

Making the Black Box Transparent: A Template and Tutorial for Registration of Studies Using Experience-Sampling Methods

Olivia J. Kirtley¹, Ginette Lafit^{1,2}, Robin Achterhof¹,
Anu P. Hiekkaranta¹, and Inez Myin-Germeys¹

¹Center for Contextual Psychiatry, Department of Neurosciences, KU Leuven, and ²Research Group
on Quantitative Psychology and Individual Differences, Department of Psychology, KU Leuven

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. These methods, however, tend to involve numerous forking paths and researcher degrees of freedom, even beyond those typically encountered with other research methodologies. Although a number of researchers working with ESM techniques are actively engaged in efforts to increase the methodological rigor and transparency of research that uses them, currently there is little routine implementation of open-science practices in ESM research. In this article, we discuss the ways in which ESM research is especially vulnerable to threats to transparency, reproducibility, and replicability. We propose that greater use of study registration, a cornerstone of open science, may address some of these threats to the transparency of ESM research. Registration of ESM research is not without challenges, including model selection, accounting for potential model-convergence issues, and the use of preexisting data sets. As these may prove to be significant barriers for ESM researchers, we also discuss ways of overcoming these challenges and of documenting them in a registration. A further challenge is that current general preregistration templates do not adequately capture the unique features of ESM. We present a registration template for ESM research and also discuss registration of studies using preexisting data.

Keywords

preregistration, reproducibility, open science, transparency, experience sampling, intensive longitudinal data

Received 4/10/19; Revision accepted 4/3/20

Some studies require a high level of experimental control, which can occur only 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, and the experience-sampling method (ESM; Hektner et al., 2007) and ecological momentary assessment (EMA; Stone & Shiffman, 1994) are two subtypes of ambulatory assessment involving participants' self-reports. The terms *ESM* and *EMA* are often used interchangeably (Trull & Ebner-Priemer, 2014), but in this article, we use *ESM* throughout. ESM involves participants completing 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; they also address issues of recall bias and increase ecological validity by measuring participants' behaviors in their daily lives (Myin-Germeys et al., 2018; Trull & Ebner-Priemer, 2014).

Corresponding Author:

Olivia J. Kirtley, Center for Contextual Psychiatry, Department of
Neurosciences, KU Leuven
E-mail: olivia.kirtley@kuleuven.be



Recent years have seen a proliferation of studies employing ESM. Although 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 because of the multilevel nature of the data (Bolger et al., 2012). Similarly, ESM research brings additional considerations regarding item selection, psychometrics, and analytic strategy (Wright & Zimmermann, 2019). Encouragingly, recent years have seen a significant elevation in interest and research energy directed toward addressing ESM's methodological and statistical issues (e.g., Eisele et al., 2020; Himmelstein et al., 2019; Houben et al., 2015; Rintala et al., 2019; Schuurman & Hamaker, 2019; Vachon et al., 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. Aside from the potential researcher degrees of freedom (Simmons et al., 2011; Wicherts et al., 2016) and data-contingent analytic decisions—the “garden of forking paths” (Gelman & Loken, 2013)—analytic flexibility can also occur simply as a function of individual differences in analytics decisions between researchers (Silberzahn et al., 2018). Many defensible analytic choices may exist for the same data set (Bastiaansen et al., 2019). Given the multitude of choices to be made 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 et al., 2019). This is particularly concerning given that poor study design and analytic flexibility are two major threats to scientific reproducibility (Munafò et al., 2017). In the broader field of psychology, open-science practices have gained popularity as a way of addressing some of these threats (Munafò et al., 2017).

ESM Research and Open-Science Practices

The field of psychological science is currently undergoing something of a renaissance resulting from the replication crisis, that is, the fact that many high-profile studies have not been successfully replicated (Klein

et al., 2018; Open Science Collaboration, 2015). Clinical psychology and psychiatry, fields in which many ESM studies are conducted, have thus far been noticeably absent from conversations around open science (Tackett et al., 2017, 2019), but this does not mean that the methods or results from clinical research are more reproducible or replicable than those from other fields. Open-science practices, including preregistration of hypotheses and analysis plans on OSF prior to data collection or analysis (Nosek et al., 2018), are initiatives aimed at promoting scientific transparency, reproducibility, and replicability. Although off to a promising start, open-science approaches—including registration and sharing of preprints, code, and materials—are only just emerging in 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 there is still some way to go before such practices become widely adopted.

Registration is, of course, not the only way in which one can improve transparency, reproducibility, and replicability. Over the years, several sets of reporting guidelines for ESM studies have been proposed (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 facilitate transparency by stimulating researchers to fully describe a study's design, analysis plan, and results in the final publication. Crucially, however, transparent reporting of ESM research in a published article does not preclude hypotheses and analysis plans from differing wildly from what was originally planned. Without a frozen, uneditable version of the initial plan for a study (i.e., a registration), such changes between plans and published research are untraceable. For example, a recent investigation of preregistered (non-ESM) studies published in the journal *Psychological Science* found that all deviated from their preregistrations, and only in a minority of cases were these deviations transparently reported (Claesen et al., 2019). The scope of reporting guidelines is only the finished product (i.e., the published article). Although we encourage ESM researchers to rigorously follow reporting guidelines (Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020), going beyond such guidelines is paramount to ensuring accountability and transparency of the entire research process, not just its product. Moreover, both the process and the product must be verifiable, and to this end, registrations and published studies can be compared using transparency checklists, such as Wicherts et al.'s (2016) researcher-degrees-of-freedom checklist.

Registration is a tool with great potential to increase transparency, reproducibility, and replicability within ESM research, but it also comes with a number of challenges. Although many of these are applicable in psychology

research more broadly—especially clinical-psychology research—the number of challenges in ESM research may make the threshold for using registration prohibitively high. Given that ESM necessitates the use of relatively complex models, issues with model nonconvergence are common, and therefore an a priori strategy for handling nonconvergence is required. Moreover, the typical ESM study involves numerous researchers focusing on multiple research questions, which poses a challenge for making a comprehensive preregistration that details all relevant analyses prior to data collection. We discuss considerations regarding model-convergence issues later in this article, but first address issues related to registration of studies using preexisting ESM data sets, as well as necessary deviations from registrations.

Registration of studies using preexisting data sets

Collecting ESM data is resource-intensive and produces large, rich data sets, 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 register a study prior to data analysis but using preexisting data. This would not conform to the strict definition of preregistration, according to which the registration occurs prior to data collection. To preclude registration of studies after data collection but before data analysis would leave vast swaths of research in a transparency “Wild West,” when arguably there is even more potential for data-dependent decision making when the data already exist (Weston et al., 2019).

Fortunately, the idea of registration of studies using preexisting data is becoming more accepted. For example, van den Akker et al. (2019) and Mertens and Kryptos (2019) have created registration templates for studies using preexisting data sets, and there are a growing number of resources and publications addressing this issue (e.g., Weston et al., 2019). Terminology is also being extended to account for this variation on the traditional preregistration format. For example, the term *postregistration* has been proposed for registrations of studies using preexisting data sets, that is, when all data have been collected but no data have been analyzed or when some data have already been analyzed and published, for example, by other members of the research team (Benning et al., 2019).

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, and

are therefore eligible to preregister the study in the traditional sense—before commencing data collection—but are aware that they themselves, collaborators, students, and other researchers will use these data to investigate numerous different research questions. It is usually unfeasible for researchers to preregister every possible hypothesis and analysis plan that will be used in the future. Additionally, creating a single large, unwieldy preregistration for many different research questions may limit the usefulness and clarity of the preregistration. In these cases, the most extensive possible initial preregistration (or even a series of preregistrations) should be made, and subsequent registrations of studies using these data can be approached as coregistrations (if data collection is ongoing) or postregistrations (if data collection has been concluded).

An additional step to prevent data-dependent decision making in postregistrations caused by knowledge of the full data set is to operate a variable checkout system, in which data sets are treated as libraries and held by a data manager or other third party and only variables specified in researchers’ registrations are “checked out” to them (Scott & Kline, 2019). When variables are released, researchers are issued a time- and date-stamped receipts for the requested variables, and these documents can also include chronological details of access to any other variables, much as an individual’s library record documents access to books. Time- and date-stamped access records will also facilitate the use of preexisting ESM data sets for Registered Reports. (For an extensive discussion of the issues around registration and transparency of studies using preexisting data sets, see van den Akker et al., 2019, and Weston et al., 2019).

Necessary deviations from a registration

A further challenge that might be particularly relevant for ESM studies is that deviations from a registered plan may be necessary. Claesen et al.’s (2019) study suggests that it is in fact commonplace for plans to change during the course of a study, but that reporting these deviations is not common. Registrations facilitate transparency in such cases. During data collection, issues that arise may require a deviation from what was originally detailed in a study’s registration. There could be an unexpected issue with a recruitment site or a technical issue with the ESM app that means participants received fewer notifications than they were supposed to as per the protocol. Or a researcher may learn that certain ESM items or instructions have not been clear to participants and may decide to change or amend these halfway through a study. Human error may also mean that a researcher forgot to specify a particular detail within the registration and realizes this only once data have been

collected. This may occur more commonly when researchers first begin to register their studies, as writing a good registration is a skill that requires time and experience to develop (Nosek et al., 2019).

One option for dealing with alterations to study plans is to make a supplementary registration. This then provides a time- and date-stamped record of when particular issues arose and which decisions were taken when. Benning and colleagues (2019) have recently proposed the term *coregistration* to describe registrations carried out during or after data collection, but prior to any data analysis. Some issues, however, may emerge only after data have been accessed, explored, or analyzed. Conceptualizing 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 the registration serves more as an open lab book. It is important to bear in mind that such a postregistration no longer enables a key preregistration 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 Registering ESM Research

We encountered some of these challenges when registering and conducting our own ESM studies using existing preregistration templates, and a further topical discussion around the utility of specialized preregistration templates for specific study designs and methodologies (Srivastava et al., 2019) led us to devise a template for registration of ESM studies, for use in preregistration, coregistration, or postregistration.

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 endeavored to make the proverbial black box transparent, by facilitating registration of ESM research with a specially adapted template. In this article, we walk through our key additions and modifications to the Preregistration Challenge template (Mellor et al., 2019), upon which our template is based. Following reviewer comments, we have also incorporated elements from the preregistration template for preexisting data (van den Akker et al., 2019), to reflect the fact that many ESM studies yield large, rich data sets, 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 postregistrations of studies using archival data (Benning et al., 2019) also brings new opportunities in this regard. Throughout this

article, we take the position that open and transparent plans are, even if imperfect, better than no plans (Nosek et al., 2019) and make suggestions for maximizing transparency in cases when deviations from the registration are necessary.

We also discuss key challenges of registering ESM studies and provide some potential solutions.

Online Resources

The ESM registration template and two completed exemplar templates are available on our OSF project page (<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, and we welcome critical feedback.

Also available at our OSF project page are R Markdown script that uses a simulation approach to calculate the number of participants required (Sample_Rationale_Number_of_Participants.RMD) and R Markdown script to illustrate the effect of serial dependency on the estimation accuracy in stationary autoregressive processes (Sample_Rationale_Temporal_Design.RMD).

A Registration Template for ESM Research

Our central considerations when devising additions to the Preregistration Challenge template (Mellor et al., 2019) were to address (a) specific characteristics of ESM studies that may affect or even preclude their replicability and reproducibility and (b) aspects that may be vulnerable to questionable research practices or analytic flexibility, particularly after data have 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 arise. In the following discussion of our template, we have included detailed descriptions of only those sections that are specific to ESM research; we do not describe sections that are also applicable to other types of studies. The two exemplar completed templates on our OSF project page illustrate how to complete these other more broadly applicable sections.

Sampling plan

In this section, we discuss subsections of the registration template that pertain to the sampling plan.

ESM data collection procedure. When conducting an ESM study, researchers must make numerous decisions regarding data collection, including decisions about the method of data

collection, sampling scheme, and use of incentives to encourage participants' engagement. 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 added a new subsection to the registration template titled "ESM data collection procedure."

The duration of ESM studies ranges from days to weeks to months, and consequently, the amount, quality, or content of data coming in may lead researchers to modify the data-collection procedure (e.g., to increase compliance). 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 easy. As these are data-dependent decisions, such modifications may unintentionally introduce researcher bias, and they may not always be reported in the final manuscript. In some cases, potential modifications can be anticipated, and decision rules can be recorded in the registration. Unanticipated modifications can be addressed in subsequent coregistrations and should definitely be reported in the article.

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 et al., 2018). Occasionally, researchers wish to extend the ESM period for all or some participants. For example, researchers running an ESM study with two groups of individuals, those with and without major depression, may find that individuals with depression exhibit reduced compliance and therefore may choose to increase the number of ESM days for that group to ensure that sufficient observations are collected. This would be an alteration that is contingent on the amount of data coming in. When it is possible to anticipate that extending the number of days may be required (e.g., because of past experience with similar studies or existing literature), a decision rule can be specified in the preregistration. Alternatively, if an extension is not anticipated, but the issue subsequently arises, then supplementing the original registration with a later coregistration would make such a data-dependent decision transparent.

Type of sampling scheme. In ESM studies, the sampling scheme refers to the timing of questionnaire prompts. The sampling scheme is dependent on the temporal dynamics of the construct that the researchers aim to measure. For example, relatively rare occurrences (e.g., alcohol consumption) are likely best measured with an event-contingent design, in which 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 random or semirandom design, in which prompts are sent out at random time points.

In designing a sampling scheme, researchers consider 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's duration (i.e. the total number of measurement occasions) and the sampling frequency (i.e., the time interval between two different measurements; Collins & Graham, 2002). The selected temporal design should be specified in the registration, and any later modifications should be detailed in a coregistration, as some decisions may inadvertently introduce bias. For instance, if researchers are interested in studying a process with high probability of occurrence during weekend days (e.g., alcohol use) but take measurements on weekdays only, then the researchers might conclude that the effect is weaker or nonexistent.

Total number and type of items (open-ended or closed-ended). Many reports of ESM studies describe only those variables that were analyzed for that specific study. Although 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 et al., 2009; Vachon et al., 2019; van Roekel et al., 2019). In this subsection of our registration template, researchers are asked to provide a general description of the total questionnaire length. A longer ESM questionnaire, especially one with more open-ended items, is a greater burden for participants, and this can reduce the compliance rate as well as the data quality (Eisele et al., 2020). When the total number of items is unknown, the potential effect of the questionnaire's length on the compliance rate is unclear. The questionnaire's length may also vary as a result of conditional branching, in which the presentation of certain items is dependent on previous responses. Additionally, researchers may choose to present items in a different random order at each prompt (Wen et al., 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 registration document. Unlike questionnaire measures, which are often subject to copyright and licensing precluding open sharing of materials (Weston et al., 2019), ESM items are not proprietary and can, therefore, be freely shared, with correct attribution to the researchers who originally created them. 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 is intended to produce an open bank of ESM

items for use in research and to facilitate assessment of their quality and their psychometric validation. 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 researchers limit the amount of time that participants have to begin responding 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 under way (e.g., if researchers receive feedback that participants are struggling to complete the questionnaire during the allotted time period). Thus, timing restrictions are important to include in the registration, and researchers may do so in this subsection.

Additional details relevant to the ESM data-collection procedure. Some additional details, including 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 ESM questionnaires. There are various instruction options that may affect compliance (Christensen, Barrett, et al., 2003), as well as motivation and data quality. These include, but are not limited to, the type of instruction (video, one-to-one, group session), the 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 instructions may be subject to change during data collection, they also represent a potential forking path. Consequently, details about instructions and briefing provided to participants should be included in this subsection of the template. 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. Although sample size is a crucial consideration for all research (Button et al., 2013), most reports of ESM studies do not describe a power calculation to justify the sample size or state a rationale for the selection of the sampling frequency (Trull & Ebner-Priemer, 2013, 2020). This represents another threat to the reproducibility of ESM studies (Munafò 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 rationale for the sample size can also be based on practical considerations, such as budget restrictions or the burden on participants. For this reason, in this section, we include some further, in-depth discussion of the key considerations for sample-size planning when registering ESM studies.

The temporal design varies considerably among published ESM studies. In psychiatry, ESM studies that assess highly variable constructs (e.g., mood) have often used 10 measurements per day for 6 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 7 consecutive days was used (Christensen, Wood, & Barrett, 2003). 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) investigated continuous-time autoregressive processes and showed that when the variability of the process is high, large time intervals between assessments negatively affect the accuracy of estimation. Therefore, justifying the selection of the temporal design on the basis of 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 nonconvergence of the statistical model. As Collins (2006) highlighted, 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 on 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 registration.

A further consideration when determining sample size for ESM research relates to the number of participants necessary to obtain accurate estimates of interindividual differences (Maas & Hox, 2005). If individual differences are likely to be large, more information is needed in order to determine an effect than is the case when heterogeneity between individuals is negligible. Researchers can justify the selection of the number of participants using power analysis (Arend & Schäfer, 2019; Bolger et al., 2012; Lane & Hennes, 2018; Raudenbush & Liu, 2001). For ESM studies including individuals from different populations (e.g., studies involving patients with different mental-health conditions), we also suggest performing power analysis to determine group size. The same applies

to ESM studies that include a higher-order grouping level, such as studies of dyads 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, and plans to sample additional subjects in case of dropouts.

A number of available resources can guide researchers in performing power analyses for general multilevel and longitudinal designs (e.g., Arend & Schäfer, 2019; Astivia et al., 2019; Bolger et al., 2012; Brandmaier et al., 2015; and Lane & Hennes, 2018). We have also produced an illustration of how to perform a simulation-based power analysis to select the number of participants, 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. Therefore, we encourage researchers to indicate in the 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 postregistering analyses of preexisting data, researchers can provide references to existing design protocols for the data set, 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). If a power calculation was conducted for a specific set of variables, but these variables are not included in other planned studies using the same data set, we also recommend conducting a power calculation for the planned analyses as though the data had not already been collected. If the available number of participants or observations within the data set falls short of providing the desired power, researchers can make a decision whether to adjust their planned analyses, or perhaps in some cases, they may decide that the available data set 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 indicating 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 number 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 et al., 2007; Myin-Germeys et al., 2018), our template asks researchers to specify measured variables first, and then manipulated variables. Researchers are asked to describe in detail only those 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 template has separate subsections for measured non-ESM, time-invariant variables and measured ESM, time-variant variables. In the latter subsection, we have included instructions to specify the response scale (e.g., Likert scale). Given the multilevel structure inherent to ESM data, researchers are asked to specify variable levels of both measured and manipulated ESM variables. As some ESM studies provide free-response options for specific items, we added an optional “Open-ended questions” subsection, 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 registration because such coding is subject to numerous researcher degrees of freedom. Within-participant ESM-level manipulations are currently less common than observational within-person ESM research; therefore, we added instructions to report both manipulated ESM and manipulated non-ESM variables in the “Manipulated variables” subsection.

In the “Indices” subsection, instructions were updated to include descriptions of how any measurements collected during or outside the ESM period will be combined into an index. Such measurements may include passive monitoring conducted via, for example, an activity tracker. Summary statistics, such as observation-level or within-person-level averages, can be formed at different levels in ESM data sets. These can also be constructed from different sets of items and can 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). Because great flexibility arises in the creation of any index, and because there are few well-validated ESM-based indices, it is highly relevant to specify which ESM items, 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 in this subsection.

Prior knowledge of the data

When a study uses preexisting data, for maximal transparency researchers should record their prior knowledge of the data set (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 that were conducted by the researchers using the same data set and that may have resulted in publications, preprints, or conference presentations, but can also include awareness of the data set from external sources, such as reports by other researchers using it. When 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-participant variability and variables that describe the between-participants variability. Moreover, because of the longitudinal structure of the data, the temporal dynamics of the target process can be modeled. Given the complexity of ESM data (i.e., missing observations, unequally spaced time points, time-varying covariates, autocorrelated observations, higher-level models, nonnormal 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 to an ESM analysis plan, here we focus on the multilevel regression model, which can be considered 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-participant 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 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-individual 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 too strict 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 in the registration.

Another example of potential forking paths is when random effects allow the modeling of nonindependence between individuals. In general, random effects are considered normally distributed (models that do not assume normality for the random effects can be found in Verbeke & Lesaffre, 1996). For instance, a model that incorporates only a random intercept and a fixed slope assumes that the mean level of the outcome differs between individuals, but the slope does not. 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 registration of data analysis that includes nested or crossed random-effects designs.

An important decision regarding the predictors included in the statistical model is which predictors are going to be set as fixed effects. This depends on the hypotheses, so it is an important a priori—and therefore registerable—decision. 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). If 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 with items 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 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 can affect the reliability of the parameter estimates. Nonconvergence in mixed-effects models arises when the structure of the variance components is complex, given the amount of available data, or when the data are highly unbalanced (Eager & Roy, 2017). It is possible to anticipate when convergence issues may arise, although not always, and there are different strategies that can be used to address this issue when 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 violate assumptions, the model does not converge, or other analytic problems arise (van den Akker et al., 2019). For instance, prior to 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-effects 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 nonconvergence issues will be addressed (e.g., by switching the optimizer used or using an alternative, prespecified, simplified model) may help to preserve the preregistration goal of allowing evaluation of a model's capacity to falsify the hypothesis being tested. Even with contingency plans for nonconvergence outlined in a registration, sometimes these plans may not work or additional unexpected convergence issues may arise. In such cases, we suggest the use of supplementary coregistrations to create a frozen record of updated plans and models, especially when other analyses specified in the registration were dependent on the nonconverging model but have not yet been conducted. These deviations would then also need to be transparently reported in the article.

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 Navarro, 2019; Pitt et al., 2002). For instance, researchers might be interested in evaluating the plausibility of the assumption of a model or whether the model is able to capture the target phenomena in a less complex

manner than another model does. 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 a 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 to different preprocessing choices, model specifications, or sets 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 (Steege et al., 2016). Table 1 describes different strategies for assessing model selection.

We also note that when the analysis plan involves estimating more complex statistical models, such as dynamic network models (Bringmann et al., 2013) or dynamic structural equation models (Asparouhov et al., 2018), the registration should take into account all necessary information to reproduce the analysis. For instance, when a model using a Bayesian approach is estimated, 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 (R Core Team, 2020), Mplus (Muthén & Muthén, 2017), Stata, JASP (The JASP Team, 2020), jamovi (The jamovi project, 2020), and SPSS. We encourage researchers to specify the software and 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 OSF.

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 related to technical problems can also be included in the analysis plan; for instance, researchers could opt not to include participants who report a technical issue. 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).

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 to perform power analysis or evaluate the effect of the temporal design. For mixed-effects models, the simulation-based approach to evaluating 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 various decisions regarding, for example, inclusion or exclusion of predictors, the structure of the random effects, the patterns of missing data, and selection of the optimizer function to evaluate model convergence (see DeBruine & Barr, 2019, for an introduction).
Stepwise selection	Mixed-effects models allow explicit modeling of the hierarchical structure of intensive longitudinal designs, and thus reduce the probability of false positives and false negatives. There are some disadvantages related to overparameterization of the random-effects structure and highly imbalanced data that might cause convergence issues in models with a singular covariance matrix for the random effects. Bates et al. (2018) proposed that researchers estimate a parsimonious random-effects structure by assessing the dimensionality of the random effects via a principal component analysis of the estimated covariance matrix of the random effects. Stepwise selection can also be performed using likelihood ratio tests. In this procedure, some terms are sequentially set to zero until a parsimonious model is achieved (Harrison et al., 2018).
Model selection based on information theory	Information theory can be used to select a model from a set of competing models. This involves ranking models using metrics such as Akaike's information criterion (AIC). Approaches that use information-based theory to perform model selection involve model averaging, performing all-subset regressions followed by application of AIC, and using Akaike weights to quantify variables' 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) performed cross-validation to select the model that best predicted affective states in studies of within-individual 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 exploratory data analyses. If the aim is to test a hypothesis of interest (i.e., confirmatory analysis), it is possible to separate the data into a training set to perform exploratory analysis and a testing set to test the hypothesis of interest (de Groot, 2014). We recommend that researchers specify the plans for exploratory analysis in the registration. For analysis of preexisting data, we encourage researchers to state any exploratory analyses that were performed prior to the registration.
Sensitivity analysis	There are different ways to test the statistical significance of a model. Model selection relies on data-analytic decisions (e.g., which variables to include, what the compliance threshold should be, and how to handle outliers). Sensitivity analysis can be used to study the effect of preprocessing the data or using different model specifications, by assessing the distribution of the estimated effects (e.g., Simonsohn et al., 2015; Steegen et al., 2016; Young & Holsteen, 2017).

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 registration of ESM studies prior to data collection, we recommend that 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 how compliance will be defined (e.g., whether missing a prompt because of technical problems counts as noncompliance). It is also important to describe and justify the thresholds for compliance that will be used to include participants in or exclude them from the analyses (Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2013). Many studies use a rule of thumb for determining the 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 that the threshold used can significantly bias model estimates and that 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 registration of ESM studies prior to data collection, we encourage researchers to report the expected compliance. For ESM studies with preexisting 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 (e.g., people are less likely to respond during working hours than at other times of day), then incorporating additional predictors that account for nonresponses (e.g., time) into the statistical model can help to reduce the bias due to missing observations (Silvia et al., 2013). Moreover, if techniques to handle missing data will be implemented (e.g., full maximum likelihood estimation or multiple imputation), the analysis plan should include detailed information about the framework for processing missing data. Broader discussions on methods for handling missing data can be found in Graham (2012) and Schafer (2001).

To reduce participants' burden, some researchers may opt for a planned missing design, in which 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 registration. (For further discussion of planned missing designs in ESM and their implications, see Silvia et al., 2014.)

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

Conclusions

We have presented a registration template for ESM research, the development of which was inspired by topical discussions around this issue (Srivastava et al., 2019) as well as our own experiences registering ESM studies using existing tools. We have also included detailed explanations and potential solutions for key challenges for the registration of studies using ESM. To guide ESM researchers further in how to approach registration, we have created two exemplar completed templates, one illustrating a preregistration prior to data collection and another illustrating a postregistration for a study using preexisting data. Many researchers are already making great strides in increasing reproducibility and transparency in 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, 2019), where much ESM research is conducted. The adoption of open-science practices in ESM research is, however, still in its elementary stages. Preregistration is

a cornerstone of open science (Nosek et al., 2018), and Bastiaansen et al. (2019) specifically suggested that it be used more in ESM research 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 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 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 registration—ought to be addressed. Issues of model selection and convergence are examples of issues often overlooked in registration guidance, as are issues related to how to fit the typical setup of ESM studies, in which many researchers work on a single data set to answer numerous research questions, into the concept of preregistration. We have 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 type of ESM study; for example, it may need to be adapted for ESM studies of experimental procedures. We designed the template for what the literature and our own experiences indicate is the modal ESM study. We also recognize that for some researchers, the list of information to specify in the template may seem extensive; however, the vast majority of the decisions must already be made as a matter of course prior to commencement of data collection. Therefore, we strongly believe that recording these decisions in a registration document does not increase researchers' burden. Nondocumentation of these decisions does not insulate ESM studies from being subject to their effects. Indeed, given the almost dizzying array of choices necessary in conducting 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 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 would be recorded transparently and to limit possibilities for analytic flexibility and researcher bias—key threats to reproducibility of results and replicability of methods (Munafò et al., 2017). Registration should not be seen as a substitute for rigorous reporting of results using existing guidelines for reporting ESM studies (Stone & Shiffman, 2002; Trull & Ebner-Priemer,

2020), however; because these guidelines pertain to the product of the research (i.e., the written report), they do not necessarily capture researcher degrees of freedom in the research process, and therefore registration has additional value. Shining a light on the scientific process reveals that it is rarely perfect. Registration does not preclude imperfection, and deviations from preregistrations appear to be common (Claesen et al., 2019), but registration does make deviations more transparent. 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” registrations are 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, p. 3).

Transparency

Action Editor: Alexa Tullett

Editor: Daniel J. Simons

Author Contributions

O. J. Kirtley conceptualized the registration template and article. O. J. Kirtley, G. Lafit, R. Achterhof, and A. P. Hiekkaranta wrote the manuscript and adapted the template. G. Lafit conducted simulations and wrote R scripts. I. Myin-Germeys provided critical revisions of the manuscript and the registration template.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

O. J. Kirtley is currently supported by a Senior Postdoctoral Fellowship from Research Foundation Flanders (FWO 1257821N). During preparation of this article, O. J. Kirtley was also supported by a postdoctoral fellowship from a Research Foundation - Flanders Odysseus grant to I. Myin-Germeys (FWO GOF8416N), which also includes the PhD studentships of R. Achterhof and A. P. Hiekkaranta and the postdoctoral fellowship of G. Lafit.

Open Practices


Open Data: not applicable

Open Materials: not applicable

Preregistration: not applicable

ORCID iDs

Olivia J. Kirtley  <https://orcid.org/0000-0001-5879-4120>

Robin Achterhof  <https://orcid.org/0000-0002-3269-2270>

Acknowledgments

We would like to thank participants in the first center for Research on Experience sampling and Ambulatory methods Leuven (REAL) Workshop (March 20–22, 2019, Leuven, Belgium) for pilot-testing the ESM preregistration template. We would also like to thank our colleagues within the Center for Contextual Psychiatry, particularly Wolfgang Viechtbauer, for

providing critical feedback on our ideas for the template during a lab-meeting presentation by G. Lafit.

The title for this article is similar to the title of a symposium at the Society for Ambulatory Assessment's conference in Syracuse, New York, June 19–22, 2019. That symposium, organized by the first and second authors, was titled “Making the Black Box Transparent: Open Science Practices in ESM Research.”

Prior Versions

The submitted and accepted manuscripts are available online at <https://psyarxiv.com/seyyq7>.

References

- Aarts, E., Dolan, C. V., Verhage, M., & van der Sluis, S. (2015). Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience*, 16, Article 94. <https://doi.org/10.1186/s12868-015-0228-5>
- Adolf, J. K., Loossens, T., Tuerlinckx, F., & Ceulemans, E. (2019). *Optimal sampling rates for reliable continuous-time first-order autoregressive modeling*. PsyArXiv. <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, Article 97. <https://doi.org/10.1186/s12874-019-0742-8>
- 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., Chow, S.-Y., de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E. L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., . . . Bringmann, L. F. (2019). *Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology*. PsyArXiv. <https://doi.org/10.31234/osf.io/c8vp7>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parasimonious mixed models*. arXiv. <https://arxiv.org/abs/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. *Journal 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). 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, Article 272. <https://doi.org/10.3389/fpsyg.2015.00272>
- Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLOS ONE*, 8(4), Article e60188. <https://doi.org/10.1371/journal.pone.0060188>
- 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. <https://doi.org/10.1037/met0000178>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376.
- 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. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, 4(1), 53–78.
- Christensen, T. C., Wood, J. V., & Barrett, L. F. (2003). 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). *Preregistration: Comparing dream to reality*. PsyArXiv. <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). *Understanding mixed effects models through data simulation*. PsyArXiv. <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 IJ 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, Article 1849. <https://doi.org/10.3389/fpsyg.2017.01849>
- 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. <https://doi.org/10.1080/02699931.2018.1524747>
- Delespaul, P. A. E. G. (1995). *Assessing schizophrenia in daily life: The experience sampling method*. Datawyse/Universitaire Pers Maastricht.
- Eager, C., & Roy, J. (2017). *Mixed effects models are sometimes terrible*. arXiv. <https://arxiv.org/abs/1701.04858>
- Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). *The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population*. PsyArXiv. <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*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Graham, J. W. (2012). *Missing data: Analysis and design*. Springer.
- Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, Article e4794. <https://doi.org/10.7717/peerj.4794>
- 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, Article 59. <https://doi.org/10.1186/s12888-018-1983-5>
- Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Sage Publications.
- 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), 432–443.
- 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), 901–930.
- Jacobson, N. C. (2020, January 14). *Compliance thresholds in intensive longitudinal data: Worse than listwise deletion: Call for action* [Symposium]. Society for Ambulatory Assessment Conference, Melbourne, Australia.
- The jamovi project. (2020). *jamovi* (Version 1.2) [Computer Software]. <https://www.jamovi.org>
- 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), Article 140. <https://doi.org/10.1186/s12874-018-0579-6>

- The JASP Team (2020). *JASP* (Version 0.14) [Computer software]. <https://jasp-stats.org>
- Kirtley, O. J., Hiekkaranta, A. P., Kunkels, Y. K., Verhoeven, D., Van Nierop, M., & Myin-Germeys, I. (2019). *The Experience Sampling Method (ESM) Item Repository*. OSF. <https://doi.org/10.17605/OSF.IO/KG376>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E., Bocian, K., Brandt, M. J., Busching, R., . . . 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. <https://doi.org/10.1177/2515245918810225>
- Krone, T., Albers, C. J., & Timmerman, M. E. (2016). Comparison of estimation procedures for multilevel AR(1) models. *Frontiers in Psychology*, 7, Article 486. <https://doi.org/10.3389/fpsyg.2016.00486>
- 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.
- Lakens, D. (2019). *The value of preregistration for psychological science: A conceptual analysis*. PsyArXiv. <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., Rifkenbark, G. G., Newton, S. D., Li, X., Kooker, J., Yomtov, D., Gambino, A. J., & 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. <https://doi.org/10.3102/1076998618776348>
- Mellor, D. T., Esposito, J., Hardwicke, T. E., Nosek, B. A., Cohoon, J., Soderberg, C. K., Kidwell, M. C., Clyburne-Sherin, A., Buck, S., DeHaven, A., & Speidel, R. (2019). *Preregistration Challenge: Plan, test, discover*. OSF. <https://osf.io/x5w7h>
- Mertens, G., & Krypotos, A.-M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, 59(1), 338–352. <https://doi.org/10.5334/pb.493>
- Morren, M., van Dulmen, S., 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., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, Article 0021. <https://doi.org/10.1038/s41562-016-0021>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Author.
- 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., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, 23, 815–818. <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, USA*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., Lewis, S. W., & 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), 472–491.
- Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6(4), 387–401.
- R Core Team (2020). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, 31(2), 226–235.
- 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). 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.
- Scott, K. M., & Kline, M. (2019). Enabling confirmatory secondary data analysis by logging data checkout. *Advances in Methods and Practices in Psychological Science*, 2(1), 45–54. <https://doi.org/10.1177/2515245918815849>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., . . . 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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. SSRN. <https://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.). SAGE Publications.
- Srivastava, S., Tullett, A. M., & Vazire, S. (Hosts). (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/>
- Steenen, 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. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Annals of Behavioral Medicine*, 16(3), 199–202.
- Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, 24(3), 236–243.
- Szollósi, 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. <https://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*, 31(12), 1386–1394. <https://doi.org/10.1037/pas0000583>
- Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, 12(5), 742–756. <https://doi.org/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. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, 9, 151–176.
- Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, 23(6), 466–470. <https://doi.org/10.1177/0963721414550706>
- 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.
- 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), Article e14475. <https://doi.org/10.2196/14475>
- van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., Hall, A. N., Kosie, J. E., Kruse, E., Olsen, J., Ritchie, S. J., Valentine, K. D., van 't Veer, A. E., & Bakker, M. (2019). *Preregistration of secondary data analysis: A template and tutorial*. PsyArXiv. <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. <https://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), Article e132. <https://doi.org/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, Article 1832. <https://doi.org/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., & IJsselstein, 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. <https://doi.org/10.1016/j.jrp.2018.03.001>