KU Leuven

Faculty of Psychology and Educational Sciences

THE EFFECT OF MISSING DATA ON THE ESTIMATION BIAS, VARIANCE, AND STATISTICAL POWER IN MULTILEVEL AUTOREGRESSIVE(1) MODELS

Master's thesis submitted for the
degree of Master of Science in
Master of Psychology: Theory and
Research by

Benjamín Šimsa

Supervisor: Prof. Dr. Eva Ceulemans
Co-supervisors: Dr. Ginette Lafit, Jordan Revol

2023

Contents

Introduction

IDK INTRO SENTENCE. In the recent years, the focus in diverse subfields of psychology has been shifting towards complexity, dynamics and within-person perspective in psychology (Hamaker, 2012). Among other things, this shift has been facilitated by the growing availability of smartphones and wearables, which allow researchers to use the Experience Sampling Method (ESM) to collect intensive longitudinal data with a high level of ecological validity (Myin-Germeys et al., 2018). Intensive longitudinal data consist of repeated (self-report) measurements nested within individual participants (Larson & Csikszentmihalyi, 2014). The use of intensive longitudinal data considerably broadens the extent of research questions psychological researchers can investigate and statistical analyses they can conduct. Importantly, the multilevel structure of intensive longitudinal data allows researchers to investigate both within-person dynamic processes, and the individual differences therein between persons (CITE).One of the new research avenues that emerged with the growing popularity of intensive longitudinal data is the study of affect dynamics (for an overview, see Houben et al. (2015)). The dynamic measures make use of the structure of intensive longitudinal data and take the fluctuating nature of affect/emotions into account.

Apart from dynamic measures such as emotional variance or instability, inertia (operationalised as autoregression in psychological studies using ESM) emerged as a dynamic measure that can contribute to the knowledge about emotional regulation (Kuppens & Verduyn, 2017). The autoregressive (AR) parameter quantifies the degree to which the value of a process at time t is influenced by the lagged variable (i. e., the process value at a previous timepoint; usually, *t-1*). In other words, the AR parameter allows us to approximate to what degree the intensity of an emotion/affect is carried onto the intensity of the emotional experience at the subsequent timepoint (for the same emotion), and to what degree the process is influenced by other factors. The part of the process value that is not explained by the lagged variable is considered to be caused by *innovation*: the ensemble of (both within- and between-person factors) that have influenced the process at a given timepoint (Ariens et al., 2020). Inertia is usually modeled using a multilevel autoregressive (MLAR) model (Koval et al., 2021). Usually, the model has two levels: the wthin-person Level 1, in which the intra-person autoregressive process is modeled, and the between-person Level 2, in which the individual differences in the autoregressive parameter and intercept are modelled. A more detailed description of the model is described in the Methods section.

Emotional inertia and psychological well-being

Although some degree of inertia is to be expected in human emotional experiences, a high level of emotional inertia (i. e., a high temporal persistence of emotional states) has been linked to psychological maladjustment (Kuppens et al., 2010). A negative emotion process with a high inertia can get caught in a self-reinforced feedback loop (also called *critical slowing down,* Leemput 2014). This causes the process to be partially resistant to both external influences and inner processes, including emotional regulation (Koval et al., 2015).

The evidence about the association between emotional inertia and the well-being/psychopathology spectrum has grown steadily over the last two decades. A recent meta-analysis indicated an association between emotioal inertia (of both positive and negative emotions) and psychological well-being/psychopathology (Houben et al., 2015). More specifically, higher emotional inertia has been linked to lower well-being and higher ocurrence of depressive symptoms(Brose et al., 2015), bipolar disorder (Mneimne et al., 2018), and lower response of depression and anxiety symptoms to cognitive-behavioral therapy (Bosley et al., 2019).

However, a recent study demonstrated that when the affect intensity (i.e., the mean of the affective process) is taken into account in the regression model, the dynamic measures such as inertia, instability and variability no longer meaningfully contribute to the prediction of well-being and psychopathology (Dejonckheere, 2019). Similarly, recent empirical studies did not show a significant association between inertia and depression/anxiety symptoms (Bosley et al., 2019) and borderline personality disorder (Houben & Kuppens, 2020). One reason for this is the statistical overlap between the emotion intensity (i. e., the process mean), inertia, and the other dynamic measures (Bos et al., 2019; Koval et al., 2013).

Problems with the estimation of inertia

In addition to the statistical overlap between dynamic measures, another reason for the limited contribution of inertia to the prediction of psychopathology/well-being is the fact that the current ways of estimating emotional inertia from ESM data do not provide sufficient precision in the estimation of inertia. Wenzel & Brose (2022) show that the neglect of estimating reliability of inertia leads to an attenuation of its relation to depressive symptoms. When reliability is taken into account via multilevel dynamic structural equation modeling and inertia is modeled as latent, the inertia of negative affect contributes to the prediction of depressive symptoms even when the emotion mean is included in the model (Wenzel & Brose, 2022). Furthermore, measurement decisions (such as the choice of items comprising the positive/negative affect composite measures), which are often made ad hoc in ESM research (Cloos et al., 2022), can lead to very different estimates of inertia.

More generally, low estimation precision will lead to the attenuation of the meaningful associations between inertia and other psychological constructs. Several simulation studies investigated the effect of varying parameters (such as number of participants and number of observations per participant) and assumption violations on measures related to the estimation precision of the MLAR(1) model. In the following sub-chapter, the evidence from simulation studies will be summarised.

Evidence from simulation studies

Jongerling et al. (2015) investigated the effect of modeling innovation variance as fixed (identical for all participants) instead of random. They found that modeling innovation as fixed when it differs across participants leads to a considerable bias in the estimation of the fixed AR effect. There is an upward bias

(overestimation) present when the correlation between the individual AR effects and individual innovation variances is positive, and vice versa. Additionally, Jongerling et al. point out that using the person-means to center the lagged predictor variable leads to a downward bias in the estimation of the fixed AR effect. The effect of person-mean centering the predictor on the estimation performance of the MLAR model was further studied by Hamaker & Grasman (Hamaker & Grasman, 2015). Their simulation study confirmed that person-mean centering leads to an underestimation of the fixed autoregressive effect, especially when the number of timepoints per participant *(T.obs)* is low. Still, they recommend using person-mean centering when one is interested in the effect of a between-person predictor on inertia.

In their simulation study comparing the maximum likelihoood and Bayesian approaches to estimating the MLAR model, Krone et al. (2016) show that the two estimation procedures have a very similar performance. Furthermore, a higher *T.obs* leads to more precise estimates, while the effect of N on the estimation performance is small. They also show that a higher variance of the random AR effects lead to a lesser estimation precision, and that the estimation bias gets smaller when the real fixed AR effect increases. Liu (2017) assessed how violating the normality of the random AR effect distribution influences the estimation performance of the MLAR model. The different distributions of the random AR effects were found to only have a small effect on the estimation performance.

TODO: Gistelinck.

While the simulation studies mentioned above provide an extensive body of evidence about the statistical properties of the MLAR model under different conditions, several questions remain unanswered. One of them is the effect of missing observations on estimation performance. The presence of missing values in intensive longitudinal dataset decreases the number of observations per participant (or, more specifically, the number of observation-pairs that can be used for the estimation of the model). As such, it can be expected that lower compliance (i. e., lower proportion of ESM beeps that the given participant answered) will worsen the estimation bias. Additionally, different patterns of missingness might have different consequences on the estimation performance. Ji et al. (2018) show that the presence of data missing completely at random (MCAR), missing at random (MAR) and not missing at random (MNAR) leads to a considerable bias in point estimated of cross-lagged ant autoregressive parameters in vector autoregressive models when list-wise deletion is used. However, no similar evidence is available about the MLAR(1) model.

As such, the goal of the present thesis is to investigate whether compliance (or the proportion of missing data) and the different patterns of missingness have an effect on estimation bias, variability, and statistical power of the multilevel AR(1) model. Additionally, the number of participants *(N),* number of timepoints per participant *(T.obs),* the simulated fixed AR effect (β_{10}) and the variance of the random slopes/AR effects ($\sigma_{\nu 1}$) will be manipulated. More details about the simulation study are provided in the Methods section. In the remainder of the Introduction, I will present the evidence about compliance in ESM studies.

Compliance and missing data in ESM studies

Due to the nature of intensive longitudinal data collection in psychology, the presence of missing data in the datasets is more of a rule than an exception. In their meta-analysis of 477 published ESM studies (with total N of 677,536), Wrzus and Nebauer (2022) found the average compliance to be 79.19% (SD = 13.64%). Importantly, both within-person, momentary factors (e. g., a participant might be less likely to answer an ESM beep during social situations, compared to when they are alone) and between-person factors could influence compliance.

On the study design level, compliance is usually higher in studies that provide financial incentives to participants (Vachon et al., 2019; Wrzus & Neubauer, 2022), while increasing questionnaire length is associated with lower compliance (Eisele et al., 2020). On the other hand, sampling frequency (i. e., the number of 'beeps' per day) was not found to be associated with compliance in both an experimental study (Eisele et al., 2020) and a large-scale meta-analysis (Wrzus & Neubauer, 2022).

The evidence about participant characteristics associated with compliance is more ambiguous. While some studies found men to have lower compliance than women Vachon et al. (2019), the meta-analysis by Wrzus & Nebauer (2022) found the evidence for gender differences in compliance to be limited. Psychotic disorders appear to be associated with lower compliance (Vachon et al., 2019), and participants with depression tend to have higher compliance (Rintala et al., 2019). Additionally, Rintala et al. (2019) report that participants with higher overall positive affect and lower overall negative affect tend to have higher compliance. Furthermore, higher compliance was found to be associated with lower overall activity stress, lower event unpleasantness, feeling less physically tired/unwell, and using less alcohol (Rintala et al., 2020). The evidence regarding the momentary factors of compliance is similarly limited. While Rintala (2019) found some momentary measures (higher positive affect, feeling less disturbed by the ESM beep, and using tobacco) to be associated with higher compliance they conclude that the majority of the momentary factors are not correlated with compliance. Even though the evidence is limited, we can expect both within- and between-person differences in compliance to arise from a complex interaction of study design, sample characteristics and momentary.

To summarise the introductory part: I have introduced the application of the multilevel autoregressive model to operationalise inertia in ESM research, summarised the simulation studies about the statistical properties of the multilevel autoregressive model, and introduced the evidence about compliance in ESM research. In the rest of the thesis, I will describe the design, results and implications of two Monte Carlo simulation studies conducted to address the knowledge gap about the effect of missing data on the estimation performance of the multilevel autoregressive model.

Methods

Multilevel AR(1) model

In this subchapter, I will describe the mathematical basis, assumptions and estimation procedures of the first-order multilevel autoregressive (MLAR(1)) model with random intercepts and random autoregressive effects, which is the focus of the simulation part of the thesis. While the notation for the model varies across different papers, the notation used by Lafit et al. (2020) will be adhered to throughout the thesis.

The MLAR model consists of two levels: the within-person Level 1 and the between-person Level 2. At level 1, described by Equation (1) (Lafit et al., 2020), each participant's first-order autoregressive process is modelled: The person-specific autoregressive parameter (inertia) γ_{1i} quantifies to what degree the process value esm_{it} of participant i at time t depends on the lagged process value $esm_{i,t-1}$. The person-specific intercept γ_{0i} represents the expected process value esm_{it} when the lagged variable $esm_{i,t-1}$ equals 0 (Jongerling et al., 2015). Following the recommendations by (Hamaker & Grasman, 2015), the predictor (lagged variable) was person-mean centered. The innovation ϵ_{it} (i.e., residuals, the part of the variance that is not explained by the lagged variable $esm_{i,t-1}$) is assumed to be independent and coming from a normal distribution with mean of 0 and variance σ_e^2 (Lafit et al., 2020). The model used in the present thesis assumes that the innovation variance is identical for all participants.

$$esm_{it} = \gamma_{0i} + \gamma_{1i} * esm_{i,t-1} + \epsilon_{it} \tag{1}$$

In the multilevel AR(1) model, the person-specific autoregressive effects 1i and the person-specific intercepts 0i are allowed to vary between participants. The Level 2 of the MLAR(1) model describes this between-person variability. The Level 2 is defined in Equation (2)). Each person-specific autoregressive effect γ_{1i} is a sum of a fixed effect β_{10} and a person-specific random effect ν_{1i}. The random effects ν_{1i} themselves come from a normal distribution with mean of 0 and variance $\sigma_{\nu 1}^2$ (Lafit et al., 2020). The same holds for the person-specific intercepts γ_{0i}: they are a sum of a fixed effect β_{00} and a random effect ν_{0i} that comes from $N(0, \sigma_{\nu 0}^2)$.

$$\begin{aligned} \gamma_{0i} &= \beta_{00} + \nu_{0i} \\ \gamma_{1i} &= \beta_{10} + \nu_{1i} \end{aligned} \tag{2}$$

Assumptions of the MLAR(1) model

In this part, the assumptions of the MLAR(1) model and the way they were taken into account in the present simulation study will be explained.

Stationarity. The MLAR(1) model is used to model stable processes in which no temporal trends are present. As such, it assumes weak stationarity: the (person-specific) process mean, innovation variance,

and autoregressive parameter are assumed to not change through the time series (Rovine & Walls, 2006). For this reason, the person-specific autoregressive effects γ_{1i} are assumed to be bounded by -1 and 1, as autoregressive effects larger than 1 (or lower than -1) cause a change in the process mean (Krone et al., 2016).

Exogeneity. TODO (see word file)

Equally spaced measurements. TODO (see word file again)

Estimation procedures for the MLAR(1) model

Initial conditions problem. TODO (see Word)

Multilevel AR(1) model and statistical power

Simulation study

The goal of the present exploratory simulation study is to assess the effects of four different patterns of missing data (data missing completely random, data missing in blocks, and two patterns of data missing dependent on process value) on estimation performance/bias, standard error and statistical power for the estimation of the fixed autoregressive effect in the MLAR(1) model.

Simulation procedure

The study followed the general principles of a Monte Carlo simulation procedure described by Lane & Hennes (2018).

Simulation conditions. Two simulation studies, Simulation A and Simulation B, were carried out to investigate the research questions. In Simulation A, no random autoregressive effects were simulated and estimated (i.e., each subject's time-series in the simulation had the same simulated autoregressive effect, and only fixed autoregressive effects were estimated). In Simulation B, random autoregressive effects were simulated and estimated (with the random effects variance set to either 0.05 or 0.1). Both random and fixed intercepts were estimated in Simulations A and B.

Simulation A followed a $4 \times 2 \times 3 \times 4 \times 3$ factorial design (yielding 288 simulation conditions in total), and Simulation B followed a $4 \times 2 \times 2 \times 4 \times 2 \times 2$ design (256 conditions in total). Each of the conditions was simulated in 1,000 simulation runs. As such, 544,000 datasets were generated (and the same number of models was estimated) in this simulation study. The manipulated variables (as well as the other parameters that remained fixed throughout all simulation conditions) are listed in Table 1 and 2.

Data generation. First, for each of the simulation conditions (i.e., combination of the parameters listed below), 1000 synthetic datasets were generated. Each dataset contained observations from N participants. A temporally dependent time-series of length $T.obs$ was generated as nested within each simulated participant via a recursive equation. Additionally, for each time-series, a burn-in period with

Table 1: Values of the manipulated parameters used in the two simulation studies

Manipulated parameter	Simulation A	Simulation B
Missingness pattern	MCAR, block, extreme-onesided, extreme-twosided	MCAR, block, extreme-onesided, extreme-twosided
Simulated fixed AR effect	0.3, 0.5, 0.7	0.3, 0.7
Variance of random AR effects	-	0.05, 0.1
Compliance	0.4, 0.6, 0.8, 1	0.4, 0.6, 0.8, 1
Number of participants (N)	20, 50	20, 50
Timepoints per participant (T.obs)	20, 50, 100	50, 100

Table 2: Parameters used for the two simulation studies.

Simulation parameter	Simulation A	Simulation B
Fixed intercept	0	0
Variance of random intercepts	3	3
Innovation variance	3	3
Correlation between random intercepts and random slopes	0	0
Significance threshold	0.05	0.05
Simulation runs per condition	1000	1000

1,000 observation was generated. The within-person error (innovation) vector ϵ_i was generated from a N(0, σ) distribution with σ set to 3 in all simulations. The fixed intercept β_{00} was set to 0 across all conditions. The random intercepts ν_{0i} for each simulated time-series were sampled from a $N(0, 3)$ distribution in both studies. In Simulation A, only fixed autoregressive effects β_{10} were simulated and manipulated, while both fixed and random autoregressive effects $\nu_1 i$ were included in Simulation B. For an overview of the values of all manipulated simulation parameters, please refer to Table 1.

Each time-series was then generated using Equation (1). The initial value was generated as a sum of the person-specific intercept γ_{0i} and the innovation ϵ_{ij}, and the following observations were calculated by multiplying the value of the time-series at *t-1* by the person-specific autoregressive effect γ_{1i} and adding the person-specific intercept γ_{0i} and the innovation ϵ_{ij}. Subsequently, after removing the burn-in datapoints, the first-order lagged version of the time-series was generated, setting the first lagged value as missing.

The non-manipulated simulation parameters ($\beta_{00}, \sigma_{\nu 0}, \sigma, \rho_\nu$) were set following a simulation design from Hamaker & Grasman (2015).

Introduction of missing values. Secondly, missing data were introduced to each of the generated datasets according to the missing data pattern and compliance of the given simulation condition. Four different missingness patterns were introduced to the data: a) data missing completely at random (MCAR); b) data missing in blocks of consecutive observations; c) lowest (100%-compliance) observations set as missing, and d) highest and lowest (100%-compliance)/2 observations set as missing.

Each of these missingness patterns correspond to a hypothetical scenario in an ESM study. The MCAR pattern assumes that the participants miss responding to beeps randomly, and each beep has the same probability of being missed, regardless on any other factors (e.g., whether the previous beep was missed, or the intensity of the emotion measured by ESM). When there is a block of missing data present, all missing observations follow each other, there is a block of non-missing values and missing values. The start and the endpoint of the missing block do not depend on the intensity of the emotion. This can correspond to a situation where a participant misses a series of beeps because they are attending a social event. For patterns c) and d), the missingness is dependent on the value of the process itself. Pattern c) represent a situation in which a participant does not respond to an ESM measure of a positive mood because they are not feeling well, while pattern d) represents a situation where a participant misses an ESM beep when they either do not feel well enough, or they feel too good to respond to their phone beeping.

It can be expected that the different missingness patterns will differ in their effects on the simulation outcomes (estimation bias, standard error, power etc.). This is because with identical proportion in missing data, datasets with different missingness patterns will have different proportions of effective observation-pairs (i.e., proportion of timepoints for which both the observation at t and the observation at t-1 are not missing) used to estimate the autoregressive effect. Figure 1 illustrates the four different missingness patterns on the same ESM time-series.

Fitting a multilevel autoregressive model. After missing values were introduced to the data, a MLAR(1) model was fitted to each of the simulated datasets using the *lme* function from the *nlme* R package (Pinheiro et al., 2022) with the value of the time-series at t as the outcome, the lagged *(t-1)* value of the time-series as the predictor, and the participant number as the grouping variable. We then extracted relevant parameters from the models that converged successfully. Missing values were treated by list-wise deletion. The restricted maximum log-likelihood method was used to estimate the model.

The predictor (lagged) variable was person-mean centered. Although person-mean centering results in an underestimation of the autoregressive effect (Hamaker & Grasman, 2015), it allows for a clearer interpretation of the within-person effects in multilevel models Hamaker & Muthén (2020).

TODO: Describe person-mean centering more in detail

Simulation outcomes. Estimation bias (MSE), the standard error of the estimation, and the statistical power to estimate the fixed autoregressive effect β_{10} were the focal outcomes of the study. Additionally, we examined the effect of the manipulated variables on the proportion of models that successfully converged and the bias in the estimation of the person-mean used for centering of the predictor (lagged) variable.

Estimation bias was computed as the difference between the real (simulated) fixed autoregressive effect β_{10} and the estimated fixed autoregressive effect $\hat{\beta_{10}}$ in each simulation run. As such, the dataset with estimation bias contained 1,000 rows per simulation condition.

Standard error (SE) and statistical power were calculated for each simulation condition (i.e., 1 row per

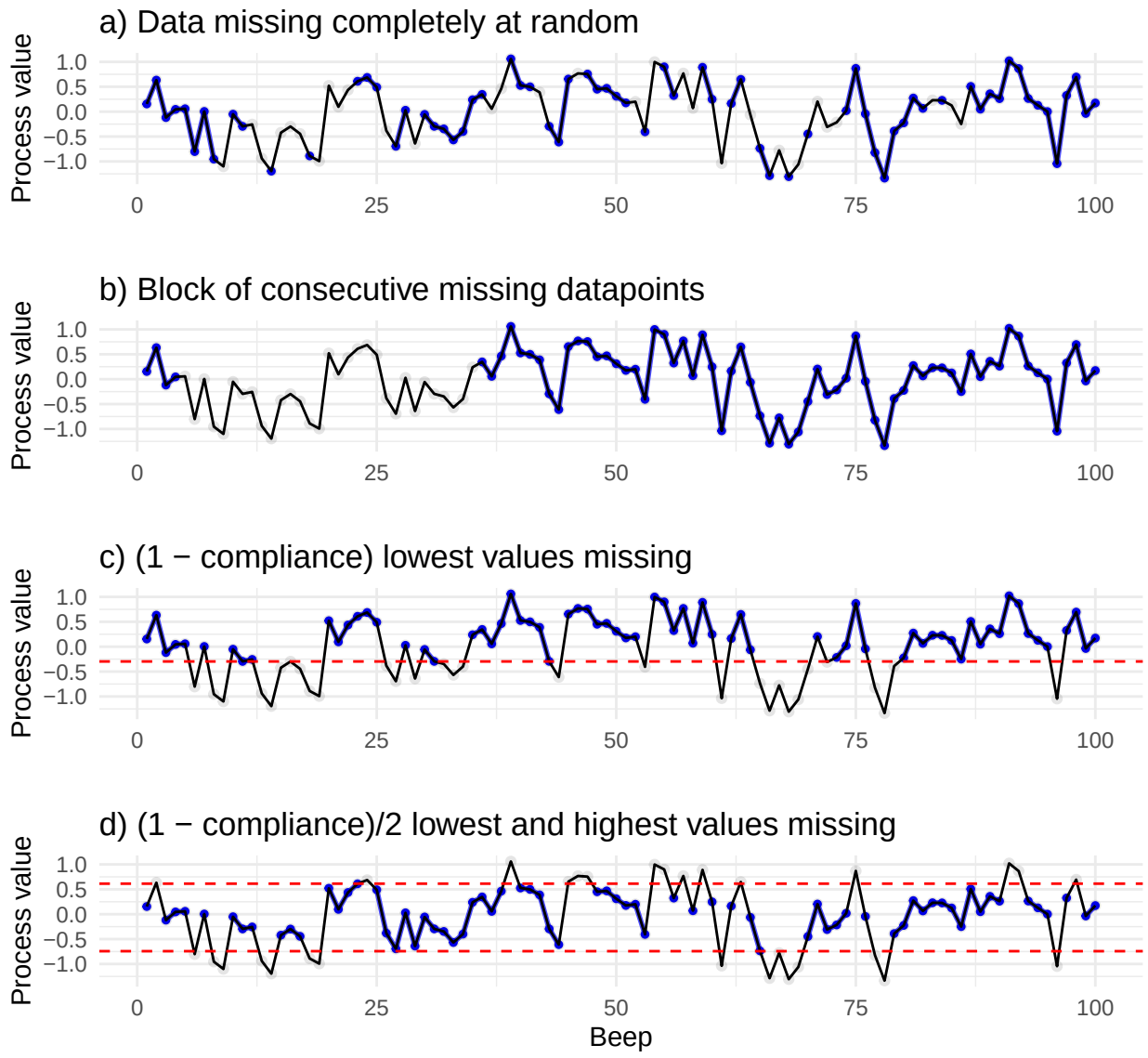

Figure 1: Illustration of the four different missingness pattern used in the simulation study. The blue dots represent observed datapoints, while the light gray dots represent the missing values.

condition). TODO: Describe how SE was calculated. Statistical power was calculated as the proportion of simulation runs (within the given simulation condition) in which the p-value for the estimated fixed autoregressive effect $\hat{\beta_{10}}$ was below the significance threshold ($\alpha = 0.05$) and the number of simulation runs that converged successfully.

The bias in the estimation of the person-mean of the time-series was computed as the average difference between the real process mean μ_i (3) and the observed person-mean $\hat{\mu_i}$ (computed after the missing data were introduced).

$$\mu_i = \beta_{00} + \nu_{0i}/1 - (\beta_{10} + \nu 1i) \tag{3}$$

Manipulated simulation parameters. Apart from the four different missingness patterns (described above), the following parameters were manipulated in bothSimulation A and Simulation B: The number of participants of each simulated study (N), the number of ESM observations within each participant (T.obs), the compliance rate (i.e., the proportion of timepoints that are not missing for each participant), and the simulated autoregressive effect.

Furthermore, in Simulation B, the variance of the random autoregressive effects was varied. The values of the manipulated variables were set considering realistic research questions in psychological research. The values of the manipulated variables for both studies are reported in Table 1.

Reproducibility and code/data availability

The simulations were conducted in R version 4.2.1 (R Core Team, 2021) . The study was conducted with emphasis on reproducibility of the results (Pawel et al., 2022). As such, we provide all data used for the reported analyses, as well as the full reproducible R code for the simulations, (including the custom functions created for the purposes of the study), and the code used to generate the plots and result tables (available at https://github.com/benjsimsa/AR-missing-simulations)..) The repository also includes a *sessionInfo* document that lists the versions of the packages used for the study. The present thesis was written using R Markdown (Allaire et al., 2022).

Additionally, the *renv* R package (Ushey, 2022) was used to set up a reproducible R environment and improve reproducibility by creating a project-local package library. For reproducible file referencing, the R package *here* (Müller, 2020) was used. For more information about the custom functions, simulation code, and the structure of the GitHub repository itself, please refer to the file README.md in the repository.

Results

Simulation A

The descriptive results for all 288 conditions included in Simulation A are reported in the appendix (TODO).

Outcome: Estimation bias (MSE)

ANOVA. We used a $4 \times 2 \times 3 \times 4 \times 3$ factorial Type I ANOVA (with estimation bias as an outcome and number of participants, number of timepoints per participant, missingness type, compliance, and the simulated fixed autoregressive effect) to assess which of the manipulated factors had a considerable influence on estimation bias. The results from every simulation run (i. e., 1,000 results per condition = 288,000 rows) were combined into a single dataset for the analysis. Given the very large sample size (which would make any difference significant) and the exploratory character of the analysis, p-values and significance thresholds were not used make inferences. Instead, we used a threshold of 0.14 for the partial ω^2, indicating a large effect size (Field et al., 2012). This cutoff will be used for all ANOVA results throughout the results section. The partial ω^2 was chosen as the less biased alternative to partial η^2 (Okada, 2013). The results and effect sizes are reported in Table 3.

Four main effects above the effect size threshold of 0.14 were found: the main effect of missingness type ($\omega^2 = 0.73$), compliance ($\omega_p^2 = 0.63$), the number of timepoints per participant ($\omega_p^2 = 0.26$), and the simulated fixed slope ($\omega_p^2 = 0.14$). Furthermore, the interaction between the missingness type and compliance ($\omega_p^2 = 0.54$) had an effect size above the cut-off.

The main effects of missingness type and compliance are visualised in Figure 2 and Figure 3 (respectively), while the interaction between missingness type and compliance are depicted in Figure 4.

Figure 2 shows that while the underestimation of the fixed slopes is fairly low (although still considerable) when the observations are missing completely at random or in block, it becomes severe when only the most extreme values (both at one side and at both sides) are missing. SOMETHING ABOUT JANNE'S PAPER! Additionally, the underestimation of the fixed slopes becomes more severe as the compliance gets lower.

The average estimation bias when compliance is 0.8 (which is very close to the average compliance of ESM studies in psychology) is -0.13. As a consequence, many estimates of inertia in psychological research could be seriously downward biased. Furthermore, the estimates are slightly biased even when compliance is 1 (i. e., there are no missing data; average bias: -0.04). This is in line with the findings about estimation bias caused by person-mean centering in multilevel autoregressive models (Hamaker & Grasman, 2015).

Zooming in on the interaction between compliance and missingness type (Figure 4) suggests that the effect of compliance on estimation bias is dramatically more severe for the two conditions in which the most extreme values of the process were set as missing (as compared to the other two conditions, i. e., data

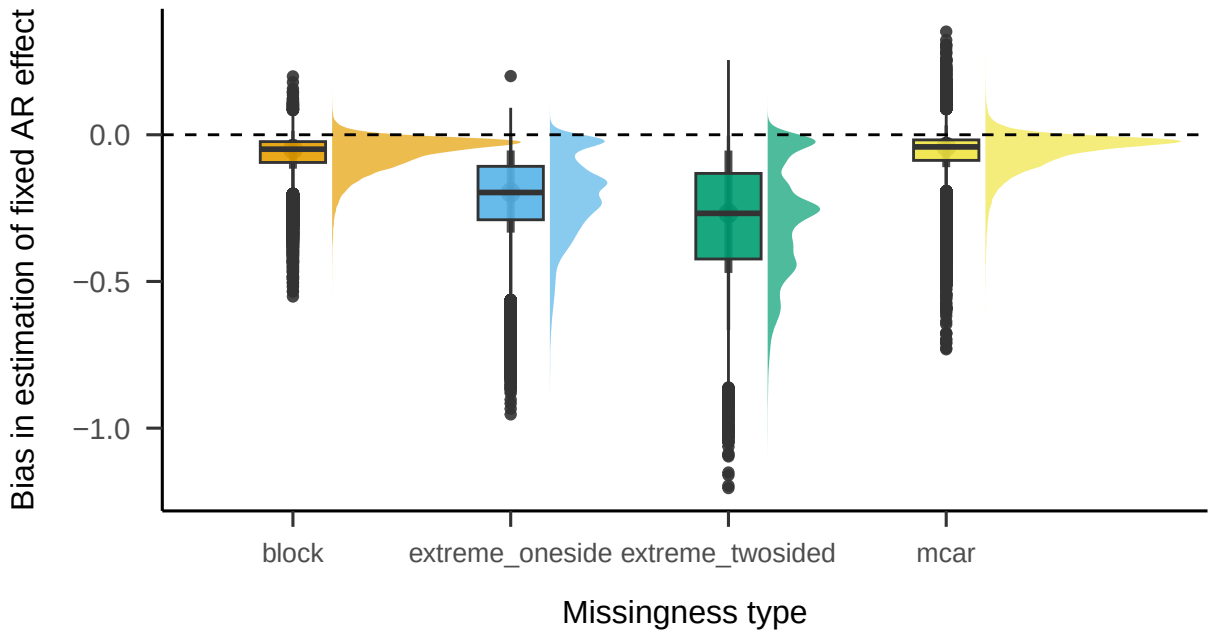

Figure 2: The effect of compliance on the bias in estimation of the fixed slopes.

MCAR and missing in blocks). In the worst-case scenario (low compliance of 0.4; the most extreme values at both sides missing), the average estimation bias was -0.48). Given that the average simulated fixed slope was 0.5, these results imply that even rather large autocorrelations can be estimated as close to 0 in studies with the combination of low compliance and a non-random missingness pattern. At the same time, the results about data MCAR and missing in blocks are encouraging. Even in a low-compliance (0.4) condition, the average estimation bias was -0.08 for the former and -0.09 for the latter.

The average estimation bias for all combinations of missingness type and compliance (averaged over the different values of the number of participants, timepoints per participant and simulated fixed slope) is reported in Table 4.

Outcome: Standard error

Descriptive statistics. The average standard errors for the different combinations of number of participants, timepoints per participant and compliance are reported in Table \ref{tab:tab_aov_se}.

ANOVA. To examine the effect of the manipulated parameters on the standard error of the estimation of the fixed slopes, we combined the results for each condition (1,000 simulation runs) into a single row. As such, the dataset used for the following analyses had one row per simulation condition (288 rows in total). A a $4 \times 2 \times 3 \times 4 \times 3$ factorial Type I ANOVA was used to analyse the data. The full ANOVA results and effect sizes are reported in Table 6.

The main effects of the number of participants ($\omega_p^2 = 0.68$), number of timepoints per participant ($\omega_p^2 = 0.68$) and compliance ($\omega_p^2 = 0.66$) crossed the cut-off for effect size.

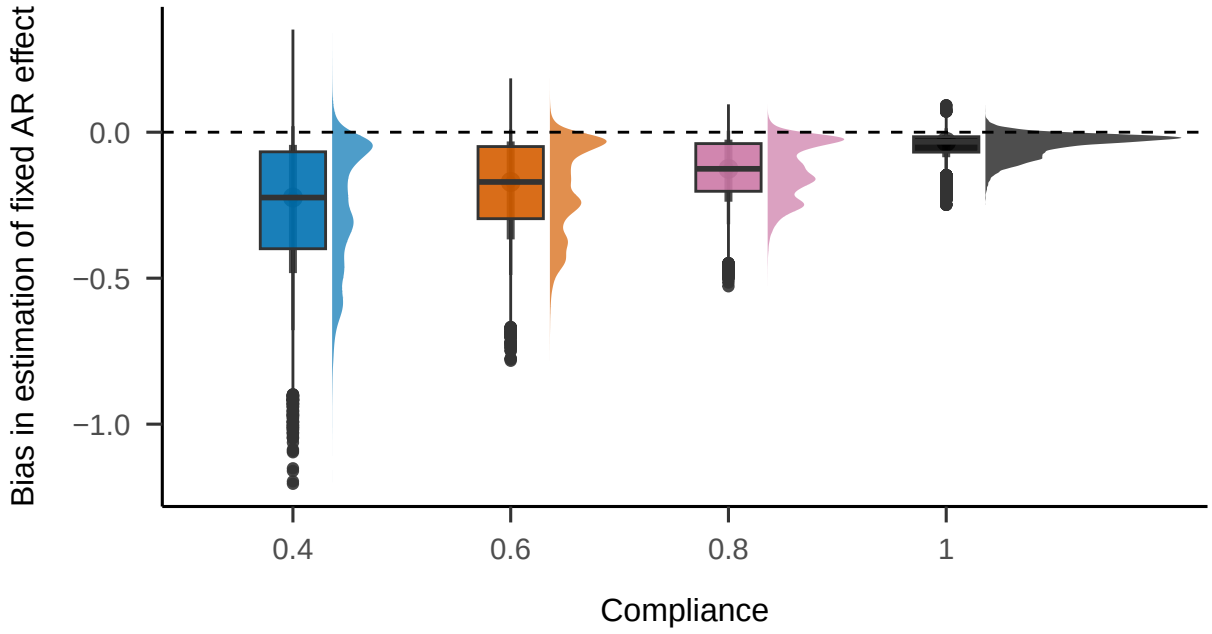

Figure 3: The effect of missingness type on the bias in estimation of the fixed slopes.

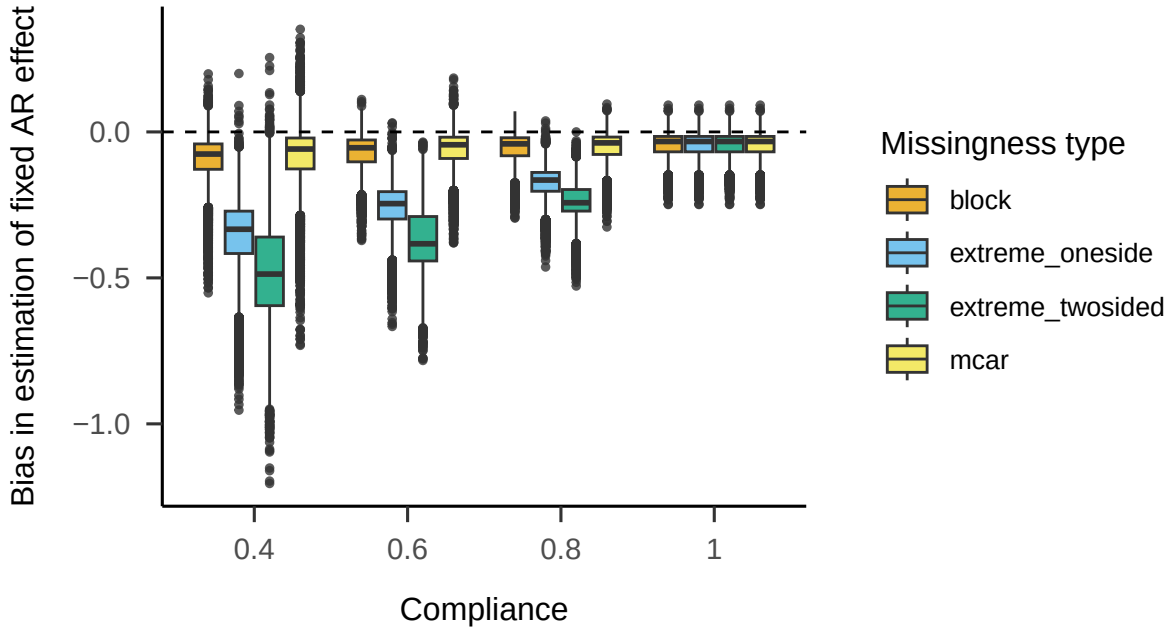

Figure 4: The effect of the interaction between missingness type and compliance on the bias in estimation of the fixed slopes.

Table 3: ANOVA results, simulation A. Outcome: Estimation bias

	Df	Sum Sq	Mean Sq	F value	p-value	Partial omega-squared
N	1	0.11	0.11	32.53	<0.001	0.00
T.obs	1	354.93	354.93	101753.95	<0.001	0.26
miss_type	3	2657.10	885.70	253921.29	<0.001	0.73
compliance	1	1706.99	1706.99	489377.69	<0.001	0.63
B1_sim	1	169.00	169.00	48449.29	<0.001	0.14
N:T.obs	1	0.02	0.02	4.40	0.0360	0.00
N:miss_type	3	0.00	0.00	0.31	0.8216	0.00
T.obs:miss_type	3	14.11	4.70	1348.32	<0.001	0.01
N:compliance	1	0.03	0.03	8.09	0.0044	0.00
T.obs:compliance	1	22.78	22.78	6529.83	<0.001	0.02
miss_type:compliance	3	1157.71	385.90	110634.77	<0.001	0.54
N:B1_sim	1	0.03	0.03	8.41	0.0037	0.00
T.obs:B1_sim	1	1.75	1.75	502.84	<0.001	0.00
miss_type:B1_sim	3	148.38	49.46	14179.70	<0.001	0.13
compliance:B1_sim	1	59.28	59.28	16994.38	<0.001	0.06
Residuals	287974	1004.48	0.00		NA	

Table 4: Simulation A. Average bias in estimation of the fixed slope for each combination of missingness type and compliance.

	Missingness type			
compliance	block	extreme_oneside	extreme_twosided	mcar
0.4	-0.09	-0.36	-0.48	-0.08
0.6	-0.07	-0.26	-0.37	-0.06
0.8	-0.05	-0.17	-0.24	-0.05
1.0	-0.04	-0.04	-0.04	-0.04

Table 5: Simulation A. Average standard error in the estimation of the fixed slope for each combination of number of participants, number of timepoints/participant, and compliance.

		Compliance			
N	T.obs	0.4	0.6	0.8	1
20	20	0.14	0.08	0.06	0.05
	50	0.07	0.05	0.04	0.03
	100	0.05	0.03	0.02	0.02
100	20	0.06	0.04	0.03	0.02
	50	0.03	0.02	0.02	0.01
	100	0.02	0.01	0.01	0.01

Table 6: ANOVA results, simulation A. Outcome: Standard error

	Df	Sum Sq	Mean Sq	F value	p-value	Partial omega-squared
N	1	0.06	0.06	625.92	<0.001	0.68
T.obs	1	0.06	0.06	621.16	<0.001	0.68
miss_type	3	0.00	0.00	14.11	<0.001	0.12
compliance	1	0.05	0.05	556.59	<0.001	0.66
B1_sim	1	0.00	0.00	21.75	<0.001	0.07
N:T.obs	1	0.01	0.01	91.92	<0.001	0.24
N:miss_type	3	0.00	0.00	2.13	0.096	0.01
T.obs:miss_type	3	0.00	0.00	1.48	0.220	0.00
N:compliance	1	0.01	0.01	82.89	<0.001	0.22
T.obs:compliance	1	0.01	0.01	114.06	<0.001	0.28
miss_type:compliance	3	0.00	0.00	13.31	<0.001	0.11
N:B1_sim	1	0.00	0.00	3.14	0.078	0.01
T.obs:B1_sim	1	0.00	0.00	1.37	0.243	0.00
miss_type:B1_sim	3	0.00	0.00	0.20	0.895	0.00
compliance:B1_sim	1	0.00	0.00	1.55	0.214	0.00
Residuals	262	0.03	0.00		NA	

Additionally, the interaction between the number of timepoints per participant and compliance ($\omega_p^2 = 0.28$), number of participants and timepoints per participants ($\omega_p^2 = 0.07$), and between the number of participants and compliance ($\omega_p^2 = 0.22$) was found.

Figure 5 depicts the interaction between the number of timepoints per participant and compliance, while Figure 6 shows the interaction between the number of participants and compliance.

Bias in the estimation of the fixed slope and bias in the estimation of the person-specific means. TODO.

Outcome: Statistical power

Descriptive statistics. The statistical power for each combination of the manipulated parameters is reported in Table (BIG TABLE TODO). As an illustration, the effect of compliance, missingness type, the number of participants and the number of timepoints per participant when the simulated fixed slope is 0.3 are visualised in Figure 11. Consistent with the results about estimation bias, statistical power is the lowest in the two conditions with the most extreme datapoints missing. For the conditions with data missing completely at random and data missing in consecutive blocks, power is very high even when the compliance is low for most conditions (except for the two conditions with T = 20).

A peculiar pattern is worth pointing out in the plot: in the two conditions with T = 20 and the most extreme data missing at both sides (green dashed line), the statistical power is higher when compliance is 0.4 compared to when compliance is 0.6. This counterintuitive result is likely due to the fact that the underestimation is the most severe when the most extreme values at both sides. As such, some of the estimates of the fixed slope will be negative, and their magnitude will be large enough for them to reach

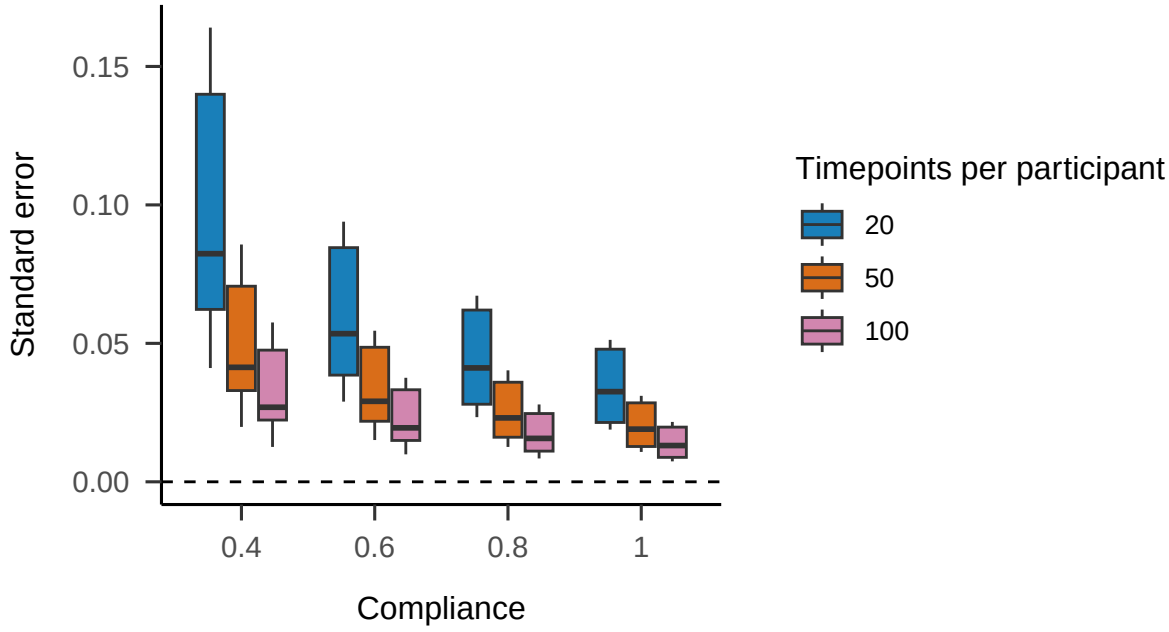

Figure 5: The effect of the interaction between number of timepoints and compliance on standard error of estimation of the fixed slopes. Simulation A.

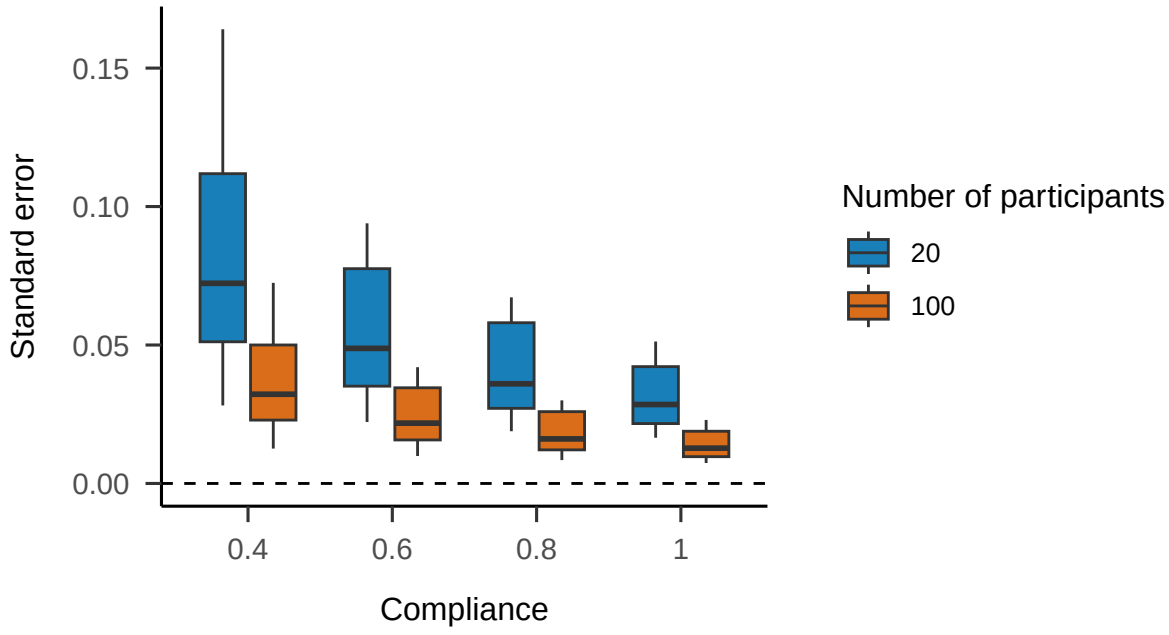

Figure 6: The effect of the interaction between number of participants and compliance on standard error of estimation of the fixed slopes. Simulation A.

statistical significance.

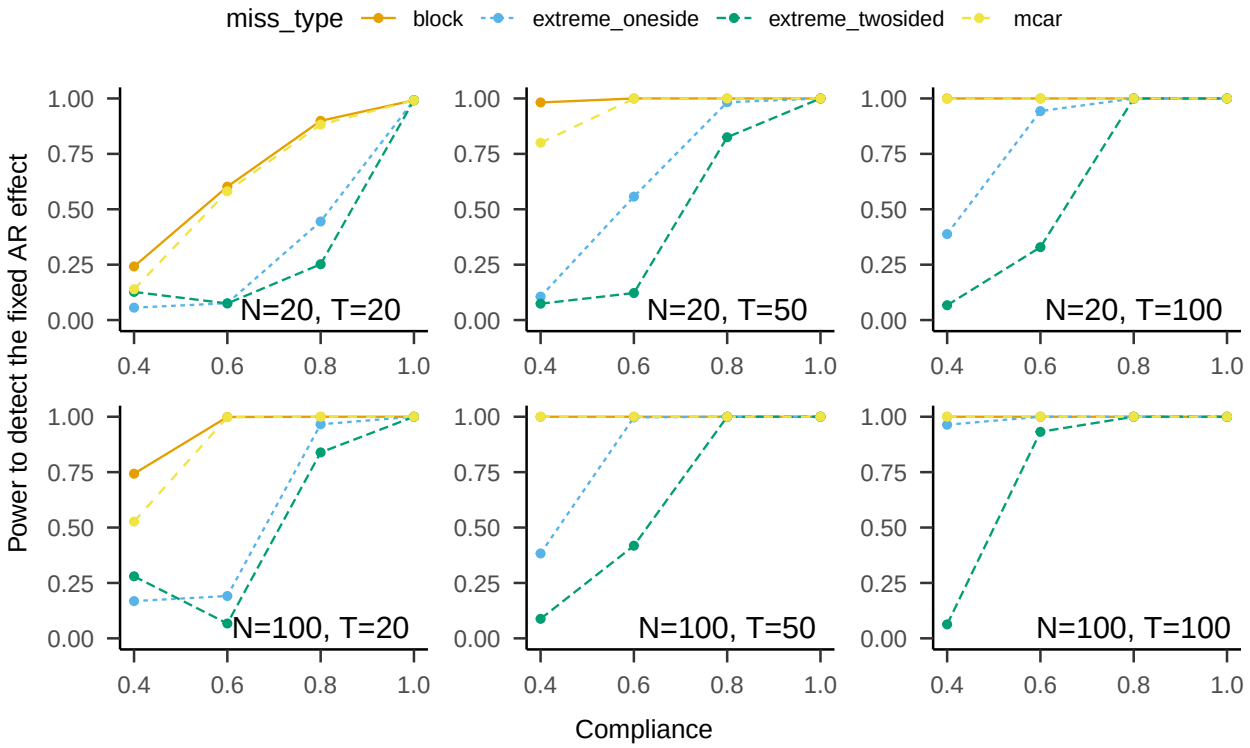

Figure 7: Simulation A. Statistical power to detect the fixed slope for all combinations of compliance, missingness type, number of participants and timepoints per participant when the simulated fixed slope is 0.3.

ANOVA. Aa $4 \times 2 \times 3 \times 4 \times 3$ factorial Type I ANOVA was used to analyse the effect of the manipulated parameters (288 conditions in total) on statistical power. The results are reported in Table 7.

Four main effect above the cut-off for the effect size were found: the effect of compliance ($\omega_p^2 = 0.43$), of missingness type ($\omega_p^2 = 0.36$), simulated fixed slope ($\omega_p^2 = 0.18$), and the effect of the number of timepoints per participant ($\omega_p^2 = 0.17$).

Table 7: ANOVA results, simulation A. Outcome: Power to detect the fixed slope

	Df	Sum Sq	Mean Sq	F value	p-value	Partial omega-squared
N	1	0.42	0.42	20.32	<0.001	0.06
T.obs	1	1.21	1.21	58.92	<0.001	0.17
miss_type	3	3.36	1.12	54.56	<0.001	0.36
compliance	1	4.47	4.47	217.19	<0.001	0.43
B1_sim	1	1.35	1.35	65.78	<0.001	0.18
N:T.obs	1	0.08	0.08	3.80	0.0524	0.01
N:miss_type	3	0.10	0.03	1.66	0.1753	0.01
T.obs:miss_type	3	0.34	0.11	5.49	0.0011	0.04
N:compliance	1	0.22	0.22	10.93	0.0011	0.03
T.obs:compliance	1	0.79	0.79	38.66	<0.001	0.12
miss_type:compliance	3	3.12	1.04	50.65	<0.001	0.34
N:B1_sim	1	0.09	0.09	4.25	0.0403	0.01
T.obs:B1_sim	1	0.24	0.24	11.50	<0.001	0.04
miss_type:B1_sim	3	0.44	0.15	7.19	<0.001	0.06
compliance:B1_sim	1	0.76	0.76	37.10	<0.001	0.11
Residuals	262	5.39	0.02		NA	

Simulation B

In Simulation B, the random AR effects were included both in the data generating procedure and in the estimated models. The variance of random AR effects ($\sigma^2_{\nu 1}$) was manipulated as an additional simulation factor (2 values: 0.05 and 0.1). For an overview of all manipulated and fixed simulation parameters, please refer to Table \ref{tab:tab_manipulated}. The descriptive results for all 256 simulation conditions are reported in the appendix (TODO). For the sake of clarity, only the results for the $\sigma^2_{\nu 1} = 0.05$ simulation condition will be reported in the following sections (and then briefly compared to the condition in which $\sigma^2_{\nu 1} = 0.1$.

Outcome: Estimation bias

To evaluate the effect of the number of participants, number of timepoints per participant, missingness type, compliance, the variance of random AR effects, and the simulated fixed autoregressive effect on the bias in the estimation of the fixed AR effect in Simulation B, a $4 \times 2 \times 2 \times 4 \times 2 \times 2$ factorial Type I ANOVA was used. The results from every simulation run (256 conditions * 1000 runs per condition) were combined into a single dataset. An identical inference criterion ($\omega^2_p \geq 0.14$) as in the analysis of Simulation A was used. The results are listed in Table 8.

Interestingly, compared to the results from Simulation A (see Table 3), the effect of *T.obs* on estimation bias ($\omega^2 = 0.03$) is much smaller and does not reach the effect size threshold. The three main effects that do reach the cut-off in Simulation B are the effect of missingness type ($\omega^2 = 0.65$), compliance ($\omega^2 = 0.48$), and the real fixed AR effect ($\omega^2 = 0.46$). Only the interaction between missingness type and

Table 8: ANOVA results, simulation B. Outcome: Estimation bias

	Df	Sum Sq	Mean Sq	F value	p-value	Partial omega-squared
N	1	0.00	0.00	1.29	0.2560	0.00
T.obs	1	28.12	28.12	8767.51	<0.001	0.03
miss_type	3	1493.35	497.78	155190.54	<0.001	0.65
compliance	1	733.78	733.78	228764.02	<0.001	0.48
sigma_v1	1	42.47	42.47	13241.70	<0.001	0.05
B1_sim	1	674.23	674.23	210200.86	<0.001	0.46
N:T.obs	1	0.07	0.07	21.59	<0.001	0.00
N:miss_type	3	0.18	0.06	19.21	<0.001	0.00
T.obs:miss_type	3	0.94	0.31	97.58	<0.001	0.00
N:compliance	1	0.00	0.00	0.02	0.8977	0.00
T.obs:compliance	1	1.80	1.80	560.85	<0.001	0.00
miss_type:compliance	3	670.81	223.60	69710.89	<0.001	0.45
N:B1_sim	1	0.02	0.02	6.91	0.0086	0.00
T.obs:B1_sim	1	0.33	0.33	101.81	<0.001	0.00
miss_type:B1_sim	3	115.05	38.35	11955.97	<0.001	0.12
compliance:B1_sim	1	40.01	40.01	12473.08	<0.001	0.05
Residuals	251520	806.77	0.00		NA	

Table 9: Simulation B. Average bias in estimation of the fixed slope for each combination of missingness type and compliance.

	Missingness type			
compliance	block	extreme_oneside	extreme_twosided	mcar
0.4	-0.10	-0.27	-0.40	-0.06
0.6	-0.08	-0.21	-0.32	-0.06
0.8	-0.07	-0.16	-0.21	-0.06
1.0	-0.06	-0.06	-0.06	-0.06

compliance ($\omega^2 = 0.45$) reached the cut-off. The interaction is visualised in Figure 8). The pattern of the interaction is very similar to the pattern of the interaction between missingness type and compliance in Simulation A (see Figure 4). Interestingly, when compared to the results of Simulation A, the average estimation bias is slightly worse for the MCAR and block missingness types and slightly less severe for the two conditions with the extreme values missing (compare Table 9 and Table 4). However, the overall conclusion remains the same as in Simulation A: there is a considerable downward estimation bias that becomes more severe the smaller the compliance rate is, and it is most severe for the condition in which the most extreme data at both sides are missing.

Outcome: Standard error

ANOVA. The results of the $4 \times 2 \times 2 \times 4 \times 2 \times 2$ factorial ANOVA used to assess the influence of the manipulated factor on the standard error are reported in table 11. Compared to Simulation A, more

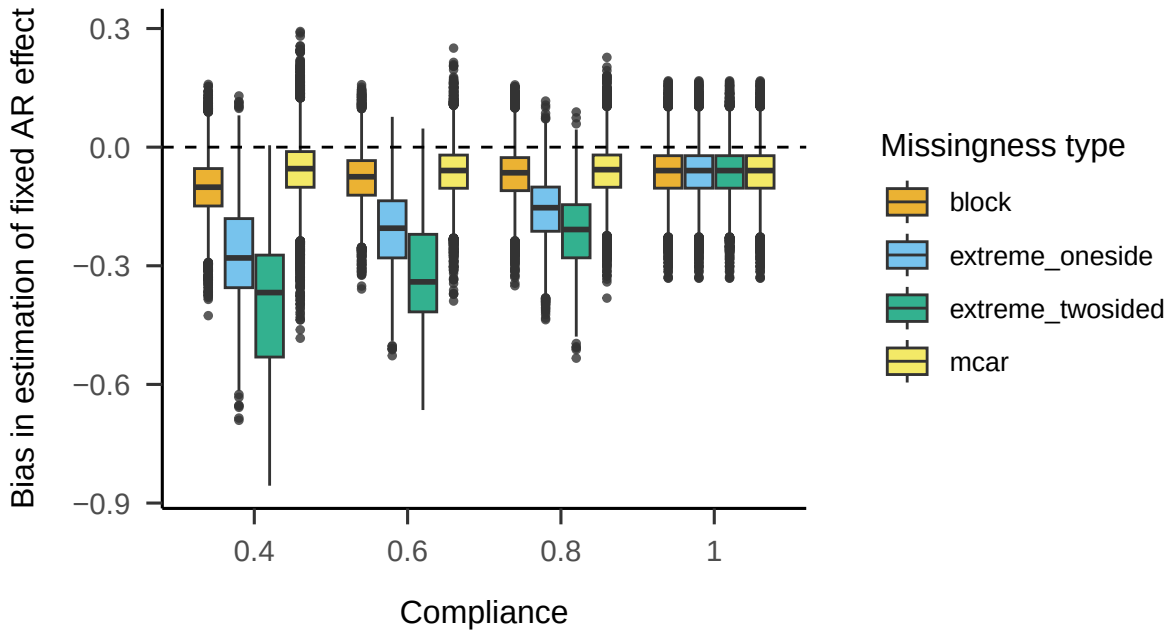

Figure 8: Simulation B: The effect of the interaction between missingness type and compliance on the bias in estimation of the fixed slopes.

main effects of the manipulated factors crossed the effect-size cut-off: the effect of number of participants ($\omega_p^2 = 0.93$), compliance ($\omega_p^2 = 0.53$), number of timepoints per participant ($\omega_p^2 = 0.33$), and the value of the simulated fixed AR effect ($\omega_p^2 = 0.28$). Additionally, the interaction between compliance and the number of timepoints per participant ($\omega_p^2 = 0.18$) crossed the threshold.

Figure 9 illustrates the main effect of N (as the most important factor) on standard error, while Figure 10 shows the interaction between compliance and $T.obs$. While the results are comparable to Simulation A, the SE is slightly higher for the same $N/T.obs$ combination for Simulation B.

Outcome: Statistical power

Descriptive statistics. The statistical power for each combination of the manipulated parameters in Simulation B is reported in Table (BIG TABLE TODO SIM2). The effects of compliance, missingness type, the number of participants and the number of timepoints per participant when the simulated fixed slope is 0.3 are shown in Figure **??**. For clarity, only the results for simulation conditions in which the $\sigma_{\nu 1}^2 = 0.1$ are visualised.

A peculiar pattern is worth pointing out in the plot: in the two conditions with T = 20 and the most extreme data missing at both sides (green dashed line), the statistical power is higher when compliance is 0.4 compared to when compliance is 0.6. This counterintuitive result is likely due to the fact that the underestimation is the most severe when the most extreme values at both sides. As such, some of the estimates of the fixed slope will be negative, and their magnitude will be large enough for them to reach

Table 10: ANOVA results, simulation B. Outcome: Standard error

	Df	Sum Sq	Mean Sq	F value	p-value	Partial omega-squared
N	1	0.0711	0.0711	3420.8496	<0.001	0.9314
T.obs	1	0.0025	0.0025	122.3644	<0.001	0.3251
miss_type	3	0.0008	0.0003	12.5061	<0.001	0.1205
sigma_v1	1	0.0029	0.0029	140.9331	<0.001	0.3570
compliance	1	0.0060	0.0060	288.9439	<0.001	0.5333
B1_sim	1	0.0021	0.0021	99.0254	<0.001	0.2801
N:T.obs	1	0.0004	0.0004	17.6323	<0.001	0.0619
N:miss_type	3	0.0001	0.0000	2.0403	0.109	0.0122
T.obs:miss_type	3	0.0002	0.0001	2.8127	0.040	0.0211
N:compliance	1	0.0007	0.0007	32.7145	<0.001	0.1118
T.obs:compliance	1	0.0012	0.0012	58.1185	<0.001	0.1848
miss_type:compliance	3	0.0008	0.0003	12.7376	<0.001	0.1226
N:B1_sim	1	0.0003	0.0003	15.0306	<0.001	0.0527
T.obs:B1_sim	1	0.0001	0.0001	3.1628	0.077	0.0085
miss_type:B1_sim	3	0.0013	0.0004	21.4078	<0.001	0.1955
compliance:B1_sim	1	0.0001	0.0001	2.4630	0.118	0.0058
Residuals	225	0.0047	0.0000		NA	

Table 11: Simulation B. Average standard error in the estimation of the fixed slope for each combination of number of participants, number of timepoints/participant, and compliance.

		Compliance			
N	T.obs	0.4	0.6	0.8	1
20	50	0.08	0.07	0.06	0.06
	100	0.06	0.06	0.05	0.05
100	50	0.04	0.03	0.03	0.03
	100	0.03	0.03	0.02	0.02

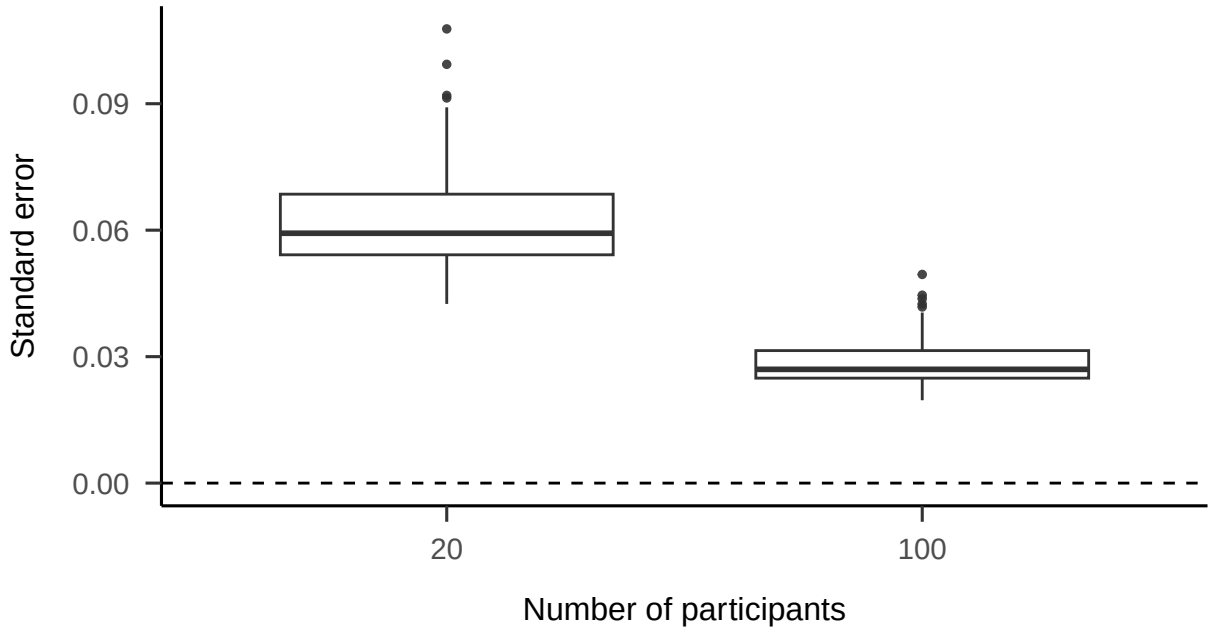

Figure 9: The effect of the number of participants on the standard error of estimation of the fixed slopes. Simulation B.

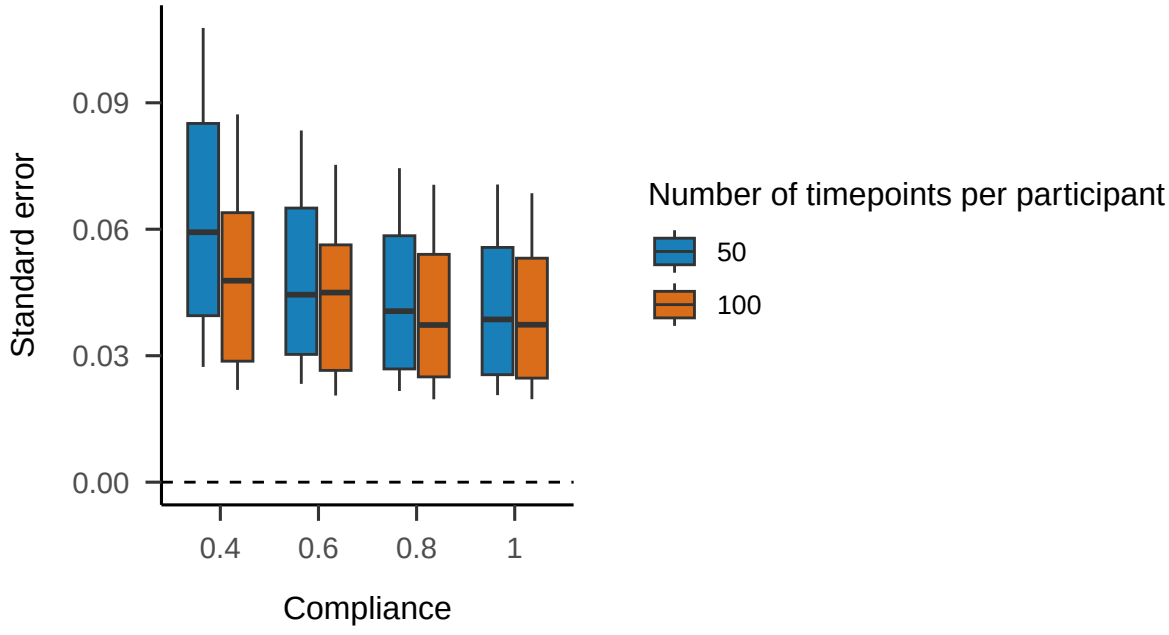

Figure 10: The effect of the interaction between number of timepoints per participant and compliance on standard error of estimation of the fixed slopes. Simulation B.

statistical significance.

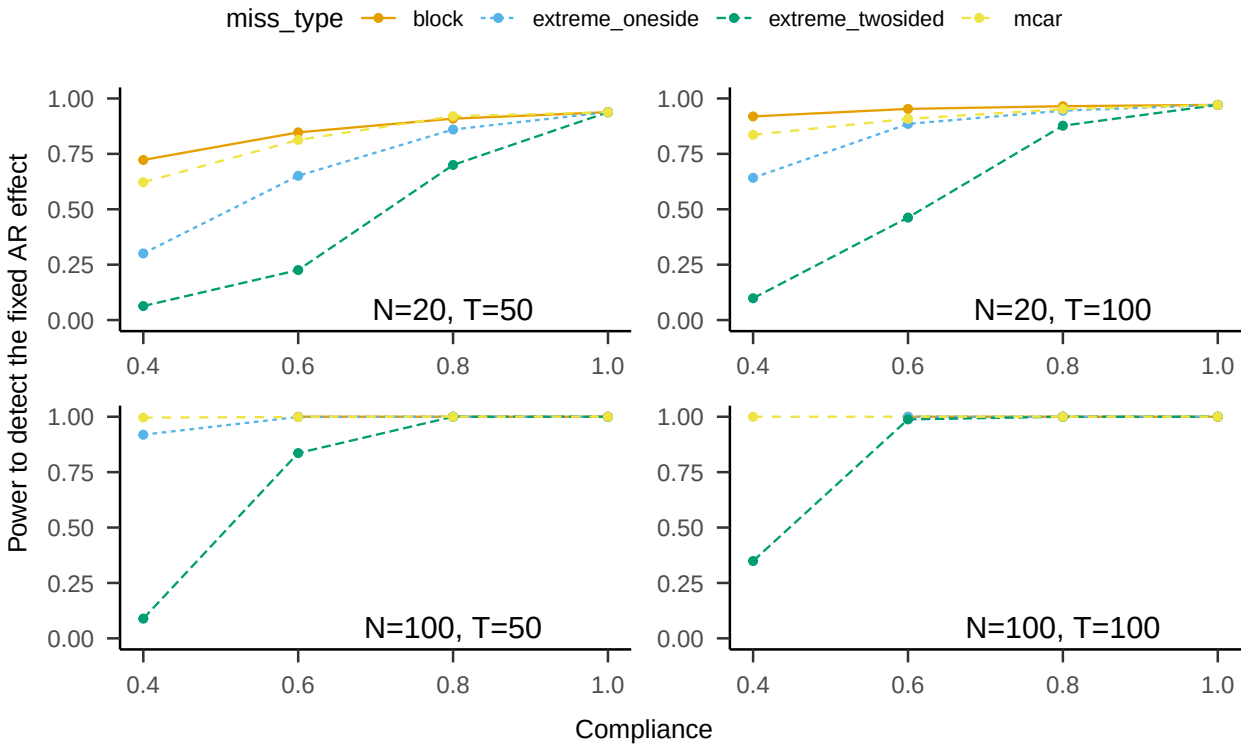

Figure 11: Simulation B. Statistical power to detect the fixed AR effect for the 4 combinations of compliance, missingness type, number of participants and timepoints per participant when the simulated fixed slope is 0.3 and the variance of random AR effects is 0.1.

ANOVA. A $4 \times 2 \times 2 \times 4 \times 2 \times 2$ factorial Type I ANOVA was used to analyse the effect of the manipulated parameters on statistical power. The results are reported in Table 12.

Missingness type ($\omega_p^2 = 0.4$) together with compliance ($\omega_p^2 = 0.38$) were the two main effects with the largest influence on the statistical power to detect the fixed AR effect. The value of the simulated AR effect has a large effect ($\omega_p^2 = 0.3$) as well. Three interactions crossed the effect size threshold: *missingness type*compliance* ($\omega_p^2 = 0.41$), *missingness type*simulated fixed AR* ($\omega_p^2 = 0.23$), and *compliance*simulated fixed AR* ($\omega_p^2 = 0.23$).

Outcome: Bias in person-mean estimation

TODO.

Table 12: ANOVA results, simulation B. Outcome: Power to detect the fixed AR effect

	Df	Sum Sq	Mean Sq	F value	p-value	Partial omega-squared
N	1	0.39	0.39	39.88	<0.001	0.13
T.obs	1	0.09	0.09	9.60	0.0022	0.03
as.factor(sigma_v1)	1	0.00	0.00	0.01	0.9124	0.00
miss_type	3	1.69	0.56	57.89	<0.001	0.40
compliance	1	1.53	1.53	157.25	<0.001	0.38
B1_sim	1	1.08	1.08	111.34	<0.001	0.30
N:T.obs	1	0.07	0.07	7.36	0.0072	0.02
N:miss_type	3	0.14	0.05	4.71	0.0033	0.04
T.obs:miss_type	3	0.02	0.01	0.75	0.5212	0.00
N:compliance	1	0.17	0.17	16.99	<0.001	0.06
T.obs:compliance	1	0.05	0.05	4.81	0.0293	0.01
miss_type:compliance	3	1.73	0.58	59.43	<0.001	0.41
N:B1_sim	1	0.16	0.16	16.50	<0.001	0.06
T.obs:B1_sim	1	0.04	0.04	4.03	0.0460	0.01
miss_type:B1_sim	3	0.76	0.25	25.98	<0.001	0.23
compliance:B1_sim	1	0.76	0.76	78.10	<0.001	0.23
Residuals	225	2.19	0.01		NA	

Discussion

We conducted two Monte Carlo simulation studies to address a knowledge gap about the influence of missing data on the estimation performance of the multilevel autoregressive model. In Simulation A, we only estimated and simulated fixed autoregressive effects (together with both fixed and random intercepts), while in Simulation B, both fixed and random autoregressive effects were simulated and estimated. Three outcomes were evaluated in both simulations: the estimation bias, standard error of the simulations, and statistical power. Four values of compliance and four missingness patterns (data MCAR, data missing in a block of consecutive observations, the highest values missing, and the highest and lowest values missing) were varied across the simulations. The other manipulated factors included the number of participants, number of timepoints per participant, the simulated value of the fixed AR effect, and the variance of the random AR effects.

The two variables related to missing data (compliance and missingness type) emerged as very important factors influencing all three outcomes. In both simulations, missingness type and compliance (and the interaction between the two) were the factors with the largest effect on the bias in the estimation of the fixed AR effect. Similarly, both missingness type and compliance had a strong influence on the statistical power to detect the fixed AR effect in both simulations. With regards to the standard error of the simulations, compliance was found to have a very large effect (more so in Simulation A than Simulation B), while the effect of missingness type was only moderate.

Our results corroborate the conclusions about the importance of the number of timepoints per

participant for precise estimation of the autoregressive effects (Hamaker & Grasman, 2015; Krone et al., 2016). In general, the estimation bias became considerably less severe as the ESM time-series length per participant increased. *T.obs* also had a large effect on statistical power. However, our simulations show that the context of missingness matters: when the compliance is low and the data are missing MCAR or in blocks, the underestimation of the fixed AR effect caused by the missing data (and the negative consequences for statistical power) gets less severe very quickly as *T.obs* increases. On the other hand, when the missingness is dependent on the value of the process itself (i. e., the most extreme observations are missing), increasing compliance appears to be more important for estimation precision and statistical power than the length of the time-series.

While Krone (Krone et al., 2016) found that estimation bias becomes smaller as the simulated fixed AR effect becomes larger, we found an opposite pattern: estimation bias was larger as the simulated fixed AR effect became larger.

Overall, there was always some degree of estimation bias present in the analyses, ranging from very severe (when *T.obs* and compliance were low, and the missingness of data was dependent on the process value) to mild (when compliance was high and the data were missing either MCAR or in block). This bias in estimation might be one of the driving forces behind the low value added by estimates of emotional inertia to the prediction of psychopathology and wellbeing, pointed out by Dejonckheere et al. (2019). Additionally, while the simulation studies did not explicitly assess the bias in the estimation of individual autoregressive effects, the results suggest that some individual differences in inertia estimates might not be caused by real differences in inertia, but due to the bias caused by missingness: for two individual participants with an identical real autoregressive parameter but different compliance and missingness patterns, the inertia estimates can vary considerably.

Our results have several implications for the design psychological research using the multilevel autoregressive model to estimate emotional inertia. First, in line with previous simulation studies, we recommend for researchers to focus on increasing *T.obs* rather than N in order to increase the statistical power and the precision of the inertia estimates. In other words, for optimal statistical performance, it is more effective to make the data collection period longer (or schedule more beeps per day) than to collect data from more participants. According to a recent meta-analysis, Secondly, while the time-series length is very important, researchers should aim to design their ESM studies in a way that will make compliance as high as possible. In line with the evidence about compliance in ESM studies, these design choices include providing financial incentives to participants (Wrzus & Neubauer, 2022) and making the individual ESM questionnaires shorter (Eisele et al., 2020). Furthermore, the results suggest that the potential presence of missing data should be accounted for in power analyses for ESM studies. In an ideal case, a researcher should have an idea about what the average compliance in their study could be an what missingness patterns might be present in the data. Of course, is not entirely feasible, as it might be difficult to estimate the average compliance, and real-life ESM data will likely include a mixture of different missing data

patterns, both at the within- and between-person level. Still, to avoid overestimating statistical power for planned studies, it is advisable to include several different missing data scenarios in the power simulations as a sensitivity check.

Limitations

First, while the two simulation studies include a wide range of scenarios and parameter combinations, our results are far from comprehensive, and they largely depend on the simulation parameters that were not varied. However, the reproducible code available in the GitHub repository (https://github.com/benjsimsa/AR-missing-simulations) provide a sufficient framework for an interested reader to rerun the analyses with different parameters and modify the code to better fit the pecularities of their own study sample and research questions.

Furthermore, the results depend on several assumptions, which were problematised as too simplified in previous studies. We assumed that the innovation variance σ was identical for all the participants, and that the random intercepts (and random slopes in Simulation B) came from a normal distribution. We only focused on normally distributed affective processes. While a normal distribution can be assumed for ESM measures of positive emotions, negative affective processes are usually heavily right-skewed in the general population (Haslbeck et al., 2022). Additionally, we assumed that the analysed ESM time-series are measured without any error; however, recent evidence shows that this is very often not the case in real-world research (Dejonckheere et al., 2022; Schuurman & Hamaker, 2019), and unreliability can lead to further attenuation of the estimated parameters (Wenzel & Brose, 2022). While person-mean centering was carried out using observed means in the simulations we conducted, different ways of person-mean centering, such as using latent person-means, might be more appropriate (Gistelinck et al., 2021).

Finally, although I took steps to make the simulations reproducible by making all code and results publicly available, using R packages *here* and *renv*, and reporting the sessionInfo for every simulation, quite a large number of packages with many dependencies were used, which might be detrimental to reproducibility.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2022). *Rmarkdown: Dynamic documents for r.* https://github.com/rstudio/rmarkdown

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Bos, E. H., Jonge, P., & Cox, R. F. A. (2019). Affective variability in depression: Revisiting the inertia–instability paradox. *British Journal of Psychology*, *110*(4), 814–827. https://doi.org/10.1111/bjop.12372

Bosley, H. G., Soyster, P. D., & Fisher, A. J. (2019). Affect Dynamics as Predictors of Symptom Severity and Treatment Response in Mood and Anxiety Disorders: Evidence for Specificity. *Journal for Person-Oriented Research*, *5*(2), 101–113. https://doi.org/10.17505/jpor.2019.09

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. https://doi.org/10.1080/02699931.2014.916252

Cloos, L. J. R., Kuppens, P., & Ceulemans, E. (2022). *Development, Validation, and Comparison of Self-Report Measures for Positive and Negative Affect in Intensive Longitudinal Research.* https://doi.org/10.31234/osf.io/5j7c6

Dejonckheere, E. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*, 17.

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment.* https://doi.org/10.1037/pas0001178

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, *29(2)*, 16.

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. https://doi.org/10.1037/1082-989X.12.2.121

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using r (2012). *Great Britain: Sage Publications, Ltd*, *958*.

Gistelinck, F., Loeys, T., & Flamant, N. (2021). Multilevel Autoregressive Models when the Number of Time Points is Small. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 15–27. https://doi.org/10.1080/10705511.2020.1753517

Hamaker, E. L. (2012). *Why researchers should think" within-person": A paradigmatic rationale.*

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? Investigating inertia with a

multilevel autoregressive model. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01492

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022). *Multimodality and Skewness in Emotion Time Series*. https://doi.org/10.31234/osf.io/qudr6

Houben, M., & Kuppens, P. (2020). Emotion Dynamics and the Association With Depressive Features and Borderline Personality Disorder Traits: Unique, Specific, and Prospective Relationships. *Clinical Psychological Science*, *8*(2), 226–239. https://doi.org/10.1177/2167702619871962

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901.

Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling Missing Data in the Modeling of Intensive Longitudinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 715–736. https://doi.org/10.1080/10705511.2017.1417046

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*(3), 334–349. https://doi.org/10.1080/00273171.2014.1003772

Koval, P., Burnett, P. T., & Zheng, Y. (2021). *Emotional Inertia: On the Conservation of Emotional Momentum* (C. E. Waugh & P. Kuppens, Eds.; pp. 63–94). Springer International Publishing. https://doi.org/10.1007/978-3-030-82965-0_4

Koval, P., Butler, E. A., Hollenstein, T., Lanteigne, D., & Kuppens, P. (2015). Emotion regulation and the temporal dynamics of emotions: Effects of cognitive reappraisal and expressive suppression on emotional inertia. *Cognition and Emotion*, *29*(5), 831–851. https://doi.org/10.1080/02699931.2014.948388

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. https://doi.org/10.1037/a0033579

Krone, T., Albers, C. J., & Timmerman, M. E. (2016). Comparison of Estimation Procedures for Multilevel AR(1) Models. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00486

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional Inertia and Psychological Maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion in Psychology*, *17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Lafit, G., Adolf, J., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2020). *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*. https://doi.org/10.31234/osf.io/dq6ky

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. https://doi.org/10.1177/0265407517710342

Larson, R., & Csikszentmihalyi, M. (2014). *The experience sampling method* (p. 2134). Springer.

Leemput, I. A. van de, Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., Nes, E. H. van, Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., Maas, H. L. J. van der, Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

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. https://doi.org/10.1111/bmsp.12096

Mneimne, M., Fleeson, W., Arnold, E. M., & Furr, R. M. (2018). Differentiating the everyday emotion dynamics of borderline personality disorder from major depressive disorder and bipolar disorder. *Personality Disorders: Theory, Research, and Treatment*, *9*(2), 192–196. https://doi.org/10.1037/per0000255

Müller, K. (2020). *Here: A simpler way to find your files.* https://CRAN.R-project.org/package=here

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. https://doi.org/10.1002/wps.20513

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129147.

Pawel, S., Kook, L., & Reeve, K. (2022). *Pitfalls and potentials in simulation studies.* http://arxiv.org/abs/2203.13076

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., Ranke, J., Team, R. C., & Team, M. R. C. (2022). *Package 'nlme'.*

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2020). Momentary predictors of compliance in studies using the experience sampling method. *Psychiatry Research*, *286*, 112896. https://doi.org/10.1016/j.psychres.2020.112896

Rovine, M. J., & Walls, T. A. (2006). Multilevel autoregressive modeling of interindividual differences in the stability of a process. *Models for Intensive Longitudinal Data*, 124147.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel

autoregressive modeling. *Psychological Methods*, *24*(1), 70–91. https://doi.org/10.1037/met0000188

Ushey, K. (2022). *Renv: Project environments*. https://CRAN.R-project.org/package=renv

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

Wenzel, M., & Brose, A. (2022). Addressing measurement issues in affect dynamic research: Modeling emotional inertia's reliability to improve its predictive validity of depressive symptoms. *Emotion*. https://doi.org/10.1037/emo0001108

Wrzus, C., & Neubauer, A. B. (2022). Ecological Momentary Assessment: A Meta-Analysis on Designs, Samples, and Compliance Across Research Fields. *Assessment*, 1–22. https://doi.org/10.1177/10731911211067538