Making the black box transparent: A pre-registration template for studies using Experience Sampling Methods (ESM)

Olivia J. Kirtley[1*], Ginette Lafit[1,2], Robin Achterhof[1], Anu P. Hiekkaranta[1], & Inez Myin-Germeys[1]

[1] Center for Contextual Psychiatry, Center for Contextual Psychiatry, KU Leuven, Department of Neuroscience, Campus Sint-Rafael, Kapucijnenvoer 33, Bus 7001 (Blok H), 3000, Leuven, Belgium.

[2] Research Group on Quantitative Psychology and Individual Differences, Department of Psychology, KU Leuven

* Corresponding author: Olivia J. Kirtley (olivia.kirtley@kuleuven.be)

Author note:

The title for this paper is similar to the title of a symposium at the Society for Ambulatory Assessment (SAA) conference (Syracuse, USA, 19th -22nd June, 2019), organised by the first and second authors, “Making the black box transparent: Open Science practices in ESM research”.

Word count: 5,444

Keywords: Pre-Registration; Reproducibility; Open Science; Experience Sampling; Power; Sample Size; Multilevel Modelling

Abstract

A growing interest in understanding complex and dynamic psychological processes as they occur in everyday life has led to an increase in studies using Ambulatory Assessment techniques, including the Experience Sampling Method (ESM) and Ecological Momentary Assessment (EMA). Whilst a number of researchers working with these techniques are currently actively engaged in efforts to increase the methodological rigor and transparency of such research, currently, there is little routine implementation of open science practices in ESM research. Pre-registration is a cornerstone of open science and as such, a key way of advancing the transparency and reproducibility of ESM research would be the availability of a specific template for the pre-registration of ESM studies. Current general templates do not adequately capture the unique features of ESM, so here we present a pre-registration template adapted for ESM research from the original Pre-Registration Challenge template and provide a walkthrough of each section. We also discuss in more detail the issues of power and sample size calculations in ESM research, a complex issue within the field, which we anticipate to be the greatest potential challenge for researchers wanting to pre-register ESM studies.

Introduction

Some studies require a high level of experimental control, which can only occur in the laboratory, whereas other research is better served by gathering data as participants go about their everyday lives. Ambulatory Assessment is the umbrella term used to refer to measurement of participants in their daily lives, of which the Experience Sampling Method (ESM; Hektner, Schmidt & Csikszentmihalyi, 2007) and Ecological Momentary Assessment (EMA; Stone & Shiffman, 1994) are two subtypes involving participant self-reports. The terms ESM and EMA are often used interchangeably (Trull & Ebner-Priemer, 2014), but in the current paper, we use ESM throughout. ESM involves participants completing a series of brief questionnaires one or more times per day - most commonly now via a smartphone app- to give *in the moment* reports regarding their thoughts, behaviours, contexts and emotions. Such techniques are ideally placed to investigate dynamic psychological processes, as well as addressing issues of recall bias and increasing ecological validity by measuring participants' behaviours in their daily lives (Myin-Germeys et al., 2018; Trull & Ebner-Priemer, 2014).

Recent years have seen a proliferation of studies employing ESM and EMA. Whilst ESM techniques undoubtedly bring numerous advantages, they are also accompanied by a myriad of complex challenges that require significant advance planning and numerous decisions on the part of the researcher. As in non-ESM studies, power and sample size calculations are required (although rarely reported (van Roekel et al., in press), but these are made more complex in ESM research due to the multilevel nature of the data (Bolger et. al., 2012). Similarly, ESM research brings additional considerations regarding item selection, psychometrics and analysis strategy (Wright & Zimmerman, 2018).

As the number of these methodological and statistical decisions increase, so too do the challenges of conducting reproducible research. Asides from the potential "researcher degrees of freedom” (Simmons, Nelson, and Simonsohn, 2011; Wicherts et al., 2016) or data-contingent analysis decisions - the “garden of forking paths” - (Gelman & Loken, 2014), analytic flexibility can also occur simply as a function of individual differences in analysis decisions between researchers (Silberzahn et al., 2018). The issue of many defensible analytic choices existing for the same dataset has also recently been highlighted in ESM research (Bastiaansen et al., 2019). Given the multitude of choices when conducting and analysing data from ESM studies, it is surprising that the first best practice guidelines for conducting ESM research (with adolescents) have only recently been developed (van Roekel, Keijsers & Chung, 2019). This is particularly concerning given that poor study design and analytic flexibility are two major threats to scientific reproducibility (Munafo et al., 2017). Encouragingly, recent years have seen a significant elevation in interest and research energy directed towards addressing methodological and statistical issues in ESM (e.g. Schuurman & Hamaker, 2019; Himmelstein, Wood & Wright, 2019; Houben, Van Den Noortgate & Kuppens, 2015; Rintala et al., 2018; Vachon, Viechtbauer, Rintala & Myin-Germeys, submitted; Wright & Zimmermann, 2018).

The field of psychological science is currently undergoing somewhat of a renaissance resulting from the Replication Crisis, where many high profile studies have been found to replicate poorly (Klein et al., 2018; Open Science Collaboration, 2015). Clinical psychology and psychiatry research, where many ESM studies are conducted, has thus far been noticeably absent from conversations around Open Science (Tackett, Brandes & Reardon, 2018; Tackett et al., 2019), but this does not equate to the methods or results from clinical research being more reproducible or replicable . Open Science practices, including pre-registration of hypotheses and analysis plans on the Open Science Framework (Nosek, Ebersole, DeHaven & Mellor, 2018), are initiatives that aim to promote scientific transparency and reproducibility. Whilst off to a promising start, the implementation of open science approaches to ESM research, including pre-registration, pre-printing and sharing of code and/or materials, are only just emerging (e.g. Dejonckheere, Kalokerinos, Bastian & Kuppens, 2018; Heininga et al., 2019, Himmelstein et al., 2019; van Roekel et al., in press; Zhang, Smolders, Lakens & Isselsteijn, 2018), and there is still some way to go before such practices become widely adopted. A recent episode of the podcast, "*The Black Goat*", with Sanjay Srivastava, Alexa Tullett and Simine Vazire, actually discussed this very issue and raised the idea of a specific pre-registration template for ESM studies. Having already encountered some challenges during pre-registration and planning of our own ESM studies[1] using existing pre-registration templates, the podcast discussion inspired us to devise a template for ESM study pre-registration.

Our central considerations when devising these additions to the general pre-registration template, were to address specific factors of ESM studies that may impact or even preclude their reproducibility, as well as aspects that may be vulnerable to questionable research practices or analytic flexibility. Some of these additions, such as questions regarding participant incentives or instructions may not seem obvious to include in a pre-registration, however these are all decisions that impact compliance (i.e., response rate), and consequently data quality. Furthermore, these are all factors which must already be determined prior to commencement of data collection. Therefore, we believe they are sufficiently relevant to include here.

Myin-Germeys et al (2009) referred to ESM as a technique for "opening the black box of daily life", however, over time it is the application of ESM itself, rather than daily life, that has remained a black box. With this in mind, we endeavor to make the proverbial black box transparent, by facilitating pre-registration of ESM research with a specially adapted template. In the current paper, we walk through the key additions and modifications to the Pre-Registration Challenge template (Mellor et al., 2019), upon which our template is based,

[1] Thus far, these have all been pre-registration of studies using pre-existing datasets.

and provide brief examples. We also include more substantive discussion of issues around sample size and power calculations, including simulations (code available via our OSF page). Open science is dynamic and resources are frequently improved as a result of rapid and interactive community feedback, therefore, we actively encourage other researchers to test the template (available at https://osf.io/2chmu/) and to provide us with critical feedback.

Additions to the OSF Preregistration Challenge pre-registration template (Mellor et al., 2019)

ESM data collection procedure

When conducting an ESM study, numerous decisions must be made regarding data collection, including the method of data collection, sampling scheme, and participant engagement incentives. Often, only a selection of these decisions are reported in the final research article, therefore we have added a new section to the pre-registration template, entitled 'ESM data collection procedure'. Learning more about these procedures improves the reproducibility of ESM studies, and gives more insight into potential influences on compliance and participant burden. In this section, researchers are asked to comprehensively describe the data collection procedure of the ESM component of their study. It is worth noting that given the richness of ESM data, most ESM datasets will be used for multiple subsequent analyses and research articles. Therefore, pre-registrations involving secondary data analysis may wish to refer to an earlier pre-registration, wherein most of the details that are required in this section are already described. For further discussion and resources for pre-registration of pre-existing data analysis, see Mertens and Krypotos (2019) and Weston, Ritchie, Rohrer and Przybylski (2019).

The following subsections are included in this section of the pre-registration form, however, not all subsections are relevant to every ESM study. Examples for each subsection can be found in the template.

Study duration (number of days)

Wide variation exists in the number of days over which ESM data are collected, as a function of expected variability in target behaviours and feasibility (Janssens, Bos, Rosmalen, Wichers, & Riese, 2018). Occasionally, researchers wish to extend the ESM period for (a subsample of) participants. Variations in the length of the ESM period can be indicated in this subsection. Study duration may affect compliance, as compliance can decrease throughout the ESM period (Rintala, Wampers, Myin-Germeys, & Viechtbauer, 2019). Conversely, a review of ESM studies in youth found no difference in compliance rates between studies of different lengths (Wen, Schneider, Stone, & Spruijt-Metz, 2017), and one recent meta-analysis found no significant effect of study duration upon compliance (Vachon et al., submitted).

Type of sampling scheme

The sampling scheme of ESM studies refers to the timing of questionnaire prompts. Generally, sampling schemes fall into three categories: Interval-contingent, event-contingent, and signal-contingent. See Christensen et al., (2003a) and Reis & Wheeler, (1991) for further information. In this subsection, researchers are asked to describe the details of their sampling scheme, including the type of sampling scheme, number of beeps per day, timing of beeps, and potential minimum windows between consecutive beeps. The sampling scheme is dependent on the temporal dynamics of the construct that it aims to measure. For example, relatively rare occurrences (e.g., alcohol consumption) are likely best measured with an event-contingent design, whereas rapid fluctuations in, for example, mood levels, might be best captured with a (semi-)random design. More details on the consequences of the decision for a specific temporal design will be discussed in the section 'Rationale for sample size: Temporal design and number of participants' below.

The aim of the temporal design is to determine the number of measurements within an individual that are necessary to obtain reliable estimations of the target phenomena. Two key components of temporal design in ESM research are the study duration, i.e. the total number of measurement occasions, and the sampling frequency, that is, the time interval between two different measurements (Collins & Graham, 2002). Decisions regarding the number of measurement occasions should take into account that the study duration should be long enough to capture the temporal dynamics of the target process. For example, the number of days in an ESM study investigating the effect that social interactions have on alcohol use, should be sufficient to capture this effect. Another consideration when setting the study duration is the periodicity or seasonality. For instance, if we are interested in studying a process with high probability of occurrence during weekend days (e.g. alcohol use) and we only measure on weekdays, then we might conclude that the effect is weaker or non-existent.

With sampling frequency, it is important to consider that the magnitude of the effect of the within person variability in a longitudinal study is closely related to the length of the interval in which the target process is measured (Timmons & Preacher, 2015). If the target phenomenon is characterized by a rapid change over time or frequent fluctuations, then shorter intervals are needed in order to capture the variability of the target phenomenon (Collins, 2006).

Given the many possible variations for the configuration of a specific sampling scheme, this may include additional details that have not been specified here. For example, signal-contingent sampling schemes may be random, with prompts truly randomly distributed throughout the measurement period, or semi-random, where the randomness of prompts is restricted, for example, by scheduling each consecutive prompt randomly within a pre-specified time block.

Total number and type of items (open-ended or closed)

Many reports of ESM studies only include a description of variables that were analysed for that specific study, and while the number of items per ESM assessment varies greatly (Janssens et al., 2018), the total number and type of items included in the ESM questionnaire are only infrequently reported (Morren, van Dulmen, Ouwerkerk, & Bensing, 2009; Vachon, et al., submitted; van Roekel, Keijsers, & Chung, in press). Consequently, the effect of the total questionnaire length on the total compliance rate is unclear. Here, researchers are asked to provide a general description of the total questionnaire length. A longer ESM questionnaire with more open-ended items signifies a greater participant burden, which may affect the compliance rate. At the same time, a longer questionnaire may improve data quality, as the measurement reactivity may be reduced when more questions are asked (Palmier-Claus et al, 2011). Questionnaire length may also vary due to conditional branching, where the presentation of certain items is dependent upon previous responses. Additionally, researchers may choose to present items in a different random order at each prompt (also referred to as ‘item rotation’; Wen, Schneider, Stone, & Spruijt-Metz, 2017). This type of information can also be described in this subsection.

Although it is beyond the scope of a pre-registration to include the full list of ESM items, we do advise making ESM materials available online to enhance transparency and reproducibility. The Experience Sampling Item Repository (Kirtley et al., 2019) is an ongoing project that aims to produce an open bank of ESM items for use in research and to quality assess and psychometrically validate these items. Researchers can consider using items from this repository as well as contributing items to make their materials open.

Time-out specifications

In order to reduce recall bias, many ESM studies limit the amount of time that participants have to respond to a questionnaire (i.e., the response window), the amount of time that participants can spend on one item, and/or the amount of time that participants may take to complete one full questionnaire. In this subsection, researchers may indicate these timing restrictions.

Additional passive monitoring.

In addition to self-report questionnaires in ESM, studies may also employ methods to gather data in the background without requiring participants to actively respond. For example, using wearable technology to measure heart rate, smartphone sensors to record location, or a background app to measure social media usage.

Hardware and software used, and prompt type specifications

Most contemporary ESM studies collect ESM data electronically. Some researchers choose to use participants' own smartphone, whereas others provide participants with a study device. Here, researchers may indicate their decisions for hardware and/or software. If ESM data are not collected digitally, this may also be indicated here.

In addition, although not yet formally investigated, it is conceivable that the type of prompt used to alert participants that a beep is available, may affect compliance (either negatively or positively). In sampling schemes where participants are prompted to respond to the questionnaire, this prompting is usually through sound/vibration. In order to gain a better sense of participant burden, it is also useful to describe the extent to which the ESM procedure allows participants to alter the nature of the prompts, or the option to turn notifications off entirely.

ESM instruction

Although fully describing the ESM instruction is beyond the scope of a pre-registration, some details regarding the manner in which participants are instructed to complete ESM questionnaires are relevant for enhancing reproducibility. The manner of instruction is likely to affect the compliance of participants (Christensen et al., 2003b). There are various instruction options that may affect compliance, motivation, and data quality. These include, but are not limited to, the type of instruction (video, one-to-one, in a group session), duration of instruction, and whether participants complete a practice questionnaire (see Palmier-Claus et al., 2011 for recommendations). Any specific instructions about how and when participants are required to complete the questionnaires are also useful.

Incentivization and increasing participant engagement

Given the relatively intensive nature of ESM data collection, maintaining participants' motivation and compliance may be challenging. Researchers often employ different methods to increase participant engagement, such as gamification, calling participants during the ESM period, providing progress updates, or making participant payment conditional on the level of participation (Palmier-Claus et al., 2011). One review of studies in youth, however, found no difference in compliance rate between studies using fixed versus incremental monetary incentives (Wen et al., 2017). In this section, researchers may give a description of their efforts to increase participant compliance.

In addition, researchers should also take into account factors that may reduce compliance, as this can affect the quality of the data (Delespaul, 1995; Palmier-Claus, et. al., 2011; Rintala, et. al., 2019). We recommend that in the pre-registration of studies based on pre-existing data, researchers report the observed compliance rate. For pre-data collection ESM studies, we encourage researchers to report the expected compliance. For instance for ESM using electronic diaries with five to seven measurements per day over a period of four to seven days, the compliance rates varies from 66% to 86%, but from 66% to 93% for paper and pencil diaries (see Rintala, et. al., 2019).

Variables

As the majority of ESM research is observational (Hektner, Schmidt, & Csikszentmihalyi, 2009; Myin-Germeys et al., 2018), and does not include manipulated variables, the order of the 'variables' section was reversed, with measured variables reported first and manipulated variables last. As in the original pre-registration template, researchers are only asked to describe variables which will be used in confirmatory analyses. In order to account for the combination of time-invariant and time-variant variables that ESM research commonly features, the 'measured variables' section was divided into 'measured non-ESM variables' and 'measured ESM variables'. In the 'measured ESM variables' section, instructions to specify the response scale of ESM variables (e.g. Likert-scale) were added. In the 'measured ESM variables' section, a field was added for reporting (where applicable) ESM item presentation order and branching were contingent on, for instance, questionnaire answers or contextual changes, such as participant-reported events. Given the multilevel structure inherent to ESM data, researchers are asked to specify variable levels in both the 'measured' and 'manipulated' ESM variables sections. As some ESM studies provide free response options for specific items, an optional section 'open-ended questions' was added, where researchers are asked to indicate how answers will be coded. Within-participant ESM level manipulations are currently less common, therefore instructions to report both manipulated ESM and manipulated non-ESM variables in the 'manipulated variables' section were added. In the 'indices' section, instructions were updated to include descriptions of how any measurements collected during or outside of the ESM period will be combined into an index, including possible passive monitoring conducted via, e.g. activity tracker etc. As summary statistics, such as within-person averages, can be formed at different levels in ESM datasets and employed as predictors, outcomes, or covariates, an instruction to specify the level of index was added. For instance, average positive mood may be calculated per beep, per day, or over a longer period of time.

Rationale for sample size: Temporal design and number of participants

Despite being a crucial consideration for all research, most ESM studies do not report a power calculation. This may be due to the relative complexity of conducting power analysis for multilevel models. For this reason, in this section we include some further, in-depth discussion of key considerations for power calculations in ESM research, along with some illustrations from simulations to highlight the critical importance of conducting power analysis for ESM studies.

The structure of ESM data allows the examination of variability in a target process, sover time, within an individual (i.e. within-subject variability), and the assessment of differences between individuals (i.e. between-subjects variability). Thus, sample size considerations in ESM studies must take into account two elements, discussed earlier in the paper: the temporal design in which the target process will be observed, and the number of participants (Collins, 2006).

The selection of the number of repeated measurements within an individual is closely related with the temporal dynamics of the target process. A measure of how consecutive measurements are linearly related over time is given by the autocorrelation. To illustrate how the autocorrelation of a dynamic process affects the power to detect an effect, we simulated data for a single individual under different scenarios. For instance, if we have one individual that is measured on T occasions, we model the outcome of interest or dependent variable, Y_t, using a first order autoregressive model:

$$Y_t = \rho Y_{t-1} + \epsilon_t$$

where ρ is the autocorrelation and indicates how past observations influence the current state of the process. The magnitude of the autocorrelation will affect the precision in the estimation of the mean of Y_t. This is due to the fact that observations are no longer independent. This can be illustrated by the effective sample size, which is defined as the number of additional observations we need to sample if we want to achieve the same precision as in the case observations are uncorrelated (Zięba, 2010):

$$T_{eff} = T(1+\rho)/(1-\rho)$$

To illustrate the effect that autocorrelation has on the width of the confidence interval of the mean, we simulate data from an autoregressive process assuming three different values for the autocorrelation: uncorrelated errors (0.0), medium autocorrelation (0.4) and large autocorrelation (0.4). We simulated 60 observations and we compute the confidence intervals for the mean using the effective sample size (see https://osf.io/2chmu/ for the R script).

Figure 1 shows the confidence intervals for the mean when varying the autocorrelation. We observed that when the autocorrelation increases the width of the confidence interval for the mean increases. This reflects the fact that when the autocorrelation is positive, then more observations are needed in order to capture the underlying variability of the process.

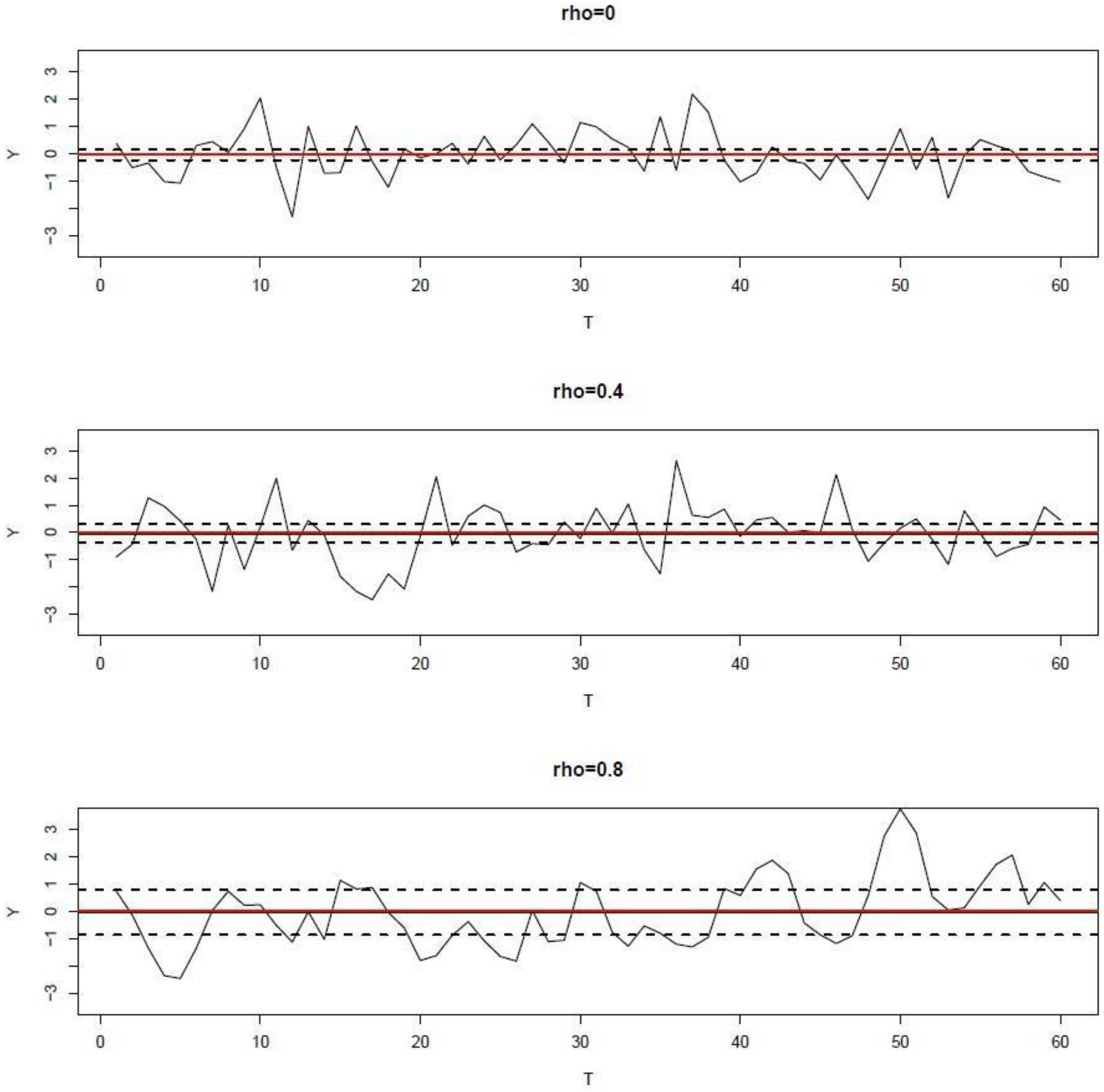

Figure 1: Effect of the effective sampling rate on the confidence intervals of the mean. We assume different autocorrelations: uncorrelated (0.0), medium (0.4) and large (0.8).

The temporal design varies considerably between published ESM studies. ESM studies in psychiatry that asses highly variable constructs (i.e. mood) generally use ten measurements per day, for six consecutive days (Myin-Germeys et. al., 2018). Conversely, a study of global self-esteem used just one measurement per day for seven consecutive days (Christensen et. al., 2003b) and a study of emotions collected three measurements per day for 90 consecutive days (Barret, 1998). Therefore, we encourage researchers to not only justify the choice of the temporal design based on theoretically-informed, expected variability of the target process over time, but also to carefully consider the importance of performing a power analysis to select the sampling rate, and the total number of measurements necessary to detect a hypothesized effect size (see Timmons & Preacher, 2015). As Collins (2006) highlights, an explicit justification of the choice of the temporal design will increase the reproducibility of longitudinal studies.

The second consideration regarding the sample size determination in ESM research relates to the number of participants necessary to obtain accurate estimates of inter-individual differences (Maas & Hox, 2005). In studies in which individual differences are large, more information is needed in order to determine an effect in comparison to a case in which heterogeneity between individuals is negligible. Therefore, once the temporal design has been established, we suggest that researchers perform a power analysis to determine the number of participants that maximize the likelihood of detecting a hypothesized effect size (Bolger et. al., 2012; Raudenbush & Liu, 2001; Snijders & Bosker, 1993). For ESM studies including individuals from different populations (e.g. studies with patients with different mental health conditions), we also suggest that power analysis is performed to determine group size. The same applies to ESM studies that include a higher order grouping level, such as dyad studies or groups under different treatments conditions.

Here, we illustrate the effect of sample size on power with a simulation experiment. We assumed that the ESM study had a duration of 6 days with 10 measurement occasions per day. Let us denote the number of individuals with N and the total number of occasions with T. We are interested in estimating the effect of a time varying predictor variable X_{it} in the outcome of interest Y_{it}. We modeled this relationship using a linear mixed effects model with a random intercept and random slope:

$$Y_{it} = \beta_{00} + \beta_{10} X_{it} + \nu_{0i} + \nu_{1i} X_{it} + \epsilon_{it}$$

we assume that the errors ϵ_{it} are serially correlated and the correlation is modeled with an AR(1) process with the autocorrelation equal to 0.3 (medium). We assume that the true intercept β_{00} is equal to one, the variable X_{it} is normally distributed with a mean of zero and a variance of one. The random effects, ν_{0i} and ν_{1i}, are bivariate normal distributed where the variance of the random intercept is one, the variance of the random slope is 0.1 and the correlation between the random effects is 0.5. We set three different scenarios for the magnitude of β_{10}: small effect (0.1), medium effect (0.3) and large effect (0.5). For each scenario, we simulated 1000 data sets and estimated the linear mixed effects model. Next, we estimated the power to test the null hypothesis thatβ_{10} is zero, for a significance level of 5% (see https://osf.io/2chmu/ for R script). Figure 2 shows the power curve when we varied the number of individuals. We observed that when the true effect in the population is small, the sample size should be at least 250 in order to achieve a power greater than 80%. For medium and large effects, the sample size should be at least 50 in order to achieve a power greater than 80%.

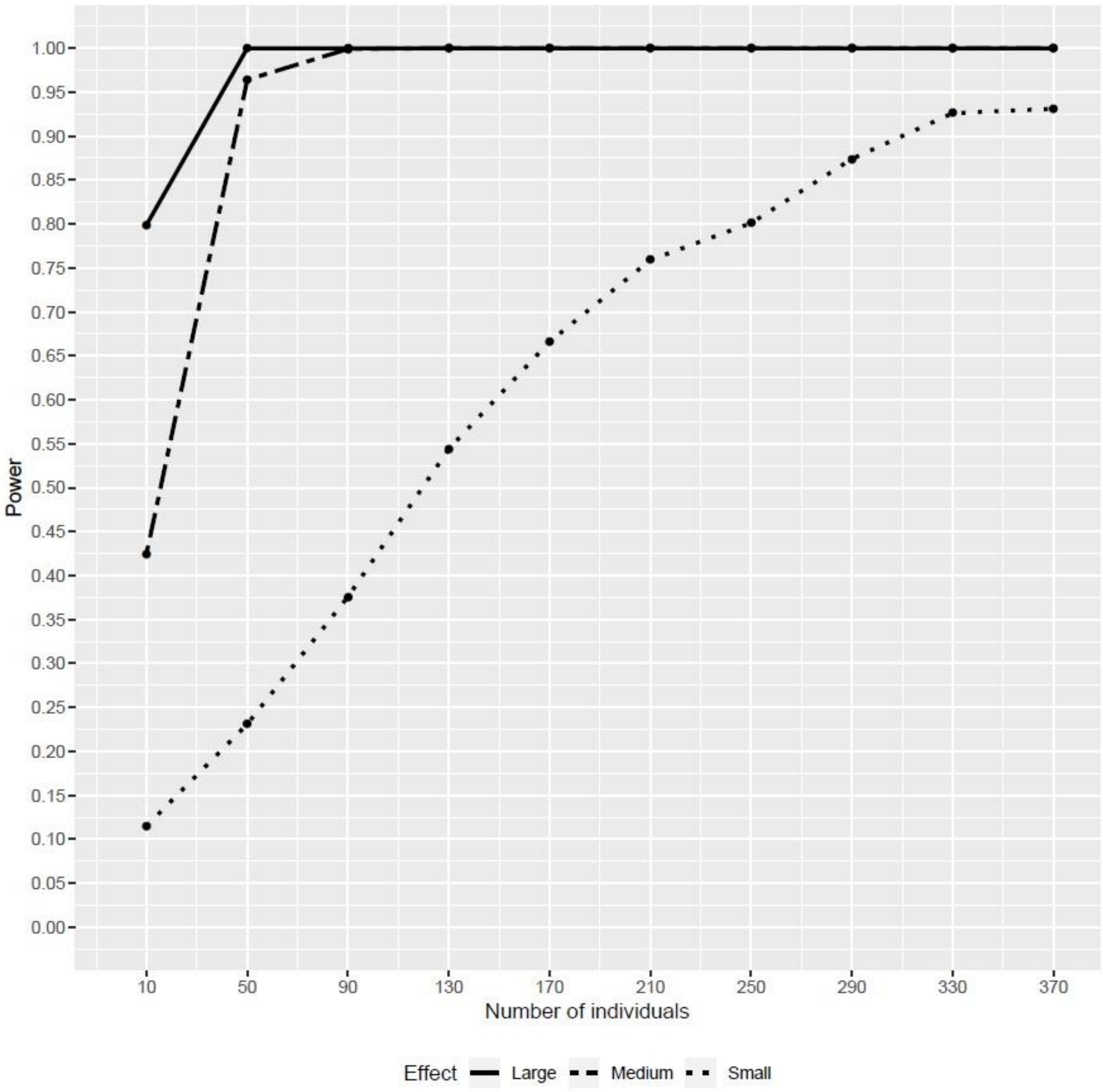

Figure 2: Statistical power for β_{10} when varying the number of participants for an ESM study with a duration of 6 days with 10 occasions per day. We assume different different effect sizes: small (0.1), medium (0.3) and large (0.5).

Analysis Plan

ESM studies produce data with a multilevel structure, in which repeated measurements are nested within days, within participants. Variables are measured at different hierarchical levels and consequently, researchers may be interested in analysing the interaction between variables that describe the within-subject variability and variables that describe the between-subject variability. Moreover, due to the longitudinal structure of the data, the temporal dynamics of the target process can be modelled. Given the complexity of ESM data (i.e. missing observations, unequally spaced time points, time varying covariates, autocorrelated observations, higher-level models, non-normal data), the most widely used statistical approach in ESM studies is the multilevel or mixed effects model (Myin-Germeys, et. al., 2018).

In order to restrict our attention to specific relevant considerations for an ESM analysis plan, we focus on the multilevel regression model. This framework can be considered as a hierarchical system of regression equations (Snijders & Bosker, 2012). The analysis plan should take into consideration the following aspects of the statistical model: (a) distribution of the outcome variable, (b) distribution of the within subject errors (c) distribution of the random effects, (d) fixed effect predictors and interactions, (e) transformations applied to time-varying explanatory variables and time invariant explanatory variables (f) inclusion of lag-dependent variables, and (g) missing data.

Whilst these considerations may seem too numerous and effortful to record as part of a pre-registration, they all represent potential “forking paths” where a high degree of analytic flexibility may be introduced into the research. For example, the distribution of the outcome variable determines the statistical model to be used in the analysis. The linear mixed effects model assumes that predictors are linearly related to the outcome variable and that the within-individual errors are independent, have equal variance and are normally distributed. These assumptions are often stringent for the analysis of ESM data. Researchers can opt to apply a transformation to the outcome variable to normalize its distribution; an important decision that should be noted in the analysis plan section of the pre-registration.

Another example is where random effects allows the modeling of non-independence between individuals. In general, random effects are considered to be normally distributed (models that do not assume normality for the random effects can be found in Verbeke & Lesaffre (1996)). For instance, a model that only incorporates a random intercept and a fixed slope assumes that the outcome mean level differs between individuals, but the slope does not differ between subjects. A model that also includes a random slope assumes that the slope varies between individuals. It has been shown that misspecification of the random effects can inflate the Type I and Type II errors (Aarts, et. al., 2015). Therefore, it is important that researchers explicitly report the structure of the random effects (e.g. if the slope is considered fixed or random).

When considering the predictors included in the statistical model an important decision is to decide which of the predictors are going to be set as fixed effects and the cross level interaction. This depends on the hypotheses, so is an important *a priori* -and therefore pre-registerable - decision for researchers to take. Researchers should also consider model complexity; models that include a large number of predictors and cross level interactions reduce the number of degrees of freedom and affect the estimated variance of the prediction errors.

Another consideration regarding the predictors are the transformations. We advise stating which are the expected transformations of the data. For example, a common practice in multilevel modeling is to center the time varying predictors using the individual mean and to center the time invariant predictors using the grand mean (Snijders & Bosker, 2012). Also, the approach used to validate a set of items, measuring a certain construct, should be explicitly stated (e.g. within-person factor analysis where items are centred per person and over the ESM period; reliability estimation using multilevel confirmatory factor analysis).

For models including a lagged variable as a predictor, it is necessary to specify the method used to account for the 'overnight lags'. For example, a common approach is to set the first beep of the day as missing (de Haan-Rietdijk, et. al. , 2017).

Finally, we note that there are many software packages to estimate multilevel models (McCoach et. al., 2018), including R, MPlus, Stata, JAMOVI and SPSS. We encourage researchers to specify the software they use and whether the default options of a function or software were used. Even better, researchers can share their statistical analysis plan and code.

Missing data

In ESM studies if an individual does not respond to a beep, then the entire set of items will be missed. In the statistical analysis plan it is important to state how missing data will be handled. For example, if there are some expected patterns of missingness (i.e. people are less likely to respond during working hours), then incorporating additional predictors that account for non responses (e.g. time) into the statistical model can help to reduce the bias due to missing observations (Silvia et. al., 2013). To reduce participant burden, some researchers may opt for a planned missing design, where participants receive a selection of items representing a particular construct, as opposed to the full set. Researchers can indicate this in the missing data section of the pre-registration. For further discussion of planned missing designs in ESM and their implications, see Silvia, Kwapil, Walsh and Myin-Germeys (2014).

Conclusions

In the current paper we have presented a pre-registration template for ESM research, the development of which was inspired by discussions on *The Black Goat* podcast. We have also included detailed explanations, accompanied by simulations, to facilitate power and sample size calculations for ESM research, to demonstrate their importance, and illustrate the implications of ignoring this. Many researchers are already making great strides in increasing reproducibility and transparency in the field of ESM research (e.g. Dejonckheere et al., 2018; Heininga et al., 2019, Himmelstein et al., 2019; van Roekel et al., in press; Zhang et al., 2018) and in clinical psychology more broadly (Tackett et al., 2017; Tackett et al., 2019), where much ESM research is conducted. The adoption of open science practices in ESM research is, however, still in its elementary stages. Pre-registration is a cornerstone of open science (Nosek et al., 2018), and its greater use in ESM research was also a specific suggestion of Bastiaansen et al (2019), to address the issue of analytic flexibility and data contingent decision-making. To this end, we hope that the availability of a template specifically tailored to ESM research will firmly embed open science practices within our field.

The template in its current state is not exhaustive and thus may not fit every single type of ESM study, for example, N=1 studies or ESM studies of experimental procedures. We designed the template for the modal ESM study, based on the literature and our own experiences. We also recognise that for some researchers, the list of information to specify in the template may seem extensive, however the vast majority of these decisions must already be taken as a matter of course prior to commencement of data collection. Therefore, we strongly believe that recording these decisions in a pre-registration document does not increase researcher burden. Non-documentation of these decisions does not insulate ESM studies from being subject to the effects of these decisions. Indeed, given the almost dizzying array of choices necessary for ESM research, being able to refer back to a locked, time-stamped record of these choices is advantageous. Our primary considerations when designing this template were to ensure that key decisions influencing reproducibility were adequately recorded and to limit possibilities for analytic flexibility- a key threat to reproducibility (Munafo et al., 2017). Beyond this, the list of specifications within the template provide an accessible overview of the components of ESM study design, that could function as a “quick start” guide for beginners in the field. Further, it could supplement work by van Roekel et al (in press) to also be used as reporting or even peer-reviewing guidelines for ESM studies, to increase the likelihood that published articles include sufficient details for studies to be replicated.

van Roekel and colleagues (in press) end their paper by highlighting the critical importance of open and transparent practices to the future of experience sampling research. In creating this template, we hope to further facilitate such open and transparent practices, as an investment in the future of experience sampling research.

Contributions:

OJK conceptualised the paper. OJK, GL, RA, APH wrote the paper and adapted the template, GL conducted simulation studies and wrote R scripts. IMG provided critical revisions on the manuscript and adapted template.

Conflicts of interest

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

Funding:

OJK is supported by a postdoctoral fellowship from an FWO Odysseus grant to IMG (FWO GOF8416N).

Acknowledgements

We would like to thank participants in the first Center for Research on Experience sampling and Ambulatory methods Leuven (REAL) Workshop (20th - 22nd March, 2019, Leuven) for piloting the ESM pre-registration template. We would also like to thank our colleagues within the Center for Contextual Psychiatry, particularly Professor Wolfgang Viechtbauer, for providing critical feedback on our ideas for the template during a lab meeting presentation by GL.

Prior versions

This manuscript has been posted online as a pre-print:

Availability of materials and code:

The ESM pre-registration template referred to within the manuscript is available online at https://osf.io/2chmu/

The file “Power_Analysis_Code_Temporal_Design.R” contains the R script to generate Figure 1. The script includes a function to simulate a variable that follows an AR(1) process and the code to compute the confidence interval for the mean using the effective sample size.

The file “Power_Analysis_Code_Number_Participants.R” contains the R script to generate Figure 2. The script includes a function to simulate data from a random slope model and the syntax to compute the power based on Monte Carlo simulations.

All R script is available via our OSF page (https://osf.io/2chmu/)

References

Aarts, E., Dolan, C. V., Verhage, M., & Sluis, S. (2015). Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience*, *16*(1), 94.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, 12(4), 579-599.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F., Boker, S. M., Ceulemans, E., Chen, M., … Bringmann, L. F. (2019, March 21). Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology. doi:10.31234/osf.io/c8vp7

Bolger, N., Stadler, G., & Laurenceau, J.-P. (2011). Power analysis for intensive longitudinal studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). New York: Guilford Press.

Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., & Kaschub, C. (2003a). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, *4*(1), 53-78.

Christensen, T. C., Wood, J. V., & Barrett, L. F. (2003b). Remembering everyday experience through the prism of self-esteem. *Personality and Social Psychology Bulletin*, *29*(1), 51-62.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology*, *57*, 505-528.

Collins, L. M., & Graham, J. W. (2002). The effect of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: Temporal design considerations. *Drug and Alcohol Dependence, 68*, 85-96.

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

Dejonckheere, E., Kalokerinos, E. K., Bastian, B., & Kuppens, P. (2018). Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition & Emotion*. doi:10.1080/02699931.2018.1524747

Delespaul, P. A. E. G. (1995). *Assessing schizophrenia in daily life: The experience sampling method*. Maastricht: Datawyse / Universitaire Pers Maastricht.

English, T., Lee, I. A., John, O. P., & Gross, J. J. (2017). Emotion regulation strategy selection in daily life: The role of social context and goals. *Motivation and Emotion*, *41*(2), 230-242.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved from: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Schuurman, N., & Hamaker, E. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70 - 91.

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., & Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry, 19*(59), doi:10.1186/s12888-018-1983-5.

Hektner, J.M, Schmidt, J.A, and Csikszentmihalyi, M. (2007). *Experience sampling method*. Thousand Oaks, California: SAGE Publications Inc.

Himmelstein, P. H., Woods, W. C., & Wright, A. G. (2018, August 16). A Comparison of Signal- and Event-Contingent Ambulatory Assessment of Interpersonal Behavior and Affect in Social Situations. doi:10.31234/osf.io/e6g3j

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.

Janssens, K. A. M., Bos, E. H., Rosmalen, J. G. M., Wichers, M. C., & Riese, H. (2018). A qualitative approach to guide choices for designing a diary study. *BMC Medical Research Methodology, 18*(1), 140. https://doi.org/10.1186/s12874-018-0579-6

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science, 1(4), 443–490. doi:10.1177/2515245918810225

Kirtley, O. J., Hiekkaranta, A. P., Kunkels, Y. K., Verhoeven, D., Van Nierop, M., & Myin-Germeys, I. (2019, April 2). The Experience Sampling Method (ESM) Item Repository. doi:10.17605/OSF.IO/KG376

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86-92.

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., ... & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics, 43*(5), 594-627.

Mellor, D. T., Esposito, J., Hardwicke, T. E., Nosek, B. A., Cohoon, J., Soderberg, C. K., … Speidel, R. (2019, February 6). Preregistration Challenge: Plan, Test, Discover. Retrieved from osf.io/x5w7h

Mertens, G., & Krypotos, A. (2019, February 20). Preregistration of secondary analyses. doi:10.31234/osf.io/ph4q7

Morren, M., Dulmen, S. van, Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: A systematic review. *European Journal of Pain*, *13*(4), 354-365.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1, 0021. doi:10.1038/s41562-016-0021

Myin- Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, *17*(2), 123-132.

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

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606. doi:10.1073/pnas.1708274114

Open Science, C. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Palmier- Claus, J. E., Myin- Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., ... & Dunn, G. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatrica Scandinavica, 123*(1), 12-20.

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

Reis, H. T., & Wheeler, L. (1991). Studying social interaction with the rochester interaction record. *Advances in Experimental Social Psychology, 24*(C), 269–318. doi:10.1016/S0065-2601(08)60332-9

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.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. doi:10.1177/2515245917747646

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review, 31*(4), 471-481.

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods, 46*(1), 41-54. doi:10.3758/s13428-013-0353-y

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as

Significant. *Psychological Science, 22*(11), 1359–1366. doi:10.1177/0956797611417632

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An Introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science*, 9(5), 552–555. doi:10.1177/1745691614543974

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237-259.

Snijders, T.A.B. & Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling* (2nd Ed). London: Sage Publishers.

Stone, A.A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral

medicine. *Annals of Behavioral Medicine*, *16*, 199-202.

Tackett, J. L., Brandes, C. M., & Reardon, K. W. (2019). Leveraging the Open Science Framework in clinical psychological assessment research. *Psychological Assessment*. doi:10.1037/pas0000583

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's Time to Broaden the Replicability Conversation: Thoughts for and From Clinical Psychological Science. *Perspectives in Psychological Science, 12*(5), 742-756. doi:10.1177/1745691617690042

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

Trull, T. J., & Ebner-Priemer, U. (2014). The Role of Ambulatory Assessment in Psychological Science. *Current Directions in Psychological Science, 23*(6), 466–470. doi:10.1177/0963721414550706

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: A systematic review and meta-analysis. Manuscript submitted for publication.

Van Roekel, E., Keijsers, L., & Chung, J. M. (In press). A Review of Current Ambulatory Assessment Studies in Adolescent Samples and Practical Recommendations. *Journal of Research on Adolescence.*

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*(433), 217-221.

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance With Mobile Ecological Momentary Assessment Protocols in Children and Adolescents: A Systematic Review and Meta-Analysis. *Journal of Medical Internet Research, 19*(4), e132. doi:10.2196/jmir.6641

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2018, July 6). Recommendations for increasing the transparency of analysis of pre-existing datasets. doi:10.31234/osf.io/zmt3q

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M. & van Assen, M. A. L. M. (2016) Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. Frontiers in Psychology, *7*:1832. doi:10.3389/fpsyg.2016.01832

Wright, A. G., & Zimmermann, J. (2018, August 5). Applied Ambulatory Assessment: Integrating Idiographic and Nomothetic Principles of Measurement. https://doi.org/10.31234/osf.io/6qc5x

Zander-Schellenberg, T., Remmers, C., Zimmermann, J., Thommen, S., & Lieb, R. (2019). It was intuitive, and it felt good: a daily diary study on how people feel when making decisions. *Cognition and Emotion*, 1-9.

Zhang, C., Smolders, K. C. H. J., Lakens, D., & Ijsselsteijn, W. A. (2018). Two experience sampling studies examining the variation of self-control capacity and its relationship with core affect in daily life. *Journal of Research in Personality, 74*, 102-113. doi:10.1016/j.jrp.2018.03.001

Zięba, A. (2010). Effective number of observations and unbiased estimators of variance for autocorrelated data-an overview. *Metrology and Measurement Systems, 17*(1), 3-16.