelfde variabele, dat wil zeggen, observaties op het huidige meetmoment worden gebruikt als voorspeller voor toekomstige observaties. Door het autoregressieve model uit te breiden naar een multilevel autoregressief model wordt het mogelijk om de herhaalde metingen van meerdere individuen tegelijk te modelleren en tevens de verschillen tussen de autoregressieve processen van de individuen te modelleren. Multilevel autoregressieve modellen worden steeds populairder binnen de psychologie, maar de methoden voor het passen van deze modellen voor psychologische data zijn nog in de ontwikkelingsfase. Het doel van deze dissertatie was om bepaalde moeilijkheden van het passen en interpreteren van multilevel autoregressieve modellen te onderzoeken, uiteen te zetten en zo mogelijk te verhelpen.

In de algemene inleiding van deze dissertatie wordt besproken waarom intensieve longitudinale data en het modelleren van individuen over de tijd essentieel is voor de psychologische wetenschap. Daarnaast wordt het idee achter het multilevel autoregressieve model uitgelegd. Hoofdstuk 2 gaat over de specificatie van de Inverse-Wishart prior kansverdeling voor de covariantiematrix van de random parameters voor het passen van het model in het Bayesiaanse framework. Hoewel er inmiddels ook frequentische technieken zijn om dit type modellen te passen, zijn er praktische redenen om te kiezen voor een Bayesiaanse insteek (naast eventuele wetenschaps- of statistischfilosofische voorkeuren). Echter, de specificatie van een Inverse-Wishart prior kansverdeling blijkt niet eenvoudig — en misspecificaties resulteren in bias in de varianties van de regressiecoëfficienten. Het probleem zit hem in het feit dat de Inverse-Wishart zeer informatief wordt als varianties dicht bij nul liggen. In Hoofdstuk 2 zetten mijn co-auteurs en ik dit probleem uiteen en presenteren we een simulatiestudie waarin we drie priorspecificaties met elkaar vergelijken die zijn voorgesteld

in de literatuur. Onze resultaten laten zien dat een priorspecificatie die gebaseerd is op de data — waarin schattingen van de varianties van de random parameters gebruikt worden in de specificatie — het beste presteert van de drie specificaties die wij hebben vergeleken, zelfs al resulteert het meerdere keren gebruiken van de data in een overschatting van de zekerheid over de schattingen. Wij adviseren dat voor ieder multilevel model, maar met name multilevel autoregressieve modellen, een sensitiviteitsanalyse uitgevoerd wordt voor de prior specificaties van de covarianties van de random parameters. We geven een voorbeeld van een dergelijke analyse in Hoofdstuk 2, uitgevoerd voor een empirische studie naar de wederkerige effecten van piekeren en positief affect.

Multivariate autoregressieve (VAR) modellen kunnen gebruikt worden om Granger-causale cross-lagged associaties vast te stellen tussen variabelen. In Hoofdstuk 3 bespreken wij hoe de sterkte van deze cross-lagged associaties vergeleken kunnen worden door de bijbehorende regressiecoefficiënten te standaardiseren. Het vergelijken van de sterkte van cross-lagged associaties wordt vaak interessant gevonden door psychologen die willen bepalen welke associaties 'causaal dominant' of 'de drijvende kracht' zijn in het psychologische proces. Gestandaardiseerde cross-lagged coefficiënten kunnen gebruikt worden om te bepalen welke predictoren de meeste unieke variantie verklaren en dus het sterkste directe effect hebben. Echter, voor multilevel modellen kunnen de parameters gestandaardiseerd worden op basis van groepsstatistieken of op basis van persoonspecifieke statistieken. Deze verschillende standaardisatiemethoden leveren verschillende resultaten op en mogelijk zelfs tegenstrijdige conclusies. In Hoofdstuk 3 stellen wij dat de parameters gestandaardiseerd zouden moeten worden op basis van persoonspecifieke statistieken. We illustreren de verschillende standaardisatiemethoden met een empirisch voorbeeld waarin we de sterkte vergelijken van de cross-lagged associaties tussen de gevoelens van competentie en moeheid van mensen met burnout. Daarnaast laten wij met dit voorbeeld zien dat het negeren van individuele verschillen in psychologische processen misleidende resultaten kan opleveren. Dit illustreert de toegevoegde waarde van multilevel autoregressieve modellen ten opzichte van meer gebruikelijke cross-lagged panel modellen die deze individuele verschillen negeren.

Een belangrijke beperking van zowel het n=1 en het multilevel autoregressieve model is dat ze negeren dat de data mogelijk meetfouten bevatten. Dit is problematisch, aangezien we er vrij zeker van kunnen zijn dat psychologische variabelen niet perfect gemeten worden en het negeren van meetfouten bias oplevert in de geschatte regressieparameters. In Hoofdstuk 4 en 5 van deze dissertatie zetten wij uiteen hoe de

n=1 en multilevel autoregressieve modellen uitgebreid kunnen worden zodat zij wel rekening houden met meetfouten. In Hoofdstuk 4 bespreken we twee uitbreidingen van het n=1 autoregressieve model waarin rekening wordt gehouden met meetfouten: Een autoregresief model met een term met witte ruis (AR+WN model) en het autoregressive moving average (ARMA) model. We vergelijken de prestaties van deze modellen en het standaard autoregressieve model (dat geen rekening houdt met meetfouten) middels een simulatiestudie. Daarnaast onderzoeken wij of deze modellen beter presteren met een frequentistische of met een Bayesiaanse schattingsprocedure. Uit de resultaten blijkt dat het Bayesiaanse AR+WN model globaal genomen het beste presteert. De Bayesiaanse schattingsprocedure blijkt met name voordelig wanneer de steekproefgroottes relatief klein zijn (i.e., minder dan 500 herhaalde metingen per persoon). We illustreren het effect van het negeren van meetfouten met een empirisch voorbeeld waarin we herhaalde metingen van de dagelijkse stemming van acht vrouwen analyseren. Uit deze analyses blijkt dat afhankelijk van de participant 30% tot 50% van de variantie in stemming een gevolg is van meetfouten of andere meetmomentspecifieke fluctuaties. Het negeren van deze fluctuaties resulteert in een substantiële onderschatting van de autoregressieve parameters.

In Hoofdstuk 5 presenteren we een bivariaat multilevel VAR model dat rekening houdt met meetfouten (VAR+WN model), ofwel een model dat rekening houdt met de betrouwbaarheid van onze metingen. Dit model is een uitbreiding van het AR+WN model uit Hoofdstuk 4. In dit model variëren de gemiddeldes, regressieparameters, innovatievarianties en covarianties en de meetfoutvarianties en covarianties over personen. Dientengevolge variëren ook de betrouwbaarheden van de metingen over personen. In Hoofdstuk 5 laten we zien dat naast dat in het VAR+WN model impliciet rekening wordt gehouden met deze betrouwbaarheden, het model ook gebruikt kan worden om schattingen van de betrouwbaarheden te verkrijgen per persoon. Daarnaast bespreken we de gevolgen van het negeren van meetfouten in het VAR model. Aangezien de betrouwbaarheden per persoon verschillen, verschillen ook de gevolgen van het negeren van meetfout per persoon. Deze gevolgen kunnen ernstig zijn: de cross-lagged parameters kunnen worden onderschat of overschat; cross-lagged associaties kunnen 'verdwijnen' uit het model (vrijwel op nul geschat worden) en er kunnen schijnverbanden opduiken tussen variabelen. In een empirisch voorbeeld waarin we de algemene positieve affect en relatie-specifieke affect van mannen en vrouwen in een relatie analyseren, vinden we een gemiddelde betrouwbaarheid over personen van ongeveer .65. Dat wil zeggen dat we, net als in Hoofdstuk 4, vinden dat een groot deel

van de variantie in de metingen een gevolg was van meetmomentspecifieke fluctuaties of meetfouten, hetgeen het belang van rekening houden met deze fluctuaties in onze modellen onderschrijft.

Ten slotte worden in het laatste hoofdstuk van dit proefschrift de voorgaande hoofdstukken samengevat en een aantal beperkingen van het autoregressieve model en suggesties voor vervolgonderzoek besproken.

Acknowledgements/Dankwoord

I'd like to take this opportunity to put my gratitude to certain people for inspiring and supporting me during my PhD project in print. I look back on my PhD project with warm feelings, and look forward to new endeavors with excitement, because of you all.

I thank the members of the evaluation commission, prof. dr. Borsboom, dr. Ceulemans, prof. dr. van der Heijden, prof. dr. de Jonge en prof. dr. de Ridder, for the time and energy they invested in evaluating my dissertation, and their roles of opposition during the defense.

I would like to thank my co-authors Raoul Grasman, Emilio Ferrer, Mieke de Boer-Sonnenschein, and Jan Houtveen for their contributions to this dissertation. I would also like to thank Conor Dolan for thinking along about projects that in the end did not become part of this dissertation.

To Herbert, my promotor: Thank you for saying the right things at the right time. Our discussions came at the right moments and helped me regain focus.

To Ellen, my supervisor: I wanted to write up a funny anecdote that somehow also conveyed my gratitude, but it didn't work. I even tried to come up with some dynamic modeling puns but it seems I'd better leave those to Oisín. Instead, some more to-the-point words of thanks: Ellen, thank you for providing a safe space for thought, ideas and debate, and for always having good humor and good sense (and for using those powers to steer me in the right directions). Thank you for being a supervisor who cares not only about delivering quality research, but also about delivering more generally well-functioning scientists. Finally, I want to thank you for attracting a group of dynamic modeling enthusiasts, which I have benefited from.

Silvia, Maryam, Oisín, Rebecca, Arnout, Nele: Thank you for inspiring dynamic modeling lab meetings, for sharing your work and thoughts, laughs, and for tolerating/participating in/starting discussions even when they seemingly were going

nowhere.

I would like to thank all my colleagues from the Research Group of Quantitative Psychology and Individual Differences at the KU Leuven for welcoming me with open arms, and for thinking aloud and having thought-provoking discussions with me about discrete and continuous time (models), assumptions, networks, and the like. Special thanks go out to Francis Tuerlinckx for making the necessary arrangements for me to come to Leuven, and for providing me with guidance and tools for learning about continuous time modeling.

Special thanks also go to Laura, for towing me around Leuven, working together with me, hosting one of my favorite dynamic modeling meetings every half year, and for being a great friend in general.

To Jesper and Maria, my favorite psychometricians: Thank you for being my paranymphs, and thank you for all our moments, rants and discussions about methods, statistics, and life.

Manon, I learned a lot with and from you as a study-partner, colleague, and friend. I hope we will carry on learning together for years to come.

To my roommates Charlotte, Joran, and Haifang, and hall-pals Gerko, Gu and Susanna: Thank you for friendly greetings early in the morning, chats, passionate discussions, and creating a wonderful working atmosphere in general.

My time as a PhD candidate would have been much less enjoyable without the weekly Bayes Meeting - thanks to all who participated. It takes a special kind of Bayes meeting to create an environment where a chair in Bayesian Statistics can admit they may actually be a frequentist.

I'd like to thank all my other colleagues from the Methodology and Statistics department at Utrecht University, IOPS, and those involved with the FSS PhD Council, for helping me become a better scientist/teacher/academic, and for diverting me from academia when necessary.

Mam: Altijd als ik aan je vroeg wat je wilde worden als je later groot was antwoordde je "oud en wijs". Een beter rolmodel kan ik mij niet wensen.

Pap: Dankjewel voor grappige verhalen over werken in het lab, het demonstreren wat er in de praktijk met onderzoek bereikt kan worden, en een schijnbaar onvoorwaardelijk vertrouwen in mijn kunnen.

Karianne: Bedankt voor er zijn! Ik vind het tof dat we samen kunnen mijmeren en morren over het leven binnen en buiten de wetenschap.

Tim: Jij was destijds de eerste die las dat ik toegelaten werd tot de bacheloropleiding Psychologie. Dankjewel voor het meebeleven van alle hoogtepunten en de minder hoge punten sindsdien. Ik ben heel blij dat we, nu ik dit schrijf, nog steeds samen teveel plezier hebben.

About the Author

Noémi Katalin Schuurman was born in Amsterdam, on June 5th 1988. She attended Vallei College 't Atrium (Amersfoort) for her preparatory university education, and graduated in 2006. She went on to study psychology at the University of Amsterdam, because she thought she would become a clinical therapist. Things turned out differently however, as she decided to specialize in psychological methods in her third bachelor year instead of clinical psychology. She felt that there was more to psychological research than met the eye, and that she might be able to uncover more with additional methods training (that, and the methods courses were the most fun to her). She quickly felt at place studying psychological methods, and obtained her bachelor's degree cum laude in 2009. She continued her studies by following the Research Master's program in Psychology at the University of Amsterdam, with a major in methodology and a minor in clinical psychology, and graduated cum laude in 2011.

During her studies, Noémi developed a strong interest in the contrasts between studying interindividual differences and intraindividual differences, dynamic modeling, and time series analysis, and worked on these topics for her internship project and master's thesis. Her supervisors at the time, Prof. Dr. Denny Borsboom, and Prof. Dr. Conor Dolan, brought her into contact with Dr. Ellen Hamaker. In 2011, Noémi started her PhD project under Dr. Hamaker's supervision at the Department of Methodology and Statistics at Utrecht University, with Prof. Dr. Herbert Hoijtink as her promotor. Since then, Noémi's research has appeared in journals such as Multivariate Behavioral Research and Psychological Methods, and she has presented her work at national and international conferences.

Besides doing research, Noémi also enjoys teaching. During her time at Utrecht University, she has taught methods and statistics to bachelor and master students, as well as to PhD candidates and other fellow researchers. She is particularly fond of providing methods and statistics consultations to students and researchers, which she started doing weekly as part of the methodology shop during her studies at the University of Amsterdam, and has continued doing as part of the consultation shop at Utrecht University. In addition, Noémi has acted as chair and methodology and statistics representative for the PhD candidate council of the Faculty of Social and

Behavioral Sciences at Utrecht University.

In September 2015, Noémi started working as a postdoctoral researcher at the Department of Methodology and Statistics at Utrecht University. Her main research interests are dynamic and idiographic modeling, multilevel modeling, Bayesian statistics, scientific integrity, and philosophy of (psychological) science.