Selection of the Number of Participants in Intensive Longitudinal Studies: A User-friendly Shiny App and Tutorial to Perform Power Analysis in Multilevel Regression Models that Account for Temporal Dependencies

Ginette Lafit*[1,2], Janne K. Adolf[1], Egon Dejonckheere[1], Inez Myin-Germeys[2], Wolfgang Viechtbauer[2,3], and Eva Ceulemans[1]

[1]Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium
[2]Department of Neurosciences, Center for Contextual Psychiatry, KU Leuven, Leuven, Belgium
[3]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, the Netherlands

This manuscript was accepted for publication in *Advances in Methods and Practices in Psychological Science* in September of 2020.

*Corresponding author: Ginette Lafit (ginette.lafit@kuleuven.be)

Abstract

In recent years the popularity of procedures to collect intensive longitudinal data, such as the Experience Sampling Method, has immensely increased. The data collected using such designs allow researchers to study the dynamics of psychological functioning, and how these dynamics differ across individuals. To this end, the data are often modeled with multilevel regression models. An important question that arises when designing intensive longitudinal studies is how to determine the number of participants needed to test specific hypotheses regarding the parameters of these models with sufficient power. Power calculations for intensive longitudinal studies are challenging, because of the hierarchical data structure in which repeated observations are nested within the individuals and because of the serial dependence that is typically present in this data. We, therefore, present a user-friendly application and step-by-step tutorial to perform simulation-based power analyses for a set of models that are popular in intensive longitudinal research. Since many studies use the same sampling protocol (i.e., a fixed number of at least approximately equidistant observations) within individuals, we assume this protocol fixed and focus on the number of participants. All included models explicitly account for the temporal dependencies in the data by assuming serially correlated errors or including autoregressive effects.

Keywords: Power analysis; Monte Carlo simulation; intensive longitudinal designs; linear mixed effect models; multilevel autoregressive models.

1 Introduction

Over the last years, psychological research has increasingly focused on investigating how complex psychological processes evolve dynamically across time within single individuals. To this end, researchers use intensive longitudinal (IL) designs and data collection methods, such as the Experience Sampling Method (ESM) (Myin-Germeys et al. 2009, 2018), in which individuals are repeatedly measured. The repeated measurements allow researchers to study dynamic aspects of psychological functioning within individuals and individual differences therein. Examples of such dynamics are emotional variability and stability and emotional inertia (Kuppens & Verduyn 2015). Individual differences in these dynamics have been consistently linked to individual differences in well-being and health (e.g., Brose et al. 2015, Dejonckheere et al. 2018, Kuppens et al. 2010).

Given the increased focus on dynamic psychological processes within individuals, it is no surprise that the recent debate on the reproducibility and transparency of psychological research (Munafò et al. 2017) has led to the development of guidelines for conducting IL research (Trull & Ebner-Priemer 2020) and the promotion of open science practices (Kirtley et al. In press). Here, we aim to continue along this path and focus on sample size planning for IL designs. In IL studies, it is common practice to use a fixed sampling schedule within individuals that is also motivated in terms of feasibility and the participants' burden. Therefore, we will focus on assessing the number of participants needed while assuming a fixed number of (at least approximately) equidistant observations within individuals. Adequate sample size planning allows to control the accuracy and power of statistical testing and modeling and is therefore of crucial importance for the replicability of empirical findings (see Ioannidis 2005, Szucs & Ioannidis 2017).

Although power analyses are often used to inform sample size planning in general (Cohen 1988), they are not yet well-established in IL research. One reason for this is that performing power calculations to select the number of participants in the context of IL studies is challenging because of the intricacies of the data (Bolger 2011, De Jong et al. 2010). First, IL data have a multilevel structure, in that repeated observations are nested within individuals. Second, observations are closer in time in comparison with traditional longitudinal designs. This likely leads to considerable temporal dependencies between data measured at adjacent observations. As we explained earlier, it is often the very purpose of an IL study to capture such temporal dependencies, as they reflect psychological dynamics that are often of inherent interest.

But not only the data structure is complicated; the applied statistical models are as well, as they should capture such dynamics and individual differences therein. First, the models have to distinguish inter-individual differences from intra-individual changes (e.g., Hamaker et al. 2015, Molenaar 2004). Multilevel regression modeling approaches offer an established way of doing this. Second, models should also take temporal dependencies into account, either to control for them or to quantify and model them. This requires that one includes either serially correlated errors or the lagged outcome variable as a predictor in the multilevel models. Although there are several resources available to perform power analyses

for multilevel models (e.g., Arend & Schäfer 2019, Browne et al. 2009, Cools et al. 2008, Green & MacLeod 2016, Hedeker et al. 1999, Landau & Stahl 2013, Lane & Hennes 2018, Mathieu et al. 2012, Raudenbush 1997, Raudenbush & Liu 2001, Snijders & Bosker 1993, Zhang & Wang 2009, Zhang 2014), these do not account for the temporal dependencies that characterize IL data.

We therefore present a user-friendly application to perform simulation-based power analyses for IL studies. The obtained power results allow informing sample size planning, by shedding light on the number of participants needed to obtain accurate and significant parameter estimates. The application was developed in R (R Core Team 2013) using the package `Shiny` (Chang et al. 2019). It covers a set of models that are widely used in the literature to study individual differences in IL studies and properly account for the temporal dependency.

In the remainder of the article, we first briefly review existing approaches to compute power in multilevel models. Next, we discuss the multilevel models that are covered by our application. In Section 4, we introduce the shiny app and discuss how it can be used for sample size planning. Afterward, using an already published data set, we illustrate how to perform sample size planning with the app. We conclude the article with a general discussion of additional considerations and possible extensions.

2 Power analyses in intensive longitudinal studies

We use statistical power as the criterion to estimate the number of participants needed in an IL study. High power is desirable because it improves the reproducibility of research findings, and prevents the overestimation of effect sizes (see Ioannidis 2005, Szucs & Ioannidis 2017). Formally, power is defined as the probability of correctly rejecting the null hypothesis, when the alternative hypothesis is true in the population under study (Cohen 1988). The power to detect an effect is therefore determined by the size of the effect in the population, the predetermined type I error rate (i.e., the significance level), and the standard error of the test statistic used. Power is higher if the population effect is larger, the type I error rate is higher, and the standard error of the test statistic is smaller. The standard error, in turn, is related to sample size, in that larger sample sizes lead to smaller standard errors. The latter explains why power analysis can inform sample size planning.

In general, two approaches can be used for performing power analysis: the analytical approach and the simulation-based approach. In the *analytical approach*, power is determined by using formulas for the standard errors of the estimated effects, expressing them as a function of the parameters of the multilevel model under study and the sample size. Using these formulas, it is possible to estimate the sample size that allows reaching a predetermined value of power (see, e.g., Cohen 1988, Hedeker et al. 1999, Moerbeek et al. 2000, 2001, Moerbeek & Maas 2005, Raudenbush 1997, Raudenbush & Liu 2001, Snijders & Bosker 1993, Wang et al. 2015). However, as holds for many other complex models, so far no analytical formulas have been derived for multilevel models that include temporal dependencies (see Arend & Schäfer 2019). Also, the analytical approach usually relies on

asymptotic estimation theory and might, therefore, be inaccurate in practice when dealing with smaller numbers of participants and measurements per participant. For example, Snijders & Bosker (1993) determined the optimal sample sizes for two-level linear models by using normal approximations for the distribution of the estimated coefficients. However, in small samples, the distribution of the estimator can be non-normal and potentially heavy-tailed, resulting in unreliable standard error estimates.

The *simulation-based approach* uses the hypothesized population model and concrete specifications of the associated parameters to generate a large number of data sets. Each of these data sets is then analyzed with the model under study and the parameter(s) of interest are tested for significance. Since the data have been randomly generated, the parameter estimates and the test results will vary across the data sets. Hence, we can compute the power as the proportion of simulated data sets for which the null hypothesis about the parameter(s) of interest has been rejected (see, e.g., Arend & Schäfer 2019, Astivia et al. 2019, Bolger 2011, Browne et al. 2009, Cools et al. 2008, Green & MacLeod 2016, Landau & Stahl 2013, Lane & Hennes 2018, Maas & Hox 2005, Mathieu et al. 2012, Zhang & Wang 2009, Zhang 2014). Performing these calculations while varying the number of participants allows us to determine the number of participants necessary to reach a predetermined amount of power (e.g., 80%). The simulation-based approach is a good alternative when analytical formulations are not available or too difficult to derive. Therefore, we adopt this approach in this paper, given the complexity of IL data and associated modeling questions.

3 Population models of interest

We focus on a set of research questions regarding IL data that can be addressed using specific multilevel regression models (Raudenbush & Bryk 2002). Figure 1 provides a graphical representation of the different models. This representation corresponds to a hypothetical dataset that is used for illustration purposes. Table 1 shows a few rows of this dataset involving individuals diagnosed with Major Depressive Disorder (MDD) and healthy controls. The participants responded to momentary questionnaires at six equidistant time points. The first column contains the participants' identification number and the second column the observation number. Moreover, the dataset includes the level-1 or time-varying variables *Affect* (for negative affect) and *Anhedonia*, which are measured at every observation. The dataset also contains two level-2 or time-invariant variables. Variable *Depression* refers to the sum score of a continuous self-report instrument on the experience of depressive symptoms assessed at baseline. Finally, *Diagnosis* is a binary variable that equals one for participants diagnosed with MDD and zero otherwise. Corresponding to Figure 1, formulas for the models are given in Table 2, while Table 3 provides an overview of the effects of interest.

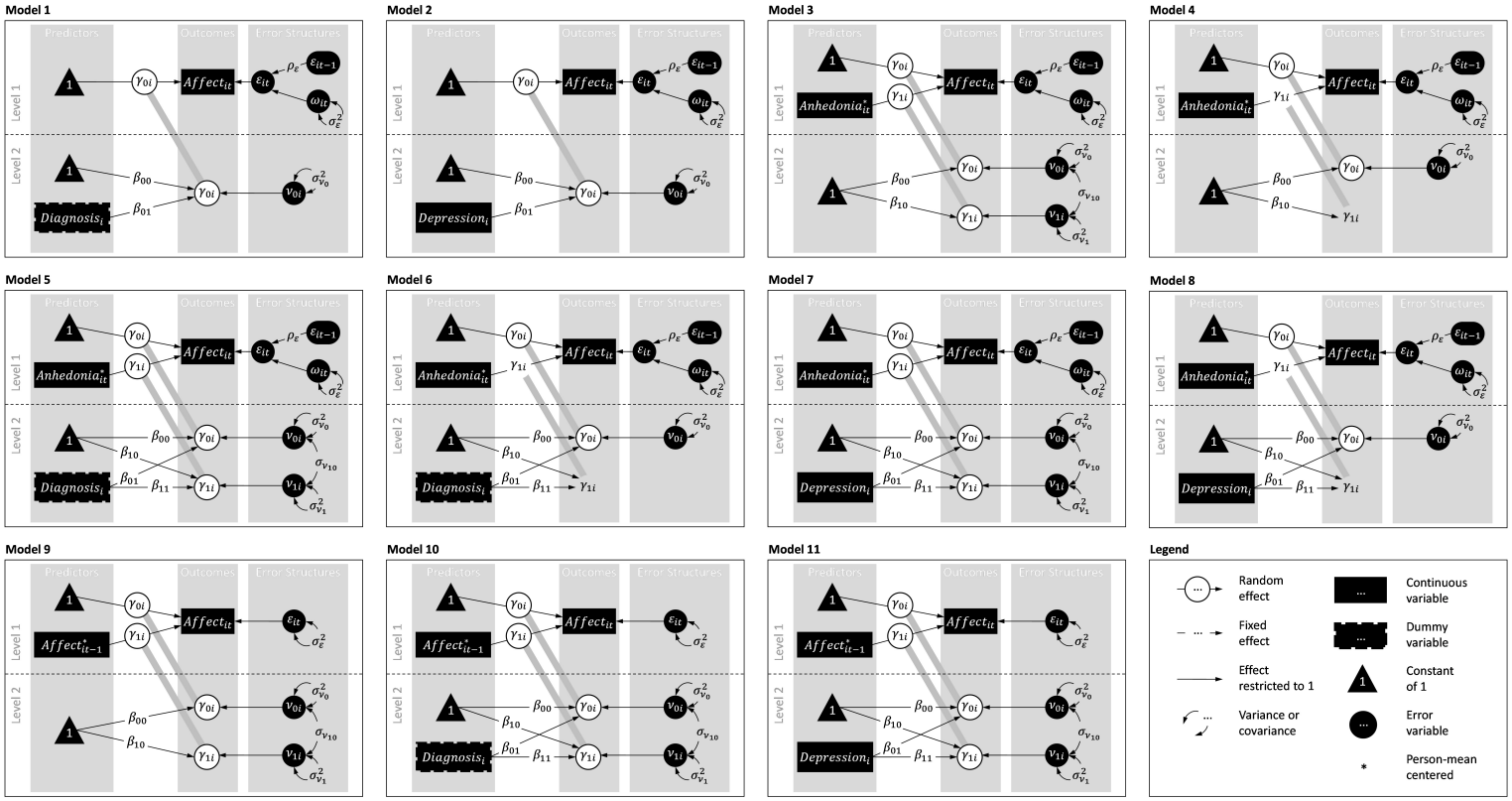

Figure 1: Graphical representation of population models of interest available in the application to perform simulation-based power analysis.

Table 1: Intensive longitudinal data structure.

PID	Observation	Negative Affect	Anhedonia	Depression	Diagnosis
1	1	28.8	42	12	1
1	2	26.0	30	12	1
1	3	27.4	22	12	1
1	4	21.4	33	12	1
1	5	14.4	23	12	1
1	6	26.6	18	12	1
2	1	16.0	19	4	0
2	2	13.2	23	4	0
2	3	9.6	12	4	0
2	4	14.4	18	4	0
2	5	8.6	10	4	0
2	6	9.2	15	4	0

Table 2: Description of models for power estimation available in the application.

	Level-1	Level-2	
		Random Intercept	Random Slope
Model 1	$\text{Affect}_{it} = \gamma_{0i} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Diagnosis}_i + \nu_{0i}$	-
Model 2	$\text{Affect}_{it} = \gamma_{0i} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Depression}_i + \nu_{0i}$	-
Model 3	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Anhedonia}_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \nu_{1i}$
Model 4	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Anhedonia}_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \nu_{0i}$	-
Model 5	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Anhedonia}_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Diagnosis}_i + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \beta_{11}\text{Diagnosis}_i + \nu_{1i}$
Model 6	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Anhedonia}_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Diagnosis}_i + \nu_{0i}$	-
Model 7	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Anhedonia}_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Depression}_i + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \beta_{11}\text{Depression}_i + \nu_{1i}$
Model 8	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Anhedonia}_{it} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Depression}_i + \nu_{0i}$	-
Model 9	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Affect}_{it-1} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \nu_{1i}$
Model 10	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Affect}_{it-1} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Diagnosis}_i + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \beta_{11}\text{Diagnosis}_i + \nu_{1i}$
Model 11	$\text{Affect}_{it} = \gamma_{0i} + \gamma_{1i}\text{Affect}_{it-1} + \epsilon_{it}$	$\gamma_{0i} = \beta_{00} + \beta_{10}\text{Depression}_i + \nu_{0i}$	$\gamma_{1i} = \beta_{10} + \beta_{11}\text{Depression}_i + \nu_{1i}$

Table 3: Overview of the effects of interest for power estimation available in the application.

Model	Time-varying Level 1 predictor		Time-invariant Level 2 predictor		Random Intercept	Random Slope	Cross-level Interaction effect
	Continuous variable	Lagged dependent variable	Dummy variable	Continuous variable			
Model 1	-	-	X	-	X	-	X
Model 2	-	-	-	X	X	-	X
Model 3	X	-	-	-	X	X	-
Model 4	X	-	-	-	X	-	-
Model 5	X	-	X	-	X	X	X
Model 6	X	-	X	-	X	-	X
Model 7	X	-	-	X	X	X	X
Model 8	X	-	-	X	X	-	X
Model 9	-	X	-	-	X	X	-
Model 10	-	X	X	-	X	X	X
Model 11	-	X	-	X	X	X	X

3.1 Group differences in mean level

Model 1 in Figure 1 estimates differences in the mean of the outcome variable, Affect, between the two groups of individuals (e.g., Heininga et al. 2019, Myin-Germeys et al.

2001, 2003). This model includes the $Affect_{it}$ value as the outcome variable for the ith individual at the tth observation and a level-2 dummy variable that indicates the diagnosis group (i.e., $Diagnosis_i$). For participants in the reference group (healthy controls), the mean level of affect equals β_{00}; for individuals diagnosed with MDD, the mean level of affect is given by $\beta_{00} + \beta_{01}$. Within both diagnosis groups, inter-individual differences in affect are modeled by the random intercept γ_{0i}. The random intercept expresses the deviation of each participant's affect level from the group-specific mean level. It is normally distributed and the standard deviation is denoted by σ_{ν_0}. To account for the likely temporal dependencies in IL data, we allow for serially correlated errors. Therefore, we assume that the level-1 errors ϵ_{it} follow a first-order autoregressive (AR(1)) process (Goldstein et al. 1994), where the correlation between two consecutive errors is denoted by ρ_ϵ, and σ_ϵ is the standard deviation of the level-1 errors.[1] To guarantee that the model is stationary (Hamilton 1994), the autocorrelation ρ_ϵ should range between -1 and 1. In model 1, the main effect of interest is β_{01} (i.e., the size of the average group difference) and we test whether it is statistically different from zero. As holds for all tests that we will discuss, the hypothesis test is two-sided and significance is evaluated with a Wald-type test statistic using a t-distribution (Snijders & Bosker 2011).

3.2 Effect of a level-2 continuous predictor on the mean level

Model 2 in Figure 1 focuses on the effect of a continuous level-2 predictor on the outcome of interest.[2] For the hypothetical dataset, we investigate whether the individual-specific depression level $Depression_i$ predicts individual differences in the mean level of $Affect_{it}$ as captured by the random intercept γ_{0i}. These random intercepts are assumed to be normally distributed with mean $\beta_{00} + \beta_{01} Depression_i$ and standard deviation σ_{ν_0}. We again assume an AR(1) structure for the level-1 errors ϵ_{it}. When testing the effect of interest β_{01} (Raudenbush & Bryk 2002), we can grand-mean center the level-2 predictor to obtain a meaningful zero point for this predictor to render the intercept interpretable (Enders & Tofighi 2007).

3.3 Effect of a level-1 continuous predictor

Next, we focus on the effect of a continuous level-1 predictor on the outcome, through Models 3 and 4 in Figure 1. For example, we might be interested to what extent $Anhedonia_{it}$ predicts $Affect_{it}$ in individuals diagnosed with MDD. Model 3 specifies a corresponding multilevel model with AR(1) level-1 errors. The mean slope of $Anhedonia_{it}$ is denoted by β_{10}, which is the parameter of interest. This model captures inter-individual differences by including a random intercept γ_{0i} and a random slope γ_{1i}. These random effects are bivariate normally

[1] The first-order autoregressive process is defined as $\epsilon_{it} = \rho_\epsilon \epsilon_{it-1} + \omega_{ij}$, where ω_{ij} is assumed to be Gaussian white noise $N(0, \sigma_\omega)$. Under this model the correlation between ϵ_{it-1} and ϵ_{it} is given by ρ_ϵ and $\sigma_\epsilon^2 = \sigma_\omega^2/(1 - \rho_\epsilon^2)$.

[2] Here and elsewhere, we use terms like 'effect' and 'influence' for brevity without implying that the associations being modelled are necessarily causal.

distributed. β_{00} then indicates the mean of the random intercepts and β_{10} the mean of the random slopes. Their standard deviations are respectively denoted by σ_{ν_0} and σ_{ν_1}. The correlation between the random effects is given by $\rho_{\nu_{01}}$ (and the covariance between the random effects is denoted by $\sigma_{\nu_{01}}$). Model 4, on the other hand, assumes that the slope of $Anhedonia_{it}$ does not vary across participants. In both models, person-mean centering the level-1 predictor is recommended since the fixed slope β_{10} then provides an estimate that only reflects the (average) within-person association between the predictor and outcome (Enders & Tofighi 2007, Raudenbush & Bryk 2002).

3.4 Group differences in the effect of a level-1 continuous predictor

Models 5 and 6 in Figure 1 are used to investigate differences between two groups of participants with respect to the association between a level-1 predictor and the outcome of interest (while assuming AR(1) errors). These models thus include the outcome $Affect_{it}$, the level-1 predictor $Anhedonia_{it}$, the level-2 variable $Diagnosis_i$, as well as a 'cross-level interaction' (Raudenbush & Bryk 2002) between the level-1 and level-2 predictors. β_{00} and $\beta_{00}+\beta_{01}$ represent the mean intercept of all individuals in the reference (healthy controls) and MDD group, respectively. The mean slope for the reference group is indicated by β_{10}, while the mean slope for the MDD group amounts to $\beta_{10} + \beta_{11}$. Therefore, the effect of interest is the difference in the mean slope between the two groups β_{11}. Model 5 includes random intercepts γ_{0i} as well as random slopes γ_{1i}. Model 6 is more restrictive and does not include random slopes.

3.5 Cross-level interaction between two continuous predictors

Models 7 and 8 in Figure 1 focus on a cross-level interaction between the continuous level-2 predictor $Depression_i$ and the continuous level-1 predictor $Anhedonia_{it}$ (e.g., Arend & Schäfer 2019), to investigate whether the level of depression (as measured at baseline) moderates the effect of anhedonia on affect. Therefore, the effect of interest is again β_{11}. As was the case for Models 5 and 6, Model 7 includes both random intercepts and slopes, whereas Model 8 assumes that the slope does not vary across participants.

3.6 Multilevel autoregressive models

Models 9 to 11 (see Figure 1) are multilevel AR(1) autoregressive models (Hamaker & Grasman 2015) that explicitly focus on the amount of temporal dependence in the outcome. In such models, the lagged outcome variable (i.e., the observed outcome at the previous measurement occasion) is included as the predictor of interest. Such autoregressive effects have been extensively studied for example in affective research (Kuppens et al. 2010). Model 9 allows us to study the mean autoregressive effect across individuals as well as individual differences therein, through β_{10} and γ_{1i}, respectively. To satisfy the stationarity assumption of the model, both effects have to range between -1 and 1. Given that temporal dependence

is now captured through the autoregressive effect, the residuals ϵ_{it} are assumed to be independent and normally distributed with zero mean σ_ϵ. Some researchers person-mean center the lagged outcome variable, although Hamaker & Grasman (2015) showed in an extensive simulation study that this results in an underestimation of β_{10}. The resulting bias will have an impact on power.

Model 10 extends Model 9 in that it allows us to estimate the difference in the mean autoregressive effect between two groups of individuals (Wang et al. 2012). The mean autoregressive effect for the reference group (healthy controls) is β_{10}, while it equals $\beta_{10}+\beta_{11}$ for the MDD group. Therefore, the effect of interest is β_{11}.

Finally, Model 11 estimates a cross-level interaction effect between a continuous level-2 predictor and the lagged outcome, to study if the level-2 predictor moderates the autoregressive effect (e.g., Brose et al. 2015, Koval et al. 2013). Consequently, β_{11} is the effect of interest. In this case, Hamaker & Grasman (2015) clearly recommend to person-mean center the lagged predictor.

4 A Shiny app to perform power analysis

In this section, we present the Shiny app *PowerAnalysisIL* that we developed to compute power as a function of the number of participants for the models described in the previous section. Figure 2 shows a screenshot of the app. The app was implemented using the `R` package `Shiny`. It is available via a git repository hosted on GitHub at `https://github.com/ginettelafit/PowerAnalysisIL`. Users can download the app and run it locally on their computer in `R` or Rstudio (RStudio Team 2015). In what follows, we describe how the app works.

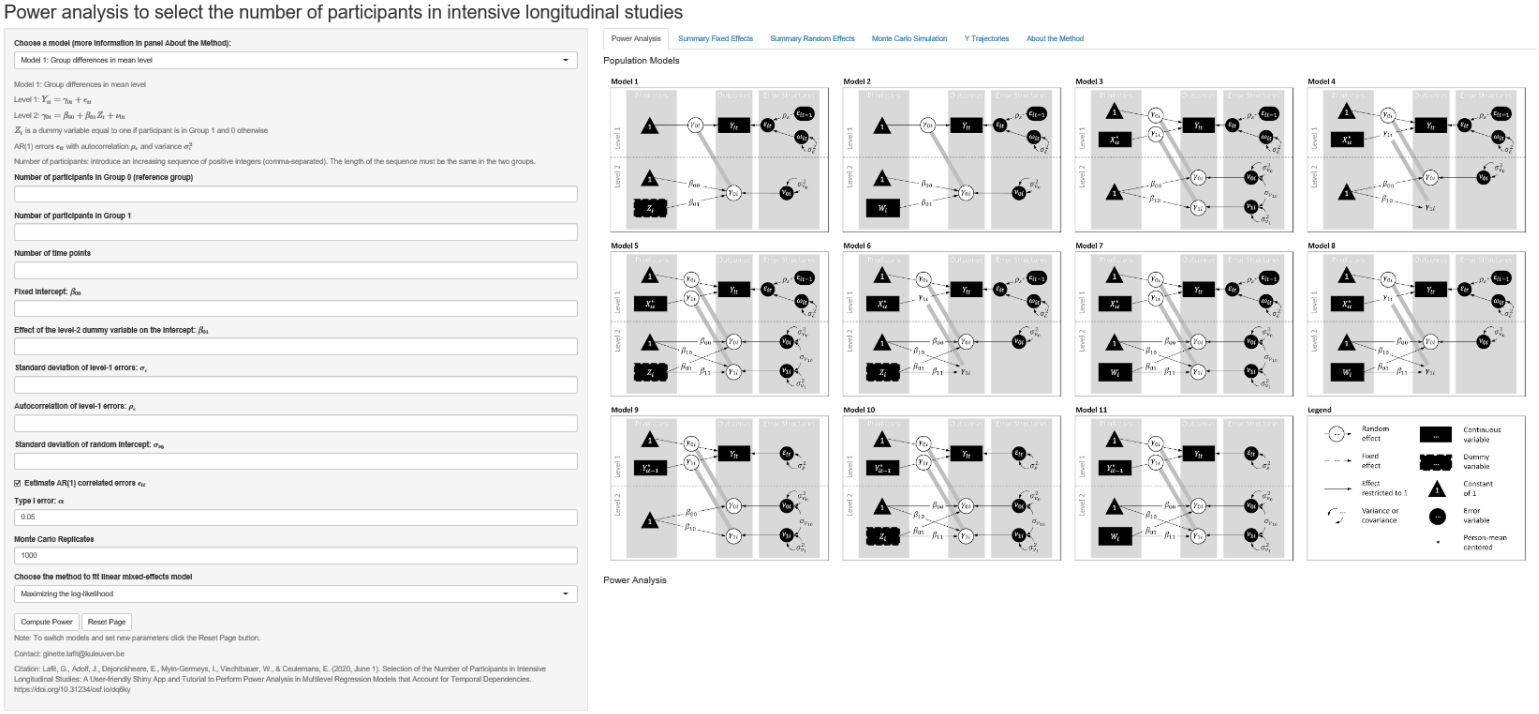

Figure 2: Screenshot of PowerAnalysisIL, a Shiny app to perform power analysis to select the number of participants in intensive longitudinal studies.

4.1 App input

First, the user indicates which multilevel model (i.e., Model 1 to 11) will be used to estimate the effect of interest and specifies plausible values for all model parameters. For instance, if one wants to focus on differences in mean affect between individuals diagnosed with MDD and healthy controls, one selects Model 1. Next, the sample sizes that should be considered in the power computations have to be provided. Returning to our example of between group differences in mean affect, the user has to set the number of participants in the reference group (i.e., healthy controls) and the corresponding number of participants diagnosed with MDD. Based on this information the software will create a level-2 dummy predictor, indicating group membership. For instance, possible sample sizes for the healthy controls and MDD group could respectively amount to 20, 30, 40, and 80 and 15, 20, 25, and 30. Then, one sets the expected number of completed equidistant observations per individual (e.g., 60). In case the selected model includes continuous level-1 or level-2 predictors, their mean and standard deviation have to be provided, assuming that they are normally distributed. For level-1 continuous predictors, one indicates whether they should be grand-mean or person-mean centered. Finally, one sets the estimation method (i.e., Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation[3]), the desired significance level α, and the number of Monte Carlo replicates in the power simulations (e.g., 1000). For Models 1 to 8, the app also allows estimating multilevel models with independent errors (i.e., assuming $\rho_\epsilon = 0$). Comparing the power of models with and without AR(1) errors makes it possible to assess the impact of temporal dependence.

4.2 Simulation

Based on this input, the app repeatedly simulates the data for each indicated sample size. For the multilevel AR models (i.e., Models 9 to 11), simply sampling the random effects from a normal distribution might yield data that are not stationary (i.e., the normal distribution does not restrict the random autoregressive effects to belong to the interval $[-1, 1]$). To guarantee stationarity, without changing the specified mean and standard deviation of the random slopes, we draw the random slopes from a Beta distribution and linearly transform them so that they fall into the interval $(-1, 1)$.[4] For each simulated dataset, the multilevel model is fitted by means of the `lme` function from the `nlme` package (Pinheiro et al. 2019) and the effect of interest is tested (i.e., two-sided Wald test). In case of convergence problems, the app shows a warning message signaling the total number of replicates that failed

[3]Both methods differ in how they estimate the variance components of the model. ML ignores the uncertainty in the estimates of the fixed effects when estimating the variance components. As a result, the estimates of the variance components are biased when the sample size is small. REML estimates unbiased variance components by taking into account the degrees of freedom of the fixed effects estimates. Raudenbush & Bryk (2002) recommend to use REML when the number of participants is small.

[4]For each individual, the random slope is generated as follow: first we draw V_i from a Beta distribution with conditional mean $\mathrm{E}(V_i|i) = \frac{1+E(\gamma_{1i}|i)}{2}$ and conditional variance $\mathrm{Var}(V_i|i) = \frac{\sigma^2_{\nu_1}}{2}$. The random slope of participant i is computed as $\gamma_{1i} = 2V_i - 1$, and the random intercept as $\gamma_{1i} = \sigma_{\nu_0}\rho_{\nu_{01}}\frac{(\gamma_{1i}-E(\gamma_{1i}|i))}{\sigma_{\nu_1}} + \sqrt{1-\rho^2_{\nu_{01}}}\mathcal{Z} + E(\gamma_{0i}|i)$, where $\mathcal{Z}$ is drawn from a standard normal distribution.

to converge. Convergence issues in multilevel models arise when the estimated covariance matrix of the random effects is singular (see, Bates et al. 2015), and might be caused by not having enough observations within participants, a small number of participants, or scaling issues (see, e.g., Clark 2020). If this happens, we recommend to evaluate the following alternatives: increasing the number of participants, increasing the number of repeated measurements per person, centering predictors, or checking the specified values of the model parameters. Finally, we note that the simulation-based approach is computationally intensive, and therefore, may demand a lot of computational time. Depending on the number of participants, the number of observations per participant, the number of Monte Carlo replicates, the population model of interest, and the operating system, it may happen that the simulation can run for multiple hours. Therefore, while performing the power analysis, the app displays a message indicating for which number of participants power is currently being computed. Moreover, users can estimate the expected number of hours necessary to perform the simulation analysis by using the option *Estimate Computational Time*[5].

4.3 App output

For the effect of interest as well as all other fixed effects included in the model, the app provides a power curve, which shows how the estimated power varies as a function of sample size (i.e., the number of participants). The estimated power is computed as the proportion of Monte Carlo replicates in which the effect was significant (at the specified α-level). Furthermore, the app presents a summary of the results for each sample size. This summary includes power and measures to evaluate the estimation performance (see, Morris et al. 2019): the average of the estimates of each fixed effect; the bias (i.e., the difference between the average of the estimates and the true value); the standard error; the (1-α)% coverage proportion, computed as the proportion of Monte Carlo replicates for which the (1-α)% confidence interval includes the true value. Moreover, summary statistics are provided for the variance components of the within-individual errors (i.e., ρ_ϵ in the AR(1) error model and σ_ϵ) and for the random effects (i.e., standard deviations σ_{ν_0}, σ_{ν_1}, and correlations between the random effects $\sigma_{\nu_{01}}$). Finally, for the largest sample size considered, density plots and box-plots of the distribution of the estimated parameters are given.

5 Applications

In this section, we illustrate how the app can be used to perform a power analysis to decide on the number of participants needed to test three different research hypotheses. For all models, the value of a large number of model parameters has to be specified. We recommend to choose these values based on data from a pilot study, or based on existing IL studies

[5]To estimate the computational time, the app conducts a power analysis using ten replicates only. Next, the run time for ten replicates is used to estimate the run time for the total number of replicates.

with similar measures and designs (see, e.g., Lane & Hennes 2018) To this end, we will use information from a clinical dataset reported on by Heininga et al. (2019).

5.1 Dataset

The dataset includes 38 individuals that have been diagnosed with MDD (score of one on the *Diagnosis* variable) and 40 control subjects (score of zero). They all participated in an ESM study of 7 days, in which they were asked to repeatedly fill in a questionnaire containing 27 items measuring various constructs, including negative affect (i.e., *Affect*; 5 items that were averaged) and *Anhedonia* (1 item). Participants answered these items on a sliding scale ranging from 'not at all' on the left (0) to 'very much' on the right (100). The questions were semi-randomly presented ten times a day between 9:30 a.m. and 9:30 p.m. within intervals of 66 minutes. Therefore, the design included 70 measurement occasions per participant. Depressive symptoms (*Depression*) were measured before the ESM testing period based on the sum score of the items of the Quick Inventory of Depressive Symptomatology (i.e., QIDS; Rush et al. 2003).

5.2 Application 1: Power to estimate the effect of a level-2 predictor

Consider a researcher who is planning a study to test the hypothesis that *Depression* is positively related to negative affect and thus wants to run Model 2 (see Figure 1). The data will be collected using an IL design, including 70 measurement occasions per individual. How many participants does she need to involve?

To perform the simulation-based power analysis, we need to specify the parameter values of the model of interest. Pilot data or the results from previous studies examining the same hypothesis can be used to obtain appropriate values. Here, we will use the clinical dataset and apply Model 2 to get estimates of these parameters. The continuous level-2 predictor, *Depression*, is centered using the grand mean. Table 4 shows the estimated parameter values. Note that estimation of this model is not part of the app (i.e., this step has to be conducted separately). In the OSF page of the project `https://osf.io/vguey/`, we show how to obtain the parameter values of Model 2 using the clinical dataset.

Table 4: Illustration 1. Estimated parameters using the clinical dataset to estimate the effect of depressive symptoms on negative affect on individuals with Major Depressive Disorder.

	Notation	Model Parameters
Number of participants	N	38
Number of time points		70
Mean of the level-2 continuous variable (Depression)	μ_W	15.70
Standard deviation of the level-2 continuous variable (Depression)	σ_W	5.00
Fixed intercept	β_{00}	43.01
Effect of the level-2 continuous variable on the level-1 intercept	β_{01}	1.50
Standard deviation of the level-1 error	σ_ϵ	12.62
Autocorrelation of the level-1 error	ρ_ϵ	0.46
Standard deviation of the random intercept	σ_{ν_0}	12.90

Step 1: App input. We select Model 2 and fill in the values of the model parameters (see Figure 3). We indicate that we want to consider the following values for the number of participants: 15, 30, 45, 60, 80, 100. We set the number of measurements within each participant to 70. We specify the fixed effects: the fixed intercept β_{00} is set to 43.01, and the effect of the level-2 continuous variable β_{01} is set to 1.50. Next, we set the standard deviation σ_ϵ and autocorrelation ρ_ϵ of the within-individual errors, given by 12.62 and 0.46, respectively. The standard deviation of the random intercept σ_{ν_0} is set to 12.90. For the variable depression, we fix the value of the mean to 15.70 and the standard deviation to 5.00. We select the options *Center the level-2 variable W* and *Estimated AR(1) correlated errors*. In this and the following illustrations, we set the Type I error α to 0.05 and the number of Monte Carlo replicates to 1000, and we choose the option *Restricted Maximum Likelihood* when specifying the estimation method. Finally, we click on *Compute Power*. Due to the computationally intensive nature of a simulation-based power analysis, it will take multiple hours to obtain the results for the three applications.

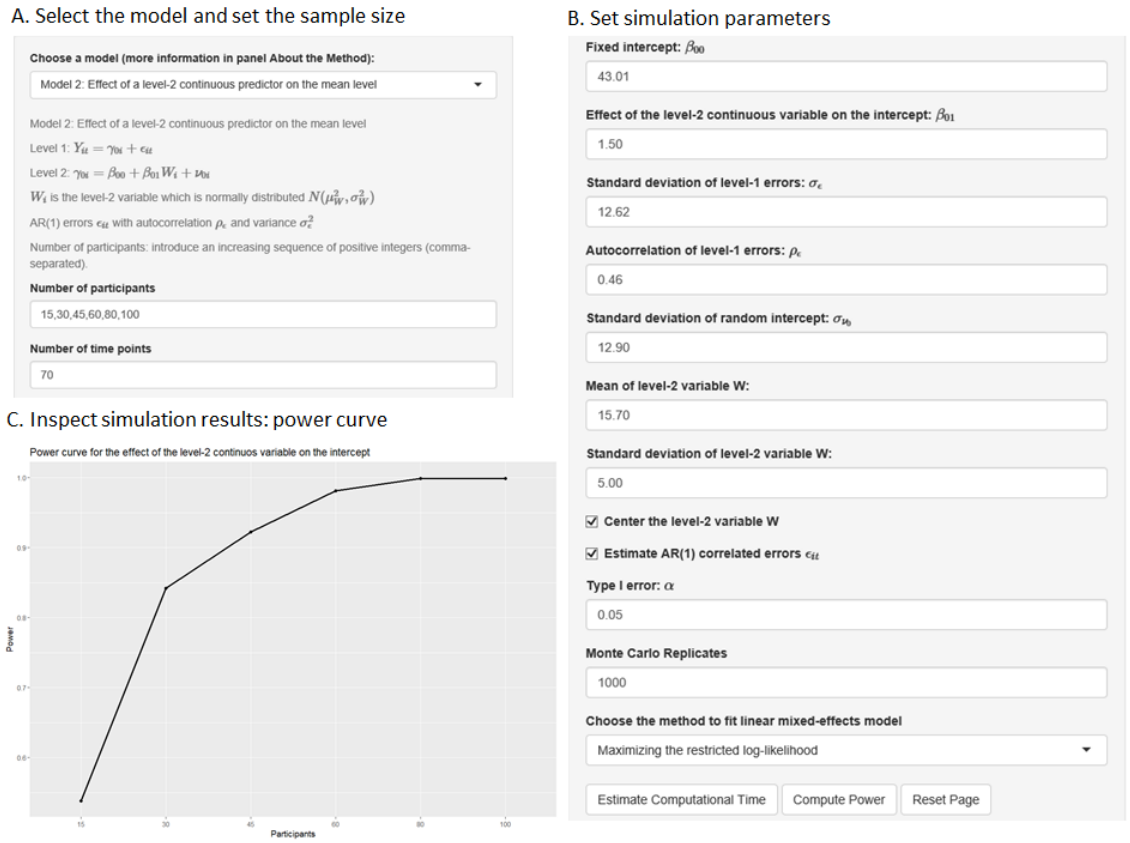

Figure 3: Illustration 1. Panel A shows a screenshot of the app, the user has to select Model 2 and set the sample size to estimate the effect of depressive symptoms on negative affect on individuals with Major Depressive Disorder. Next, the user has to set the value of the parameters of Model 2 (Panel B). Panel C shows the Power curve to estimate the effect of depression on negative affect.

Step 2: Visualize the power curve and inspect app output. The app provides the power curves as a function of the sample sizes indicated above. Figure 3 shows the estimated power curve to test the effect of depression on negative affect. We observe that when the number of participants is 15, the power for the effect of interest (i.e., $\beta_{01} = 1.50$) is 53.8%. This result implies that in only 538 out of the 1000 simulated datasets, the null hypothesis that depression does not have a significant effect on negative affect was rejected. We observe that when the number of participants increases, the power increases as well. Specifically, power larger than 80% is achieved when the number of participants is greater than 30.

The app also provides information about the distribution of the estimates of the fixed and random effects across the Monte Carlo replicates. Figure 4 shows the summary statistics

for the fixed effects. We see, for instance, that the coverage rate for β_{01} is close to 95%, indicating a satisfactory estimation of the 95% confidence interval. The app also calculates the power for the fixed intercept, although this is of little interest here.

	True value	Mean	Std.error	Bias	(1-alpha)% Coverage	Power
Fixed intercept N 15	43.01	43.0228	0.1089	0.0128	0.890	1.000
Fixed intercept N 30	43.01	42.8869	0.0755	-0.1231	0.902	1.000
Fixed intercept N 45	43.01	43.1233	0.0622	0.1133	0.941	1.000
Fixed intercept N 60	43.01	42.9715	0.0551	-0.0385	0.940	1.000
Fixed intercept N 80	43.01	43.0246	0.0460	0.0146	0.948	1.000
Fixed intercept N 100	43.01	43.0340	0.0408	0.0240	0.947	1.000
Effect of the level-2 continuous variable on the intercept N 15	1.50	1.5116	0.0229	0.0116	0.922	0.538
Effect of the level-2 continuous variable on the intercept N 30	1.50	1.5052	0.0159	0.0052	0.904	0.842
Effect of the level-2 continuous variable on the intercept N 45	1.50	1.4894	0.0133	-0.0106	0.947	0.922
Effect of the level-2 continuous variable on the intercept N 60	1.50	1.5095	0.0113	0.0095	0.941	0.981
Effect of the level-2 continuous variable on the intercept N 80	1.50	1.4923	0.0096	-0.0077	0.946	0.999
Effect of the level-2 continuous variable on the intercept N 100	1.50	1.5053	0.0086	0.0053	0.946	0.999

Figure 4: Illustration 1. Summary of the fixed effect across 1000 Monte Carlo replicates to estimate the effect of depressive symptoms on negative affect on individuals with Major Depressive Disorder.

5.3 Application 2: Power to detect the effect of a level-1 predictor

Now we turn to the effect of a level-1 predictor, anhedonia, on negative affect for individuals diagnosed with MDD, and thus to Model 3. To set the values of the model parameters, we again analysed the clinical dataset and obtained the results as shown in Table 5.

Table 5: Illustration 2. Estimated parameters using the clinical dataset to estimate the effect of anhedonia on negative affect on individuals with Major Depressive Disorder.

	Notation	Model Parameters
Number of participants	N	38
Number of time points		70
Mean of the level-1 continuous variable (anhedonia)	μ_X	51.60
Standard deviation of the level-1 continuous variable (anhedonia)	σ_X	23.70
Fixed intercept	β_{00}	42.90
Fixed Slope	β_{01}	0.13
Standard deviation of the level-1 error	σ_ϵ	12.00
Autocorrelation of the level-1 error	ρ_ϵ	0.43
Standard deviation of the random intercept	σ_{ν_0}	15.00
Standard deviation of the random slope	σ_{ν_1}	0.12
Correlation between the random intercept and the random slope	$\rho_{\nu_{01}}$	0.003

Step 1: App input. We select Model 3 and set the sample size. We will evaluate the power for the following numbers of participants: 15, 20, 30, 40, 60, 100, restricting the number of measurements within participants to 70. Subsequently, we specify the associated parameter values (see Figure 5). The fixed intercept β_{00} is 42.90 and the fixed slope β_{10} is 0.13. The standard deviation of the level-1 errors is 12, and the autocorrelation is 0.43. The standard deviation of the random intercept and random slope are 15.00 and 0.12, respectively. The correlation between the random effects is 0.003. The mean and standard deviation of the level-1 variable are 51.60 and 23.70, respectively. To guarantee that the fixed slope reflects the (average) within-person association between anhedonia and negative affect, we select the option to person-mean center the level-1 variable (i.e., *Person-center level-1 variable X*). Finally, to account for temporal dependencies, we choose the option *Estimate AR(1) correlated errors*.

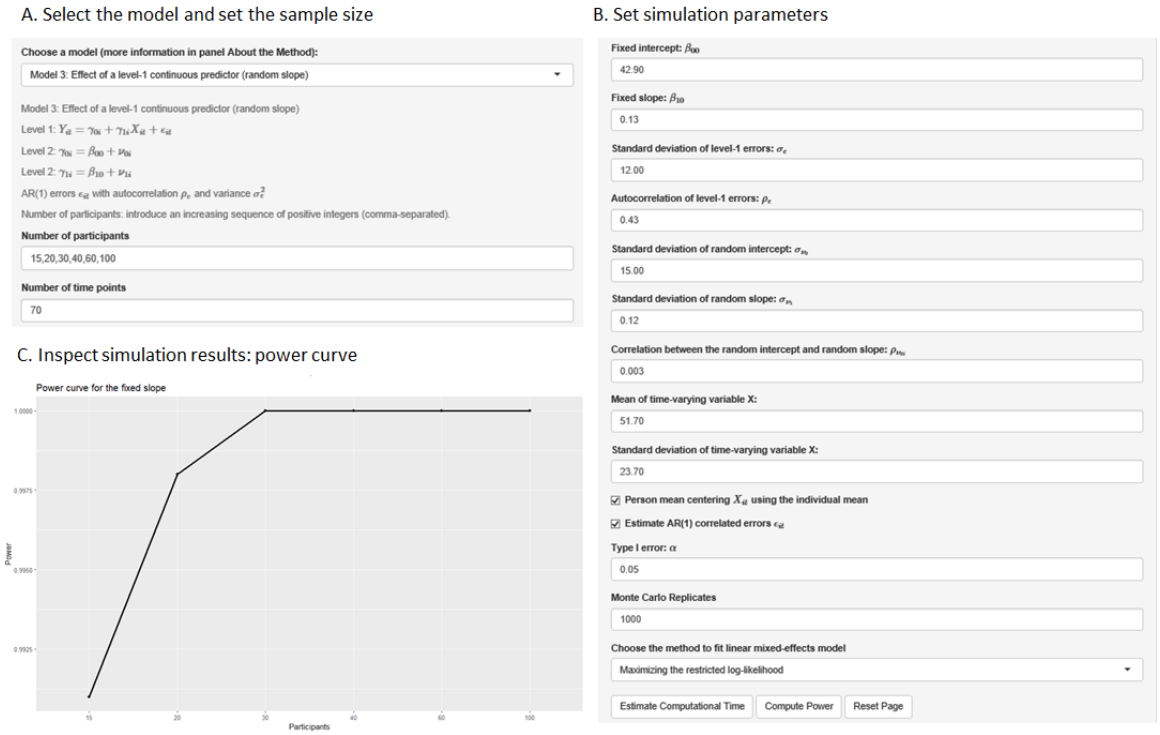

Figure 5: Illustration 2. Panel A shows a screenshot of the app, the user has to select Model 3 and set the sample size to estimate the effect of anhedonia on negative affect on individuals with Major Depressive Disorder. Next, the user has to set the value of the parameters of Model 3 (Panel B). Panel C shows the Power curve to estimate the effect of anhedonia on negative affect.

Step 2: Visualize the power curve and inspect app output. From the power

curve in Figure 5, we conclude that power is larger than 99% when there are more than 15 participants. Summary statistics of the fixed effects can be found in Figure 6. Figure 7 shows the summary statistics of the estimated parameters of the standard deviation and autocorrelation of the level-1 errors and the standard deviation and correlation between the random effects. We observe that when the number of participants increases, the bias of the estimates of σ_{ν_0}, σ_{ν_1} and $\rho_{\nu_{01}}$ diminished. Figure 8 shows the distribution of the estimated parameters across the Monte Carlo replicates when the number of participants is 100. We observe that when the number of participants is 100, the estimates of σ_{ν_0} and σ_{ν_1} are slightly negatively biased.

	True value	Mean	Std.error	Bias	(1-alpha)% Coverage	Power
Fixed intercept N 15	42.90	42.9567	0.1179	0.0567	0.939	1.000
Fixed intercept N 20	42.90	42.9296	0.1059	0.0296	0.934	1.000
Fixed intercept N 30	42.90	42.8422	0.0900	-0.0578	0.941	1.000
Fixed intercept N 40	42.90	42.9985	0.0754	0.0985	0.946	1.000
Fixed intercept N 60	42.90	42.9181	0.0596	0.0181	0.953	1.000
Fixed intercept N 100	42.90	42.9109	0.0471	0.0109	0.942	1.000
Fixed slope N 15	0.14	0.1413	0.0011	0.0013	0.925	0.991
Fixed slope N 20	0.14	0.1385	0.0009	-0.0015	0.940	0.998
Fixed slope N 30	0.14	0.1407	0.0008	0.0007	0.932	1.000
Fixed slope N 40	0.14	0.1392	0.0006	-0.0008	0.941	1.000
Fixed slope N 60	0.14	0.1407	0.0005	0.0007	0.944	1.000
Fixed slope N 100	0.14	0.1402	0.0004	0.0002	0.946	1.000

Figure 6: Illustration 2. Summary of the fixed effect across 1000 Monte Carlo replicates to estimate the effect of anhedonia on negative affect on individuals with Major Depressive Disorder.

	True value	Mean	Std.error	Bias
Standard deviation of the level-1 error N 15	12.000	11.9892	0.1179	0.0567
Standard deviation of the level-1 error N 20	12.000	12.0194	0.1059	0.0296
Standard deviation of the level-1 error N 30	12.000	11.9970	0.0900	-0.0578
Standard deviation of the level-1 error N 40	12.000	11.9944	0.0754	0.0985
Standard deviation of the level-1 error N 60	12.000	11.9999	0.0596	0.0181
Standard deviation of the level-1 error N 100	12.000	12.0045	0.0471	0.0109
Autocorrelation of the level-1 error N 15	0.430	0.4292	0.0009	-0.0008
Autocorrelation of the level-1 error N 20	0.430	0.4304	0.0009	0.0004
Autocorrelation of the level-1 error N 30	0.430	0.4292	0.0007	-0.0008
Autocorrelation of the level-1 error N 40	0.430	0.4292	0.0006	-0.0008
Autocorrelation of the level-1 error N 60	0.430	0.4293	0.0005	-0.0007
Autocorrelation of the level-1 error N 100	0.430	0.4302	0.0004	0.0002
Standard deviation of the random intercept N 15	15.000	14.7090	0.0921	-0.2910
Standard deviation of the random intercept N 20	15.000	14.7163	0.0792	-0.2837
Standard deviation of the random intercept N 30	15.000	14.8964	0.0646	-0.1036
Standard deviation of the random intercept N 40	15.000	14.9527	0.0543	-0.0473
Standard deviation of the random intercept N 60	15.000	14.9555	0.0437	-0.0445
Standard deviation of the random intercept N 100	15.000	14.9156	0.0348	-0.0844
Standard deviation of the random slope N 15	0.120	0.1181	0.0009	-0.0019
Standard deviation of the random slope N 20	0.120	0.1159	0.0008	-0.0041
Standard deviation of the random slope N 30	0.120	0.1182	0.0006	-0.0018
Standard deviation of the random slope N 40	0.120	0.1188	0.0005	-0.0012
Standard deviation of the random slope N 60	0.120	0.1198	0.0004	-0.0002
Standard deviation of the random slope N 100	0.120	0.1194	0.0003	-0.0006
Correlation between the random intercept and the random slope N 15	0.003	-0.0097	0.0095	-0.0127
Correlation between the random intercept and the random slope N 20	0.003	-0.0160	0.0083	-0.0190
Correlation between the random intercept and the random slope N 30	0.003	-0.0125	0.0065	-0.0155
Correlation between the random intercept and the random slope N 40	0.003	0.0124	0.0058	0.0094
Correlation between the random intercept and the random slope N 60	0.003	0.0035	0.0045	0.0005
Correlation between the random intercept and the random slope N 100	0.003	0.0048	0.0037	0.0018

Figure 7: Illustration 2. Summary of the standard deviation and autocorrelation of the level-1 errors, and standard errors and correlation of the random effects across 1000 Monte Carlo replicates.

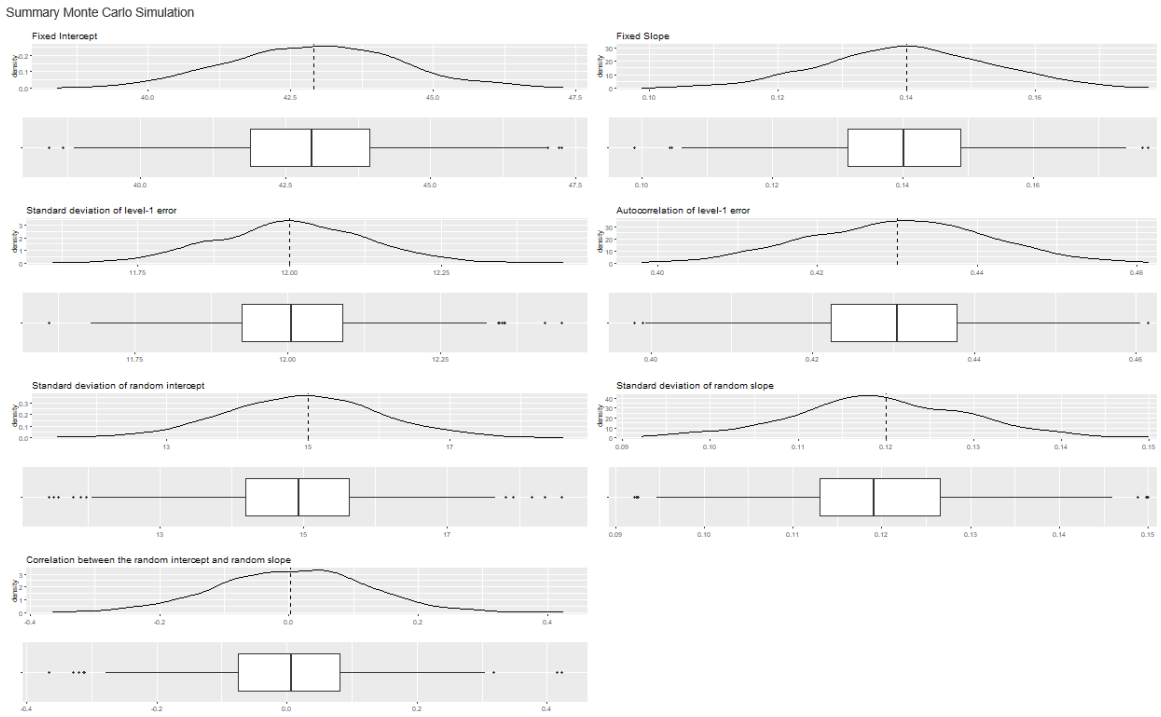

Figure 8: Illustration 2. Distribution of the estimate parameters across 1000 Monte Carlo replicates when the number of participants is 100. Dashed lines are the true model parameters.

5.4 Application 3: Power to detect the differences in the autoregressive effects between two groups

Finally, we focus on the difference in the autoregressive effect of negative affect between individuals diagnosed with MDD and control subjects and thus on Model 10. As in the previous examples, we use the clinical dataset to obtain estimates of the parameter values, shown in Table 6.

Table 6: Illustration 3. Estimated parameters using the clinical dataset to estimate differences in the autoregressive effect of negative affect between individuals with Major Depressive Disorder and control subjects.

	Notation	Model Parameters
Number of participants in Group 0 (i.e., reference group)	N_0	40
Number of participants in Group 1	N_1	38
Number of time points		70
Fixed intercept	β_{00}	10.20
Difference in the fixed intercept between the reference group and group 1	β_{01}	32.40
Fixed slope (i.e., autoregressive effect)	β_{10}	0.20
Difference in the fixed slope between the reference group and group 1	β_{11}	0.10
Standard deviation of level-1 errors	σ_ϵ	8.80
Standard deviation of the random intercept	σ_{ν_0}	11.50
Standard deviation of the random slope	σ_{ν_1}	0.16
Correlation between the random intercept and the random slope	$\rho_{\nu_{10}}$	0.265

Step 1: App input. We select *Model 10: Multilevel AR(1) model - Group differences in the autoregressive effects*. The number of participants in the reference group (i.e., healthy controls) and the number of participants in Group 1 (i.e., MDD) are set to 20, 40, 60, 80, 100, 200 and 250, respectively, and the number of measurements within participants to 70. We specify the parameter values as follows (see Figure 9): The fixed intercept β_{00} is 10.20 and the difference in the fixed intercept between the two groups (β_{01}) is 32.40. The autoregressive effect β_{10} is 0.20. The difference in the autoregressive effect between the two groups β_{11} is 0.10. The standard deviation of the level-1 errors is 8.80. The standard deviation of the random intercept and random slope are 11.50 and 0.16, respectively. The correlation between the random effects is 0.265. We person-mean center the lagged outcome variable.

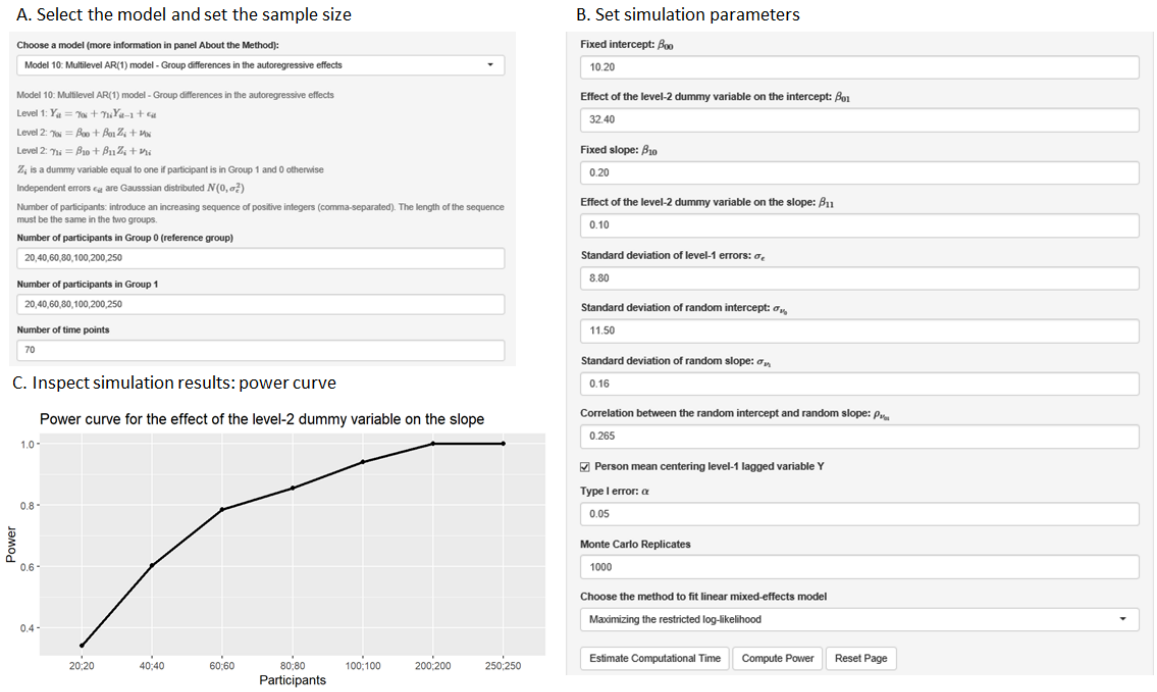

Figure 9: Illustration 3. Panel A shows a screenshot of the app, the user has to select Model 10 and set the sample size to estimate differences in the autoregressive effect of negative affect between individuals with Major Depressive Disorder and control subjects. Next, the user has to set the value of the parameters of Model 10 (Panel B). Panel C shows the Power curve to estimate differences in the autoregressive effect of negative affect between the two groups.

Step 2: Visualize the power curve and inspect app output. Figure 9 shows the estimated power curve. The power to test the difference in the autoregressive effect (β_{11}) between the two groups is larger than 80% when there are 80 participants diagnosed with MDD and 80 healthy controls. In Figure 10, we observe that there is a downward bias in the estimated value of the fixed slope in the reference group β_{10}. Furthermore, when the number of participants increases, the 95% coverage proportion of the fixed slope diminishes. This is related to the bias in the estimate of the fixed slope and narrower confidence intervals (i.e., smaller standard errors) when the sample size increases. This result is in line with Hamaker & Grasman (2015) who showed for this model that the estimated fixed slope is negatively biased when the lagged dependent variable is person-mean centered.

	True value	Mean	Std.error	Bias	(1-alpha)% Coverage	Power
Fixed intercept N(Group=1) 20 N(Group=0) 20	10.2	10.0914	0.0809	-0.1086	0.947	0.971
Fixed intercept N(Group=1) 40 N(Group=0) 40	10.2	10.1745	0.0568	-0.0255	0.949	1.000
Fixed intercept N(Group=1) 60 N(Group=0) 60	10.2	10.2157	0.0448	0.0157	0.964	1.000
Fixed intercept N(Group=1) 80 N(Group=0) 80	10.2	10.1874	0.0382	-0.0126	0.956	1.000
Fixed intercept N(Group=1) 100 N(Group=0) 100	10.2	10.2246	0.0362	0.0246	0.945	1.000
Fixed intercept N(Group=1) 200 N(Group=0) 200	10.2	10.2123	0.0247	0.0123	0.956	1.000
Fixed intercept N(Group=1) 250 N(Group=0) 250	10.2	10.1893	0.0221	-0.0107	0.965	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 20 N(Group=0) 20	32.4	32.5153	0.1131	0.1153	0.952	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 40 N(Group=0) 40	32.4	32.4853	0.0786	0.0853	0.956	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 60 N(Group=0) 60	32.4	32.3073	0.0637	-0.0927	0.963	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 80 N(Group=0) 80	32.4	32.4412	0.0565	0.0412	0.954	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 100 N(Group=0) 100	32.4	32.3820	0.0503	-0.0180	0.952	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 200 N(Group=0) 200	32.4	32.3501	0.0360	-0.0499	0.951	1.000
Effect of the level-2 dummy variable on the intercept N(Group=1) 250 N(Group=0) 250	32.4	32.4265	0.0320	0.0265	0.960	1.000
Fixed slope N(Group=1) 20 N(Group=0) 20	0.2	0.1796	0.0014	-0.0204	0.920	0.979
Fixed slope N(Group=1) 40 N(Group=0) 40	0.2	0.1790	0.0010	-0.0210	0.889	1.000
Fixed slope N(Group=1) 60 N(Group=0) 60	0.2	0.1779	0.0008	-0.0221	0.852	1.000
Fixed slope N(Group=1) 80 N(Group=0) 80	0.2	0.1785	0.0007	-0.0215	0.829	1.000
Fixed slope N(Group=1) 100 N(Group=0) 100	0.2	0.1783	0.0006	-0.0217	0.793	1.000
Fixed slope N(Group=1) 200 N(Group=0) 200	0.2	0.1781	0.0004	-0.0219	0.653	1.000
Fixed slope N(Group=1) 250 N(Group=0) 250	0.2	0.1785	0.0004	-0.0215	0.586	1.000
Effect of the level-2 dummy variable on the slope N(Group=1) 20 N(Group=0) 20	0.1	0.0950	0.0019	-0.0050	0.950	0.342
Effect of the level-2 dummy variable on the slope N(Group=1) 40 N(Group=0) 40	0.1	0.0965	0.0014	-0.0035	0.935	0.603
Effect of the level-2 dummy variable on the slope N(Group=1) 60 N(Group=0) 60	0.1	0.0969	0.0011	-0.0031	0.945	0.786
Effect of the level-2 dummy variable on the slope N(Group=1) 80 N(Group=0) 80	0.1	0.0948	0.0010	-0.0052	0.946	0.856
Effect of the level-2 dummy variable on the slope N(Group=1) 100 N(Group=0) 100	0.1	0.0953	0.0009	-0.0047	0.943	0.941
Effect of the level-2 dummy variable on the slope N(Group=1) 200 N(Group=0) 200	0.1	0.0972	0.0006	-0.0028	0.951	1.000
Effect of the level-2 dummy variable on the slope N(Group=1) 250 N(Group=0) 250	0.1	0.0961	0.0006	-0.0039	0.944	1.000

Figure 10: Illustration 3. Summary of the fixed effect across 1000 Monte Carlo replicates to estimate differences in the autoregressive effect of negative affect between individuals with Major Depressive Disorder and control subjects.

6 Discussion

Intensive longitudinal designs allow studying within-person psychological dynamics. When multiple participants are included in an IL study, multilevel models are a powerful approach to capture these within-person processes as well as inter-individual differences therein. When planning IL studies, it is obviously essential to collect a sufficient amount of data to ensure reliable estimates and sufficient power. In this paper, we focused on the number of participants that are needed to obtain sufficient statistical power for testing hypotheses about specific parameters of the multilevel models that are popular in IL studies. These power questions cannot be addressed by existing software for standard multilevel models, as standard models do not account for temporal dependencies in the outcome variable. Therefore, we presented a Shiny app developed in `R` to compute power for models with an AR(1) error structure or with the lagged outcome variable as a predictor in a simulation-based

way. The app yields power curves that show how estimated power varies as a function of the number of participants. In the following, we will discuss limitations of the current version of the Shiny app as well as potential extensions.

6.1 Accommodating uncertainty about the hypothesized model parameters

Using simulation-based power analysis for multilevel models is challenging, in that users have to specify all the parameter values of the population model of interest. Following Lane & Hennes (2018) and Maxwell et al. (2008), we recommend to base these values on a literature review, on data from a pilot study (as we did by means of the clinical dataset), or on previously conducted studies with similar measures and designs. Having said that, we acknowledge that the second and third approaches may imply that data are used from a small or unrepresentative sample which may produce biased estimates as input for the power analysis (e.g., Albers & Lakens 2018). Therefore, a more robust power-calculation approach would account for uncertainty regarding the hypothesized model parameters. This can be achieved by performing a sensitivity analysis in which the values of the model parameters are varied to some extent (e.g., Lane & Hennes 2018, Wang & Rhemtulla 2020). This way one can assess whether and to which extent using different possible parameter values influences the obtained power results. We note however that the current version of the app cannot display power curves as a function of sets of different plausible parameter values. Therefore, users have to perform a sensitivity analysis by conducting separate power analyses for each set of parameter values.

6.2 Selection of the numbers of measurement occasions and persons

When applying multilevel modeling to IL data, the obtained power is a function of both the number of measurement occasions and the number of participants. In this paper, we targeted the number of participants, however, and kept the number and spacing of the measurement occasions fixed. While this worked well for the research questions that we considered in this paper (i.e., we consider a relatively high number of measurement occasions), it is important to note that for other research questions increasing the number of measurement occasions might be called for. It makes, for instance, sense that when inter-individual differences in within-person effects are of interest, the number of measurement occasions should be high as well. Indeed, earlier work of de Haan-Rietdijk et al. (2017), Krone et al. (2016), Liu (2017), Schultzberg & Muthén (2018), and Timmons & Preacher (2015) has demonstrated the effect that the number and spacing of the measurement occasions can have on estimation accuracy of multilevel approaches for IL data. Thus, how to best plan for adequate power depends on where power vulnerabilities are (see e.g., Lane & Hennes 2018).

The question thus is what users have to do when they are also interested in studying how the number of measurement occasions might impact power. While one cannot get power curves for that from the app, a relatively simple solution consists of repeatedly simu-

lating with different numbers of measurement occasions while keeping the vector of sample sizes fixed. However, adding more participants or more measurements per participant may come with different costs and burdens for both researchers and participants. Therefore, researchers designing IL studies might be interested in balancing both sample size components to optimize power and minimize costs and participant burden. One way to achieve this is to obtain a set of combinations (i.e., of the number of participants and the number of measurement occasions per participant) that yield equal power. Next, one selects the combination that optimizes budgetary restrictions or feasibility. We, however, note that the current version of the app does not allow the users to obtain such a set of combinations that produce equivalent power. We, therefore, recommend the reader to refer to Brandmaier et al. (2015), Moerbeek (2011), and von Oertzen (2010), for a broader discussion on this topic.

6.3 Other remarks and future extensions

In the current tutorial, we illustrated how to use the app to estimate the number of participants to answer three specific research questions. For each research question, we focused on computing power for a single (fixed) effect. Yet, the app also provides the power curve for all other fixed effects included in the model. Therefore, in studies that involve testing multiple fixed effects, the number of participants should be large enough to detect all these effects with high power.

Even though the app is already quite extensive and includes no less than eleven models, many other models could be included still. For instance, in many applications, the objective is to assess the significance of the random effects. This is not possible in the current version of the app. As another example, we now focused on two-levels models in which repeated measurements are nested within individuals. In the future, the proposed approach can be extended to three-level models (i.e., occasions nested within days, which in turn are nested within individuals). Three-level models are especially relevant if the dynamics under study differ systematically across days. Ignoring these differences could affect the reliability of the estimated results (de Haan-Rietdijk et al. 2016) and consequently the power.

We also highlight that the proposed app simulates and analyzes data under the assumption that the measurement occasions are equally spaced and contain no missing data. In IL research, participants might not respond at some measurement occasions or during night breaks (e.g., Fuller-Tyszkiewicz et al. 2013, Santangelo et al. 2014, Stone et al. 2003). While missingness might sometimes occur completely at random, it might be systematic in other cases (e.g., associated with certain affective states or certain times or contexts), which can lead to unreliable estimates (Courvoisier et al. 2012). To account for this, it would be useful to extend the simulation approach to study the effect of different types of missing data and attrition on power. For instance, when data can be assumed to be completely missing at random, users could simply specify the expected number of completed measurement occasions. Studying the effects of other mechanisms of missingness is more involved however, because the missing data mechanism has to be fully specified in order to simulate data.

Finally, we would like to highlight that power is not the only criterion to base sample size

selection on. Aside from maximizing the likelihood that a hypothesized effect in a population is detected, researchers might, for instance, be interested in increasing the precision of an estimate by controlling the width of the confidence interval of interest (e.g., Maxwell et al. 2008). Bearing in mind that sample size planning is important for two related objectives power and precision, our simulation-based approach could be extended in this direction, allowing users to additionally select the sample size that yields a targeted confidence interval width.

7 Conclusion

The current study introduced a Shiny app to select the number of participants in IL designs. The application performs simulation-based power analysis to detect effects in multilevel models. We hope that the application contributes to good research practices by allowing rigorous sample size planning for IL studies, which is of crucial importance to increase the reliability and replicability of psychological research.

Author Contributions

GL, JKA, WV, EC conceptualized the application. GL, JKA, EC conceptualized the tutorial. GL developed the Shiny app. GL, ED analyzed and interpret the clinical dataset. GL, JKA, EC wrote the manuscript. ED, IMG, WV provided critical revisions on the manuscript and Shiny app.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

The supplemental material for this paper is available at: https://osf.io/vguey/.

Funding

The research presented in this article was supported by research grants from the Fund for Scientific Research-Flanders (FWO, Project No. G.074319N) and from the Research Council of KU Leuven (C14/19/054) awarded to Eva Ceulemans.

Acknowledgements

We would like to thank our colleagues within the Research Group of Quantitative Psychology and Individual Differences and the Center for Contextual Psychiatry, particularly Leonie Cloos, for testing the app and providing useful feedback.

Prior versions

This manuscript has been posted online as a pre-print: https://psyarxiv.com/dq6ky/.

References

Albers, C. & Lakens, D. (2018), 'When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias', *Journal of Experimental Social Psychology* **74**, 187–195.

Arend, M. G. & Schäfer, T. (2019), 'Statistical power in two-level models: A tutorial based on Monte Carlo simulation.', *Psychological Methods* **24**(1), 1–19.

Astivia, O. L. O., Gadermann, A. & Guhn, M. (2019), 'The relationship between statistical power and predictor distribution in multilevel logistic regression: a simulation-based approach', *BMC Medical Research Methodology* **19**(1), 97–117.

Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. (2015), 'Parsimonious mixed models', *arXiv preprint arXiv:1506.04967* .

Bolger, N. (2011), Power analysis for intensive longitudinal studies, *in* N. Bolger, G. Stadler & J.-P. Laurenceau, eds, 'Handbook of research methods for studying daily life', Guilford, New York, pp. 285–301.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C. & Lindenberger, U. (2015), 'LIFESPAN: a tool for the computer-aided design of longitudinal studies', *Frontiers in Psychology* **6**, 272.

Brose, A., Schmiedek, F., Koval, P. & Kuppens, P. (2015), 'Emotional inertia contributes to depressive symptoms beyond perseverative thinking', *Cognition and Emotion* **29**(3), 527–538.

Browne, W. J., Lahi, M. G. & Parker, R. M. (2009), 'A guide to sample size calculations for random effect models via simulation and the MLPowSim software package', *Bristol, United Kingdom: University of Bristol* .

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J. et al. (2019), *Shiny: Web Application Framework for R*. R package version 1.3.2.

Clark, M. (2020), 'Michael clark: Convergence problems'.
URL: *https://m-clark.github.io/posts/2020-03-16-convergence/*

Cohen, J. (1988), 'Statistical power analysis for the behavioral sciences. 2nd'.

Cools, W., Van den Noortgate, W. & Onghena, P. (2008), 'ML-DEs: A program for designing efficient multilevel studies', *Behavior Research Methods* **40**(1), 236–249.

Courvoisier, D. S., Eid, M. & Lischetzke, T. (2012), 'Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics.', *Psychological Assessment* **24**(3), 713–720.

de Haan-Rietdijk, S., Kuppens, P. & Hamaker, E. L. (2016), 'What's in a day? a guide to decomposing the variance in intensive longitudinal data', *Frontiers in Psychology* **7**, 891.

de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L. & Hamaker, E. L. (2017), 'Discrete- vs. continuous-time modeling of unequally spaced experience sampling method data', *Frontiers in Psychology* **8**, 1849.

De Jong, K., Moerbeek, M. & Van der Leeden, R. (2010), 'A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects', *Psychotherapy Research* **20**(3), 273–284.

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B. & Kuppens, P. (2018), 'The bipolarity of affect and depressive symptoms.', *Journal of Personality and Social Psychology* **114**(2), 323–341.

Enders, C. K. & Tofighi, D. (2007), 'Centering predictor variables in cross-sectional multilevel models: a new look at an old issue.', *Psychological Methods* **12**(2), 121–138.

Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M. & Mills, J. (2013), 'Does the burden of the experience sampling method undermine data quality in state body image research?', *Body Image* **10**(4), 607–613.

Goldstein, H., Healy, M. J. & Rasbash, J. (1994), 'Multilevel time series models with applications to repeated measures data', *Statistics in Medicine* **13**(16), 1643–1655.

Green, P. & MacLeod, C. J. (2016), 'SIMR: an R package for power analysis of generalized linear mixed models by simulation', *Methods in Ecology and Evolution* **7**(4), 493–498.

Hamaker, E. L. & Grasman, R. P. (2015), 'To center or not to center? investigating inertia with a multilevel autoregressive model', *Frontiers in Psychology* **5**, 1492.

Hamaker, E. L., Kuiper, R. M. & Grasman, R. P. (2015), 'A critique of the cross-lagged panel model.', *Psychological Methods* **20**(1), 102–116.

Hamilton, J. D. (1994), *Time series analysis*, Vol. 2, Princeton New Jersey.

Hedeker, D., Gibbons, R. D. & Waternaux, C. (1999), 'Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups', *Journal of Educational and Behavioral Statistics* **24**(1), 70–93.

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J. & Kuppens, P. (2019), 'The dynamical signature of anhedonia in major depressive disorder: positive emotion dynamics, reactivity, and recovery', *BMC Psychiatry* **19**(1), 59.

Ioannidis, J. P. (2005), 'Why most published research findings are false', *PLoS Medicine* **2**(8), e124.

Kirtley, O., Lafit, G., Achterhof, R., Hiekkaranta, A. P. & Myin-Germeys, I. (In press), 'Making the black box transparent: A template and tutorial for (pre-) registration of studies using Experience Sampling Methods (ESM)', *Advances in Methods and Practices in Psychological Science* .

Koval, P., Pe, M. L., Meers, K. & Kuppens, P. (2013), 'Affect dynamics in relation to depressive symptoms: Variable, unstable or inert?', *Emotion* **13**(6), 1132–1141.

Krone, T., Albers, C. J. & Timmerman, M. E. (2016), 'Comparison of estimation procedures for multilevel AR(1) models', *Frontiers in Psychology* **7**, 486.

Kuppens, P., Allen, N. B. & Sheeber, L. B. (2010), 'Emotional inertia and psychological maladjustment', *Psychological Science* **21**(7), 984–991.

Kuppens, P. & Verduyn, P. (2015), 'Looking at emotion regulation through the window of emotion dynamics', *Psychological Inquiry* **26**(1), 72–79.

Landau, S. & Stahl, D. (2013), 'Sample size and power calculations for medical studies by simulation when closed form expressions are not available', *Statistical Methods in Medical Research* **22**(3), 324–345.

Lane, S. P. & Hennes, E. P. (2018), 'Power struggles: Estimating sample size for multilevel relationships research', *Journal of Social and Personal Relationships* **35**(1), 7–31.

Liu, S. (2017), 'Person-specific versus multilevel autoregressive models: Accuracy in parameter estimates at the population and individual levels', *British Journal of Mathematical and Statistical Psychology* **70**(3), 480–498.

Maas, C. J. & Hox, J. J. (2005), 'Sufficient sample sizes for multilevel modeling.', *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* **1**(3), 86–92.

Mathieu, J. E., Aguinis, H., Culpepper, S. A. & Chen, G. (2012), 'Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling.', *Journal of Applied Psychology* **97**(5), 951–966.

Maxwell, S. E., Kelley, K. & Rausch, J. R. (2008), 'Sample size planning for statistical power and accuracy in parameter estimation', *Annual Review of Psychology* **59**, 537–563.

Moerbeek, M. (2011), 'The effects of the number of cohorts, degree of overlap among cohorts, and frequency of observation on power in accelerated longitudinal designs', *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* **7**(1), 11–24.

Moerbeek, M. & Maas, C. J. (2005), 'Optimal experimental designs for multilevel logistic models with two binary predictors', *Communications in Statistics—Theory and Methods* **34**(5), 1151–1167.

Moerbeek, M., van Breukelen, G. J. & Berger, M. P. (2000), 'Design issues for experiments in multilevel populations', *Journal of Educational and Behavioral Statistics* **25**(3), 271–284.

Moerbeek, M., Van Breukelen, G. J. & Berger, M. P. (2001), 'Optimal experimental designs for multilevel logistic models', *Journal of the Royal Statistical Society: Series D (The Statistician)* **50**(1), 17–30.

Molenaar, P. C. (2004), 'A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever', *Measurement* **2**(4), 201–218.

Morris, T. P., White, I. R. & Crowther, M. J. (2019), 'Using simulation studies to evaluate statistical methods', *Statistics in Mdicine* **38**(11), 2074–2102.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J. & Ioannidis, J. P. (2017), 'A manifesto for reproducible science', *Nature Human Behaviour* **1**(1), 0021.

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W. & Reininghaus, U. (2018), 'Experience sampling methodology in mental health research: new insights and technical developments', *World Psychiatry* **17**(2), 123–132.

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P. & van Os, J. (2009), 'Experience sampling research in psychopathology: opening the black box of daily life', *Psychological Medicine* **39**(9), 1533–1547.

Myin-Germeys, I., Peeters, F., Havermans, R., Nicolson, N., DeVries, M., Delespaul, P. & Van Os, J. (2003), 'Emotional reactivity to daily life stress in psychosis and affective disorder: an experience sampling study', *Acta Psychiatrica Scandinavica* **107**(2), 124–131.

Myin-Germeys, I., van Os, J., Schwartz, J. E., Stone, A. A. & Delespaul, P. A. (2001), 'Emotional reactivity to daily life stress in psychosis', *Archives of General Psychiatry* **58**(12), 1137–1144.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. et al. (2019), *Linear and Nonlinear Mixed Effects Models*. R package version 3.1-141.

R Core Team (2013), 'R: A language and environment for statistical computing'.

Raudenbush, S. W. (1997), 'Statistical analysis and optimal design for cluster randomized trials.', *Psychological Methods* **2**(2), 173–185.

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

Raudenbush, S. W. & Liu, X.-F. (2001), 'Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change.', *Psychological Methods* **6**(4), 387–401.

RStudio Team (2015), *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R. et al. (2003), 'The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-c), and self-report (QIDS-sr): a psychometric evaluation in patients with chronic major depression', *Biological Psychiatry* **54**(5), 573–583.

Santangelo, P., Bohus, M. & Ebner-Priemer, U. W. (2014), 'Ecological momentary assessment in borderline personality disorder: a review of recent findings and methodological challenges', *Journal of Personality Disorders* **28**(4), 555–576.

Schultzberg, M. & Muthén, B. (2018), 'Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling', *Structural Equation Modeling: A Multidisciplinary Journal* **25**(4), 495–515.

Snijders, T. A. & Bosker, R. J. (1993), 'Standard errors and sample sizes for two-level research', *Journal of Educational Statistics* **18**(3), 237–259.

Snijders, T. A. & Bosker, R. J. (2011), *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, Sage.

Stone, A. A., Broderick, J. E., Schwartz, J. E., Shiffman, S., Litcher-Kelly, L. & Calvanese, P. (2003), 'Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction', *Pain* **104**(1-2), 343–351.

Szucs, D. & Ioannidis, J. P. (2017), 'Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature', *PLoS Biology* **15**(3), e2000797.

Timmons, A. C. & Preacher, K. J. (2015), 'The importance of temporal design: How do measurement intervals affect the accuracy and efficiency of parameter estimates in longitudinal research?', *Multivariate Behavioral Research* **50**(1), 41–55.

Trull, T. J. & Ebner-Priemer, U. W. (2020), 'Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices.', *Journal of Abnormal Psychology* **129**(1), 56–63.

von Oertzen, T. (2010), 'Power equivalence in structural equation modelling', *British Journal of Mathematical and Statistical Psychology* **63**(2), 257–272.

Wang, C., Hall, C. B. & Kim, M. (2015), 'A comparison of power analysis methods for evaluating effects of a predictor on slopes in longitudinal designs with missing data', *Statistical Methods in Medical Research* **24**(6), 1009–1029.

Wang, L. P., Hamaker, E. & Bergeman, C. (2012), 'Investigating inter-individual differences in short-term intra-individual variability.', *Psychological Methods* **17**(4), 567–581.

Wang, Y. A. & Rhemtulla, M. (2020), 'Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial', *Advances in Methods and Practices in Psychological Science* .

Zhang, Z. (2014), 'Monte Carlo based statistical power analysis for mediation models: Methods and software', *Behavior Research Methods* **46**(4), 1184–1198.

Zhang, Z. & Wang, L. (2009), 'Statistical power analysis for growth curve models using SAS', *Behavior Research Methods* **41**(4), 1083–1094.