Check for updates

Assessment
2025, Vol. 32(2) 206–223
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911241286868
journals.sagepub.com/home/asm
Sage

The Effect of Different Construct Operationalizations, Study Duration, and Preprocessing Choices on Power-Based Sample Size Recommendations in Intensive Longitudinal Research

Ginette Lafit[1,*], Jordan Revol[1,*], Leonie Cloos[1], Peter Kuppens[1], and Eva Ceulemans[1]

Abstract

To shed light on the dynamics of psychological processes, researchers often collect intensive longitudinal (IL) data by asking people to repeatedly report on their momentary experiences in daily life. Two important decisions when designing an IL study concern the number of persons and the number of measurement occasions to be included. These sample size decisions are ideally based on statistical power considerations. When conducting statistical power analysis, it is necessary to provide the value of the effect size of interest as well as of all other model parameters. In IL research, these values have to be based on previous studies. This implies that these values are subject to large heterogeneity due to differences in study design and preprocessing choices. This between-study heterogeneity can severely impact power-based sample size recommendations. In this article, we introduce an approach to investigate the impact of study design and pre-processing of previous studies and to determine a recommended sample size to account for this impact. We demonstrate how to use this approach to investigate the effect of different construct operationalizations, study duration, and preprocessing choices. This approach paves the way for more thoughtful and robust sample-size decisions.

Keywords

intensive longitudinal designs, statistical power, multilevel modeling, construct operationalization, data preprocessing

Intensive longitudinal (IL) designs have become indispensable for investigating how psychological processes fluctuate in daily life (Kuppens et al., 2022; Trull et al., 2015). The data for IL designs are typically collected using experience sampling methods (ESM, Myin-Germeys et al., 2018) in which repeated measurements are taken from the same individuals over a period of time, which are then usually analyzed with multilevel regression models that account for the serial dependence in the data (Goldstein et al., 1994). These models allow to characterize within-person processes, and how they differ between persons. They have for example been used to study emotional inertia (Kuppens et al., 2010), or emotional responses to positive and negative events (Dejonckheere et al., 2018), improving the understanding of psychopathology symptoms in daily life (Bringmann et al., 2013; Myin-Germeys et al., 2009), and for unraveling dynamic aspect of personality (Hofmans et al., 2019).

The popularity of IL research has, however, also exposed the multitude of decisions that have to be made concerning study design (Eisele et al., 2022; Janssens et al., 2018; Kaurin et al., 2023; Kirtley et al., 2021; Trull & Ebner-Priemer, 2020). Obviously, such choices have consequences for the findings, their replicability, and their comparability across studies (Weermeijer et al., 2022). Among the first choices when designing a study is to determine the number of persons and the number of measurement occasions to include. When the goal of a study is to test statistical hypothesis, these sample size decisions are ideally based on statistical power considerations (Lakens, 2022).

[1]KU Leuven, Belgium

*These authors contributed equally to this work.

Corresponding Author:
Ginette Lafit, Methodology of Educational Sciences Research Group, Faculty of Psychology and Educational Sciences, KU Leuven, Andreas Vesaliusstraat 2—bus 3762, Leuven 3000, Belgium.
Email: ginette.lafit@kuleuven.be

In an a-priori power analysis, statistical power is determined by the population effect size of interest, the predetermined type I error rate (i.e., the significance level), and the standard error of the test statistic used, which is a function of the true values of the parameters of the statistical model and the sample size. Given that the true model parameter values and thus the true effect size in the population are unknown, statistical power is computed based on assumed values for the effect size. Ideally, researchers can set the effect size to the smallest effect size that is considered theoretically or practically interesting (SESOI). The SESOI can be determined using expert opinions (Riesthuis et al., 2022), Cohen's benchmarks for small, medium, and large standardized effect sizes (Cohen, 2013), or anchors-based methods to quantify the minimal detectable difference (Anvari & Lakens, 2021; McGlothlin & Lewis, 2014). In IL research, determining the SESOI can be challenging as theories about relationships between daily life time-varying constructs may not be well-developed (e.g., see the studies by Gabriel et al., 2019; Schreuder et al., 2020), standardized effect sizes are difficult to obtain for statistical models that account for the nested structure of IL data[1] (Rights & Sterba, 2019), and anchor-based methods have mainly been developed for before/after intervention studies (Anvari & Lakens, 2021). Thus, when is not possible to determine the SESOI, researchers can estimate the value of the model parameters using data from previous studies (Lakens, 2022; Lane & Hennes, 2018). This practice is specifically relevant when the goal of the new study is to conduct a direct replication (i.e., when the same protocol of the original study is used in the new study) or conceptual replications (i.e., the same hypothesis or theory is tested with a different study protocol; e.g., see the studies by Derksen & Morawski, 2022; Nosek & Errington, 2017).

Using estimates of the model parameters based on previous studies is an adequate solution for conducting an a priori power analysis when: (a) previous studies are sufficiently similar to the to-be-designed study, (b) the risk of publication bias is low (e.g., analyses have been preregistered), and (c) the sample size is sufficiently large to ensure that model parameters are estimated accurately (for a discussion, see the work of Albers & Lakens, 2018; Gelman & Carlin, 2014; Lakens, 2022). However, as it has been shown in psychological research in general, effect sizes are often subject to large heterogeneity due to between-subject study variation, which can severely compromise power-based sample size recommendations (McShane & Böckenholt, 2014; Ward & Kemp, 2019). Therefore, assessing how previous IL study characteristics influence sample size recommendation based on power analysis is a crucial question. In fact, IL studies are characterized by large heterogeneity with respect to the items used to measure time-varying constructs and by methodological heterogeneity in sampling protocols.

A prime example of study heterogeneity due to differences in measurements of daily life constructs can be found in IL studies measuring affect. In these studies, it is common to select specific emotions items ad hoc to construct positive or negative affect scores (Brose et al., 2020; Cloos et al., 2023; Horstmann & Ziegler, 2020; Wright & Zimmermann, 2019). Yet, which items are used and how they are combined differs across studies (Dejonckheere et al., 2018; Eisele et al., 2021a, 2022; Weermeijer et al., 2022), leading to different effect size estimates. For example, Abel and Minor (2023) conducted a meta-analysis examining deficits in consummatory social pleasure in schizophrenia and found a significant variation in the empirical findings across ESM studies. This heterogeneity was largely explained by the different items used to measure affect. Another meta-analysis investigating the relationship between emotion regulation and affective experiences showed that the large heterogeneity in the results across studies was potentially moderated by different operationalizations of affect when using composite scores (Boemo et al., 2022). Moderation effects of operationalizations of momentary affect were also found in a meta-analysis on daily social interactions and within-person variability in affect (Liu et al., 2019), where some studies constructed affect scores using items derived from the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988), while other studies distinguished between high- and low-arousal affective states. The heterogeneity in affect measures is relevant for IL research in general given the prevalence of these measures in IL studies. For instance, a recent meta-analysis on compliance across research fields in ambulatory assessment found that 66% of the included studies focused on emotions, and 23.4% on mental health (Wrzus & Neubauer, 2023), where measures of affect are commonly used (e.g., see the work by Bell et al., 2024; Fernández-Álvarez et al., 2024; Joseph et al., 2021).

Another source of heterogeneity in results from previous studies is related to differences in the total number of assessments and participants between studies (Boemo et al., 2022). These numbers are determined by the sampling protocol and preprocessing decisions conducted on the collected data (Eisele et al., 2021b; Weermeijer et al., 2022). With respect to the sampling protocol, in a meta-analysis on design of ESM studies across research fields, Wrzus and Neubauer (2023) found noticeable differences in the reported total number of assessments across studies, with an average study duration of 12.4 days ($SD = 16.38$, 2–180), an average of 6.53 ($SD = 5.02$) assessments per day, and an average of

51.11 (SD = 45.29) total number of assessments. Even though previous research has found no evidence for smaller within-person relationships as a function of study duration (Fuller-Tyszkiewicz et al., 2013) or the sampling protocol (Hasselhorn et al., 2021), the degree in which the total number of assessments influences estimates of the model parameters and, consequently, power-based sample size recommendations has not yet been studied.

Similarly, before analyzing IL data, researchers often have to make multiple preprocessing decisions which can influence estimates of the model parameters (Revol et al., 2024). Two of the most common decisions involve when to exclude measurement occasions within participants due to too short or too long response times, or when to discard participants, based on low compliance, that is, a too low proportion of answered to measurement occasions (e.g., see the study by Eisele et al., 2021b; Weermeijer et al., 2022). Regarding the exclusion of measurement occasions based on delayed responses, Eisele et al. (2021b) found that the mean level and variance of affect measures and the association between affect and contextual stress were significantly related to the size of the response delay. These results indicate that the exclusion of delayed responses may alter the estimated value of the model parameters, and therefore influence sample size recommendations based on these estimates. Regarding the exclusion of participants based on their compliance level, a common practice when analyzing IL data is to select participants whose compliance is equal or higher than a pre-specified threshold (e.g., exclude participants whose compliance is lower than 50%, Vachon et al., 2019). Even though Weermeijer et al. (2022) found no differences in the results of hypothesis testing when investigating within-person relationships in momentary affect due to using different compliance thresholds, the impact of these design choices on the strength of associations has not yet been investigated.

In sum, when basing power analyses on results from previous studies to determine the value of the model parameters which are needed to calculate statistical power, the outcome could be heavily affected by particular elements of study design and/or data preprocessing (which can be intended or not, sometimes convenience based, or even random). Knowing the impact of such decisions on the effect size estimates when using data from previous studies and how these estimates impact power-based sample size recommendations could significantly help to guide well-powered IL research.

The Current Study

In this article, we present an approach to investigate how the heterogeneity in study design and preprocessing choices influence power-based sample size recommendations when using data from previous IL studies to estimate the value of the model parameters. Specifically, we implement a safe-guard power analysis (Perugini et al., 2014) which allows to determine the recommended sample that is needed to account for imprecise estimates of the effect size of interest when using data from previous studies. By comparing the recommended sample size from the traditional power analysis and the safe-guard power analysis, we determine how much power-based sample size recommendations vary when considering differences in study design and preprocessing decisions, and therefore, how robust these recommendations are. As a result, the proposed approach provides a range of sample size recommendations that leads to high statistical power reflecting the uncertainty in effect size estimates due to the variation in study design and preprocessing decisions.

Using the data of a recent study (Cloos et al., 2023) that includes distinct types of items to measure positive and negative affect (i.e., standard discrete emotions and unipolar single items), we showcase the proposed approach by creating scenarios aiming to illustrate different decisions related to study design and preprocessing choices. Importantly, those scenarios could lead to different estimates of the model parameters. Since we use the same dataset, we can directly link differences in parameter estimates, their uncertainty, and power-based sample size recommendations to the three factors mentioned above. Specifically, we assess differences in power-based sample size recommendations across (a) 12 different operationalizations of momentary positive and negative affect; (b) differences in the study duration; (c) data-exclusion procedures related to response time and compliance. To showcase the approach, we investigate associations between momentary positive and negative affect and a person-level predictor (depression). We also provide open-source code and a tutorial to implement the proposed approach for power-based sample size planning.

Method

The Prior Dataset

To demonstrate how uncertainty due to the use of a previous study can impact power-based sample size recommendations, we are going to use data from a recent study (Cloos et al., 2023). The sample of this study consists of 161 students and volunteers recruited through the universities' participant management system, advertisement, and social media platforms, with a mean age of 22.41 years (SD = 6.96), an approximately equal male (51%) to female (48%) gender ratio, with two participants not identifying as either (1%). Eligible

participants (i.e., Dutch-speaking, own smartphone) were invited to a video conference information session where they were briefed about the ESM procedure, compliance, and informed consent. During the briefing, they filled in several demographics and baseline questionnaires and the informed consent. Participants also selected their preferred 12-hour time frame for receiving ESM prompts.

A 2-week ESM period started the day after. A signal contingent sampling scheme with semi-random intervals (˜72 minutes between notifications) was used, where participants were prompted 10 times a day within their selected 12-hour time frame to fill in 30-item ESM questionnaires (with 28 different affective items, see the study by Cloos et al., 2023) on their personal smartphones. Once the participants received a notification, they had 30 minutes to respond. If no response was given within 5 minutes, a reminder was sent. Unanswered questionnaires expired 30 minutes after the reminder notification was sent or were overwritten by the next questionnaire. M-Path (Mestdagh et al., 2022) was used for the ESM data collection. Participants could earn up to 50€ (or eight credits for research participation required for first-year psychology students). The reward was based on overall compliance so that 75% compliance was sufficient to earn the full reward, but each decrease of 10% was associated with a deduction of 1 credit/6€. The study was approved by the Social and Societal Ethics Committee—KU Leuven, Belgium. Details about the complete procedure can be found online documentation of the study (Cloos et al., 2023; https://osf.io/485qy/).

Open Science Practices

All analyses of the current study were registered on the Open Science Framework (OSF) after data collection using the template for registration of ESM studies (Kirtley et al., 2021). Registration happened after data collection, but prior to data access (https://osf.io/nxasv), in line with what has been referred to as post-registration (Benning et al., 2019). Deviations from the registration are included in the Supplemental Material. All relevant code and materials are made available online on the OSF project for this study: https://osf.io/rvkte/ and GitLab repository https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0148925/ope_and_preprocess_affect_on_power. In the GitLab repository, we also included a tutorial to showcase how researchers can conduct the proposed analysis using their own dataset.

Measures[2]

Baseline Measures. Depression was assessed with the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977; validated in Dutch by Van de Velde et al., 2011). The scale includes 20 items measured on a four-point Likert-type scale ranging from 1 to 4 for each item (1 = *Rarely or None of the Time*, 2 = *Some or Little of the Time*, 3 = *Moderately or Much of the time*, 4 = *Most or Almost All the Time*; 4 items are reverse coded). A sum score of the items was calculated ($\omega = 0.90$).

ESM Measures. The full ESM questionnaire can be found at https://osf.io/485qy/. At each notification, participants received 30 items. The first 28 items were affect items that were presented in random order. The last two items were event items that were introduced after the affect items in random order. All items were rated on a scale from 0 to 100.

Positive affect was measured with eight positive emotion items that were selected to reflect the four emotions: happiness, vigor, love, and calmness, with one high-intensity item (happy; energetic; loving; relaxed) and one low-intensity item (satisfied; alert; caring; calm), respectively. All items were measured at the momentary level by asking participants to rate the statement "How [emotion item] do you feel at the moment?" on a scale of 0 = *Not at all [emotion item]* to 100 = *Very [emotion item]*. In addition, two single-items were included to measure positive affect at each moment. The first item asks participants to determine "How positive do you feel right now?" on a scale from 0 = *Neutral* to 100 = *Very positive* (= neutral anchor). The second item asks participants to answer the question "To what extent do you experience positive emotions right now?" on a scale with the anchor points 0 = *I don't experience positive emotions* to 100 = *I experience very strong positive emotions* (= absent anchor).

Negative affect was measured with 14 negative emotion items that were selected to reflect seven emotions: sadness, fatigue, shame, guilt, anger, anxiety/fear, and stress, with one high-intensity item (depressed; exhausted; ashamed; guilty; angry; anxious; stressed) and one low-intensity item (gloomy; tired; shy; regretful; irritated; concerned; nervous), respectively. All items were measured at the momentary level by asking participants to rate the statement "How [emotion item] do you feel at the moment?" on a scale of 0 = *Not at all [emotion item]* to 100 = *Very [emotion item]*. Two additional single-item measures of negative affect were included at each moment asking participants to determine "How negative do you feel right now?" on a scale from 0 = *Neutral* to 100 = *Very negative* and to answer the question "To what extent do you experience negative emotions right now?" on a scale with the anchor points 0 = *I don't experience negative emotions* to 100 = *I experience very strong negative emotions* (= absent anchor).

Table 1. Steps of Preprocessing, Data Analysis, and Sample Size Selection.

Step	Description
Data preprocessing	Create three types of datasets varying the study duration of the prior dataset Select measurement occasions based on response delayed Select participants based on average compliance Compute PA and NA measures under different operationalizations
Data analysis	For each preprocessed data and corresponding construct operationalization estimate the 2 models Retrieve the value of the model parameters for each of the 2 models
Sample size selection	For each research question and a predefined number of measurements occasions (i.e., 70 or 140), select the number of persons (*N*) that reach a statistical power equal to or higher than 90% using the analytic approach developed by Lafit et al. (2024) Validate the results of the analytic approach by conducting a simulation-based power analysis using the selected number of persons *N* and the predefined number of measurements occasions

Table 2. Operationalization of Momentary Positive Affect Scores.

Variable name	Description
PAMean	Average of the 8 positive affect items (happy, energetic, loving, relaxed, satisfied, alert, caring, and calm) at each measurement occasion.
PAMax	Maximum value of the 8 positive affect items (happy, energetic, loving, relaxed, satisfied, alert, caring, and calm) at each measurement occasion.
PAHighMean	Average of the 4 high-intensity positive affect items (happy, energetic, loving, relaxed).
PALowMean	Average of the 4 low-intensity positive affect items (satisfied, alert, caring, calm).
neutrPA	Single item measure of positive affect "How positive do you feel?" (0 = *Neutral*, 100 = *Very positive*).
absPA	Single item measure of positive affect "To what extent do you experience positive emotions?" (0 = *I don't experience positive emotions*, 100 = *Very positive*).

Multiverse Analysis Using the Prior Dataset

To investigate the impact of uncertainty in using data from a previous study when taking into consideration heterogeneity in affect operationalizations, study duration, and preprocessing choices on sample size recommendations, we conduct a multiverse analysis (Steegen et al., 2016). To run our multiverse analysis, we follow three steps. First, we created multiple prior datasets that reflect the different affect operationalizations, study duration, and preprocessing choices, using the data from Cloos et al. (2023). Next, we fitted two models of interest to each of these prior datasets. Finally, we used the resulting sets of model parameters to obtain sample size recommendations. Table 1 summarizes these steps.

Prior Datasets: Mimicking Different Affect Operationalizations, Study Duration, and Preprocessing Choices. First, we apply some preprocessing steps on the data from Cloos et al. (2023) following the framework proposed by Revol et al. (2024). We started by checking that all measurement notifications were sent at the correct time. We excluded observations that were sent outside the 12-hour time frame and adjacent observations with less than a minute between them.

In the second preprocessing step, we computed 12 alternative momentary affect scores based on Cloos et al. (2023; see Tables 2 and 3 for an overview). First, we considered four single-item scores: two for positive affect, and two for negative affect. Next, the 22 discrete emotion items were combined into four composite scores of positive affect and four composite scores of negative affect. The four composite scores for positive affect were defined as (a) PAMean computed as the average of the eight positive affect items at each measurement occasion ($\omega_{within} = 0.78$, $\omega_{between} = 0.83$); (b) PAMax computed as the maximum value of the eight positive affect items at each measurement occasion; (c) PAHighMean computed as the average of the four high-intensity positive affect items ($\omega_{within} = 0.72$, $\omega_{between} = 0.79$); and (d) PALowMean computed as the average of the four low-intensity positive affect items ($\omega_{within} = 0.51$, $\omega_{between} = 0.61$). Similarly, the four composite scores for negative affect were defined as (a) NAMean computed as the average of the 14 negative affect items at each measurement occasion ($\omega_{within} = 0.84$, $\omega_{between} = 0.95$); (b) NAMax computed as the maximum value of the 14 negative affect items at each measurement occasion; (c) NAHighMean computed as the average of the 4 high-intensity negative affect items

Table 3. Operationalization of Momentary Negative Affect Scores.

Variable name	Description
NAMean	Average of the 14 negative affect items (depressed, exhausted, ashamed, guilty, angry, anxious, stressed, gloomy, tired, shy, regretful, irritated, concerned, and nervous) at each measurement occasion.
NAMax	Maximum value of the 14 negative affect items (depressed, exhausted, ashamed, guilty, angry, anxious, stressed, gloomy, tired, shy, regretful, irritated, concerned, and nervous) at each measurement occasion.
NAHighMean	Average of the 7 negative high-intensity negative affect items (depressed, exhausted, ashamed, guilty, angry, anxious, stressed).
NALowMean	Average of the 7 negative low-intensity negative affect items (gloomy, tired, shy, regretful, irritated, concerned, nervous).
neutrNA	Single item measure of negative affect "How negative do you feel?" (0 = *Neutral*, 100 = *Very negative*).
absNA	Single item measure of negative affect "To what extent do you experience negative emotions?" (0 = *I don't experience positive emotions*, 100 = *Very negative*).

($\omega_{within} = 0.71$, $\omega_{between} = 0.90$); and (d) NALowMean computed as the average of the 4 low-intensity negative affect items ($\omega_{within} = 0.70$, $\omega_{between} = 0.90$).

In the third preprocessing step, we create three types of datasets varying the study duration of the prior dataset. Specifically, we consider the following three scenarios: (a) a study consisting of 4 days of ESM, resulting in a dataset with 40 measurement occasions per participant; (b) a study including 1 week of ESM (7 days), resulting in a dataset with 70 measurement occasions per participant; and (c) a study including 2 weeks of ESM (14 days), resulting in a dataset with 140 measurement occasions per participant. For the first scenario, we used the first 4 days, and for the second scenario, the first week.

For each of the datasets created in the third preprocessing step, we exclude some of the measurement occasions based on response delay. To do so, we first computed the mean and standard deviation (*SD*) of the response delay across all participants.[3] Response delay was calculated by subtracting the scheduled time of a beep (as programmed in the application) from the actual time of filling in the questionnaire. For each participant, we set a measurement occasion as missing if the response delay is equal to or higher than: (a) Mean (delay) + 1.5 $\times$ *SD*; (b) Mean (delay) + 3 $\times$ *SD*.

Finally, we excluded participants from the sample whose compliance rate was lower or equal to the following thresholds: 30%, 50%, 60%, and 70%. Compliance was defined as the number of daily ESM questionnaires that participants both completed and were not excluded based on response delay, divided by the total number of questionnaires (i.e., 140). For example, if the compliance threshold was set to 70%, then we excluded participants whose compliance rate was lower or equal to 70%. This way, we obtained 24 (i.e., 3 study durations times 2 response delay thresholds times 4 compliance rates) prior datasets that each contained the 6 alternative momentary positive and negative affect scores (see Tables 2 and 3).

Fitting the Model of Interest to the 24 Prior Datasets. These 24 prior datasets and different operationalizations of momentary affect were used to estimate multilevel regression models, to investigate the following research questions[4]:

1. Associations between momentary positive affect and the person-level variable depression;
2. Associations between momentary negative affect and the person-level variable depression.

To investigate these research questions, for the i-th person ($i = 1, \ldots, N$) and the t-th number of measurement occasions ($t = 1, \ldots, T$), we used the following linear mixed-effects model:

$$Affect_{it} = \beta_{00} + \beta_{01} Depression_i + \nu_{0i} + \epsilon_{it},$$

where the Level 2 predictor (Depression) was grand-mean centered. In this model, the coefficient β_{00} represents the mean level of affect when the depression level of the i-th person is equal to the average level in the population. β_{01} represents differences in the mean level of positive affect due to the person-level predictor, Depression, such that if β_{01} is positive, the average affect score is larger for persons with higher depression. Further between-person differences in the mean of the outcome are modeled by the random effect ν_{0i}, which is normally distributed with mean zero variance denoted by $\sigma^2_{\nu_0}$. In this model, we account for the serial dependency that characterizes IL designs by assuming that the Level 1 errors follow an autoregressive (AR(1)) process (Goldstein et al., 1994):

$$\epsilon_{it} = \rho_\epsilon \epsilon_{i,t-1} + \epsilon_{it},$$

where the correlation between two consecutive errors, denoted by ρ_ε and ε_{it}, is a Gaussian error term with variance σ^2_ε. Under this model, the variance of the Level 1 errors σ^2_ϵ is given by $\sigma^2_\varepsilon/(1 - \rho^2_\varepsilon)$. To guarantee that this model is stationary, the autocorrelation should range

between –1 and 1 (Hamilton, 2020). Finally, it is assumed that the random effects and the Level 1 errors are independent. We also note that for conducting the power analysis, we need to specify a distribution for the Level 2 predictor (e.g., see the article by Lafit et al., 2021). We assume that the Level 2 predictor (Depression) follows a normal distribution with mean μ_{Dep} and variance σ^2_{Dep}, as the mean and variance are usually descriptive statistics that are available in previous studies. However, the generalizability of the results presented in the next section should be evaluated with respect to the model assumptions.

The models were estimated using restricted maximum likelihood (Raudenbush & Bryk, 2002). All analyses were conducted using R version 4.0.3 (R Core Team, 2020). Linear mixed-effects models were estimated using the R package nlme (Pinheiro, 2009). We also used the functionality of dplyr (Wickham et al., 2022), tidyr (Wickham & Girlich, 2022), stringr (Wickham, 2019), lubridate (Grolemund & Wickham, 2011), data.table (Dowle & Srinivasan, 2021), and MASS (Venables & Ripley, 2002) to generate and handle the datasets. Finally, ggplot2 (Wickham, 2016), ggh4x (van den Brand, 2024), ggpubr (Kassambara, 2020), gridExtra (Auguie, 2017), scales (Wickham et al., 2023), and viridis (Garnier et al., 2021) were used to create the figures.

Determination of Power-Based Sample Size Recommendations

Next, we investigated how different operationalizations of momentary affect, study duration, and preprocessing choices for excluding measurement occasions and participants based on response delay and compliance impacted sample size recommendations when the goal was to achieve sufficient statistical power (e.g., equal to or higher than 90%) when investigating the research questions introduced in the previous section. Based on typical affect dynamic studies (Vachon et al., 2019), the number of repeated measurement occasions for conducting the power analysis (T) was hereby set to either 70 or 140.

Specifically, for each research question, we used the 288 (24 datasets times 12 construct operationalizations) sets of estimated model parameters to determine the number of participants (N) necessary to reach a statistical power equal or higher than 90%. To this end, we implemented a grid search algorithm that works as follows. First, the algorithm computes statistical power for $N = 30$ and the specified value of T (i.e., 70 or 140), using the analytic approach proposed by Lafit et al. (2024). This approach yields an asymptotic approximation of the standard error of the effect of interest, based on the specified values of the model parameters and N and T. This approximation can then be used to compute the statistical power of a two-sided Wald-type test of the effect of interest at a prespecified α level (which we set equal to 0.05). Compared to simulation-based power analysis, this analytic approach does not require simulating data and fitting a statistical model to these data, making it much faster. In the next step of the algorithm, N is either increased or decreased as a function of the distance between the computed power score and the target value (i.e., 90%). Specifically, when the computed power is far from the target value, the algorithm generates a sequence of values of N with large increments (or decrements) between subsequent values. On the other hand, when the computed power is near the target, the sequence includes small increments (or decrements). The algorithm stops when statistical power is equal to or higher than 90% (with a tolerance of 2%). In the algorithm, we set the minimum N to 20 and the maximum N to 1,000. Thus, it could occur that the algorithm stopped when statistical power was higher than 92% when $N = 20$. On the contrary, if the algorithm reached the maximum $N = 1,000$, statistical power could be below 90%.

Finally, to validate the results obtained using the analytic approach, we re-computed statistical power using a simulation-based approach (e.g., see the study by Lafit et al., 2021). In this approach, a large number of datasets were simulated (i.e., 1,000 datasets) according to the respective model (see section Determination of Sample Size Recommendations) and specified values of model parameters. Each of these datasets was then used to fit the respective multilevel model and test the hypothesis of interest using a two-sided Wald-type test at the prespecified α level. Finally, statistical power was estimated as the percentage of generated datasets in which the null hypothesis of interest was rejected at the selected α level. If the statistical power computed with the simulation-based approach did not reach the target value of statistical power (i.e., 90% with a tolerance of 2%), N was optimized following the procedure described for the analytical approach. Statistical power computed with the simulation-based approach should be very similar to the analytically computed statistical power. More information about the grid search algorithm can be found in the GitLab repository of the project.

Quantifying the Uncertainty in Sample-Size Recommendations Due to Imprecise Effect Size Estimates. Finally, to evaluate the impact of estimation uncertainty of the effect of interest on sample size recommendations, we conduct a safe-guard power analysis following the approach proposed by Perugini et al. (2014). In this approach, a lower-bound estimate of the effect size is used to calculate statistical power for a replication study. By using a

more conservative value of the effect size, a safe-guard power analysis helps to protect against the overestimation of effect sizes when using data from prior studies. Therefore, we conducted a safe-guard power analysis as follows. First, we calculate the 95% confidence interval of the associated effect of interest. We use the lower bound of the 95% confidence interval (i.e., the closest bound to zero) as the estimate of the true effect of interest to compute the number of persons, given the prespecified number of measurements occasions, necessary to reach a statistical power equal or higher than 90% using the analytic approach. This result was again validated using the simulation-based approach for the selected N. Next, we calculate the relative absolute difference in the estimated effect size of interest when using the lower bound of the 95% confidence interval as:

$$\Delta Estimate = \left| \frac{\hat{\beta}_{01,LB} - \hat{\beta}_{01}}{\hat{\beta}_{01}} \right| \times 100,$$

where $\hat{\beta}_{01}$ is the point estimate of the effect size of interest, and $\hat{\beta}_{01,LB}$ is the lower bound of the 95% confidence interval as the estimate of the true effect of interest. Finally, we calculate the relative increase in the sample size recommendation due to imprecise effect size estimates as:

$$\Delta N = \frac{N_S - N_P}{N_P} \times 100,$$

where N_P denotes the power-based recommended number of participants when the point estimate of the effect size of interest is used. Meanwhile, N_S indicates the selected number of persons based on the safe-guard power analysis. ΔN represents the percentage increase in the recommended number of participants due to imprecise effect size estimates and can be interpreted as the increase in the sample size recommendation needed to prevent underpowered studies due to imprecise estimation of the main effect of interest (e.g., see the study by Perugini et al., 2014).

Results

Prior Datasets: Mimicking Different Affect Operationalizations and Preprocessing Choices

Descriptive statistics of the different affect operationalizations are included in the GitLab repository of the project. Figure 1 summarizes the sample size of the 24 created prior datasets, reflecting different study durations and preprocessing choices regarding response delay and compliance. Logically, we observe that the stricter the compliance threshold and the response delay

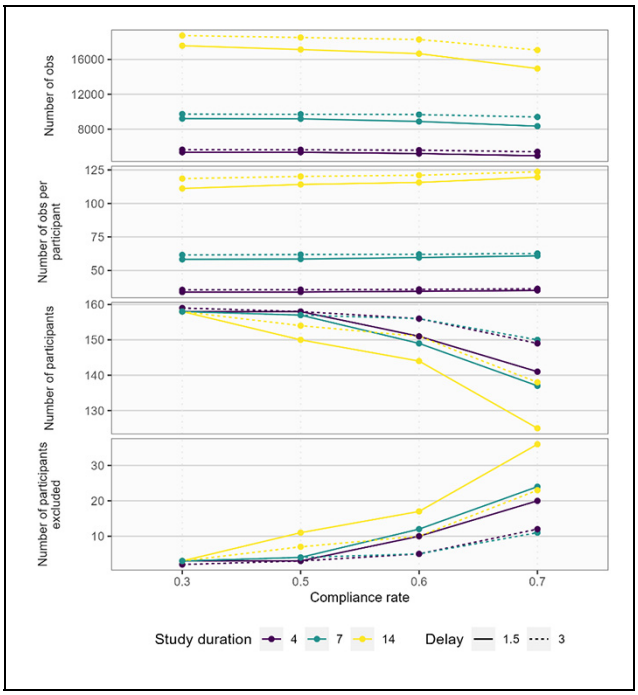


Figure 1. Data Quantity Across Preprocessing Criteria.
Note. From top to bottom: (a) total number of observations; (b) average of the number of observations per participant; (c) number of participants; (d) number of participants excluded from the sample.

threshold, the lower the number of participants and measurement occasions included in the sample. For example, when the study duration was 14 days, the threshold for excluding delayed responses is 1.5, the compliance threshold is 0.70, and the number of participants is below 130. Meanwhile, if the threshold for including delayed responses is 3 and the compliance threshold is 0.5, the number of participants is above 150. Finally, we also observe that when the study duration is 14 days and the threshold for excluding delayed observations is 1.5, the number of excluded observations increases. This effect is even larger when the compliance threshold is set to 70%. This pattern can be partially explained by compliance levels decreasing over days (see preprocessing report in the Supplemental Material).

Fitting Models of Interest to the 24 Prior Datasets

Figures 2 and 3 display the estimated parameter values of the effect of interest and their 95% confidence interval across the different affect operationalizations, study duration, and preprocessing choices. For each research question, we observe that the estimated effects of interest differ in the same direction from zero. For example, the estimated association between momentary positive affect and the person-level variable depression is

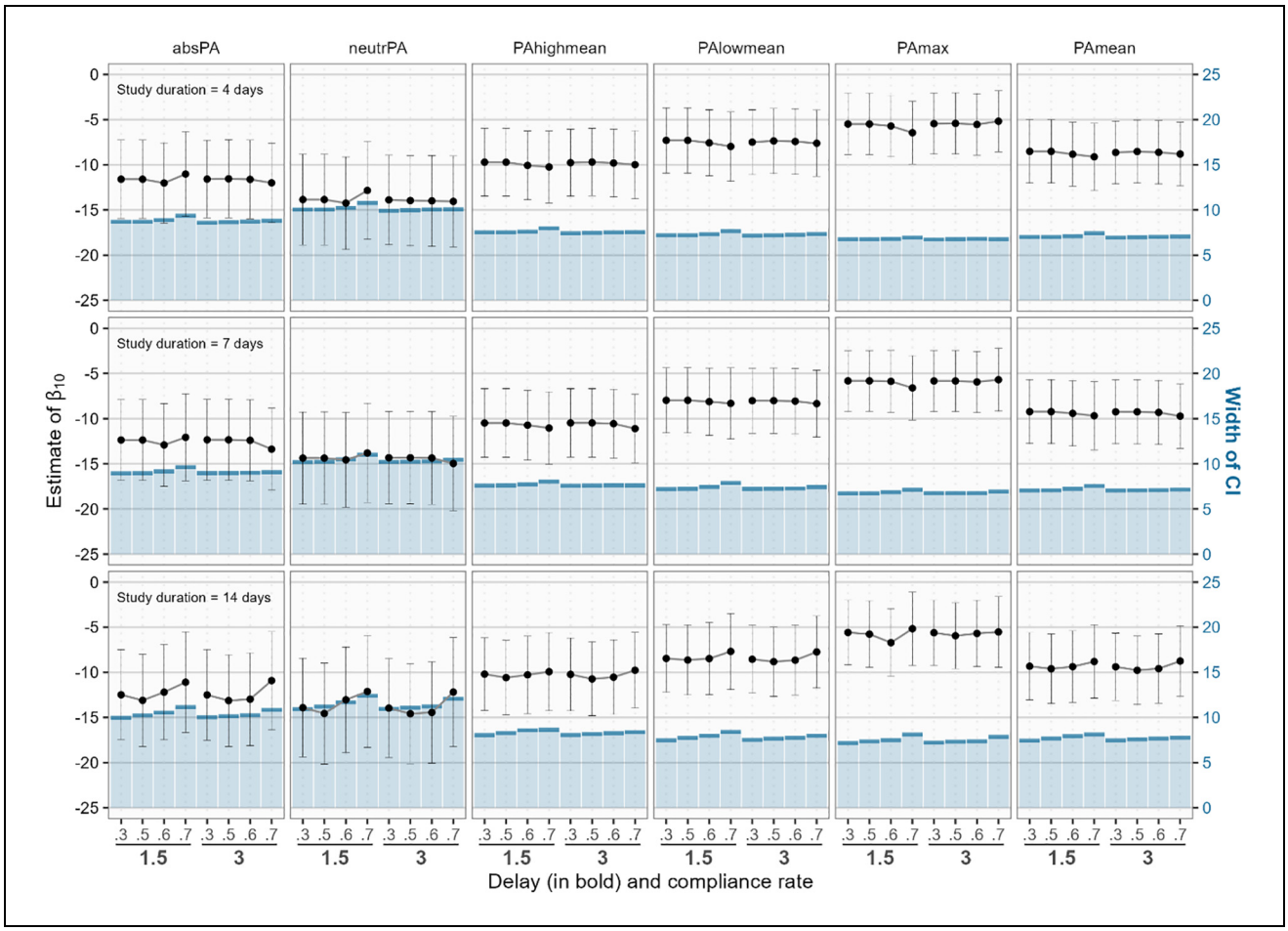


Figure 2. Estimated Value of the Effect of Interest, 95% Confidence Interval, and Width of the 95% Confidence Interval (Right Y-Axis) of Hypothesis 1 Investigating the Association Between Momentary Positive Affect and the Person-Level Variable Depression Across Affect Operationalizations, Study Duration, and Preprocessing Choices.

negative across different affect operationalizations and preprocessing choices (see Figure 2). However, there are differences in the absolute size of the effect and the width of the 95% confidence interval. For Model 1, we can observe that the width of the 95% confidence intervals is larger when absPA and neutrPA are used to operationalize positive affect. In terms of the effect of study duration and preprocessing, we observe no meaningful differences in the size of the effect of interest and the width of the 95% confidence interval. Results of the multiverse analysis for Model 2 indicate that the largest differences in the size of the effect of interest and the uncertainty are observed when negative affect is operationalized using NAMax. We also observe a small impact of preprocessing choices on the width of the 95% confidence interval when the threshold to exclude delayed observation is set to 1.5 and the compliance threshold is 70%. This shows that using strict thresholds to exclude observations and participants has an impact on estimation uncertainty as the actual sample size use for the analysis decreases.

Determination of Sample Size Recommendations

Figure 4 shows the selected sample size using the analytic and simulation-based approaches when considering the two research questions, the number of repeated measurement occasions (i.e., 70 or 140), construct operationalization, study duration, and preprocessing choices. From a total of 1,152 results,[5] there were 11 (0.96%) cases in which the difference in the selected N between the analytic approach and the simulation-based approach was larger than 10 persons. These results indicate differences between the analytic and simulation-based approaches. These differences are explained by the fact that the analytic approach uses the values of the model parameters to compute statistical power, whereas the simulation-based approach uses the values of the

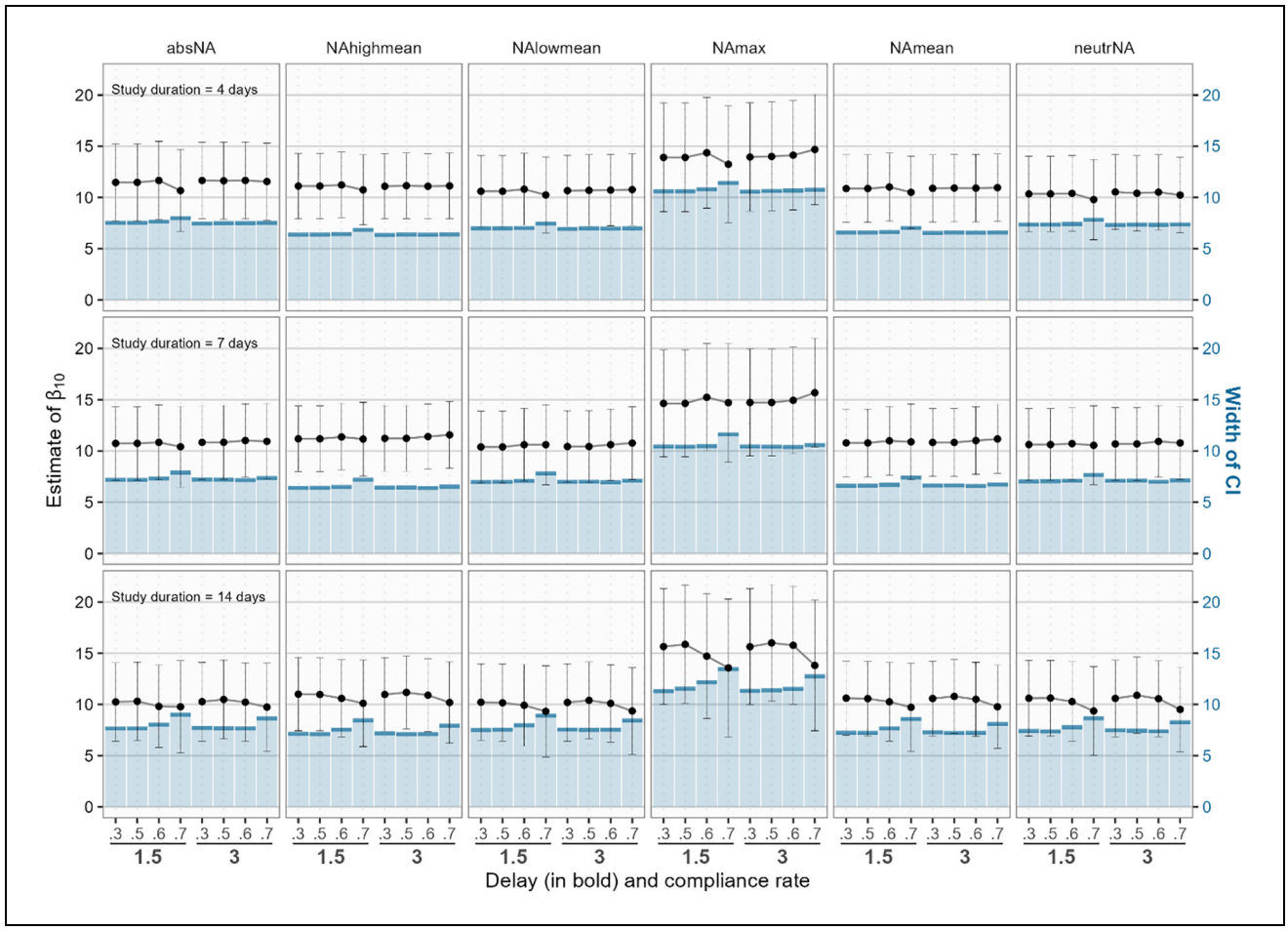


Figure 3. Estimated Value of the Effect of Interest, 95% Confidence Interval, and Width of the 95% Confidence Interval (Right Y-Axis) of Hypothesis 2 Investigating the Association Between Momentary Negative Affect and the Person-Level Variable Depression Across Affect Operationalizations, Study Duration, and Preprocessing Choices.

model parameters to generate data. These data are later used to fit the statistical model, and statistical power is computed using the estimated values of the model parameters, which are influenced by sample size. When the sample size is limited (small N or T), the estimates of the model parameters using the generated datasets can be biased impacting the estimated statistical power. As a consequence, the sample size recommendation between the two approaches may differ. This has been extensively studied in the work by Lafit et al. (2024).

Effect of Construct Operationalization on Sample Size Recommendations

To investigate the effect on power-based sample size recommendations due to differences in construct operationalizations, we consider the two other factors, study duration and preprocessing choices fixed, and we show the results when varying the construct operationalizations. Thus, the study duration was set to 14 days, and the delayed and compliance threshold were set to 1.5% and 30%, respectively. Table 4 displays the recommended number of persons for each research question across the different affect operationalizations. The table also includes the recommended N when using the lower bound of the 95% confidence interval of the effect of interest (i.e., N_S), and the relative increase in the sample size due to uncertainty in the estimation of the effect of interest (i.e., ΔN). Note that we only show the results obtained when setting the number of measurement occasions per person to 70. The results when the number of measurement occasions is set to 140 are included in the Supplemental Material. These results paint a similar picture but the recommended N is slightly lower.

First, we focus on the results of Model 1. We observe that N_P ranges from 71 to 181 participants across the different operationalizations of positive affect, with

estimated effects close to zero leading to larger sample sizes. Specifically, PAMax leads to a smaller effect size estimates and consequently larger N_P. Table 4 also shows the selected number of participants (N_s) when using the lower bound of the 95% confidence interval. The results indicate that more than the double of the number of participants is needed to reach a statistical power of 90% when taken into account the uncertainty due to imprecise effect size estimates. Furthermore, we observe that differences in the strength of the association between depression and positive affect and differences in sample size recommendations are not symmetric. Specifically, even moderate estimation uncertainty can have large impact on the power-based sample size recommendations. This is evident when comparing ΔEstimate and ΔN across the construct operationalizations. ΔEstimate ranges from 39.25% (neutrPA) to 64.16% (PAMax). Meanwhile, ΔN ranges from 146.58% (neutrPA) to increases larger than 452.49% (PAMax).

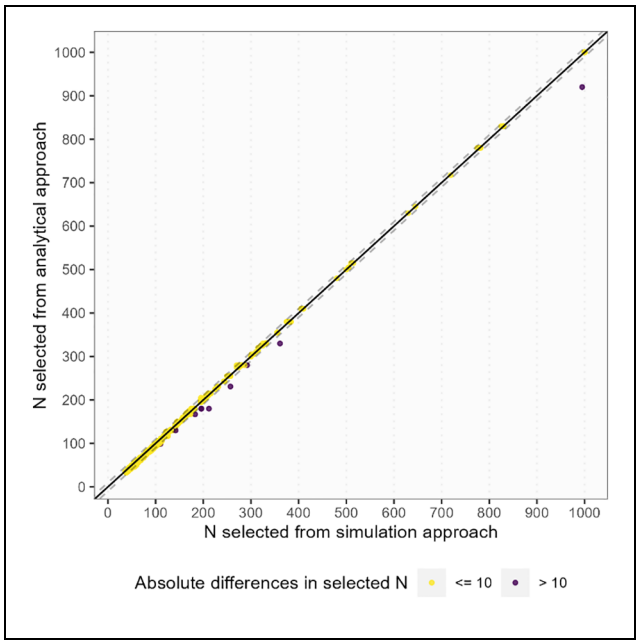


Figure 4. Comparison of the Selected Number of Persons to Achieve Statistical Power Equal to or Higher Than 90% Using the Analytic (X-Axis) and Simulation-Based Approach (Y-Axis) Across the Different Research Questions, Number of Measurement Occasions, Construct Operationalizations, Study Duration, and Preprocessing Choices.
Note. Dashed lines indicate differences between the two approaches of $N = 10$

We now turn to the results of Model 2. In Table 4, we observe that less variation in the effect size estimates across the difference operationalizations of negative affect leads to smaller differences in power-based sample size recommendations (N_P). Similar to Model 1, we observe that the safe-guard sample size recommendations (N_S) suggest to increase N_P more than 100%, and differences in the effect size estimates and sample size recommendations are also non-symmetric, with ΔEstimate taking moderate values around 35%, and ΔN increasing more than 100%.

To sum up, we observe that differences in the absolute size of the estimated effect and in the width of the

Table 4. The Effect of Construct Operationalization on the Selected Number of Participants to Reach Statistical Power of 90%..

Models	Estimate	95% CI	Width 95% CI	ΔEstimate (%)	N_P	N_S	ΔN (%)
Model 1:							
absPA	−12.48	[−17.47; −7.50]	9.97	39.90	76	195	156.58
neutrPA	−13.91	[−19.38; −8.45]	10.93	39.25	73	180	146.58
PAHighMean	−10.20	[−14.22; −6.18]	8.04	39.41	71	195	174.65
PALowMean	−8.46	[−12.20; −4.73]	7.47	44.09	90	280	211.11
PAMax	−5.58	[−9.17; −2.00]	7.17	64.16	181	>1,000	>452.49
PAMean	−9.33	[−13.05; −5.60]	7.45	39.98	75	201	168.00
Model 2:							
absNA	10.24	[6.41; 14.07]	7.66	37.40	68	165	142.65
neutrNA	10.60	[6.90; 14.31]	7.41	34.91	58	133	129.31
NAHighMean	11.00	[7.43; 14.57]	7.14	32.45	49	99	102.04
NALowMean	10.21	[6.46; 13.96]	7.50	36.73	65	155	138.46
NAMax	15.64	[9.99; 21.29]	11.30	36.13	60	147	145.00
NAMean	10.61	[6.99; 14.23]	7.24	34.12	57	125	119.30

Note. Statistical power is computed assuming that the number of measurement occasions is 70, and the true effect size is based on prior datasets varying construct operationalizations of affect, while study duration is set to 14 days, and delayed and compliance thresholds set to 1.5% and 30%, respectively. Width 95% CI is calculated as the difference between the upper and lower bound of the 95% confidence interval, ΔEstimate(%) denotes the relative absolute difference (in percentage) in the estimated effect size of interest when using the lower bound of the 95% confidence interval, N_P denotes the power-based recommended number of participants when the point estimate of the effect size of interest is used, N_S indicates the selected number of persons based on the safe-guard power analysis, and ΔN(%) represents the percentage increase in the recommended number of participants due to imprecise effect size estimates.

Table 5. The Effect of Study Duration on the Selected Number of Participants to Reach Statistical Power of 90%.

Models	Estimate	95% CI	Width 95% CI	ΔEstimate (%)	N_P	N_S	ΔN(%)
Model 1:							
4 days ESM	−8.51	[−12.02; −5.01]	7.01	41.13	74	210	183.78
7 days ESM	−9.23	[−12.75; −5.71]	7.04	38.14	67	165	146.27
14 days ESM	−9.33	[−13.05; −5.60]	7.45	39.98	75	201	168.00
Model 2:							
4 days ESM	10.87	[7.57; 14.16]	6.59	30.27	45	80	77.78
7 days ESM	10.79	[7.48; 14.09]	6.61	30.58	45	88	95.56
14 days ESM	10.61	[6.99; 14.23]	7.24	34.12	57	125	119.30

Note. Statistical power is computed assuming that the number of measurement occasions is 70, and the true effect size is based on prior datasets varying study duration, while affect measures are operationalized as the average of affect items, and delayed and compliance thresholds set to 1.5% and 30%, respectively. Width 95% CI is calculated as the difference between the upper and lower bound of the 95% confidence interval, ΔEstimate(%) denotes the relative absolute difference (in percentage) in the estimated effect size of interest when using the lower bound of the 95% confidence interval, N_P denotes the power-based recommended number of participants when the point estimate of the effect size of interest is used, N_S indicates the selected number of persons based on the safe-guard power analysis, and ΔN(%) represents the percentage increase in the recommended number of participants due to imprecise effect size estimates.

associated 95% confidence interval, largely determine the recommended sample size, with estimated effects close to zero leading to very large sample sizes. Similarly, the results when using the bounds of the confidence interval indicate that the sample size can vary considerably as a function of the uncertainty of the estimated effect of interest, with a more pronounced variability in N_P when lower limits of the 95% confidence interval are closer to zero. Finally, ΔN can also be used to quantify the uncertainty of power-based sample size recommendations across different construct operationalizations, such that operationalizations associated with larger increases in the safe-guard sample size recommendations are related to affect operationalizations with lower effect size estimates (e.g., see PAMax) or higher estimation uncertainty (e.g., see NAMax).

Effect of Study Duration on Sample Size Recommendations. Table 5 shows the results of the recommended number of persons to reach a statistical power of 90% to test Hypotheses 1 and 2 when varying the study duration of the prior datasets. We focus on the results when setting the number of measurement occasions per persons for the to-be-planned study to 70. We further imposed that negative and positive affect were operationalized using the average of the corresponding affect items and that delayed and compliance thresholds were set to 1.5% and 30%, respectively. The rest of the results can be found in the Supplemental Material.

With respect to the estimated effect of interest when using prior datasets, we can observe for Model 1 that longer study duration is associated with larger effect size estimates, but also with an increase in the width of the 95% confidence interval. For Model 2, we observe that the value of the effect size estimate does not increase with study duration but estimation uncertainty does. This result can be explained by the fact that compliance decreases over time. Zooming into sample size recommendations, we do not observe meaningful differences in $\boldsymbol{N_P}$ due to the study duration of the prior dataset for the two models under study. However, we observe differences in the recommended number of participants in the safe-guard power analysis ($\boldsymbol{N_S}$). For the first research question, the prior study with a duration of 7 days is associated with a lower $\boldsymbol{N_S}$ and $\boldsymbol{\Delta N}$, whereas for the second research question, using a prior dataset with 4 days results in lower safe-guard sample size recommendation. These results suggest that if data quality and quantity in ESM studies worsen over time, the uncertainty in parameter estimates of prior studies may increase impacting power-based sample size recommendations.

Effect of Preprocessing Choices on Sample Size Recommendations. Table 6 shows the results of the recommended number of persons to reach a statistical power of 90% to test Hypotheses 1 and 2 when varying the preprocessing choices in the prior dataset. We consider the scenario in which negative and positive affects were operationalized using the average of the corresponding affect items and the ESM study lasted two weeks. Once again, we focus on the results when setting the number of measurement occasions per person in the power analysis to 70. The rest of the results are included in the Supplemental Material.

Concerning the effect of preprocessing choices on the recommended number of persons, we observe that using high-compliance thresholds negatively impacts the estimated value of the effect size of interest and increases the width of the 95% confidence interval. This can be

Table 6. The Effect of Different Preprocessing Choices on the Selected Number of Participants to Reach Statistical Power of 90%.

Models	Estimate	95% CI	Width 95% CI	ΔEstimate (%)	N_P	N_S	ΔN(%)
Model 1:							
Delay; Compliance							
1.5; 30%	−9.33	[−13.05; −5.60]	7.45	39.98	75	201	168.00
1.5; 50%	−9.61	[−13.45; −0.77]	7.68	39.96	74	195	163.51
1.5; 60%	−9.37	[−13.34; −5.41]	7.93	42.26	71	215	202.82
1.5; 70%	−8.80	[−12.87; −4.74]	8.13	46.14	73	245	235.62
3.0; 30%	−9.38	[−13.12; −5.65]	7.47	39.77	73	201	175.34
3.0; 50%	−9.77	[−13.56; −5.98]	7.58	38.79	69	180	160.87
3.0; 60%	−9.59	[−13.42; −5.75]	7.67	40.04	70	195	178.57
3.0; 70%	−8.75	[−12.63; −4.87]	7.76	44.34	74	230	210.81
Model 2:							
Delay; Compliance							
1.5; 30%	10.61	[6.99; 14.23]	7.24	34.12	57	125	119.30
1.5; 50%	10.56	[6.95; 14.18]	7.23	34.28	55	118	114.55
1.5; 60%	10.25	[6.42; 14.09]	7.67	37.46	57	141	147.37
1.5; 70%	9.71	[5.42; 14.01]	8.59	44.28	67	215	220.90
3.0; 30%	10.58	[6.94; 14.22]	7.28	34.40	56	122	117.86
3.0; 50%	10.79	[7.17; 14.40]	7.23	33.46	52	114	119.23
3.0; 60%	10.50	[6.88; 14.12]	7.24	34.48	55	125	127.27
3.0; 70%	9.77	[5.72; 13.82]	8.10	41.45	63	181	187.30

Note. Statistical power is computed assuming that the number of measurement occasions is 70, and the true effect size is based on prior datasets varying the delay and compliance threshold, while study duration is set to 14 days, and affect measures operationalized as the average of affect items. Width 95% CI is calculated as the difference between the upper and lower bound of the 95% confidence interval, ΔEstimate(%) denotes the relative absolute difference (in percentage) in the estimated effect size of interest when using the lower bound of the 95% confidence interval, N_P denotes the power-based recommended number of participants when the point estimate of the effect size of interest is used, N_S indicates the selected number of persons based on the safe-guard power analysis, and ΔN(%) represents the percentage increase in the recommended number of participants due to imprecise effect size estimates.

explained by the fact that using more stringent compliance thresholds impacts the number of participants included in the study (see Figure 1), and therefore the strength of the association between the Level 2 predictor (depression) and momentary affect. However, we observe that N_P does not seem to be severely impacted by this. However, we observe that the higher the compliance threshold, the larger the variability of the selected N computed using the lower bound of the 95% confidence interval of the effect of interest (i.e., N_S). This is related to the impact of the compliance threshold on the width of the confidence interval (i.e., stricter compliance threshold leads to wider confidence intervals due to the lower number of participants included in the sample). Concerning the effect of using different thresholds for flagging delayed observations, we observed that using more stringent threshold for selecting delayed observations (i.e., 1.5) in combination with more stringent compliance thresholds results in larger uncertainty in sample size recommendations, measured by ΔN.

Discussion

In this article, we proposed a novel approach to quantify the uncertainty of using data from previous studies when conducting power-based sample size selection in IL designs. Specifically, we focused on the scenario in which researchers cannot specify the SESOI and rely on data from previous IL studies to estimate the effect size of interest and other relevant model parameters. The proposed approach can be used to quantify the variation in power-based sample size recommendations due to three factors: (a) how affect constructs were operationalized; (b) study duration of the prior dataset; and (c) preprocessing choices related to the exclusion of measurement occasions based on response delay and exclusion of participants based on compliance levels. The approach also allows to investigate how much the sample size has to increase to accommodate for the uncertainty in the estimation of the effect size of interest. Using a previous IL dataset, we focused on the recommended number of persons for investigating typical research questions in affect dynamic research, given a predefined number of measurement occasions. To quantify the uncertainty in sample size recommendations based on imprecise estimates of the effect size of interest, we implemented a safe-guard power analysis (Perugini et al., 2014). In this analysis, the lower bound of a 95% confidence interval of the effect of interest was used to calculate the sample size needed to obtain a statistical power of 90%. The difference between the number of persons obtained with the safe-guard power analysis

and the power analysis using the point estimate of the effect size of interest determined the additional number of participants (i.e., ΔN) needed to avoid conducting a severely underpowered study due to imprecise estimates of the effect of interest. ΔN was then used to compare how robust power-based sample size recommendation were with respect to three aforementioned factors.

With respect to the effect of construct operationalization of affect on power-based sample size recommendations, results showed that ΔN can be used to determine which operationalizations led to more imprecise estimates of the effect size of interest, and therefore, to higher sample size to reach a predefined target for statistical power. Results also showed that even small to moderate differences in effect size estimates due to estimation uncertainty can lead to large differences in sample size recommendations. This has implications when using data from previous studies, and there is a large heterogeneity in the items and operationalizations of time-varying constructs. In this scenario, ΔN can be calculated for each study to compare how robust sample size recommendations are. The results of analyses when varying study duration (i.e., number of days) provided evidence on how data quality and quantity impact the uncertainty in the estimation of the effect of interest based on the prior dataset, and consequently ΔN. Results showed that estimation uncertainty (i.e., measured by the width of the 95% confidence interval) may increase with study duration. This can be explained by compliance levels being negatively associated with study duration. This effect can also worsen when using stringent compliance thresholds (e.g., 70%) for selecting participants to include in the analysis. This higher uncertainty has a direct impact on the sample size recommended by the safe-guard power analysis, meaning that a larger number of persons would be needed to prevent underpowered studies due to overestimation of the effect size of interest. A similar result was found when investigating the effect of using more strict compliance thresholds to select participants in combination with more conservative thresholds to select delayed observations. These results suggested that if data quantity or quality is negatively impacted by study duration or preprocessing choices, then not only the uncertainty in parameter estimation but also the uncertainty in sample size recommendations increases.

In addition, the proposed approach also allows to identify which decisions related to study design or preprocessing choices do not have a significant impact in the uncertainty of sample size recommendations. As we showed in the empirical application, there were some combinations of preprocessing choices that resulted in similar sample size recommendations (and also similar ΔN), showing that despite the differences in the absolute value of the point estimates and their uncertainty across different preprocessing choices, ΔN for a predefined number of measurement occasions remained unchanged across different scenarios.

Practical Recommendations

The results of our study provide guidelines on how researchers can investigate and quantify how certain they can be about the sample sizes that are obtained using statistical power as a criterion, based on prior datasets. As we have shown sample size recommendations can, in some cases, vary significantly as a function of affect operationalization, study duration, or preprocessing; in other cases, the recommendations are unaffected by these decisions. In the former cases, researchers should proceed very cautiously with their new study and adopt the same preprocessing choices and affect operationalizations that led to the recommended sample size. Moreover, if the sample sizes that result from using the lower bound of the confidence interval strongly differ (i.e., ΔN is large), this is best interpreted as a warning sign that the prior data are probably not that helpful to inform sample size planning as there is high uncertainty in the estimation of the main effect of interest.

The apparent influence of the affect operationalizations and preprocessing choices also highlights the importance of transparent reporting practices in IL research (Trull & Ebner-Priemer, 2020). Researchers should be further encouraged to include information on the affect operationalizations they used, going from exact wording of the item, the response scale, to the computation of the affect score. Moreover, they should also document which compliance threshold was used to exclude participants or other types of preprocessing decisions that were made before conducting the analysis. Finally, it would be good practice to provide detailed information on which construct operationalization and preprocessing choices were used when conducing a power analysis.

Limitations

Several limitations need to be considered when interpreting the current findings. First, we only focused on the effect of affect operationalizations, study duration, and two specific preprocessing choices related to response delay and compliance. In practice, there are multiple other design choices (e.g., the type of sampling scheme, the length of the questionnaire), or preprocessing decisions (e.g., exclusion of measurement occasions due to careless responding) that may influence the estimated values of the model parameters. However, such

additional degrees of freedom can in principle easily be added to the multiverse analysis.

Second, we showcased how differences in study design and preprocessing choices impact sample size recommendation using data from a recent study. We consider this to be a common scenario when meta-analyses are not available, and researchers have only available data from a single IL study to obtain the value of the effect size of interest and other model parameters. The presented approach can be used when data from studies included in a meta-analysis are available. Then it would be possible to obtain the values of the model parameters for each of the studies and compare how much power-based sample size recommendations change.

Third, we only investigated the effect of estimation uncertainty about the effect size of interest, assuming that the value of the rest of the model parameters remains fixed. However, imprecise estimation of other model parameters (e.g., variance components, or distribution of the Level 2 predictor) can also impact the estimate of the effect of interest and its uncertainty. In this regard, the proposed approach can also be extended to investigate how robust sample size recommendations are when varying the value of other model parameters. In this regard, we note that if we fix the value of the effect size of interest but assume different values for the other model parameters, the standard error of the estimate of the effect of interest will change, and consequently ΔN. For the model used in this article, Lafit et al. (2024) derived exact formulas for the standard error of the fixed effect estimates. These formulas can be used to determine how and to which degree uncertainty in parameter estimation and ΔN change when varying the value of model parameters, such as the variance components.

In the empirical application, we only considered one type of research question and a specific way of investigating those. Indeed, sample size recommendations based on statistical power may differ depending on the specific hypothesis of interest and the statistical model used to study this hypothesis. For example, in this article, we modeled the serial dependence that characterized IL data by using autoregressive errors. Recent findings suggest that the way serial dependence is taken into account in a statistical model has implications for the interpretation and estimation of the model parameters (Ariens et al., 2023). Yet, the proposed methodology can be easily extended to different types of research questions and/or statistical models. Indeed, different modeling approaches for answering the same question could in principle also be compared in the multiverse analysis.

Finally, in this article, we were not interested in identifying which affect operationalization is better to test a specific research question. Considerations of the validity of affective measures are out of the scope of the present article. Notwithstanding, assessing which affect measure should be used in a study is a central design decision that not only depends on the specific research question and analysis but also on theoretical considerations (e.g., see the work by Cloos et al., 2023; Horstmann & Ziegler, 2020).

Conclusion

The proposed study aimed to shed light on the effect of different operationalizations of momentary affect, study duration, and data preprocessing choices on sample size selection based on statistical power when using data from previous studies to determine the value of the true effect size of interest. Due to a lack of general recommendations on which operationalization to use or on how to best preprocess IL data, it is crucial to investigate to which extent such differences influence statistical power, and consequently, the replicability of findings in daily life research.

Author Contributions

The authors made the following contributions. Ginette Lafit: Conceptualization, Formal Analysis, Methodology, Supervision, Writing—Original Draft Preparation, Writing—Review & Editing; Jordan Revol: Conceptualization, Formal Analysis, Methodology, Visualization, Software, Writing—Review & Editing; Leonie Cloos: Conceptualization, Methodology, Data curation, Writing—Review & Editing; Peter Kuppens: Conceptualization, Funding acquisition, Writing—Review & Editing. Eva Ceulemans: Conceptualization, Methodology, Funding acquisition, Supervision, Writing—Review & Editing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research presented in this article was supported by research grants from the Fund for Scientific Research-Flanders (FWO; Project No. G0C9821N) awarded to E. Ceulemans, from the Research Council of KU Leuven (C14/19/054; iBOF/21/090) awarded to E. Ceulemans and P. Kuppens, and Start-up financing (STG/23/050) of KU Leuven awarded to G. Lafit.

ORCID iDs

Ginette Lafit https://orcid.org/0000-0002-8227-128X
Jordan Revol https://orcid.org/0000-0001-5511-3617
Leonie Cloos https://orcid.org/0000-0003-1550-7405

Supplemental Material

Supplemental material for this article is available online.

Notes

1. For multilevel models, there is no unifying approach for computing standardized effect sizes (Rights & Sterba, 2019). For example, a popular measure to summarize the magnitude of an effect size in simple linear regression analysis is the R-squared, which provides a measure of the proportion of the variance explained by the model. Since, the R-squared takes on values from the interval [0,1], it can be compared across studies with similar designs. However, for multilevel models, the different approaches to compute R-squared measures suffer from methodological pitfalls (see the article by Rights & Sterba, 2019 for an overview). Specifically, these measures do not fully partition the model implied variance in within- and between-cluster variance.
2. Only measures used in the current manuscript are described in detail. For a full overview of the questionnaires administered at baseline and follow-up assessments, we refer the reader to the article by Cloos et al. (2023)
3. We computed the mean and standard deviation of response delay across all participants to avoid including responses with long response delay due to between-person differences in the distribution of the response delay.
4. In this study, we have also investigated power-based sample size recommendations for four additional research questions including association between positive and negative affect and the person-level variable neuroticism and within-person associations between momentary positive and negative affect and event intensity. The results of the analyses for these research questions are included in the Supplemental Material.
5. There were 1,152 sample size recommendations across 576 sets of estimated model parameters and 2 values of the main effect of interest (i.e., point estimate and upper bound of the 95% confidence interval).

References

Abel, D. B., & Minor, K. S. (2023). Social pleasure in daily life: A meta-analysis of experience sampling studies in schizophrenia. *Schizophrenia Research*, *260*, 56–64.

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.

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159.

Ariens, S., Adolf, J. K., & Ceulemans, E. (2023). One does not simply correct for serial dependence. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 807–821. https://doi.org/10.1080/10705511.2023.2173203

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics* (R Package Version 2.3). https://CRAN.R-project.org/package = gridExtra

Bell, I. H., Eisner, E., Allan, S., Cartner, S., Torous, J., Bucci, S., & Thomas, N. (2024). Methodological characteristics and feasibility of ecological momentary assessment studies in psychosis: A systematic review and meta-analysis. *Schizophrenia Bulletin*, *50*(2), 238–265.

Benning, S. D., Bachrach, R. L., Smith, E. A., Freeman, A. J., & Wright, A. G. (2019). The registration continuum in clinical science: A guide toward transparent practices. *Journal of Abnormal Psychology*, *128*(6), 528–540.

Boemo, T., Nieto, I., Vazquez, C., & Sanchez-Lopez, A. (2022). Relations between emotion regulation strategies and affect in daily life: A systematic review and meta-analysis of studies using ecological momentary assessments. *Neuroscience & Biobehavioral Reviews*, *139*, 104747.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLOS ONE*, *8*(4), Article e60188.

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion*, *20*(4), 677–699.

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *35*(3), 189–204.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

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.

Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication." *Perspectives on Psychological Science*, *17*(5), 1490–1505.

Dowle, M., & Srinivasan, A. (2021). *data.table: Extension of "data.frame"* (R Package Version 1.14.2). https://CRAN.R-project.org/package = data.table

Eisele, G., Lafit, G., Vachon, H., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2021a). Affective structure, measurement invariance, and reliability across different experience sampling protocols. *Journal of Research in Personality*, *92*, 104094.

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). 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), 136–151.

Eisele, G., Vachon, H., Myin-Germeys, I., & Viechtbauer, W. (2021b). Reported affect changes as a function of response delay: Findings from a pooled dataset of nine experience sampling studies. *Frontiers in Psychology*, *12*, Article 580684.

Fernández-Álvarez, J., Colombo, D., Penedo, J. M. G., Pierantonelli, M., Baños, R. M., & Botella, C. (2024). Studies of social anxiety using ambulatory assessment: Systematic review. *JMIR Mental Health*, *11*(1), Article e46593.

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.

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods*, *22*(4), 969–1006.

Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., & Scherer, C. (2021). *Rvision–Colorblind-friendly color maps for R* (R Package Version 0.6.2). https://doi.org/10.5281/zenodo.4679423

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

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.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*, 1–25.

Hamilton, J. D. (2020). *Time series analysis*. Princeton University Press.

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*, *54*, 1541–1558.

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (2019). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment*, *31*(4), 432–443.

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality*, *34*(6), 1037–1059.

Janssens, K. A., Bos, E. H., Rosmalen, J. G., Wichers, M. C., & Riese, H. (2018). A qualitative approach to guide choices for designing a diary study. *BMC Medical Research Methodology*, *18*, 1–12.

Joseph, N. T., Jiang, Y., & Zilioli, S. (2021). Momentary emotions and salivary cortisol: A systematic review and meta-analysis of ecological momentary assessment studies. *Neuroscience & Biobehavioral Reviews*, *125*, 365–379.

Kassambara, A. (2020). *ggpubr: "ggplot2" based publication ready plots* (R Package Version 0.4.0). https://CRAN.R-project.org/package = ggpubr

Kaurin, A., King, K. M., & Wright, A. G. (2023). Studying personality pathology with ecological momentary assessment: Harmonizing theory and method. *Personality Disorders: Theory, Research, and Treatment*, *14*(1), 62–72.

Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920924686.

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

Kuppens, P., Dejonckheere, E., Kalokerinos, E. K., & Koval, P. (2022). Some recommendations on the use of daily life methods in affective science. *Affective Science*, *3*(2), 505–515.

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920978738.

Lafit, G., Artner, R., & Ceulemans, E. (2024). Enabling analytical power calculations for multilevel models with autocorrelated errors through deriving and approximating the precision matrix. *Behavior Research Methods*, *56*, 8105–8131.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267.

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, H., Xie, Q. W., & Lou, V. W. (2019). Everyday social interactions and intra-individual variability in affect: A systematic review and meta-analysis of ecological momentary assessment studies. *Motivation and Emotion*, *43*, 339–353.

McGlothlin, A. E., & Lewis, R. J. (2014). Minimal clinically important difference: Defining what really matters to patients. *Journal of the American Medical Association*, *312*(13), 1342–1343.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625.

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kuppens, P., & Dejonckheere, E. (2022). *m-Path: An easy-to-use and flexible platform for ecological momentary assessment and intervention in behavioral research and clinical practice*. https://doi.org/10.31234/osf.io/uqdfs

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.

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*, Article e23383.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332.

Pinheiro, J. (2009). *nlme: Linear and nonlinear mixed effects models* (R Package Version 3.1-96). https://CRAN.R-project.org/package = nlme

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401.

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

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revol, J., Carlier, C., Lafit, G., Verhees, M., Sels, L., & Ceulemans, E. (2024). Preprocessing ESM data: A step-by-step framework, tutorial website, R package, and reporting templates. *Advances in Methods and Practices in Psychological Science*. Advance online publication. https://doi.org/10.31234/osf.io/hnu2t

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*, *36*(1), 203–215.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338.

Schreuder, M. J., Groen, R. N., Wigman, J. T., Hartman, C. A., & Wichers, M. (2020). Measuring psychopathology as it unfolds in daily life: Addressing key assumptions of intensive longitudinal methods in the TRAILS TRANS-ID study. *BMC Psychiatry*, *20*, 1–14.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

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.

Trull, T. J., Lane, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emotion Review*, *7*(4), 355–361.

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), Article e14475.

van den Brand, T. (2024). *ggh4x: Hacks for "ggplot2"* (R Package Version 0.2.8). https://CRAN.R-project.org/package = ggh4x

Van de Velde, S., Levecque, K., & Bracke, P. (2011). Flanders versus The Netherlands: Focus on differences between depressive symptoms in men and women measured on the basis of CES-D8. *Journal of Psychiatry*, *53*(2), 73–82.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Ward, M., & Kemp, S. (2019). The probability of conceptual replication and the variability of effect size. *Methods in Psychology*, *1*, 100002.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063.

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods*, *54*, 2981–2992.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations* (R Package Version 1.4.0). https://CRAN.R-project.org/package = stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A grammar of data manipulation* (R Package Version 1.0.8). https://CRAN.R-project.org/package = dplyr

Wickham, H., & Girlich, M. (2022). *tidyr: Tidy messy data* (R Package Version 1.2.0). https://CRAN.R-project.org/package = tidyr

Wickham H, Pedersen T, Seidel D (2023). scales: Scale Functions for Visualization. R package version 1.3.0, https://github.com/r-lib/scales, https://scales.r-lib.org

Wright, A. G., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, *31*(12), 1467–1480.

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846.