Making the black box transparent: A template and tutorial for (pre-)registration of 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:

This manuscript is a pre-print. It is a revised version, which addresses reviewers' comments on the original version.

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: 8,276

Keywords: Preregistration; Reproducibility; Open Science; Transparency; Experience Sampling; Intensive longitudinal data

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). There are, however, numerous “forking paths” and researcher degrees of freedom, even beyond those typically encountered with other research methodologies. Whilst a number of researchers working with ESM techniques are 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. In the current paper, we discuss the ways in which ESM research is especially vulnerable to threats to transparency, reproducibility and replicability. We propose that greater use of (pre-)registration, a cornerstone of open science, may address some of these threats to the transparency of ESM research. (Pre-)registration of ESM research is not without challenges, including model selection, accounting for potential model convergence issues and the use of pre-existing datasets. As these may prove to be significant barriers to (pre-)registration for ESM researchers, we also discuss ways of overcoming these challenges and of documenting them in a (pre-)registration. A further challenge is that current general templates do not adequately capture the unique features of ESM. Here we present a (pre-)registration template for ESM research, adapted from the original Pre-Registration Challenge (Mellor et al., 2019) and pre-registration of pre-existing data (van den Akker et al., 2020) templates, and provide examples of how to complete this.

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 the 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, behaviors, 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' behaviors 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; Trull & Ebner-Priemer, 2020; van Roekel et al., 2019), 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, 2019). Encouragingly, recent years have seen a significant elevation in interest and research energy directed towards addressing methodological and statistical issues in ESM (e.g. Eisele et al., 2020; Himmelstein, Woods & Wright, 2019; Houben, Van Den Noortgate, & Kuppens, 2015; Rintala et al., 2018; Schuurman

& Hamaker, 2019; Vachon, Viechtbauer, Rintala, & Myin-Germeys, 2019; Wright & Zimmermann, 2019). Some of these advances present ESM researchers with new choices, which all represent key decisions for a study and as such, potential points of variation in methodological and statistical approaches within as well as between studies.

ESM research and forking paths

As the number of these methodological and statistical decisions increases, so too do the challenges of conducting transparent, reproducible, and replicable research. Asides from the potential “researcher degrees of freedom” (Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016) or data-contingent analysis decisions - the “garden of forking paths” - (Gelman & Loken, 2013), 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 analyzing 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). In the broader field of psychology, open science practices have gained popularity as a way of addressing some of these threats (Munafo et al., 2017).

ESM research and open science practices

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, 2019; 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 prior to data collection and/or analysis (Nosek, Ebersole, DeHaven, & Mellor, 2018), are initiatives that aim to promote scientific transparency, reproducibility, and replicability. 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., 2019; Zhang, Smolders, Lakens & Ijsselsteijn, 2018), and there is still some way to go before such practices become widely adopted.

Pre-registration is, of course, not the only way in which one can improve transparency, reproducibility and replicability. Several papers over the years have set out reporting guidelines for ESM studies (Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020; van Roekel et al., 2019), yet these are adhered to with varying degrees of rigor in the published literature, even among “top-tier” journals (Trull & Ebner-Priemer, 2020). Reporting guidelines do go some way to facilitating transparency by stimulating researchers to fully describe a study’s design, analysis plan, results, etc. in the final publication. Crucially, however, transparent reporting of ESM research in a published paper does not preclude the fact that hypotheses and analysis plans may differ wildly from what was originally planned. Without a frozen, uneditable version of the initial plan for a study (i.e. a pre-registration) such changes between plans and published research are untraceable. As an illustration of this, a recent investigation of pre-registered (non-ESM) studies published in the journal Psychological Science found that all deviated from their pre-registrations and only in a minority of cases were these deviations transparently reported within the manuscript (Claesen, Gomes, Tuerlinckx & Vanpaemel, 2019). The scope of reporting guidelines is only the finished product (i.e. the published paper). Whilst we encourage ESM researchers to rigorously report their studies following reporting guidelines (Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020), it is paramount that we go beyond reporting guidelines if we are to ensure accountability and transparency of the entire research *process*, not just its *product*. Moreover, we must also ensure that both the process and the product are verifiable and to this end, pre-registrations and published studies can be compared using transparency checklists, e.g. Wicherts et al's (2016) researcher degrees of freedom checklist.

Pre-registration is a tool with great potential to increase transparency, reproducibility and replicability within ESM research, but also comes with a number of challenges. Although many of these are also applicable in psychology research more broadly - especially clinical psychology research - the number of challenges in ESM research may make the threshold for using pre-registration prohibitively high. Given that ESM is a technique that necessitates the use of relatively complex models, issues with model non-convergence are common, and therefore an *a priori* strategy for handling this is required. Moreover, the typical ESM study involves numerous researchers focusing on multiple research questions, posing a challenge for making a comprehensive pre-registration that details all relevant analyses prior to data collection. Considerations regarding model convergence issues are discussed later in the paper, but in the following sections we address issues of (pre-)registration using pre-existing ESM datasets, as well as necessary deviations from (pre-)registrations.

(Pre-)registration of studies using pre-existing datasets.

Collecting ESM data is resource-intensive and produces large, rich datasets, which are often used by many researchers within a team to investigate different primary research questions, as well as later on for secondary data analysis. With ESM research, therefore, the chances are high that researchers may wish to (pre-)register a study prior to data analysis but using pre-existing data. This would not conform to the “strict” definition of pre-registration, where registration occurs prior to data collection. To preclude registration of studies post-data collection but pre-data analysis would leave vast swathes of research in a transparency “Wild West”, when arguably there is even more potential for data-dependent decision-making when data already exist (Weston, Ritchie, Rohrer, & Przybylski, 2019).

Fortunately, the idea of (pre-)registration for pre-existing data is becoming more accepted. Van den Akker et al (2019) and Mertens and Krypotos (2019) have created (pre-)registration templates for pre-existing datasets and there are a growing number of resources and publications addressing this issue (e.g. Weston, Ritchie, Rohrer, & Przybylski, 2019). Terminology is also being extended to account for this variation on the traditional pre-registration format. For example, the term “post-registration” has been proposed for registrations of studies using pre-existing datasets, where all data have been collected but no data have been analyzed or where some data have already been analyzed and published, e.g. by other members of the research team (Benning, Bachrach, Smith, Freeman, & Wright, 2019).

(Pre-)registration of ESM studies with multiple primary research questions.

In another likely and challenging scenario, researchers may be in the process of setting up an ESM study, therefore eligible to pre-register - as opposed to co-register or post-register - this work ahead of data collection commencing, but are aware that they themselves, collaborators, students, etc. will use these data to investigate numerous different diverse research questions. It is usually unfeasible for the researcher to pre-register every possible hypothesis and analysis plan that will be used in the future. Additionally, creating a single large unwieldy pre-registration for many different research questions may also limit the usefulness and clarity of the pre-registration. In these cases, the most extensive possible initial pre-registration (or even a series of pre-registrations) should be made and subsequent registrations using these data approached as co-registrations (if data collection is ongoing) or post-registrations (if data collection has been concluded).

An additional step to prevent data-dependent decision making in post-registrations caused by knowledge of the full dataset, is to operate a variable check-out system, where datasets are treated as libraries and held by a data manager or other third party and only variables specified in a researchers' registration are "checked out" to them (Scott & Kline, 2019). When variables are released, researchers will be issued with a time and date-stamped receipt for the requested variables and this can also include details of historic access to any other variables, similar to an individual's library record. Time- and date-stamped access records will also facilitate the use of pre-existing ESM datasets for Registered Reports. For an extensive discussion of the issues around the (pre-)registration of and transparency in studies using pre-existing datasets, see van den Akker et al. (2020) and Weston et al. (2019).

Necessary deviations from a (pre-)registration.

A further challenge that might be particularly relevant for ESM studies, is that deviations from a (pre-)registered plan may be necessary. Claesen et al.'s (2019) study suggests that it is in fact commonplace for plans of any study to change during the course of a study, but that reporting these deviations is not common. The existence of (pre-)registrations subsequently facilitates transparency when plans deviate from publications. It is possible that issues arise during data collection that require a deviation from what was originally detailed in a study's (pre-)registration. This could be an unexpected issue with a recruitment site, for example, a technical issue with the ESM app that means participants received fewer notifications than they were supposed to as per the protocol, or where during data collection, a researcher learns that certain ESM items or instructions were not clear to participants and decides to change or amend items or instructions halfway through a study. Human error may also mean that a researcher has forgotten to specify a particular detail within their (pre-)registration and only realizes this once data have been collected. This may occur more commonly when researchers first begin with (pre-)registration, as writing a good pre-registration is a skill which requires time and experience to develop (Nosek et al., 2019).

An option of dealing with study alterations is to make a supplementary registration. This then provides a time-and-date-stamped record of when particular issues arose during a study and which decisions were taken when. Benning and colleagues (2019) have recently proposed the term “co-registration” to describe registrations carried out during or after data collection, but prior to any data analysis. Some issues, however, may only emerge *after* data have been accessed, explored, or analyzed. In conceptualizing pre-registration as a continuum (Benning et al., 2019), we feel there is still value in transparently documenting changes to analysis plans in supplementary registrations, even when data have been accessed. In this case, the function of the registration is altered and serves more as an open lab book. Importantly, here the “post-registration” no longer enables a key pre-registration goal of evaluating a statistical test’s capacity to falsify a hypothesis (Lakens, 2019; Nosek et al., 2018; Wagenmakers et al, 2012).

Addressing the challenges of (pre-)registering ESM research

Having already encountered some of the challenges described above when (pre-)registering and conducting our own ESM studies using existing pre-registration templates, a further topical discussion around the utility of specialized pre-registration templates for specific study designs and methodologies (Srivastava, Tullett & Vazire, 2019) led us to devise a template for ESM study (pre-)registration.

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. Following reviewer comments, we have also incorporated elements from the pre-registration template for pre-existing data (van den Akker et al., 2019), to reflect the fact that many ESM studies yield large, rich datasets, which are also used by other researchers for primary analysis and, over time, for secondary analysis. The expanding conceptualization of the “registration continuum” to include “post-registrations” of studies using archival data (Benning et al., 2019) also brings new opportunities in this regard. Throughout the paper, we have taken the position that open and transparent plans are, even if imperfect, better than no plans (Nosek et al., 2019) and have made suggestions for maximizing transparency in cases where deviations from the (pre-)registration are necessary, e.g. through “co-registration” (Benning et al., 2019).

To help guide researchers, we provide two examples of completed ESM registration templates: a pre-registration (Example 1) and a post-registration (Example 2), available online **(https://osf.io/2chmu/).** 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 we welcome critical feedback. Within our paper, we also discuss key challenges of ESM study (pre-)registration and provide some potential solutions.

A (pre-)registration template for ESM research

Our central considerations when devising these additions to the Preregistration Challenge template (Mellor et al., 2019) were to address: (1) specific factors of ESM studies that may impact or even preclude their replicability and reproducibility, and (2) aspects that may be vulnerable to questionable research practices or analytic flexibility, particularly after data has already been (partially) accessed. ESM studies allow for much flexibility in the measurement of data, the construction of variables, and the analysis of these variables. At every stage of data collection and analysis, researcher degrees of freedom may (intentionally or unintentionally) arise. In the following walkthrough of the template, we have included detailed descriptions of only those sections that are specific to ESM research and not sections of the template that are also applicable to other types of studies. Examples of how to complete these other more broadly applicable sections can be found in the two exemplar templates provided.

Sampling plan

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 (Trull & Ebner-Priemer, 2020). In order to increase transparency about these decisions from the outset of a study, we have therefore added a new section to the (pre-)registration template, entitled ‘ESM data collection procedure’.

The duration of ESM studies ranges from days to weeks to months and consequently, it is possible that modifications to the data collection procedure may be made, e.g. to increase compliance, that are contingent on the amount, quality, or content of participants’ data coming in. With the development of more advanced mobile applications and researcher interfaces for ESM studies, monitoring incoming data and changing important elements of data collection during the ESM period has become increasingly easier. As these are data-dependent decisions, such modifications may unintentionally introduce researcher bias, and may not always be reported in the final manuscript. In some cases, potential modifications can be anticipated and decision rules recorded in the pre-registration. Unanticipated modifications can be addressed in subsequent co-registrations and should definitely be reported in the paper.

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 behaviors and feasibility (Janssens, Bos, Rosmalen, Wichers, & Riese, 2018). Occasionally, researchers wish to extend the ESM period for (a subsample of) participants. For example, researchers running an ESM study with two groups of individuals, with and without major depression, may find that individuals with depression exhibit reduced compliance and therefore the researchers choose to increase the number of ESM days for that group to ensure sufficient observations are collected. This would be an alteration that is contingent on the amount of data coming in. Where it is possible to anticipate that extending the number of days may be required e.g. based upon past experience with similar studies or existing literature, a decision rule for this can be specified in the pre-registration. Alternatively, if this is not anticipated, but the issue subsequently arises, then supplementing the original pre-registration with a later co-registration would make such a data-dependent decision transparent.

Type of sampling scheme.

The sampling scheme of ESM studies refers to the timing of questionnaire prompts. 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, where participants can fill out prompts once a specific event or behavior has occurred. On the other hand, relatively rapid fluctuations in, for example, mood might be best captured with a (semi-)random design, where prompts are sent out at random time points.

The aim of the temporal design is to determine the number of measurements within an individual that are necessary to obtain reliable estimates 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, which is, the time interval between two different measurements (Collins & Graham, 2002). The selected temporal design should be specified in the pre-registration and any later modifications in a co-registration, as some decisions may inadvertently introduce bias. 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.

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

Many reports of ESM studies only include a description of variables that were analyzed for that specific study. 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., 2019; van Roekel, Keijsers, & Chung, 2019). Here, researchers are asked to provide a general description of the total questionnaire length. A longer ESM questionnaire, especially one with more open-ended items, signifies a greater participant burden, which can reduce the compliance rate as well as the data quality (Eisele et al., 2020). Without knowledge of the total number of items, the potential effect of the total questionnaire length on the compliance rate is unclear. The 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.

We ask researchers to include the full list of ESM items as an appendix at the end of the (pre-)registration document. Unlike questionnaire measures, which are often subject to copyright and licensing precluding “open materials” sharing (Weston et al., 2019), ESM items are not proprietary and can, therefore, be freely shared, with correct attribution to the researchers who originally created the item(s). Making ESM items open and tracking down the citation of record for particular items can be facilitated by making use of the Experience Sampling Item Repository (Kirtley et al., 2019). This ongoing open science project 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. Such time-out specifications should have a theoretical rationale, as they ideally directly relate to the temporal dynamics of the constructs that are assessed. They are also highly relevant for the replicability of a study, and may potentially be modified once a study is underway, for example, if researchers receive feedback that participants are struggling to complete the questionnaire during the allotted time-period. Consequently, researchers may amend the time-out period. As such, timing restrictions are important to include in the pre-registration, and researchers may do so in this subsection.

ESM instruction.

Some details regarding the manner in which participants are instructed to complete ESM questionnaires are relevant for enhancing reproducibility. At baseline, participants need to be instructed or trained in a standardized manner so that they are able to respond properly to the ESM questionnaires. There are various instruction options that may affect compliance (Christensen et al., 2003a), as well as 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). Information about how participants were instructed to complete the ESM questionnaire is crucial for being able to reproduce a study's methods and as these may be subject to change during data collection, also represent a potential forking path. As such, details about participant instructions and briefing should be provided. Any instructions included within the actual ESM questionnaire (on the phone) should be listed together with the ESM items.

Rationale for sample size: Temporal design and number of participants.

Despite being a crucial consideration for all research (Button, et al., 2013), most ESM studies do not report a power calculation to justify the sample size or state a rationale for the selection of the sampling frequency (Trull & Ebner-Priemer, 2020; Trull & Ebner-Priemer, 2013). This represents another threat to the reproducibility of ESM studies (Munafo et al., 2017). The structure of ESM data allows the examination of the variability of a target process over time, within as well as between individuals (Hofmans, et al., 2019). Considerations regarding sample size, therefore, must account for both the temporal design in which the target processes will be observed and the number of participants. The sample size rationale can also be based on practical considerations, e.g. budget restrictions or limiting participant burden. For this reason, in this section, we include some further, in-depth discussion of the key considerations for sample size planning when pre-registering ESM studies.

The temporal design varies considerably between published ESM studies. In psychiatry, ESM studies that assess highly variable constructs (e.g. mood) have often used ten measurements per day, for six consecutive days (Myin-Germeys et al., 2018). Conversely, in a study of a more stable construct, global self-esteem, just one measurement per day for seven consecutive days was used (Christensen et al., 2003b). A study's temporal design is closely related to the information necessary to obtain reliable estimates of within-individuals dynamics (e.g., Krone, et al., 2016, 2017; Liu, 2017; Raudenbush, & Liu, 2001; Schultzberg & Muthén, 2018; Timmons & Preacher, 2015). For example, Adolf et al. (2019) and de Haan-Rietdijk et al. (2017) investigate continuous-time autoregressive processes, and show that when the variability of the process is high, larger time intervals between assessments negatively impact the estimation accuracy. Therefore, justifying the selection of the temporal design based on the properties of the statistical model being studied, as well as taking into consideration expected missingness, will increase the accuracy of the estimates. Furthermore, this will also reduce the likelihood of non-convergence issues when estimating the statistical model. As Collins (2006) highlights, an explicit justification of the choice of the temporal design will increase the reproducibility of longitudinal studies. Researchers may also base their target sample size upon existing sampling protocols or theoretical considerations and in these cases, we recommend that researchers explicitly state this as their sample size rationale and provide references to these protocols or studies in the (pre-)registration.

The second consideration when determining sample size for ESM research relates to the number of participants necessary to obtain accurate estimates of inter-individual differences (Maas & Hox, 2005). In studies where individual differences are likely to be large, more information is needed in order to determine an effect relative to instances where heterogeneity between individuals would be negligible. Researchers can justify the selection of the number of participants based on power analysis (Arend & Schäfer, 2019; Bolger et. al., 2011; Lane & Hennes, 2018; Raudenbush & Liu, 2001). 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 treatment conditions. The rationale for the selection of the number of participants can also contain information related to the feasibility of sampling participants from specific populations, budgetary restrictions, as well as plans to sample additional subjects in case of dropouts.

A number of resources exist to guide researchers in performing power analyses for general multilevel and longitudinal designs, including Arend and Schäfer (2019); Astivia et al. (2019); Bolger et al. (2011); Brandmaier et al. (2015); and Lane and Hennes (2018). We have also produced an illustration of how to perform a simulation-based power analysis to select the number of participants, when explicitly accounting for the dependency of occasions within an individual, and made this available online (https://osf.io/2chmu/).

ESM studies frequently include numerous variables with the intention that a wide variety of hypotheses will be tested. In this respect, we encourage researchers to indicate in the (pre-)registration whether the temporal design has been selected to study the dynamics of a specific set of variables. Researchers can then specify for which hypotheses power analyses were conducted.

Finally, when post-registering analyses of pre-existing data, researchers can provide references to existing design protocols for the dataset, as well as any additional information related to the rationale for the sample size determination (e.g. whether a power analysis was conducted to select the number of participants). With pre-existing datasets, it may be the case that a power calculation was conducted in relation to a specific set of variables, but that these variables are not included in other studies using the same dataset. In these cases, we would also recommend conducting a power calculation for the planned analyses as though data had not already been collected. Then if the available number of participants or observations within the dataset falls short of this, researchers can make a decision whether to adjust their planned analyses or perhaps in some cases, that the available dataset will not yield sufficient power to conduct the planned analyses and therefore should not be used.

Stopping rule

When there is little control over recruitment in any study, researchers may wish to implement a 'stopping rule' that represents the sample size that must be achieved before data collection is terminated. This rule will usually be based on a power analysis that calculates the required sample size to find an effect. In power analyses for ESM studies, there is the additional requirement of a minimum amount of measurements per person to reach a certain level of power. If this threshold is not met for any given participant, researchers may wish to extend the ESM period to collect more data until the threshold is met. Such a stopping rule would be valuable to indicate in this section of the template and relevant decision rules, e.g. about extending data collection, can be specified here.

Variables

As the majority of ESM research is observational (Hektner, Schmidt, & Csikszentmihalyi, 2007; Myin-Germeys et al., 2018), researchers are first asked to specify measured variables, then manipulated variables. Researchers are only asked to describe in detail variables that will be used in confirmatory analyses, but are required to provide a full list of the ESM items as well. 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/time invariant variables' and 'measured ESM variables/time-variant variables'. In the 'measured ESM variables/time-variant variables' section, instructions to specify the response scale of ESM variables (e.g. Likert-scale) were added. 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. Some indication of how open-ended answers are coded is relevant to include in the (pre-)registration, since such coding is subject to numerous researcher degrees of freedom. 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. Summary statistics, such as observation-level or within-person-level averages, can be formed at different levels in ESM datasets. These can also be constructed from different sets of items and be employed as predictors, outcomes, or covariates. As scoring options expand with increasingly complex designs, the likelihood of score construction becoming a forking path also increases (Wicherts et al., 2016). Since large flexibility exists in the creation of any index, and since there are few well-validated ESM-based indices, it is highly relevant to specify which ESM items and at which level, are used for the construction of new variables. For instance, average positive mood may be calculated per notification, per day, or over a longer period of time. Information regarding the methods used to determine whether items can be combined, e.g. multilevel factor analyses, should also be reported here.

Prior knowledge of the data.

As previously discussed in this paper, and consistent with the idea of a registration continuum (Benning et al., 2019), studies using pre-existing data can be post-registered. In these cases, for maximal transparency researchers should record their prior knowledge of the datasets (van den Akker et al., 2019), as any knowledge of the data can lead researchers to make data dependent decisions, and consequently introduce further researcher degrees of freedom into the process. This knowledge can be from previous analyses conducted by the researcher using the same dataset, which may have resulted in publications, pre-prints, conference presentations, etc., but also includes awareness of the dataset from external sources, e.g. papers by other researchers using the same dataset. Where applicable, the references to these sources should be provided.

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 analyzing 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 **errors**), 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 considerations of specific relevance for an ESM analysis plan, here 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 (Bolker, et al., 2009): (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.

These considerations may seem numerous and effortful to record as part of a (pre-)registration, but as they all represent potential “forking paths” where a high degree of analytic flexibility may be introduced into the research, they are essential. For example, the distribution of the within-individuals errors 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 or assume that the errors are non-Gaussian distributed: an important decision that should be noted in the analysis plan section of the (pre-)registration.

Another example is where random effects allow the modeling of non-independence between individuals. In general, random effects are considered as normally distributed (models that do not assume normality for the random effects can be found in Verbeke and 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 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, if the random effects are allowed to be correlated). The same suggestions apply to the (pre-)registration of data analysis that includes nested or crossed random effect designs.

When considering the predictors included in the statistical model, an important decision is which predictors are going to be set as fixed effects. This depends on the hypotheses, so is an important *a priori* - and therefore pre-registerable - decision for researchers to take. Furthermore, if the model includes time-varying and time-invariant predictors, the analysis plan should state whether the model includes cross-level interaction effects.

Another consideration regarding the predictors is related to the transformations applied to the variables. We advise stating which transformations of the data are expected. For example, a common practice in multilevel modeling is to center the time-varying predictors using the individual’s mean and to center the time-invariant predictors using the grand mean (Snijders & Bosker, 2012). Where a set of ESM items measuring a certain construct will be validated, this should be explicitly stated along with the approach (e.g. within-person factor analysis where items are centered 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 also necessary to specify the method used to account for the ‘overnight lags’. For example, a common approach is to set the first notification of the day as missing (de Haan-Rietdijk et. al., 2017).

Model complexity and convergence issues.

In the (pre-)registration, researchers should also consider how to evaluate 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 (Barr, et al., 2013; Matuschek, et al., 2017). This can result in a mixed-effects model that fails to converge or affect the reliability of the parameter estimates. Non-convergence in mixed-effects models arises when the structure of the variance components is complex, given the amount of available data or when the data is highly unbalanced (Eager & Roy, 2017). It is possible to anticipate when convergence issues may arise, although not always exhaustively and there are different strategies that can be used to address this issue when (pre-)registering analysis plans. To address issues related to model complexity, we added a section to the template where researchers are requested to explain what they will do when data violates assumptions, the model does not converge, or other analytical problems arise (van den Akker et al., 2019). For instance, prior to (pre-)registration, researchers can evaluate the complexity of the planned models using simulation-based approaches (e.g., DeBruine & Barr, 2019). An alternative strategy involves evaluating models with parsimonious random effect structures. Bates et al. (2018) proposed using Principal Component Analysis to select the random effects structure in a linear mixed-effects model. Alternatively, iterative hypothesis testing procedures can be used to select the random effects structure (Cheng, et al., 2010; Harrison, et al., 2018; Müller, et al., 2013). A description of the method that will be used to select a parsimonious random effects structure can be included within the analysis plan.

Specification of how non-convergence issues will be addressed, for example by switching the optimizer used or using an alternative (pre-specified) simplified model, may help to preserve the (pre-)registration goal of allowing evaluation of a model's capacity to falsify the hypothesis being tested. Even with contingency plans for non-convergence outlined in a (pre-)registration, there may be some cases where these plans do not work or additional unexpected convergence issues arise. Here, we would suggest the use of supplementary co-registrations to create a frozen record of updated plans and models, especially where other analyses specified in the (pre-)registration were dependent upon the non-converging model but have not yet been conducted. These deviations would then also need to be transparently reported in the paper.

Model selection and robustness.

An important question that arises when describing the analysis plan is how to select from different competing explanations of the data. Different criteria can be used to perform model selection (see Pitt, et al., (2002) and Navarro (2019)). For instance, researchers might be interested in evaluating the plausibility of the assumption of a model, or if the model is able to capture the target phenomena in a less complex manner. Different strategies have been proposed to select the model with the best predictive accuracy. For example, the data set can be separated into a training and testing set. The training set is used to perform exploratory analysis and the testing set is held back and used to perform confirmatory analysis (de Groot, 2014). More sophisticated techniques involve using cross-validation (Bulteel, et al., 2018).

A more general concern is how to assess the robustness of scientific findings (Weston et al., 2019), which implies investigating the sensitivity of statistical findings under different data pre-processing choices, model specifications, or inclusion or exclusion of covariates. There are several approaches that can be used to conduct a sensitivity analysis, examples include specification curve analysis (Simonsohn, et al., 2015; Young & Holsteen, 2017) and multiverse analysis (Steegen, et al., 2016). Table 1 describes different strategies that can be applied to assess model selection.

We also note that when the analysis plan involves estimating more complex statistical approaches, such as dynamic network models (Bringmann, et al., 2013), dynamic structural equation models (Asparouhov, et al., 2018), the (pre-)registration should take into account all necessary information to reproduce the analysis. For instance, when estimating a model using a Bayesian approach, the distribution of the parameters as well as the priors can be described in the analysis plan.

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 whether the default options of a function or software were used. Even better, researchers can share their statistical analysis plan and code, using platforms such as GitHub and the Open Science Framework.

Data exclusion and missing data.

In ESM studies, there are many factors that might affect the quality of the collected data, including compliance and technical issues. For instance, if an individual does not respond to a notification, then the entire set of items within an observation will be missed. Additionally, participants might drop out of the study or technical problems may render observations from certain days unusable. Exclusion criteria due to technical problems can also be included in the analysis plan, for instance, researchers could opt not to include participants that report a technical issue. In this respect, the analysis plan should include all the information necessary to define whether a unit will be excluded from the analysis. Poor specification of data exclusion decision rules can represent a major researcher degree of freedom (Wicherts et al., 2016).

Compliance.

Low compliance and thus factors influencing compliance can reduce the quality of the data (Delespaul, 1995; Eisele et al., 2020; Palmier-Claus, et al., 2011; Rintala, et al., 2019). For pre-data collection pre-registration of ESM studies, we recommend researchers state decision rules related to participant dropouts (e.g., whether observations prior to the dropout will be included in the analysis). In addition, researchers should state in the pre-registration how compliance will be defined (e.g., whether missing prompts due to technical problems count as non-compliance). It is also important to describe and justify the thresholds for compliance that will be used to include or exclude participants within the analyses (Stone & Shiffman, 2002; Trull

& Ebner-Priemer, 2013). Many studies use a rule of thumb for determining compliance threshold- often that participants must complete a minimum of 30% of prompts (Delespaul, 1995)- yet this is subject to much debate and recent work suggests this can significantly bias model estimates and it is optimal to include all available observations (Jacobson, 2020). Compliance thresholds, if not specified *a priori*, represent another forking path, as they may be adjusted post-hoc in order to maximize available data. For pre-data collection pre-registration of ESM studies, we encourage researchers to report the expected compliance. For ESM studies with pre-existing data, researchers can report information about participant dropouts and compliance levels (e.g., overall compliance, compliance for different types of reports, the mean level of compliance, range of compliance across participants).

Handling of missing data and outliers.

In the statistical analysis plan, it is important to state how missing data and outliers 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). Moreover, in case that techniques to handle missing data are implemented (e.g., fully maximum likelihood estimation or multiple imputation), the analysis plan should include detailed information about the framework to process missing data. A broader discussion on methods to handle missing data can be found in Graham (2012) and Schafer (2001).

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).

Finally, the analysis plan should include considerations on how the statistical outliers will be defined and how they will be treated in the analysis plan. A practical discussion on how to incorporate the information related to outliers in the (pre-)registration template can be found in van den Akker et al. (2019). In the (pre-)registration, researchers should also include the expected sample size that will be used in the data analysis. For studies using pre-existing datasets, any information related to the expected pattern of missingness or outliers should be included in the (pre-)registration.

Conclusions

In the current paper, we have presented a (pre-)registration template for ESM research, the development of which was inspired by topical discussions around this issue (Srivastava, Vazire & Tullett, 2019) as well as our own experiences of (pre-)registering ESM studies using existing tools. We have also included detailed explanations and potential solutions for key challenges for the (pre-)registration of studies using ESM. To guide ESM researchers further in how to approach registration, we have created two exemplar templates, illustrating a pre-registration and post-registration. 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., 2019; 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.

That being said, our own experiences of registering ESM studies as well as discussion stemming from thought-provoking reviewer comments on an earlier version of this manuscript, have highlighted that the greatest barrier to the uptake of (pre-)registration is not the lack of a specific template. Rather, it is the lack of clear guidance regarding how key ESM methodological and statistical decisions - which must necessarily be recorded in a (pre-)registration- ought to be addressed. Issues of model selection and convergence are examples of these, as well as how to fit the typical setup of ESM studies, with many researchers working on a single dataset to answer numerous research questions, into the concept of (pre-)registration. To this end, within the current paper, we have also discussed approaches that may be taken to address these and other challenges.

The template in its current state is not exhaustive and thus may not cover decisions for every single type of ESM study, for example, ESM studies of experimental procedures. We designed the template for the modal ESM study, based on the literature and our own experiences. We also recognize 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. There are also some threats to reproducibility that (pre-)registration does not solve, for example, issues of weak theorizing and poor correspondence between theories and the statistical models that are supposed to map onto them (Szollosi et al., 2020).

Our primary considerations when designing this template were to ensure that key decisions influencing reproducibility were recorded transparently and to limit possibilities for analytic flexibility and researcher bias - key threats to reproducibility of results and replicability of methods (Munafo et al., 2017). (Pre-)registration should not be seen as a substitute for rigorous reporting of results using existing ESM study reporting guidelines (Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020), however, as these guidelines pertain to the product of the research, i.e. the paper, they do not necessarily capture researcher degrees of freedom in the research process, therefore (pre-)registration has additional value. In shining a light on the process, it becomes evident that the scientific process is rarely perfect. (Pre-)registration does not preclude imperfection and deviations from a (pre-)registration appear to be a common occurrence (Claesen et al., 2019), but (pre-)registration does make deviations more transparent. (Pre-)registration is a scientific skill that must be developed and refined with experience, as is the case with other scientific skills. In the early stages of developing this skill, deviations and “messy” (pre-)registrations are more likely, but we believe that for advancing transparency within ESM research “some plans are better than having no plans, and sharing those plans in advance is better than not sharing them.” (Nosek et al., 2019; p3).

Contributions:

OJK conceptualized the paper. OJK, GL, RA, APH wrote the paper and adapted the template. GL conducted simulations 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: https://psyarxiv.com/seyq7

Disclosure of data, materials and online resources

The ESM pre-registration template and the two completed exemplar templates referred to within the manuscript are available online at https://osf.io/2chmu/

The file “Sample_Rationale_Number_of_Participants.RMD” contains the R Markdown script to calculate the number of participants required, using a simulation approach.

The file “Sample_Rationale_Temporal_Design.RMD” contains the R Markdown script to illustrate the effect of serial dependency in the estimation accuracy of the mean level of stationary autoregressive processes.

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.

Adolf, J., Loossens, T., Tuerlinckx, f., & Ceulemans, E. (2019, November 25). Optimal fixed sampling rates for reliable continuous-time first-order autoregressive modeling. https://doi.org/10.31234/osf.io/5cbfw

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological methods, 24*(1), 1-19.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359-388.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255-278.

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

Benning, S. D., Bachrach, R. L., Smith, E. A., Freeman, A. J., & Wright, A. G. C. (2019). The registration continuum in clinical science: A guide toward transparent practices. J*ournal of Abnormal Psychology, 128*(6), 528–540. https://doi.org/10.1037/abn0000451

Bolger, N., Stadler, G., & Laurenceau, J.-P. (2012). 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.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution, 24*(3), 127-135.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C., & Lindenberger, U. (2015). LIFESPAN: A tool for the computer-aided design of longitudinal studies. *Frontiers in Psychology, 6*, 272.

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), e60188.

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR (1) based models do not always outpredict AR (1) models in typical psychological applications. *Psychological methods, 23*(4), 740-756, doi: 10.1037/met0000178.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365.

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., & Muller, K. E. (2010). Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in medicine, 29*(4), 504-520.

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.

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & vanpaemel, w. (2019, May 9). Preregistration: Comparing Dream to Reality. https://doi.org/10.31234/osf.io/d8wex

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.

DeBruine, L. M., & Barr, D. J. (2019, June 2). Understanding mixed effects models through data simulation. https://doi.org/10.31234/osf.io/xp5cy

de Groot, A. D. (2014). The meaning of “significance” for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta psychologica, 148,* 188-194.

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*, *33*(5), 1076-1083. 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.

Eager, C., & Roy, J. (2017). Mixed effects models are sometimes terrible. arXiv preprint arXiv:1701.04858.

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020, February 20). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. https://doi.org/10.31234/osf.io/zf4nm

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

Graham, J. W. (2012). Missing data: Analysis and design. New York, NY: Springer.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., ... & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ, 6*, e4794.

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. C. (2019). A comparison of signal- and event-contingent ambulatory assessment of interpersonal behavior and affect in social situations. *Psychological Assessment, 31*(7), 952–960. https://doi.org/10.1037/pas0000718

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.

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.

Jacobson, N. (2020, 14 January). Compliance Thresholds in Intensive Longitudinal Data: Worse than Listwise Deletion. Talk given at the Society for Ambulatory Assessment (SAA) conference, Melbourne, Australia.

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

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

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

Krone, T., Albers, C. J., & Timmerman, M. E. (2017). A comparative simulation study of AR (1) estimators in short time series. *Quality & Quantity, 51*(1), 1-21.

Krone, T., Albers, C. J., & Timmerman, M. E. (2016). Comparison of estimation procedures for multilevel AR (1) models. *Frontiers in Psychology, 7*, 486.

Lakens, D. (2019, November 18). The Value of Preregistration for Psychological Science: A Conceptual Analysis. https://doi.org/10.31234/osf.io/jbh4w

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

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

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305-315.

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. and Krypotos, A.-M. (2019). Preregistration of Analyses of Preexisting Data. *Psychologica Belgica, 59*(1), pp.338–352. DOI: http://doi.org/10.5334/pb.493

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.

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science, 28*(2), 135-167.

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.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior, 2*(1), 28-34.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., … Vazire, S. (2019, August 14). Preregistration Is Hard, And Worthwhile. https://doi.org/10.1016/j.tics.2019.07.009

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 Collaboration (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.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. Psychological review, 109(3), 472.

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.

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.

Schafer, J. L. (2001). Multiple imputation with PAN. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 357–377). Washington, DC: American Psychological Association.

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

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

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 notifications 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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. Available at SSRN: https://ssrn.com/abstract=2694998 or http://dx.doi.org/10.2139/ssrn.2694998

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

Srivastava, S., Tullett, A. M., & Vazire, S. (2019, February 20). Our Best Episode Ever (No. 53) [Audio podcast episode]. In The Black Goat. http://www.theblackgoatpodcast.com/posts/our-best-episode-ever/

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

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*(3), 236-243.

Stone, A.A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, *16*, 199-202.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences, 24*(2), 94 - 95. doi.org/10.1016/j.tics.2019.11.009

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. 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., & 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

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, 9, 151-176.

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. *Journal of Medical Internet Research, 21*(12):e14475

Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., … Bakker, M. (2019, November 20). Preregistration of secondary data analysis: A template and tutorial. https://doi.org/10.31234/osf.io/hvfmr

Van Roekel, E., Keijsers, L., & Chung, J. M. (2019). A Review of Current Ambulatory Assessment Studies in Adolescent Samples and Practical Recommendations. *Journal of Research on Adolescence, 29*(3), 560-577. doi.org/10.1111/jora.12471

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.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

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. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science, 2*(3), 214-227

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. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment, 31*(12), 1467–1480. https://doi.org/10.1037/pas0000685

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research, 46*(1), 3-40.

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

Table 1: Approaches for Assessing Model Selection in Intensive Longitudinal Data Analysis

Method	Description
Simulation	Simulation procedures can be used to generate data from a statistical model to study estimation accuracy, as well as performing power analysis or evaluating the effect of the temporal design for intensive longitudinal designs. For mixed-effect models, the simulation-based approach to evaluate model complexity implies specifying the model parameters (i.e., fixed effects, distribution of the variance components and predictors) and then using this model to generate data for the outcome of interest. This procedure can be used to evaluate different decisions such as inclusion or exclusion of predictors, the structure of the random effects, patterns of missing data, selection of the optimizer function to evaluate model convergence (see DeBruine & Barr (2019) for an introduction).
Stepwise selection	Mixed-effect models allow to explicitly model the hierarchical structure of intensive longitudinal designs, and thus reduces the probability of false positives and false negatives. There are some disadvantages related to over-parameterization of the random effect structure or highly imbalanced data that might cause convergence issues due to models with a singular covariance matrix for the random effects. In order to estimate a parsimonious random effect structure, Bates et al. (2018) proposed to assess the dimensionality of the random effects by performing a principal component analysis to the estimated covariance matrix of the random effects. Other procedures involve performing stepwise selection using Likelihood Ratio Tests, this procedure performs model selection by sequentially setting some terms to zero until a parsimonious model is achieved (Harrison, et al., 2018).
Information-theory	Information theory can be used to select a model from a set of competing models, this involves creating a ranked of models using metrics such as AIC. Approaches that use information based theory to perform model selection involve model averaging, performing all-subset regressions followed by AIC, using Akaike weights to quantify variable importance (see Harrison, et al. (2018) for a broader discussion).
Cross-validation	Cross-validation can be used to evaluate how well a proposed model predicts new or unseen data. This procedure involves splitting the data set into a training data set and a testing data set. The model parameters are estimated using the training set, and the estimated parameters are used to estimate the prediction errors using the testing set. For example, Bulteel, et al. (2018) performs cross-validation to select the model that best predicts affective states when studying within-individuals dynamics. Moreover, cross-validation can be used to study the effect of excluding

	or transforming predictors.
Exploratory data analysis	In certain situations, such as violations of model assumptions (i.e., the errors are not Gaussian distributed), researchers might engage in data exploratory analyses. If the aim is to carry confirmatory testing, it is possible to separate the data to perform exploratory analysis and confirmatory analysis (de Groot, 2014). We recommend that researchers enumerate the plans for exploratory analysis in the pre-registration. For secondary analysis data, if applicable, we encourage researchers to state the exploratory analyses that have been performed prior to the pre-registration.
Sensitivity analysis	When analyzing data there are different ways to test the statistical significance of a model. Model selection relies on data analytic decisions (e.g., which variable to include, setting the compliance threshold or how to handle outliers). Sensitivity analysis can be used to study the effect of data pre-processing or different model specifications to assess the distribution of the estimated effects (e.g., Simonsohn, et al., 2015; Steegen, et al., 2016; Young & Holsteen, 2017).