General Article

Advances in Methods and Practices in Psychological Science
July-September 2024, Vol. 7, No. 3, pp. 1–27

DOI: 10.1177/25152459241267912
www.psychologicalscience.org/AMPPS

So You Want to Do ESM? 10 Essential Topics for Implementing the Experience-Sampling Method

Jessica Fritz[1,2,3], Marilyn L. Piccirillo[4], Zachary D. Cohen[5], Madelyn Frumkin[6,7], Olivia Kirtley[8], Julia Moeller[9], Andreas B. Neubauer[10], Lesley A. Norris[11], Noémi K. Schuurman[12], Evelien Snippe[13], and Laura F. Bringmann[14]

[1]Department of Psychology, Osnabrück University, Osnabrück, Germany; [2]Department of Psychology, Philipps-University of Marburg, Marburg, Germany; [3]Department of Psychiatry, University of Cambridge, Cambridge, England; [4]Department of Psychology, University of Washington, Seattle, Washington; [5]Department of Psychology, University of Arizona, Tucson, Arizona; [6]Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, Missouri; [7]Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School, Cambridge, Massachusetts; [8]Center for Contextual Psychiatry, Department of Neuroscience, KU Leuven, Leuven, Belgium; [9]Department of Education, Leipzig University, Leipzig, Germany; [10]Department of Developmental Psychology and Research Methods, Institute of Psychology, RWTH Aachen University, Aachen, Germany; [11]Warren Alpert Medical School, Department of Psychiatry and Human Behavior, Brown University, Providence, Rhode Island; [12]Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands; [13]University Medical Center Groningen, Department of Psychiatry, University of Groningen, Groningen, the Netherlands; and [14]Department of Psychometrics and Statistics, University of Groningen, Groningen, the Netherlands

Abstract

The experience-sampling method (ESM) captures psychological experiences over time and in everyday contexts, thereby offering exciting potential for collecting more temporally fine-grained and ecologically valid data for psychological research. Given that rapid methodological developments make it increasingly difficult for novice ESM researchers to be well informed about standards of ESM research and to identify resources that can serve as useful starting points, we here provide a primer on 10 essential design and implementation considerations for ESM studies. Specifically, we (a) compare ESM with cross-sectional, panel, and cohort approaches and discuss considerations regarding (b) item content and phrasing; (c) choosing and formulating response options; (d) timescale (sampling scheme, sampling frequency, survey length, and study duration); (e) change properties and stationarity; (f) power and effect sizes; (g) missingness, attrition, and compliance; (h) data assessment and administration; (i) reliability; and (j) replicability and generalizability. For all 10 topics, we discuss challenges and—if available—potential solutions and provide literature that can serve as starting points for more in-depth readings. We also share access to a living, web-based resources library with a more extensive catalogue of literature to facilitate further learning about the design and implementation of ESM. Finally, we list topics that although beyond the scope of our article, can be relevant for the success of ESM studies. Taken together, our article highlights the most essential design and implementation considerations for ESM studies, aids the identification of relevant in-depth readings, and can thereby support the quality of future ESM studies.

Keywords

experience-sampling method, ESM, ecological momentary assessment, EMA, ambulatory assessment, AA, daily life methods, intensive longitudinal data, ILD

Received 11/3/23; Revision accepted 6/26/24

Corresponding Author:
Jessica Fritz, Department of Psychology, Osnabrück University, Osnabrück, Germany
Email: jf585@cantab.ac.uk

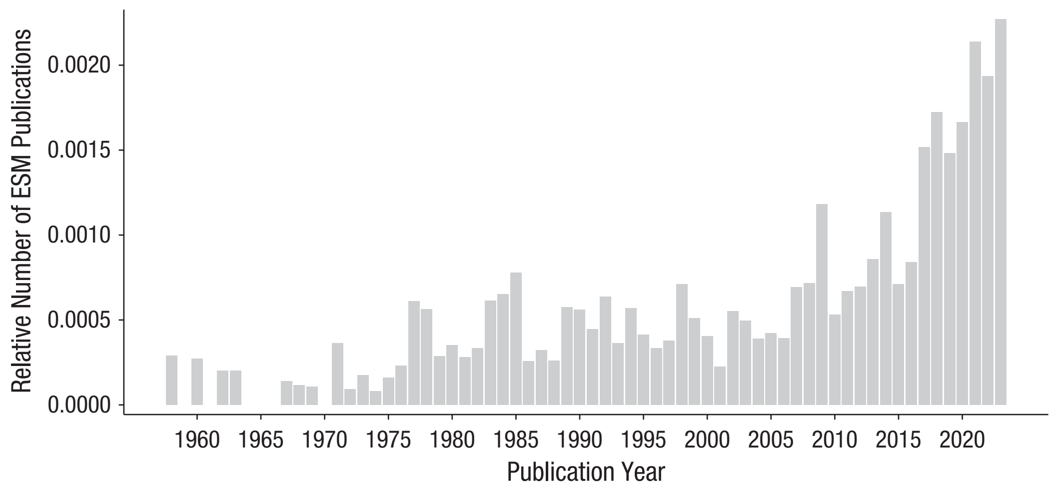

Fig. 1. The relative number of experience-sampling-method (ESM) publications on Web of Science (WoS). This figure shows the number of ESM publications on WoS per year since 1958 compared against the number of all psychological publications on WoS per year since 1958. For the absolute numbers of ESM and of all psychological publications on WoS, see Appendix B.

The experience-sampling method (ESM) is "a structured diary method in which subjects are asked in normal daily life to report their thoughts, feelings and symptoms, and also the context" (Myin-Germeys et al., 2009, p. 1533). "The longitudinal character of the ESM enables the study of dynamic patterns over time, thus capturing the film rather than a snapshot of daily life reality" (Myin-Germeys et al., 2009, p. 1539). ESM has also been referred to by other terms, including "ecological momentary assessment," "ambulatory assessment," "real-time data capture," "intensive longitudinal," and "time-series" research (Trull & Ebner-Priemer, 2009). Although some authors have used these terms interchangeably, others have suggested that the terms are associated with slightly different assessment techniques (e.g., ambulatory assessment is often used when physiological data are collected; Trull & Ebner-Priemer, 2009). Throughout, we use the term "ESM" to describe these methods.

Psychological research has been criticized for its limited ecological validity—a type of validity that denotes the degree to which research results can be generalized to the everyday context (Moeller, Dietrich, et al., 2023). This is an understandable criticism given that (quantitative) psychological research is usually either limited in the number of observations (e.g., observational studies with cross-sectional, panel, or cohort designs[1]), the setting (e.g., tightly controlled experimental lab studies), or both. Studies with limited observations or settings do not necessarily capture the dynamic and context-dependent nature of the human experience and can therefore not automatically be generalized to the everyday context. As a method that can effectively capture psychological experiences over time and in various everyday contexts, ESM offers exciting potential to mitigate this limitation (Trull & Ebner-Priemer, 2020).

In recent years, there has been a noteworthy surge in ESM publications, which seems to surpass a mere general increase in publications (see Fig. 1): More than a quarter of all ESM articles listed on Web of Science (372 of 1,269; 29.31%) were published in only the past 3 years (for details of our literature search strategy, see Appendix A). The growing use of ESM in psychological research is often attributed to the simultaneous advances in digital technologies that facilitate ESM research (e.g., Kuppens et al., 2022). By using smartphones, smartwatches, or other wearables, researchers can assess people's behaviors, thoughts, and feelings with an increasing ease and temporal resolution (e.g., collecting subjective self-report or objective biophysiological or mobile-sensing data[2]; e.g., Kuppens et al., 2022). These developments provide researchers with unique opportunities to investigate complex psychological processes that evolve dynamically both within a single individual and at large scale in populations.

Because technological advancements of the past years moved ahead at a rapid pace, it is unsurprising that the ESM field continuously progressed into new generations of tools and methods. However, these developments make it increasingly difficult for novice ESM researchers to be well informed about standards of ESM research and to identify resources that can serve as useful starting points for the design and implementation of ESM studies. Therefore, we provide a primer (foremostly) for novice ESM researchers on essential design and implementation considerations. This article originated from a panel discussion[3] with experienced ESM researchers, during which we put a special focus on ESM design and implementation topics that are often not included in peer-reviewed articles (e.g., because of their perceived tangential nature regarding the content focus of the article or because of word counts). Most of the topics that we discuss in this article are to some degree similar for studies that are more limited in the number of observations (e.g., cross-sectional, panel, or cohort studies) and/or the setting (e.g., lab experiments). However, because of the high sampling frequency and the emphasis on ecological validity and temporal processes, some design and conduction aspects are unique to ESM (e.g., ESM-specific sampling schemes), and many require tailoring to the ESM context (e.g., power, reliability, and replicability).

We direct the reader elsewhere for (a) reporting guidelines for ESM studies (e.g., Liao et al., 2016; Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020), (b) reviews of methods and practices for ESM research (e.g., Bolger et al., 2003; Trull & Ebner-Priemer, 2020), (c) discussions on current and future applications of ESM (e.g., Mestdagh & Dejonckheere, 2021; Myin-Germeys et al., 2009, 2018), (d) analytical considerations (e.g., Hamaker & Wichers, 2017), (e) insights into conducting ESM studies in specific psychological subdisciplines (e.g., affective sciences: Kuppens et al., 2022; personality psychology: Kaurin et al., 2023), and (f) in-depth handbooks on how to conduct ESM research (e.g., Bolger & Laurenceau, 2013; Conner & Mehl, 2012; Hektner et al., 2007; Myin-Germeys & Kuppens, 2021). Instead, we provide a brief overview of the following 10 topics that we believe are essential for the quality of ESM design and implementation, discuss current challenges, and—if available—highlight potential solutions:

1. study type and research questions;
2. item content and phrasing;
3. choosing and formulating response options;
4. timescale (sampling scheme, sampling frequency, survey length, and study duration);
5. change properties and stationarity;
6. power and effect sizes;
7. missingness, attrition, and compliance;
8. data assessment and administration;
9. reliability;
10. replicability and generalizability.

For each of these topics, we additionally provide a selection of literature as starting points for more in-depth readings. Moreover, we have compiled a living, web-based resources library[4] with a more extensive catalogue of literature to facilitate further learning about the design and implementation of ESM studies. We hope that the resources library will be continually edited and improved by the vibrant ESM community so that it provides an ever more extensive and up-to-date overview of the available literature.

Study Type and Research Questions

In ESM studies, data are typically collected at least once per day in the participants' everyday lives (Trull & Ebner-Priemer, 2014). Over several days, weeks, or months, this results in so-called intensive longitudinal data (Hamaker & Wichers, 2017). ESM enables researchers to distinguish between within-subjects and between-subjects effects (e.g., Voelkle et al., 2014). To explain the differences between these effects, we use the relationship between rumination and anxiety as a running example.

The between-subjects approach compares multiple people against one another, thereby examining effects based on *interindividual* variance (Schuurman, 2023). This approach answers research questions such as "Do people who tend to ruminate more than others tend to also have higher levels of anxiety than others?" and necessarily requires data from multiple participants but can be applied to data from just one time point (e.g., cross-sectional data). For the between-subjects approach, researchers can therefore rely on data from either (a) many participants on one time point (e.g., cross-sectional data; although this comes with notable assumptions because cross-sectional data are technically not pure measures of between-persons associations; for details, see Hamaker, 2023), (b) many participants over a few time points (e.g., prospective panel or cohort data; Hamaker & Wichers, 2017), or (c) many participants over many time points (i.e., intensive longitudinal data).

The within-subjects approach typically[5] compares one person or multiple people against themselves over time, thereby examining *intraindividual* variance over time. This approach answers research questions such as "Does the average person tend to ruminate more than usual when she has a higher level of anxiety than usual?" and does not necessarily require data from multiple participants but requires data from multiple time points (for exceptions, see Note 5 and Schuurman, 2023). For the within-subjects approach, researchers can rely on data from either (a) many participants over several time points (e.g., prospective panel or cohort data; Hamaker & Wichers, 2017), (b) many participants over many time points (i.e., intensive longitudinal data), or (c) a few participants over many time points (i.e., multiple sets of single-case time-series data or intensive longitudinal data).

A specific case of the within-subjects research approach is when only one participant is assessed over many time points, which is also referred to as "*N*-of-1"

Table 1. Comparing Features of Cross-Sectional, Panel, Cohort, and ESM Studies

Study type	Cross-sectional	Panel or cohort[a]	ESM
Dynamic		(x)	x
Between-subjects research questions	(x)	x	x
Within-subjects research questions	(x)	x	x
N-of-1 research questions			x

Note: Parentheses "(x)" indicate the presence of noteworthy assumptions. ESM = experience sampling method.
[a]As the terms panel study and (single or accelerated) cohort study are inconsistently used and defined in the literature, we note that we specifically refer to longitudinal designs that assess a group of participants at several (mostly 2 to 10) time points (as opposed to assessing different groups of people at several time points, such as in repeated cross-sectional designs; Lugtig & Smith, 2019).

approach (or single-case, idiographic, or person-specific approach). Imagine our participant of interest is named Beth; this approach would answer research questions such as "Does Beth tend to ruminate more than usual when she has a higher level of anxiety than usual?" The *N*-of-1 version of the within-subjects approach is particularly fruitful when one either cannot recruit multiple participants, for example, because the construct or context of interest is rare (e.g., Wichers & Groot, 2016), or when the data are supposed to aid person-specific interventions (e.g., personalized feedback in clinical care; Bos et al., 2022; Bringmann et al., 2021; Piccirillo et al., 2019; Wright & Woods, 2020; Zuidersma et al., 2020).[6] In sum, because ESM studies collect intensive longitudinal data, they can more effectively capture the dynamics (i.e., temporal properties) of psychological experiences than cross-sectional, panel, and cohort studies and can also answer person-specific research questions (see also Table 1).

Last but not least, ESM studies are—by virtue of temporally fine-grained assessments—uniquely suited to answer research questions regarding the timescales of psychological experiences in real life and real time. For example, ESM studies are uniquely suited to answer questions such as "How strongly are past rumination levels (an hour, a few hours, or a day ago) associated with current anxiety levels?"; "How long does it generally take for people to return to their usual level of anxiety?"; "How stable are rumination levels (over an hour, a few hours, or a day)?"; "In which order do anxiety and rumination levels change over time?"; and many more.

Taken together, the foremost challenge for the researcher is to determine whether the question of interest is indeed best addressed with an ESM study. One critical consideration therein is that because of the need for frequent and time-sensitive assessments, ESM studies can be more burdensome for the participants than cross-sectional, panel, or cohort studies. This is especially the case for *N*-of-1 approaches because those depend solely on the number of observations collected for that one specific person (in other words, while in panel, cohort, and multiple-participant ESM analyses statistical power can be achieved through both the number of participants and the number of time points, in fully idiographic *N*-of-1 analyses, power can be achieved solely through the number of time points at which the respective person under examination has provided data; e.g., Mansueto et al., 2023).

Starting Points for In-Depth Readings

- Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.
- Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1
- Schuurman, N. K. (2023). *A "within/ between problem" primer: About (not) separating within-person variance and between-person variance in psychology*. PsyArXiv. https://doi.org/10.31234/osf.io/7zgkx
- Wright, A. G. C., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology, 16*(1), 49–74. https://doi.org/10.1146/annurev-clinpsy-102419-125032

Item Content and Phrasing

After having determined that a research question is best addressed using ESM, the next step is to choose the relevant items or assessment sources to capture the variables of interest (e.g., subjective self-report or other-report items, objective data such as heart rate or smartphone sensor data). In this article, we specifically focus on self-report (or other-report) items. Given the dearth of measures developed specifically for ESM, researchers frequently create their own ESM items—either by adapting items from existing retrospective self-report questionnaires or by generating entirely new items—to assess the construct of interest. To keep ESM questionnaires short, the construct of interest is often assessed with a single item (Horstmann & Ziegler, 2020; J. Song et al., 2023; Wright & Zimmermann, 2019). Together with the lack of expert consensus regarding how best to determine the psychometric properties of

ESM items (e.g., reliability and validity; Dejonckheere et al., 2022; see also the Reliability section), this means that psychometrically validated ESM questionnaires are a rarity (Wright & Zimmermann, 2019). A number of sources (e.g., Eisele et al., 2021; Myin-Germeys et al., 2018; Palmier-Claus et al., 2011, 2019) provide guidance on item design, but there are no empirically supported "gold-standard" criteria for developing good ESM items. Currently, such "gold-standard" criteria are being developed (Eisele et al., 2024) as part of the ESM Item Repository (Kirtley et al., 2024; www.esmitemrepository.com), an open-science initiative that aims to increase the transparency and validity of ESM items. However, to date, ESM researchers are left in the uncomfortable position of making decisions about item content and phrasing with little empirical evidence to support those decisions (e.g., Eisele et al., 2021).

As described above, there are numerous unanswered questions regarding ESM items. Typical questions regarding item design relate to (a) item comprehensibility and psychometric properties, (b) the number and selection of items, and (c) the time frame that the items are assessing. Often, the first challenge is to ensure participants understand the items because confusing or illogical wording may increase participant burden and decrease compliance, thus reducing data quality (Eisele et al., 2022). The second challenge is to determine an appropriate length for the questionnaire. Shorter ESM questionnaires can sometimes result in better compliance and data quality (Eisele et al., 2022). However, researchers must ensure that developing short questionnaires does not come at the expense of validity. Focus groups or cocreation sessions with the target population can help with the operationalization of the construct of interest, such as with choosing the relevant content and an appropriate number of items (Soyster & Fisher, 2019; Vogt et al., 2004). Moreover, piloting the chosen or developed ESM items is recommended, ideally with the inclusion of a qualitative component (e.g., a "think-aloud study"; Ericsson & Simon, 1998), to guide further item refinement. Both those endeavors are especially necessary when new items or measures are being developed. A thorough briefing of participants before the commencement of the study creates the opportunity for aiding participants' understanding of items and is useful for improving adherence to the sampling protocol (Rintala et al., 2021)—this endeavor is relevant for all ESM studies regardless of using new or already established items.

Another important consideration when designing or selecting ESM items is whether the time frame of reference for the item (e.g., "since the last beep" or "right now") is appropriate (Deakin et al., 2022). Simply adding "right now" to an item that has been adapted from a retrospective trait-type questionnaire does not make the item suitable for ESM studies (Myin-Germeys et al., 2018). Researchers should also be mindful that although it may be desirable to capture certain behaviors, such as self-harm, "in the moment," this may realistically not be possible. Therefore, either asking about these behaviors "since the last beep" or assessing them only contingent to when they occur (i.e., event-contingent assessments that are initiated by the participant) may be a better option. Alternatively, if it is essential for the design of the study to assess equally spaced time windows, one could ask about these behaviors "in the last day," or when real-time anchoring is at the core of the study, one could ask for the behaviors "in the last hour" (see e.g., Kuppens et al., 2022; Medland et al., 2020).

In sum, unknowns regarding the construction of ESM items will be resolved only when the field devotes substantive empirical research to this topic. More knowledge and better guidance will not happen overnight but will be expedited by researchers increasing transparency about how and why items or measures have been selected, modified, or created—an essential step that all ESM researchers can adopt in their research workflow today. Moreover, to further facilitate the availability and refinement of validated ESM items, researchers are encouraged to submit their validated items to the ESM Item Repository (Kirtley et al., 2024; www.esmitemrepository.com).

Starting Points for In-Depth Readings

- Eisele, G., Kasanova, Z., & Houben, M. (2021). Questionnaire design and evaluation. In I. MyinGermeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 71–89). KU Leuven. https://www.kuleuven.be/samenwerking/real/real-book
- Kirtley, O. J., Eisele, G., Kunkels, Y. K., Hiekkaranta, A. P., Van Heck, L., Pihlajamäki, M., Kunc, B., Schoefs, S., Kemme, N. D. F., Biesemans, T., & Myin-Germeys, I. (2024, July 2). *The experience sampling method (ESM) item repository*. PsyArXiv. https://doi.org/10.17605/OSF.IO/KG376

Choosing and Formulating Response Options

Once the items have been selected, the next step is to choose and formulate response options. ESM researchers typically use either a Likert scale or a visual-analogue scale (VAS). A Likert scale consists of multiple response options that, for example, reflect the extent of agreement, frequency, or severity (e.g., 1 = *not at all*, 4 = *moderate*, 7 = *very much*). The number of response options can vary but often ranges from three to 10 and generally assesses data that are seen as discrete (Eisele et al., 2021). In contrast, a VAS often ranges from 0 to 100 (e.g., 0 = *not at all*, 50 = *moderate*, 100 = *very much*) and assesses data that are seen as continuous (Eisele et al., 2021). However, this distinction is not clear-cut. For example, when a Likert scale has seven or more answer options, it often starts to resemble a continuous scale, and in a similar vein, a VAS can be used in a rather discrete way when participants use only some parts of the scale (e.g., only the outer ends and middle part of the scale) or when researchers discretize the scale post hoc into ranges. ESM researchers sometimes also use binary (or dichotomous) response scales, which is a specific case of discrete data that have only two response options (e.g., 1 = *yes* and 0 = *no*). Finally, ESM researchers can choose to use open-text boxes as answer format so that participants can note down their individual answer in a nonstructured or semi-structured way.

For ESM researchers who want to proceed quantitatively, the first challenge is that they are left with the dilemma of having to choose between a binary, a Likert, and a VAS response format, which all come with noteworthy caveats. The upside of binary-response scales is that they are rather simple and efficient and can nonetheless capture relevant information for psychological constructs for which a more fine-grained assessment does not automatically result in more informative findings (e.g., Fisher, 2023; for an example on presence/absence vs. the severity of adverse experiences, see Schlechter et al., 2021). The downside of binary-response scales is, however, that the resulting data per definition do not follow a normal distribution, and not all analytic techniques for intensive longitudinal data can deal with this. Therefore, Likert scales or VASs are often preferred over binary scales. When deciding for a Likert scale, one generally risks the chance of overdiscretizing the answer options, which can lead to an insufficient variance. When deciding for a VAS, it is not always clear how to interpret response values and small differences therein. For example, researchers do not have a good idea of what a score of 76 really means or what the participant thinks it means. Moreover, a difference of 73 and 76 could be a meaningful difference but could equally just be measurement error because of the width of the participant's finger (Boring et al., 2012; Le et al., 2018; Mayer et al., 2018). That said, researchers can, however, also not be sure that all participants interpret the response options on Likert scales in the same way, particularly not if some response options have no label. From other fields of research, there are some indications that VASs might lead to more endorsement of the end points of the rating scale (i.e., the extreme responses) than Likert scales (Studer, 2012), and generally speaking, Likert scales have been suggested to be most user-friendly (Couper et al., 2006; Eisele et al., 2021). Therefore, although data assessed with VASs tend to be continuous—which is often considered advantageous for statistical analyses—VASs do not seem to provide a clear advantage to Likert response scales (e.g., Haslbeck et al., 2023; Müssig et al., 2022; Simms et al., 2019).

A second challenge that ESM researchers face, potentially regardless of using Likert scales or VASs, is that the number and location of anchor points may well matter. For example, evidence suggests that changes in scale anchors can alter response distributions (Matejka et al., 2016). And a third challenge is that the participants' interpretation of the response options may change over time, suggesting (no or) low longitudinal measurement invariance, which limits the validity of the items.

Even though not much is known yet about what an optimal response scale looks like for ESM items, the general advice is to keep the response scale consistent throughout a (construct-specific) set of ESM items (Eisele et al., 2021). Moreover, we recommend to use at least one of the following three strategies, which likely result in a more homogeneous and reliable use of the response-scale format: (a) piloting the chosen response-scale format with the target population (e.g., in form of a "think-aloud study"; Ericsson & Simon, 1998), (b) briefing participants before the commencement of the study regarding the response-scale format (e.g., in form of written or video instructions), and (c) examining the response distributions of the items to retrospectively increase understanding of how the scale has been used. It may also be useful to include an open-text box alongside quantitative response options so that participants can additionally record individual answers. Using additional open-ended answer formats can be particularly useful when research questions or items are exploratory (van Roekel et al., 2019) and when nonvalidated items or response options are used and researchers want to better understand response distributions and response shifts.

Starting Points for In-Depth Readings

- Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review, 24*(2), 227–245. https://doi.org/10.1177/0894439305281503
- Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2022). *Multimodality and skewness in emotion time series*. PsyArXiv. https://doi.org/10.31234/osf.io/qudr6

- Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5421–5432). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858063
- Schwartz, C. E., Sprangers, M. A., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology & Health, 19*(1), 51–69. https://doi.org/10.1080/0887044031000118456

Timescale (Sampling Scheme, Sampling Frequency, Survey Length, and Study Duration)

Inherent to designing ESM items and response scales are timescale decisions. Timescale decisions usually comprise (a) the sampling scheme (i.e., a fixed, random, semirandom, or event-contingent scheme), (b) the sampling frequency (i.e., the number of assessments per day, also sometimes referred to as "sampling density"), (c) the prompt depth and survey length (i.e., the number of items and the time it takes to assess them), and (d) the study duration (Dejonckheere & Erbas, 2021; Kaurin et al., 2023). Regarding the sampling scheme, fixed schemes consist of preselected assessment time points, whereas random schemes consist of assessments on randomly selected time points. Semirandom schemes consist of assessments that occur at a random time within a predetermined time frame, to ensure that each day has an equal number of assessments (and, on average, equal time windows between assessments; Janssens et al., 2018). Fixed sampling schemes are sometimes also referred to as "interval-contingent schemes," and (semi)random schemes are also referred to as "signal-contingent schemes." In these schemes, assessments are initiated by prompts (Himmelstein et al., 2019). In contrast, event-contingent schemes consist of assessments that take place whenever the event of interest happens (e.g., social interactions; Dawood et al., 2020; Himmelstein et al., 2019; Stadel et al., 2024), that is, assessments are typically initiated by the study participants.

Besides the sampling scheme, timescale decisions regarding sampling density, prompt depth, and study duration—also referred to as "3Ds" (see Kaurin et al., 2023)—vary widely across ESM studies because timescale decisions should align with the theory and research question of interest (Bolger & Laurenceau, 2013; Collins, 2006). Kaurin and colleagues (2023) showed in their meta-analysis that less dense samplings protocols have usually deeper prompts (i.e., longer surveys) and that more dense sampling protocols have usually a shorter study duration. Hence, the 3Ds are related to each other across studies because ESM researchers tend to decrease one when increasing another. Further discussions on timescale decisions can be found in Conner and Lehman (2012), Dejonckheere and Erbas (2021), Himmelstein and colleagues (2019), Kaurin and colleagues (2023), Leertouwer and colleagues (2021), Shiffman and colleagues (2008), Stadel and colleagues (2024), and van Roekel and colleagues (2019).

A first challenge regarding timescales in ESM studies is to choose a sampling scheme. Most sampling-scheme choices have both advantages and disadvantages. Consider social interactions as an example: Fixed or (semi) random schemes may be advantageous when beginning and end points of an event are not clearly identifiable or when events happen (nearly) continuously (e.g., Himmelstein et al., 2019), such as when participants work in settings where they interact with other people close to nonstop (e.g., preschool teachers, hairdressers, general practitioners), so that fixed or (semi)random sampling can help to ensure both a manageable and a sufficient number of assessments. On the other hand, event-contingent schemes may be advantageous for more infrequent events, such as social interactions in representative samples of college students, so that the event-contingent sampling can help record social interactions at the actual frequency and time at which they occur, increasing the ecological validity (e.g., Himmelstein et al., 2019). Although widespread lore or intuition suggests that event-contingent schemes require less retrospection than fixed or (semi)random schemes, the little empirical data that exist on this topic suggest that this might not (always) be the case (e.g., Himmelstein et al., 2019; Stadel et al., 2024).

Unless one chooses an event-contingent scheme, a second major challenge is to determine an appropriate sampling frequency for the construct of interest, because the eventual results depend in large part on the timing of the assessments (see Collins & Graham, 2002; Haslbeck & Ryan, 2022). For example, the strength of temporal (lagged) associations has been demonstrated to differ as a function of the length of the time interval between assessments (also referred to as the "lag problem"; Bolger & Laurenceau, 2013; Collins & Graham, 2002; Dormann & Griffin, 2015; Gollob & Reichardt, 1987; Hamaker & Wichers, 2017; Jacobson et al., 2019; Kuppens et al., 2022; Shiffman et al., 2008; Stone & Shiffman, 2002). Thus, the frequency of assessments needs to be based on the assumed fluctuation speed of the construct of interest (or on the speed of its interaction with other relevant constructs of interest). A third and related challenge is that the timing of assessments should be representative (Conner & Lehman, 2012; Dejonckheere & Erbas, 2021; Stone & Shiffman, 2002), which is particularly challenging when contextual factors influence the assessment (e.g., assessing tiredness right after lunch) or when the assessment time is associated with the measured outcome (e.g., assessing tiredness in a fixed design every day during the late evening).

A fourth challenge is that even though ESM designs reduce recall bias, this bias may still occur to some extent when using retrospective questions (e.g., affect during the past day) in contrast to concurrent questions (e.g., affect right now; Neubauer, Scott, et al., 2020). Designs with retrospective questions are therefore sometimes referred to as "ecological retrospective assessment" (Leertouwer et al., 2021). Likewise, recall bias may increase when the construct of interest is assessed over a time span during which it is likely to fluctuate meaningfully (e.g., appetite level during the past day; Conner & Mehl, 2012; Kuppens et al., 2022; Shiffman et al., 2008). To address these issues, ESM researchers can increase the sampling frequency of fixed or (semi)random assessments, implement periods of measurement bursts into these assessment schemes (for more details on this strategy, see Ram et al., 2014), or use event-contingent schemes. Event-contingent sampling is especially useful when the timescale of the construct of interest is both unknown and relevant (e.g., assessing the amount of alcohol consumed right after the event as opposed to at the end of the day).

A fifth challenge is that the depth of the prompts should be sufficiently comprehensive to capture the construct of interest validly without having a completion time that burdens the participants and decreases their compliance. Here, ESM researchers could consider conducting a pilot study to qualitatively assess study burden and to quantitatively assess compliance. Moreover, in some situations, planned-missingness designs can help to systematically reduce the number of items per prompt in such a way that eventually enough data for all items are available to be able to infer the relevant estimates (for more details, see the section on missingness, attrition, and compliance; or Silvia et al., 2014). Finally, a sixth challenge is that the duration of the study will need to be attuned to (a) a period (or periods) with sufficient occurrences of the construct of interest (e.g., Helmich et al., 2021), (b) the required statistical power, (c) an acceptable level of participant burden, (d) financial resources and researcher availability, and if applicable, (e) the cyclicity of the construct of interest (e.g., week vs. weekend, seasonal, periodic, or event-related effects, such as stress among college students likely being higher around exam times than before or after exam times; Fritz et al., 2021).

Starting Points for In-Depth Readings

- 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. https://doi.org/10.1016/S0376-8716(02)00217-X
- Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 71–89). KU Leuven. https://www.kuleuven.be/samenwerking/real/real-book
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Change Properties and Stationarity

When designing the timescale of an ESM study, one aspect that is often overlooked is that the properties of the construct of interest can change systematically—which we hereafter refer to as "change." Change differs from momentary (i.e., timepoint-specific) fluctuations—which we hereafter refer to as "fluctuations," "variations," or "deviations." Change can occur in many ways. First, the mean level can change—for instance, a participant can become systematically sadder or happier (or a combination of both). Such changes are not just fluctuations, variations, or deviations from the participant's mean, but are structural changes of the construct's mean on average over time. Second, the variance of a construct can change—for example, a participant's sadness or happiness can become systematically more variable over time. Third, the association between variables can change systematically—for example, sadness and happiness can be highly related at the beginning of the study but may have no association at the end of the study. Fourth, how well an ESM item can explain itself—also known as "autocorrelation" or "inertia"—can change systematically; for example, the degree to which sadness at one time point can explain sadness at the next time point can change over the course of the study (Bringmann et al., 2017; Koval & Kuppens, 2012).[7] And to make matters even more complex, over time, participants can begin to give a different meaning to the construct or its response options, which erroneously gives the appearance of observed change in the statistics of the construct (e.g., in the mean level; e.g., McNeish et al., 2021). Data in which the statistical properties of a process, such as the mean, the variance, or the autocorrelation, do not change over time—or in other words, data with no (or very little) change—are also called "stationary data" (Haslbeck & Ryan, 2022; Molenaar, 2004).

So what does this mean for the design of an ESM study? The first and foremost challenge is to decide whether one is interested in change and to then make appropriate choices regarding the timescale and the analytic method (Bringmann et al., 2022). If one is not interested in measuring change itself, it is critical to reduce external influences that may induce change, such as psychosocial treatment, exams, or holidays. Therefore, one could incorporate a prescreener into the study protocol to check before the study starts whether participants go on holidays or are very likely to encounter influential experiences (e.g., a wedding when the participant takes part in a study on mood). Collecting high-frequency assessments during a brief period of time may also increase the chance of capturing approximately stationary data. However, collecting truly stationary psychological data can be complex, given the dynamic nature of many psychological constructs (Bringmann et al., 2022). If one is interested in assessing change itself, or in other words, "nonstationary ESM data" (i.e., means, variances, associations, or autocorrelations change systematically over time), a longer study duration covering a period during which participants are assumed to change would be advantageous (or necessary; e.g., Helmich et al., 2021).

Planning for change is particularly important for choosing the analytic approach. Currently, many of the commonly used analytic approaches (e.g., linear multilevel models, dynamic structural equation models, or vector-autoregressive models) assume stationarity (Ariens et al., 2020). Given that there are only a few approaches that can deal well with nonstationarity, modeling nonstationary data represents a much needed (yet complex) area of future study. Note also that it matters what kind of change one would like to model. For example, modeling approaches used for detecting change in the autocorrelation require more observations than approaches used for detecting change in the mean (Bringmann et al., 2017).

Starting Points for In-Depth Readings

- Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research, 137*, Article 110191. https://doi.org/10.1016/j.jpsychores.2020.110191
- Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy, 149*, Article 104011. https://doi.org/10.1016/j.brat.2021.104011
- Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*(1), 505–528. https://doi.org/10.1146/annurev.psych.57.102904.190146

Power and Effect Sizes

Alongside carefully planning design features related to items and timescale, it is important to ensure adequate statistical power, that is, to ensure that sufficient observations are collected that can be manipulated via the number of time points and/or the number of participants.[8] Generally speaking, power is defined as the probability of rejecting a statistical null hypothesis when it is truly false or more colloquially, detecting an effect as statistically significant that is "truly there." Power is an important and often underestimated parameter for empirical research. Studies with low power that fail to detect true effects are problematic for determining the validity of a given construct. Some simulation studies and resulting recommendations for sample sizes and number of time points for ESM studies exist (see e.g., Lane & Gates, 2017; Mansueto et al., 2023; Schultzberg & Muthén, 2017). However, because these recommendations are highly context specific (e.g., type and complexity of the analytic model, timescale of the relevant constructs, and their interactions), we believe that a power analysis conducted for the specific analytic model and content at hand would provide much more decisive guidance. There are simulation-based tools available that can help ESM researchers to estimate the number of participants and/or the number of time points needed to obtain sufficient power (e.g., Bolger et al., 2012; Lafit et al., 2021, 2022; Schultzberg & Muthén, 2017). However, we want to stress that calculating the statistical power for ESM studies is rather challenging (given the complexity of the time-series models) and sometimes even impossible. Many of the software resources that are available do not account for temporal dependencies, and those that do are so far available for only a limited number of analytic approaches (Lafit et al., 2021). Moreover, conducting a power analysis requires researchers to think clearly about the anticipated or postulated effect size, because an appropriate effect-size expectation is crucial to ensuring that the power calculation is accurate. Therefore, measures of effect sizes play a pivotal role in the advancement of ESM research.

The first challenge that needs to be overcome regarding power considerations is a shift in how ESM researchers approach statistical power. ESM researchers frequently rely on design decisions of prior work (e.g., "We've always recruited 80 participants and administered surveys for 14 days"), which were often determined by pragmatic considerations and not by using power analyses. Pragmatic considerations, such as the participant burden and the amount of financial resources available, are, of course, crucial, and it should be the norm to report them. Yet pragmatic considerations should not deter researchers from conducting and reporting power analyses (even if the power analysis resulted in an insignificant target effect).

A second notable challenge is that ESM researchers may not have a set expectation about the effect size, which is required when calculating statistical power. Here, we recommend that researchers either choose the smallest effect size that they still consider meaningful and want to be able to detect, or that they search the literature for similar studies and base expectations of effects on prior studies. That said, turning to the literature may leave ESM researchers disappointed because effect sizes are often not reported in a standardized metric, which is problematic for conducting power analyses. For instance, researchers might use a Monte Carlo simulation approach (e.g., using the tools developed by Lafit and colleagues, 2021, or Muthén & Muthén, 2017) and estimate the power for a regression coefficient of .10. But if this regression coefficient is unstandardized, a regression coefficient of .10 can be anything from a tiny to a huge standardized effect depending on the scale of the independent and dependent variable. To date, resources to compute and interpret standardized effect sizes are available (e.g., Bulteel et al., 2016; Jaeger et al., 2017; Rights & Sterba, 2019; Schuurman et al., 2016) but are not consistently used. Therefore, if ESM researchers would start to routinely report standardized effects, the field could together move the quality of ESM research forward. And last but not least, if a suitable power analysis seems truly implausible, we recommend using guidance from existing simulation studies (where applicable) or to at least openly report about the alternative (pragmatic) considerations taken and their potential implications.

Starting Points for In-Depth Readings

- 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). The Guilford Press.
- Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*(6), 1086–1105. https://doi.org10.1080/02664763.2016.1193725
- Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920978738
- Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods, 24*(3), 309–338. https://doi.org/10.1037/met0000184
- Schultzberg, M., & Muthén, B. (2017). 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. https://doi.org/10.1080/10705511.2017.1392862

Missingness, Attrition, and Compliance

A topic that is closely related to statistical power is how to address missing data. "The best solution to the missing data problem is not to have any," as the saying goes (Allison, 2001, p. vi). In ESM studies, this goal is usually unattainable because assessing participants throughout their everyday activities almost assuredly means that participants will miss some of the many ESM prompts. In some situations, the participants may simply be unable (e.g., a bus driver at work) and in other situations unwilling (e.g., during a cinema visit) to respond to a prompt. Typically, researchers administer ESM prompts during certain time windows, after which they expire, to make sure that responses are recorded at the intended time and context—meaning that participants miss prompts if they are not able to complete the survey within the appropriate window. If prompts did not expire, following prompts would pile up on previously unanswered ones, thus making it increasingly unlikely for researchers to receive valid responses. Even though many ESM studies are much more intensive for participants than cross-sectional, panel, or cohort studies, reviews of the ESM literature found average compliance rates of 70% to 80% (e.g., de Vries et al., 2021; Jones et al., 2019; Wen et al., 2017; Williams et al., 2021; Wrzus & Neubauer, 2023). However, recent reviews also revealed that about 30% to 50% of ESM studies do not report compliance rates (de Vries et al., 2021; Wrzus & Neubauer, 2023), suggesting that true rates of compliance may be lower than the reported average rates of 70% to 80%. Thus, the key message is that it is unavoidable to plan for missing prompts when designing an ESM study. When planning for missing prompts, it is important to differentiate between randomly missed prompts (e.g., prompts not being received because of technological failures) and systematically missed prompts that relate to constructs of interest (e.g., patients with depressive symptoms who tend to oversleep in the morning could systematically miss early morning prompts), because systematic missingness in particular can bias the data (see Schreuder et al., 2022).

When designing an ESM study, there are generally speaking four approaches that can address missingness. First, researchers should develop a plan to increase compliance. Second, researchers can implement planned-missingness designs. Third, researchers can plan the analysis of missing-data patterns. And fourth, researchers can plan the statistical handling of missing responses based on observed responses. Regarding the first approach, researchers should set out a plan for increasing participant compliance when designing the ESM study. However, besides some mixed evidence for the suggestion that longer questionnaires are significantly associated with reduced compliance (Eisele et al., 2022; Hasselhorn et al., 2021; Williams et al., 2021), there is so far only limited empirical evidence regarding the impact of study design on compliance (see Reiter & Schoedel, 2024; van Berkel et al., 2020). The majority of the advice for increasing compliance in ESM studies is anecdotal. For example, ESM researchers attempt to increase compliance by (a) training the participants regarding the study protocol, (b) compensating based on compliance (e.g., paying participants per response or offering a "bonus" for reaching a certain compliance threshold), (c) implementing suspension windows (i.e., participants can register times at which they are unavailable to prevent getting prompts during those times), (d) making ESM studies overall as user-friendly as possible, and (e) troubleshooting directly with participants if compliance appears problematic (e.g., Burke et al., 2017; Trull & Ebner-Priemer, 2020).

The second approach—planned missing-data designs—is an approach that is implemented before the conduction of the study. It allows researchers to administer all items but to systematically reduce the number of items per prompt in such a way that eventually enough data for all items are available to be able to infer the relevant estimates (Silvia et al., 2014). This is possible because such missing data are missing completely at random and thus unrelated to the construct at hand and to other constructs (Little & Rubin, 2002; Silvia et al., 2014). However, although such a design reduces the assessment burden, unintended additional missing data are likely. Moreover, this approach seems predominantly useful when multiple items are used to assess one construct, and it requires the usage of specific analytic procedures (e.g., latent variable models) to compensate for the intended missing data, which are not (yet) available for all time-series analyses (see e.g., Silvia et al., 2014).

The third approach—planning the examination of missing-data patterns—can include analyses that help to disentangle whether missingness is completely at random (MCAR) or missing at random (MAR; some variables are associated with the missingness; e.g., Enders, 2010; Little, 1988). Although tests to identify MCAR in cross-sectional data are readily available (Enders, 2010; Little, 1988; Rubin, 2018), we were unsuccessful in finding similar methods for ESM data. Moreover, MAR pattern can technically be examined with correlational analyses that enable the identification of factors associated with missingness patterns; however, because of the wealth of temporally dependent data, even this seemingly simple endeavor can quickly prove challenging. Thus, although knowing whether missingness is MCAR, MAR, or missing not at random informs about the implications of the missing data, obtaining such knowledge is not straightforward for ESM data. It may, however, be worthwhile to implement event-related suspension windows and/or open-text boxes that can help to gather information on why responses have been missed.

For the fourth approach—handling missing prompts statistically—researchers need to be aware of two caveats. The first caveat is of a theoretical nature because one needs to decide whether it makes sense to impute situational states (i.e., impute values for missing prompts) with statistical techniques. Among experts, this seems still to be an open debate. According to some ESM researchers, ESM prompts are highly context-specific and fluctuating in nature, which is why they argue that inferring (or estimating) values for missing prompts based on values from earlier or later prompts would be illogical. Other ESM researchers do, however, recommend to use techniques such as missing data imputation, full information maximum likelihood (FIML) estimation, or Kalman filtering (Ji et al., 2018; McNeish & Hamaker, 2020). The second caveat regarding treating missing data is of a statistical nature, because the relevant techniques (e.g., multiple imputation, FIML, or Kalman filtering; Ji et al., 2018; McNeish & Hamaker, 2020), although readily available, are not necessarily straightforward for ESM data.

Starting Points for In-Depth Readings

- Ji, L., Chow, S. M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 715–736. https://doi.org/10.1093/iwcomp/iwaa019
- 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. *Behavioral Research Methods, 46*(1), 41–54. https://doi.org/10.3758/s13428-013-0353-y
- Williams, M., Lewthwaite, H., Fraysse, F., Gajewska, A., Ignatavicius, J., & Ferrar, K. (2021). Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: Systematic review and meta-analysis. *Journal of Medical Internet Research, 23*(3), Article e17023. https://www.jmir.org/2021/3/e17023
- Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment, 30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Data Assessment and Administration

After design decisions for an ESM study have been finalized, decisions regarding data assessment and administration need to be made. Because of the frequency with which participants are assessed, ESM data sets are often much larger than data sets from cross-sectional (panel or cohort) studies. Accordingly, participants have not only an identifying ID but also identifying time stamps that encode the time structure of the data. Given the wealth of rapidly incoming data, it is wise to carefully choose a data-collection software and to plan its implementation early on.

When deciding on an ESM software for the data collection, some initial questions should be considered, such as: What is the budget? Does the software need to be compatible for both computer and mobile screens? Should the study use an internet- or (offline-capable) app-based software? Does the software need to be able to handle complex ESM designs (e.g., with item-order randomization)? Is the data storage in line with the data-protection regulations that apply to the study? Should notifications be sent as text messages, emails, or push notifications (or something alike)? Is a feedback module required? What level of technical support is needed (e.g., around-the-clock support)? Because there is a large and consistently evolving range of software that can be used for the assessment of ESM data, it would go beyond this article to provide an exhaustive discussion of the options. However, the interested reader can find an overview that compares various features of relevant software tools in form of a crowd-sourced table, “ESM & Mobile Sensing Solutions Feature Table”[9] (this effort was initiated and is led by Arslan et al., 2018). Moreover, several articles provide guidance regarding choosing ESM software (e.g., Henry et al., 2024; Trull & Ebner-Priemer, 2020; Weermeijer et al., 2021), and some also provide insights into using wearables for the passive collection of physiological and movement data (see De Calheiros Velozo et al., 2024; Mehl et al., 2023; Vaid & Harari, 2019). Deciding on the ESM software is, however, only the initial challenge regarding data assessment.

The myriad of challenges that ESM researchers face when assessing and administering their data are beyond the scope of what could be covered in this article, but we want to briefly provide a few fundamental suggestions for (novice) ESM researchers. First, if participants change time zones, the assessment software should ideally be able to accommodate those changes, because trying to correct time zones retrospectively can not only be difficult, but can also, in the worst case, bias the data (note that similar problems can occur if the study is conducted in countries that switch winter and summer time on different days). Second, when planning the data assessment, it is helpful to plan for regular compliance checks to be able to intervene on time if necessary (see also the section above on missingness, attrition, and compliance). Third, open-text fields can be a useful addition to the data-assessment form, because these can help during data collection to spot whether participants experience technical issues they might report there. And last but not least, given the wealth of data and the complexity of their structure, it is advisable to carefully plan how variables should be coded, stored, and (ideally) shared in a trusted (public or protective-access) repository.

Starting Points for In-Depth Readings

- Arslan, R. C., Tata, C. S., & Walther, M. P. (2018). *ESM & mobile sensing solutions: Feature table*. https://comparison-to.formr.org
- Weermeijer, J., Kiekens, G., & Wampers, M. (2021). Experience sampling platforms. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 103–117). KU Leuven. https://www.kuleuven.be/samenwerking/real/real-book

Reliability

Many of the topics discussed above determine (part of) the validity of ESM measurements. Reliability is another aspect that affects validity. Reliability concerns the consistency of measurements: If the true value of a construct does not vary and researchers accordingly get (approximately) the same result each time that they replicate their assessments, then they have reliable measurements. If random noise is present in addition to the true value of the construct and researchers accordingly get a (substantially) different result each time that they replicate their assessments, then they have unreliable measurements. Importantly, reliability concerns only random errors because measurements with a high amount of consistent error could still be perfectly reliable (albeit clearly invalid; Lord & Novick, 1968; Revelle, 2023). In the context of psychology, the consensus is that measurements will never be perfectly reliable. If unaccounted for, an imperfect reliability will bias the results of one's analyses: the less reliable, the more bias (e.g., see Buonaccorsi, 2010; Schuurman & Hamaker, 2019). Hence, an important step in validating ESM items (or entire scales) is to evaluate their reliability. Low reliability should be accounted for in the analyses; however, given that statistical corrections are always imperfect, too, it would be best to develop as reliable ESM measures as possible (Yang et al., 2022).

Considering and accounting for reliability in ESM studies comes with notable challenges. Psychology has a strong history of evaluating the reliability of measurements in the context of between-subjects research, but (validation) results from between-subjects research cannot be expected to generalize to within-subjects research (see Hamaker, 2012; Molenaar, 2004; Schuurman, 2023). Moreover, it is so far an unresolved question whether reliability methods for between-subjects research (e.g., Cronbach's alpha coefficients) are suitable in the ESM context without adaptions. One reason why this seems unlikely is that between-subjects reliability methods usually treat within-subjects fluctuations as measurement error. Some reliability methods have, however, specifically been developed for or adapted to the ESM context. For example, reliability methods for multilevel models can be applied to intensive longitudinal data with multiple subjects (e.g., multilevel alpha and omega; Castro-Alvarez et al., 2024; Geldhof et al., 2014; Nezlek, 2017; Revelle & Condon, 2018). One potential downside of these approaches for ESM research is that many of them disregard the dependency (dynamics) among repeated measures. Another downside of many current reliability methods is that they rely on multiple-item measures, whereas ESM research often relies on single-item measures (see former section, Item Content and Phrasing). Recently, two test–retest reliability methods have been introduced for estimating reliabilities for single-item measures (the measurement errors multilevel autoregressive [MEAR] model approach by Schuurman & Hamaker, 2019, and the immediate test–retest method by Dejonckheere et al., 2022). Furthermore, ESM studies are often (to some extent) idiographic, and most of the current methods do not take into account that within-subjects reliability may differ from person to person (Dejonckheere et al., 2022; Lord & Novick, 1968; Schuurman & Hamaker, 2019).[10] Accordingly, person-specific reliability methods (e.g., dynamic extensions of omega that model dependency among the true scores over time, e.g., Castro-Alvarez et al., 2022; H. Song & Ferrer, 2012; or the immediate test–retest method and the MEAR model approach described above, Dejonckheere et al., 2022; Schuurman & Hamaker, 2019) are necessary for ESM studies that consider person-specific results. And finally, reliability may also change over time. Therefore, establishing measurement invariance across both persons and time is another key challenge in evaluating the reliability of ESM items (Adolf et al., 2014; McNeish et al., 2021; Vogelsmeier et al., 2023). Although tackling these challenges is not easy, a worthwhile step forward is to consistently study and report on the behavior of ESM measures (e.g., via the ESM item repository: Kirtley et al., 2024; www.esmitemrepository.com). That way, researchers can learn from each other and progressively develop the required measures and reliability methods for ESM research.

Starting Points for In-Depth Readings

- Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment, 34*(12), 1138–1154. https://doi.org/10.1037/pas0001178
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91. https://doi.org/10.1037/a0032138
- Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91. https://doi.org/10.1037/met0000188

Replicability and Generalizability

The crisis of replicability and generalizability within psychological sciences (Ioannidis, 2012, 2022) extends to ESM research. Replicability refers broadly to finding similar results as an original study, when using the same methods to examine novel data in *the same context*. Generalizability refers to finding similar results as an original study, when using the same methods to examine novel data in *a new context*, such as in a different country, with a different therapy modality, or in a different school (see Moeller, Dietrich, et al., 2023). Because research on the replication crisis has revealed that many psychological findings could not be replicated because of artifacts caused by flawed research methods, much effort has been invested to improve the trustworthiness of psychological science. Nevertheless, many challenges for replicability and generalizability remain, and some of them are unique or particularly aggravated in ESM research compared to cross-sectional, panel, or cohort research. For example, ESM studies have historically often (a) employed small sample sizes (typically because of a relatively higher participant burden in combination with limited research-related resources), (b) used items or measures that have not been validated for ESM research (typically because of the need of momentary and short surveys), and (c) been conducted with novel data-collection tools or data-analyses procedures, resulting in many studies being the first of their kind (Myin-Germeys et al., 2018). Moreover, ESM studies usually set out to examine constructs that are time- and context-specific and are likely to differ between individuals—a level of heterogeneity that cannot be fully solved by methodological innovations (Bryan et al., 2021). And finally, publications based on ESM data often lack important details on the design and conduction of the ESM study that would be necessary to replicate the study, such as participant instructions, incentivization, or data cleaning and preparation (Kaurin et al., 2023). Taken together, a combination of missing study details, design-related shortcomings, the time and context specificity of ESM studies, as well as rapid innovations impede future replication efforts and hamper generalization efforts.

Although the challenge of replicability and generalizability of ESM research should not be underestimated, the good news is that every ESM researcher can contribute to an enhanced replicability in simple but meaningful ways already at the designing (and implementation) stage. First, one can preregister the study protocol (see Kirtley et al., 2021, for an ESM registration template). Various topics that are necessary for such a preregistration have been discussed in this article. Although many topics that are reported in a preregistration are similar for cross-sectional, panel, and cohort data, some topics are specific to ESM studies or at least more pertinent for a registration of an ESM study, including (a) sampling scheme (i.e., fixed, random, or event-related), (b) sampling frequency, (c) survey depth (i.e., number of items and completion time), (d) timescale nature of the items (here specifically, time-variant vs. time-invariant variables), and (e) incentivization and increasing participant compliance (cf. Kirtley et al., 2021). There are two additional topics that were beyond the scope of our article but are relevant for a preregistration and should be considered at the designing stage of the ESM study: (a) the response window, that is, how much time participants get to respond to a prompt before it expires (sometimes also referred to as "time-out specification;" e.g., Deakin et al., 2022), and if applicable, (b) passive monitoring (e.g., wearable data) and its linkage with self-report data (see e.g., Myin-Germeys et al., 2009; Niemeijer et al., 2023). Second, ESM researchers can publish a protocol paper (Kaurin et al., 2023) to facilitate the replicability of the study. Several checklists exist for the reporting of design-related features of ESM studies that can aid both the fine-tuning of design choices and a more complete reporting of details relevant for replication efforts (e.g., Kaurin et al., 2023; Liao et al., 2016). Alternatively, researchers can, of course, when the study is completed, publish an article with a comprehensive and fully replicable methods section (if needed, with additional online resources that can be released on public or protective-access repositories, such as OSF: https://osf.io/). Another valuable way toward improving the replicability and generalizability of ESM studies would be to conduct collaborative data collections across different research labs and contexts (see e.g., Dora, Piccirillo, Foster, Arbeau, et al., 2023)—a consideration that should ideally be made early at the designing stage of the study, to be able to increase the harmonization of procedures across labs and contexts as much as possible. The analyses of such context-diverse data sets could then be performed by many independent analysts (see e.g., Bastiaansen et al., 2020) in a theoretically replicable way, so that sources of variation can be teased apart and can inform on how the field can better ensure the replicability and generalizability of ESM studies in the future. And last, ESM researchers can contribute to the development of novel solutions for the replicability and generalizability of ESM research by joining efforts of relevant initiatives such as the ManyMoments project (https://bit.ly/ManyMoments; Moeller, 2023; Moeller, Bergmann, et al., 2023; Moeller, Dietrich, et al., 2023; Moeller, Langener, et al., 2023).

Starting Points for In-Depth Readings

- Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S.-M., 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. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research, 137*, Article 110211. https://doi.org/10.1016/j.jpsychores.2020.110211
- Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920924686
- Moeller, J., Bergmann, C., Bringmann, L., Bastiaansen, J., Schmiedek, F., Loderer, K., Riediger, M., Pekrun, R., & the ManyMoments Consortium (2023). *ManyMoments—Improving replicability of experience sampling method studies in multi-lab collaborations*. (Manuscript in preparation).
- Moeller, J., Dietrich, J., Neubauer, A. B., Brose, A., Kühnel, J., Dehne, M., Jähne, M. F., Schmiedek, F., Bellhäuser, H., Malmberg, L.-E., Stockinger, K., Riediger, M., & Pekrun, R. (2022). *Generalizability crisis meets heterogeneity revolution: Determining under which boundary conditions findings replicate and generalize*. PsyArXiv. https://doi.org/10.31234/osf.io/5wsna

Discussion and Concluding Remarks

This article was aimed at providing an overview of collective wisdom on designing and implementing ESM studies, current problems therein, and potential solutions. The reader may have noticed that although a lot can already be learned about how to best design an ESM study (for highlights, see Table 2), there are at least equally many design choices for which researchers currently do not have sufficient knowledge. For an overview

Table 2. Relevant Design Considerations for ESM Studies at a Glance

Study type and research questions
• The research question of interest needs to be both suitable for and best addressed with an ESM study design. • Because of frequent and time-sensitive assessments, ESM studies can be more burdensome for participants than cross-sectional, panel, or cohort studies (especially with *N*-of-1 designs).
Item content and phrasing
• Focus groups or cocreation sessions can inform the operationalization of the construct of interest (e.g., choosing content and number of items). • Shorter ESM questionnaires can sometimes result in better compliance and data quality, but that is not always the case and should not come at the expense of validity. • Piloting ESM questionnaires can prevent confusing or illogical wording as well as the inclusion of redundant items, and may therefore reduce participant burden, increase compliance, and improve data quality. • Training participants before the study can increase understanding of the items and may thereby reduce participant burden, increase compliance, and improve data quality. • The time frame for the item (i.e., fixed, e.g., “in the last day”; [semi]random, e.g., “right now” or “since the last beep”; or event-contingent, e.g., “right now”) should be aligned with the research aim.

(continued)

Table 2. *(continued)*

Choosing and formulating response options
• Likert or VAS response formats are often preferred over dichotomous response formats because not all analytic techniques for ESM can deal with binary responses. • VAS response formats do not seem to have a clear advantage to Likert response formats, which have been suggested to be most user-friendly. • Response formats should be consistent throughout a (construct-specific) set of ESM items. • If response formats are not validated, they should ideally be pilot-tested, participants should be briefed about the usage (via text or video) before the study, and response distributions and shifts should retrospectively be explored. • Additional open-text boxes can be useful to record person-specific answers and to identify reasons for shifts in response distributions or missingness.
Timescale (sampling scheme, sampling frequency, survey length, and study duration)
• The sampling scheme needs to match the research aim and should be suitable for the context at hand. • Sampling scheme and sampling frequency need to align with the (assumed) fluctuation speed of the construct of interest (or its interrelations) because this can affect the quality of the results. • The timing of assessments should be representative and should ideally not fully depend on the time of day or contextual factors (e.g., assessing tiredness only in the late evening). • Recall bias can be addressed through increasing the sampling frequency of fixed or (semi)random assessments, via measurement bursts, or by using event-contingent schemes. • A pilot study can inform about burden and compliance to enable a more beneficial choice of sampling frequency weighed against prompt depth (survey length and completion time) and study duration. • The duration of the study needs to balance sufficient assessments of the construct of interest (e.g., regarding power, variability, and cyclicity) with feasibility considerations (e.g., participant burden and finances).
Change properties and stationarity
• It should be carefully considered whether one is interested in measuring change itself (i.e., means, associations, or other statistical properties that change systematically over time, resulting in nonstationary data). o If yes, the study duration should (be long enough to) cover a period during which participants are expected to change. o If no, external influences that may induce change should be reduced (e.g., by prescreening for potential external influences and/or by keeping the study period brief). • It is useful to keep in mind that to date, only a few analytic approaches exist that can deal well with nonstationary ESM data.
Power and effect sizes
• Simulation studies are available that provide recommendations for sample size; however, such guidance is highly context specific and should be used cautiously. • Power should be calculated based on the smallest effect sizes considered meaningful or based on previous reports of standardized effect sizes. • Pragmatic considerations, such as participant burden and financial resources, are crucially important but should not deter researchers from conducting and reporting power analyses. • Results of power analyses should also be reported when the target effects are insignificant. • When a power analysis seems truly implausible, existing simulation studies may provide guidance, and alternative (pragmatic) considerations and their potential implications should be transparently reported.
Missingness, attrition, and compliance
• Strategies to increase compliance can for example include piloting ESM studies to improve the user experience, financial incentives, training participants in the study procedures, allowing for suspension windows, and motivation enhancement conversations with participants during periods of low compliance. • Planned missingness designs allow researchers to collect a subset of the items per prompt while eventually still being able to obtain relevant estimates for all items. • Suspension windows, open-text boxes, and the like can help to determine the type of missing-data patterns and may help to identify factors that are associated with missingness. • If one is content assuming that data of missed prompts can be statistically compensated for (e.g., via imputation, FIML, Kalman filtering), the use of relevant statistical techniques—which although readily available are not always straightforward—should be planned.

(continued)

Table 2. *(continued)*

Data assessment and administration
• The ESM software for data collection should be chosen carefully to ensure that it suits one's study requirements well (e.g., costs, data protection regulations, compatibility with mobile phones, offline capability, prompt notification options, feedback capabilities, around-the-clock support, time-zone changes, or other study-specific design choices). • Variables should be logically encoded, and a plan for storing, downloading, and sharing the temporally dense data should be set out.
Reliability
• Reliability methods from between-subjects research (e.g., Cronbach's alpha coefficients) are not necessarily suited for the ESM context; therefore, reliability methods should be used that are tailored to ESM data (e.g., multilevel alpha and omega). • Special attention to the choice of reliability methods is required when working with single-item measures (e.g., use the MEAR modeling approach or the immediate test–retest method) or person-specific analyses (e.g., use dynamic extensions of omega) because not all ESM-tailored reliability methods can be applied in these contexts.
Replicability and generalizability
To accelerate replicability, generalizability and meta-science, ESM researchers are encouraged to do the following: • report their methods transparently and in an as replicable way as possible (including sufficient detail on the sample, items, timescale, compliance, software, and analytic procedure), for example, via preregistrations, protocol papers, or thorough methods sections; • submit validated items to the ESM Item Repository (www.esmitemrepository.com); • report standardized effect sizes (to aid subsequent power analyses); • share their code and data (when possible) on trusted repositories; • engage in (cross-lab) collaborations for data collection, (meta-)analysis, and the interpretation of findings.

Note: ESM = experience-sampling method; VAS = visual analogue scale; FIML = full information maximum likelihood; MEAR = measurement errors multilevel autoregressive.

of exactly these issues, we refer the reader to the article "A Momentary Assessment on the Future of ESM Research" (Piccirillo et al., 2024), which proposes a research agenda of empirical work that could be conducted to advance the design of ESM studies.

There are several topics that were beyond the scope of our article. First, we did not report on theory development, including important considerations such as the relevance of context or testing meaningful hypotheses (read more on this in Kuppens et al., 2022). Second, we did not discuss challenges related to measurement reactivity, that is, when behavior, thought pattern, or mood change merely because of completing the ESM assessments (Myin-Germeys et al., 2018; Truijens et al., 2023). Third, we addressed neither analytic methods nor analytic caveats of intensive longitudinal data, such as overnight effects (i.e., gaps in the collected data between an evening and the next morning prompt) or cyclical effects (e.g., week vs. weekend, seasonal, periodic, or event-related effects). And fourth, we did not discuss common data-related biases of intensive longitudinal data that can be dealt with statistically, such as zero inflation if the construct of interest (e.g., an emotion or an event) does only rarely occur (Shao et al., 2023) or initial elevation bias when initial data have systematically higher values than successive data (see Kuppens et al., 2022; Shrout et al., 2018).

Although it is impossible to cover the full scope and complexity of designing an optimal ESM study in one article, we provide an overview of 10 crucial design- and implementation-related topics and a living, web-based resources library (see Note 5) with a more extensive list of relevant literature to facilitate further learning. Therefore, we believe that our work contributes to the quality of future ESM research through bringing areas of caution, essential consideration, and helpful readings to our readers' attention.

Appendix A

Material to reproduce Figure 1

Step 1a: search query for the experience-sampling method (ESM) publication records of Web of Science (WoS). "ESM" OR "experience sampling method" OR "EMA" OR "ecological momentary assessment" OR "IL" OR "intensive longitudinal" OR "time-series" OR "timeseries" OR "time series" OR "ambulatory assessment" OR "real time data capture" OR real-time data capture" OR "daily life methods" OR daily-life methods" (Title) AND Article OR Book OR Book Chapter OR Data Paper OR Discussion OR Early Access OR Editorial Material OR Letter OR Meeting Summary OR Proceedings Paper OR Review OR Software Review (Document Type) AND English OR Dutch

OR German (Language) AND Psychology (Web of Science Categories)

Step 1b (alternative for 1a): search link for the ESM publication records of WoS. https://www.webofscience.com/wos/woscc/summary/3cd90184-f89c-4cfc-8c56-d589a7bded1c-ab980ba7/relevance/1

Step 2: follow these steps on WoS to retrieve the data set for ESM publication records. (click) Analyze results → (select) Publication years → (select and insert) Showing the "maximum" number out of "all" entries → (click) Select all → (choose) All data rows → (click) Download table

Step 3a: search query for all psychological publication records of WoS. Article OR Book OR Book Chapter OR Data Paper OR Discussion OR Early Access OR Editorial Material OR Letter OR Meeting Summary OR Proceedings Paper OR Review OR Software Review (Document Type) AND English OR Dutch OR German (Language) AND Psychology (Web of Science Categories)

Step 3b (alternative for 3a): search link for all psychological publication records of WoS. https://www.webofscience.com/wos/woscc/summary/8d13299a-0d53-43d2-8c69-657cc4acf6aa-ab9903e2/relevance/1

Step 4: follow these steps on WoS to retrieve the data set for all psychological publication records. (click) Analyze results → (select) Publication years → (select and insert) Showing the "maximum" number out of "all" entries → (click) Select all → (choose) All data rows → (click) Download table

Step 5: code for the reproduction of the figures, to be used in R studio. Note that you need to insert file directories and file names where indicated:

```
*****************************************************************************

#loading libraries
library(dplyr)
library(tidyr)
library(ggplot2)

#load in ESM publication records data, and convert the data
   into usable format
ESMyears <- read.delim("specify the file directroy & name",
   sep = "")
ESMyears$PubNumbers <- ESMyears$Years
ESMyears$PubYears <- ESMyears$Publication
ESMyears_ordered <- ESMyears[order(ESMyears$PubYears),]
ESMyears_ordered$PubYears <- as.numeric(ESMyears_
   ordered$PubYears)
ESMyears_ordered$type <- "ESM"

#plot the ESM publication records per year
ESM_plot <- ggplot(
  data = ESMyears_ordered,
  mapping = aes(x = PubYears, y = PubNumbers)
) +
  geom_col(
    size = 0.5,
    fill = "grey"
  ) +
  labs(
    x = "\npublication year",
    y = "number of ESM publications\n"
  ) +
  scale_x_continuous(
    breaks = seq(from = 1960, to = 2020, by = 5)
  ) +
  scale_y_continuous(
    expand = expansion(mult = 0.025)
  ) +
  theme_minimal() +
  theme(
    panel.grid.minor = element_blank(),
    panel.grid.major = element_blank(),
    axis.title = element_text(size = 14, family = "serif"),
    axis.text.x = element_text(size = 12, angle = 45, hjust =
       1, family = "serif"),
    axis.text.y = element_text(size = 12, family = "serif"),
    axis.line = element_line(colour = "black", size = 0.5),
    axis.ticks = element_line(colour = "black", size = 0.5)
  )
ESM_plot
ggsave(plot = ESM_plot,
    filename = "choose a file directory & name for the plot",
    device = cairo_pdf,
    width = 8, height = 4, units = "in")
ggsave(plot = ESM_plot,
    filename = "choose a file directory & name for the plot",
    device = png, type = "cairo", dpi = 300,
    width = 8, height = 4, units = "in")

#load in all psychology publication records data, and convert
   the data into usable format
ESMyears_all <- read.delim("specify the file directroy &
   name ", sep = "")
ESMyears_all$PubNumbers <- ESMyears_all$Years
ESMyears_all$PubYears <- ESMyears_all$Publication
ESMyears_all_ordered <- ESMyears_all[order(ESMyears_
   all$PubYears),]
ESMyears_all_ordered$PubYears <- as.numeric(ESMyears_
   all_ordered$PubYears)
ESMyears_all_ordered <- ESMyears_all_
   ordered[(ESMyears_all_ordered$PubYears)>=1958,] # this
   line of codes removes all data before 1958, to cover the
   same time frame as available for the ESM data
ESMyears_all_ordered$type <- "ALL"
```

```
#bind the two datasets together
ESM_years_data <- bind_rows(
   ESMyears_all_ordered %>% select(PubYears,
    PubNumbers, type),
   ESMyears_ordered %>% select(PubYears, PubNumbers,
    type)
)
#plot the number of ESM publication records (black) in
    comparison to all psychological publication records (grey),
    per year
comparison_plot <- ggplot(
   data = ESM_years_data %>%
      mutate_at("type", factor, levels = c("ESM", "ALL"),
          ordered = TRUE),
   mapping = aes(x = PubYears, y = PubNumbers, fill = type)
) +
   geom_col(
      size = 0.5
   ) +
   labs(
      x = "\npublication year",
      y = "number of publications\n"
   ) +
   scale_x_continuous(
      breaks = seq(from = 1960, to = 2020, by = 5)
   ) +
   scale_y_continuous(
      expand = expansion(mult = 0.025)
   ) +
   scale_fill_manual(
      values = c("ALL" = "lightgrey", "ESM" = "black")
   ) +
   theme_minimal() +
   theme(
      panel.grid.minor = element_blank(),
      panel.grid.major = element_blank(),
      axis.title = element_text(size = 14, family = "serif"),
      axis.text.x = element_text(size = 12, angle = 45, hjust =
          1, family = "serif"),
      axis.text.y = element_text(size = 12, family = "serif"),
      axis.line = element_line(colour = "black", size = 0.5),
      axis.ticks = element_line(colour = "black", size = 0.5)
   )
comparison_plot
ggsave(plot = comparison_plot,
      filename = "choose file directory & name",
      device = cairo_pdf,
      width = 8, height = 4, units = "in")
ggsave(plot = comparison_plot,
      filename = "choose file directory & name",
      device = png, type = "cairo", dpi = 300,
      width = 8, height = 4, units = "in")

#convert the merged datasets into a wide format
ESM_years_data_wide <- pivot_wider(ESM_years_data,
    names_from = "type", values_from = "PubNumbers",
    id_cols = "PubYears")

#plot the relative number of ESM publication records, i.e.,
    ESM publication records divided by all psychological
    publication record
ESM_years_data_wide$PubNumbers_ratio <- ESM_years_
    data_wide$ESM / ESM_years_data_wide$ALL
ratio_plot <- ggplot(
   data = ESM_years_data_wide,
   mapping = aes(x = PubYears, y = PubNumbers_ratio)
) +
   geom_col(
      size = 0.5,
      fill = "grey"
   ) +
   labs(
      x = "\npublication year",
      y = "relative number of ESM publications"
   ) +
   scale_x_continuous(
      breaks = seq(from = 1960, to = 2020, by = 5)
   ) +
   scale_y_continuous(
      expand = expansion(mult = 0.025)
   ) +
   theme_minimal() +
   theme(
      panel.grid.minor = element_blank(),
      panel.grid.major = element_blank(),
      axis.title = element_text(size = 14, family = "serif"),
      axis.text.x = element_text(size = 12, angle = 45, hjust =
    1, family = "serif"),
      axis.text.y = element_text(size = 12, family = "serif"),
      axis.line = element_line(colour = "black", size = 0.5),
      axis.ticks = element_line(colour = "black", size = 0.5)
   )
ratio_plot
ggsave(plot = ratio_plot,
      filename = "choose file directory & name",
      device = cairo_pdf,
      width = 8, height = 4, units = "in")
ggsave(plot = ratio_plot,
      filename = "choose file directory & name",
      device = png, type = "cairo", dpi = 300,
      width = 8, height = 4, units = "in")

#calculate sum of all ESM publications since 1958
372/sum(ESMyears_ordered$PubNumbers) # = 0.2931442

******************************************************************
```

Appendix B

Figures showing the absolute numbers of experience-sampling method (ESM) publications and of all psychological publications

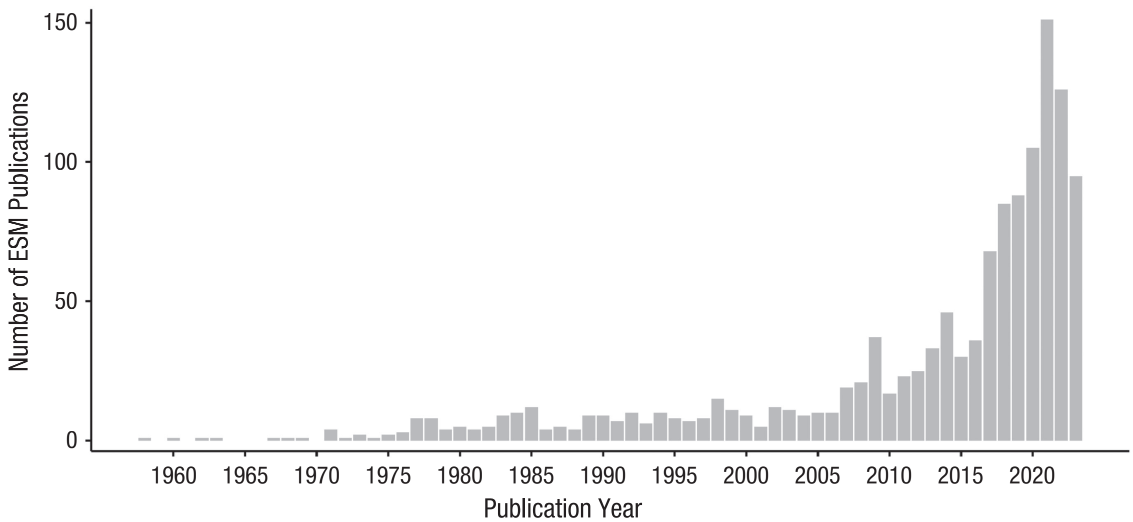

Fig. B1. The absolute number of experience-sampling-method (ESM) publication records on Web of Science (WoS) per year since 1958.

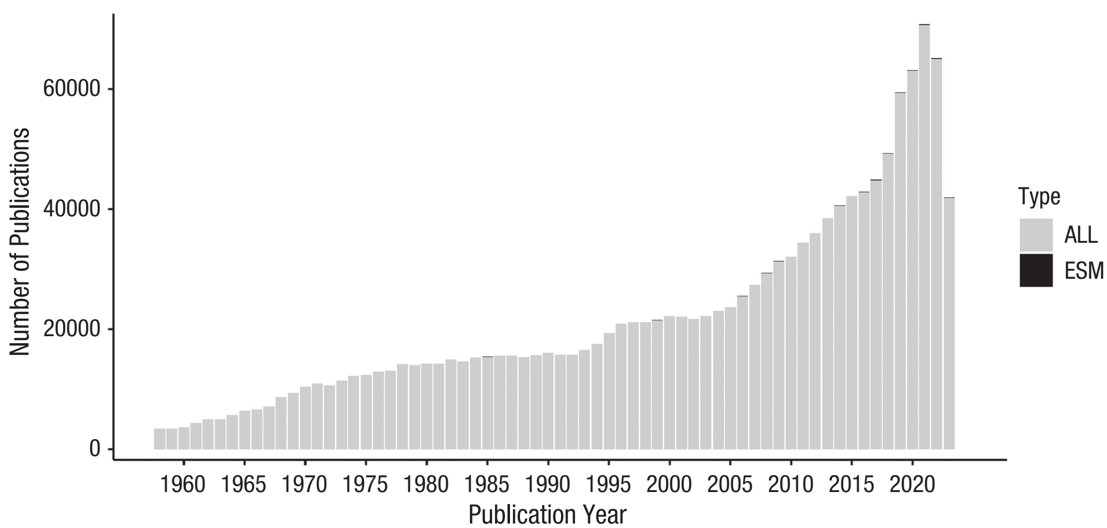

Fig. B2. The absolute number of both experience-sampling method (ESM) and all psychological publication records on Web of Science (WoS) per year since 1958. Black = ESM publication records; gray = all psychological publication records.

Transparency

Action Editor: Pamela Davis-Kean
Editor: David A. Sbarra
Author Contributions

J. Fritz and M. L. Piccirillo share first authorship.

Jessica Fritz: Conceptualization; Investigation; Writing – original draft; Writing – review & editing.

Marilyn L. Piccirillo: Conceptualization; Investigation; Writing – original draft; Writing – review & editing.

Zachary D. Cohen: Investigation; Writing – review & editing.

Madelyn Frumkin: Investigation; Writing – review & editing.

Olivia Kirtley: Investigation; Writing – review & editing.

Julia Moeller: Investigation; Writing – review & editing.
Andreas B. Neubauer: Investigation; Writing – review & editing.
Lesley A. Norris: Investigation; Writing – review & editing.
Noémi K. Schuurman: Investigation; Writing – review & editing.
Evelien Snippe: Investigation; Writing – review & editing.
Laura F. Bringmann: Conceptualization; Investigation; Writing – original draft; Writing – review & editing.

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

J. Fritz was supported by a Wolfson College Junior Research Fellowship. M. L. Piccirillo was supported by the National Institute of Alcohol Abuse and Alcoholism (T32AA00455, Larimer; K99AA029459, M. L. Piccirillo). M. Frumkin was supported by a fellowship award from the National Institute of Mental Health (F31 MH124291, M. Frumkin). O. Kirtley was supported by a Senior Postdoctoral Fellowship from Research Foundation Flanders (FWO 1257821N). J. Moeller was supported by a Jacobs Foundation Early Career Research Fellowship and a grant by the German Research Foundation (451682742). A. Neubauer was supported by a grant by the German Research Foundation (NE 2480/1-1). L. Norris was supported by a fellowship award from the National Institute of Mental Health (F31 MH123038, L. Norris). E. Snippe was supported by a ZonMw Off Road grant (Project 451001029). L. F. Bringmann was supported by the Netherlands Organization for Scientific Research (Veni Grant; NWOVeni191G.037).

Open Practices

This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID IDs

Jessica Fritz https://orcid.org/0000-0002-6342-0892
Marilyn L. Piccirillo https://orcid.org/0000-0002-4616-5180
Zachary D. Cohen https://orcid.org/0000-0002-4883-1028
Olivia Kirtley https://orcid.org/0000-0001-5879-4120
Andreas B. Neubauer https://orcid.org/0000-0003-0515-1126
Laura F. Bringmann https://orcid.org/0000-0002-8091-9935

Acknowledgments

We thank Michael Mullarkey, who was a member of the panel discussion from which the article resulted. The submitted manuscript was submitted to the OSF preprint server (https://osf.io/fverx/).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459241267912

Notes

1. Because the terms "panel study" and (single or accelerated) "cohort study" are inconsistently used and defined in the literature, we note that we specifically refer to longitudinal designs that assess a group of participants at several (mostly two to 10) time points (as opposed to assessing different groups of people at several time points, such as in repeated cross-sectional designs; Lugtig & Smith, 2019).
2. In this article, we focus predominantly on subjective self-report assessments, but many of our recommendations will also apply to objective biophysiological (e.g., heart rate) or mobile-sensing (e.g., GPS tracking) data. For recommendations that are specific to passive sensing, we refer the reader elsewhere (e.g., Harari et al., 2017; Jacobson et al., 2020; Mehl et al., 2023; Niemeijer et al., 2023).
3. The public panel discussion, which we organized in response to the growing attention to ESM research, included seven speakers with long-standing experience in ESM research. The corresponding slides and videos are provided on our OSF page: https://osf.io/4a723/. The panel discussion was split in two events.
4. https://docs.google.com/document/d/1fAYQqIc4bUpO-FqjSvoGABEQfeZfMjib2I5UM9xgNRc/edit.
5. Note that under specific circumstances, within-subjects effects can also be based on assessments from just one time point. One way to do this is through assessing within-subjects focused questions (Eid et al., 1999). For example, instead of asking participants about their current level of anxiety, one can ask about their current level of anxiety compared with their usual level of anxiety, thereby directly assessing within-subjects effects. For further examples for estimating within-subjects effects based on data from just one time point, see Schuurman (2023).
6. Moreover, it is possible to apply hybrid within- and between-subjects approaches that first examine separate idiographic models for each person (i.e., within-subjects) and then compare the idiographic results across persons (i.e., between-persons) to identify patterns that generalize across people versus those that do not (e.g., see the group iterative multiple model estimation approach by Gates et al., 2023).
7. Note that there are several other statistical properties (e.g., the rank order) that can also change systematically, but this goes beyond the scope of what can be covered in this article.
8. For completeness, we want to stress that power can usually also be increased through reducing the complexity of the chosen analytic method or the number of variables included therein.
9. https://docs.google.com/spreadsheets/d/18R9x9Qbl9tADJGpJBJID_T9EWZeQ_4W3OFdn3iKRU7U/edit#gid=204277638.
10. Some do take into account that within-subjects reliability differs from person to person, but assume that error variances are equal from subject to subject and that only the true scores vary (e.g., Neubauer, Voelkle, et al., 2020).

References

Adolf, J., Schuurman, N. K., Borkenau, P., Borsboom, D., & Dolan, C. V. (2014). Measurement invariance within and between individuals: A distinct problem in testing the equivalence of intra- and inter-individual model structures. *Frontiers in Psychology*, *5*, Article 883. https://doi.org/10.3389/fpsyg.2014.00883

Allison, P. D. (2001). *Missing data*. Sage.

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

Arslan, R. C., Tata, C. S., & Walther, M. P. (2018). *ESM & mobile sensing solutions: Feature table*. https://comparison-to.formr.org

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F. J., Boker, S. M., Ceulemans, E., Chen, M., Chow, S. M., 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. (2020). Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, Article 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*, 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. The Guilford Press.

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). The Guilford Press.

Boring, S., Ledo, D., Chen, X. A., Marquardt, N., Tang, A., & Greenberg, S. (2012). The fat thumb: Using the thumb's contact size for single-handed mobile interaction. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 39–48). https://doi.org/10.1145/2371574.2371582

Bos, F. M., Von Klipstein, L., Emerencia, A. C., Veermans, E., Verhage, T., Snippe, E., Doornbos, B., Hadders-Prins, G., Wichers, M., & Riese, H. (2022). PErsonalized Treatment by Real-time Assessment (PETRA): User-centered development of a web-application for personalized diaries in psychiatric care. *JMIR Mental Health*, *9*(8), Article e36430. https://doi.org/10.2196/36430

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy*, *149*, Article 104011. https://doi.org/10.1016/j.brat.2021.104011

Bringmann, L. F., Date, C. V. D. V., Wichers, M., Riese, H., & Stulp, G. (2021). ESMvis: A tool for visualizing individual experience sampling method (ESM) data. *Quality of Life Research*, *30*, 3179–3188. https://doi.org/10.1007/s11136-020-02701-4

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*, *22*, 409–425. https://doi.org/10.1037/met0000085

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using raw VAR regression coefficients to build networks can be misleading. *Multivariate Behavioral Research*, *51*, 330–344. https://doi.org/10.1080/00273171.2016.1150151

Buonaccorsi, J. P. (2010). *Measurement error*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420066586

Burke, L. E., Shiffman, S., Music, E., Styn, M. A., Kriska, A., Smailagic, A., Siewiorek, D., Ewing, L. J., Chasens, E., French, B., Mancino, J., Mendez, D., Strollo, P., & Rathbun, S. L. (2017). Ecological momentary assessment in behavioral research: Addressing technological and human participant challenges. *Journal of Medical Internet Research*, *19*(3), Article e77. https://doi.org/10.2196/jmir.7138

Castro-Alvarez, S., Bringmann, L. F., Back, J., & Liu, S. (2024). *The many reliabilities of psychological dynamics: An overview of statistical approaches to estimate the internal consistency reliability of intensive longitudinal data*. PsyArXiv. https://doi.org/10.31234/osf.io/qyk2r

Castro-Alvarez, S., Tendeiro, J. N., de Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022). Mixed-effects trait-state-occasion model: Studying the psychometric properties and the person–situation interactions of psychological dynamics. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 438–451. https://doi.org/10.1080/10705511.2021.1961587

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. https://doi.org/10.1146/annurev.psych.57.102904.190146

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*, S85–S96. https://doi.org/10.1016/s0376-8716(02)00217-x

Conner, T. S., & Lehman, B. (2012). Getting started: Launching a study in daily life. In M. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 89–107). The Guilford Press.

Conner, T. S., & Mehl, M. R. (Eds.). (2012). *Handbook of research methods for studying daily life*. The Guilford Press.

Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, *24*(2), 227–245. https://doi.org/10.1177/0894439305281503

Dawood, S., Hallquist, M. N., Pincus, A. L., Ram, N., Newman, M. G., Wilson, S. J., & Levy, K. N. (2020). Comparing signal-contingent and event-contingent experience sampling ratings of affect in a sample of psychotherapy outpatients. *Journal of Psychopathology and Behavioral Assessment*, *42*(1), 13–24. https://doi.org/10.1007/s10862-019-09766-7

De Calheiros Velozo, J., Habets, J., George, S. V., Niemeijer, K., Minaeva, O., Hagemann, N., Herff, C., Kuppens, P., Rintala, A., Vaessen, T., Riese, H., & Delespaul, P. (2024). Designing daily-life research combining experience sampling method with parallel data. *Psychological Medicine*, *54*(1), 98–107. https://doi.org/10.1017/S0033291722002367 de Vries, L. P., Baselmans, B. M. L., & Bartels, M. (2021). Smartphone-based ecological momentary assessment of well-being: A systematic review and recommendations for future studies. *Journal of Happiness Studies*, *22*(5), 2361–2408. https://doi.org/10.1007/s10902-020-00324-7

Deakin, E., Ng, F., Young, E., Thorpe, N., Newby, C., Coupland, C., Craven, M., & Slade, M. (2022). Design decisions and data completeness for experience sampling methods used in psychosis: Systematic review. *BMC Psychiatry*, *22*, Article 669. https://doi.org/10.1186/s12888-022-04319-x

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment*, *34*(12), 1138–1154. https://doi.org/10.1037/pas0001178

Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 71–89). KU Leuven.

Dora, J., Piccirillo, M. L., Foster, K. T., Arbeau, K., Armeli, S., Auriacombe, M., Bartholow, B., Beltz, A. M., Blumenstock, S. M., Bold, K., Bonar, E. E., Braitman, A., Carpenter, R. W., Creswell, K. G., De Hart, T., Dvorak, R. D., Emery, N., Enkema, M., Fairbairn, C., . . . King, K. M. (2023). The daily association between affect and alcohol use: A meta-analysis of individual participant data. *Psychological Bulletin*, *149*(1–2), 1–24.

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological Methods*, *20*(4), 489–505. https://doi.org/10.1037/met0000041

Eid, M., Schneider, C., & Schwenkmezger, P. (1999). Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *European Journal of Personality*, *13*, 283–306.

Eisele, G., Hiekkaranta, A. P., Kunkels, Y. K., aan het Rot, M., van Ballegooijen, W., Bartels, S. L., Bastiaansen, J. A., Beymer, P. N., Bylsma, L., Carpenter, R., Ellison, W. D., Fisher, A. J., Forkmann, T., Frumkin, M., Fulford, D., Naragon-Gainey, K., Greene, T., Heininga, V. E., Jones, A., . . . Kirtley, O. J. (2024, June 6). *ESM-Q: A consensus-based quality assessment tool for experience sampling method items*. PsyArXiv. https://doi.org/10.31234/osf.io/sjynv

Eisele, G., Kasanova, Z., & Houben, M. (2021). Questionnaire design and evaluation. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology* (pp. 71–90). Center for Research on Experience Sampling and Ambulatory Methods.

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Fisher, A. (2023, October 4). *The promise and possibility of discrete data for emotion-related research* [Keynote presentation]. Emotion 2023: Tilburg University, Tilburg, The Netherlands.

Fritz, J., Stochl, J., Kievit, R. A., van Harmelen, A.-L., & Wilkinson, P. O. (2021). Tracking stress, mental health, and resilience factors in medical students before, during, and after a stress-inducing exam period: Protocol and proof-of-principle analyses for the RESIST cohort study. *JMIR Formative Research*, *5*(6), Article e20128. https://doi.org/10.2196/20128

Gates, K. M., Chow, S.-M., & Molenaar, P. C. (2023). *Intensive longitudinal analysis of human processes*. Chapman & Hall.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*(1), 80–92. https://doi.org/10.2307/1130293

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Hamaker, E. L. (2023). The curious case of the cross-sectional correlation. *Multivariate Behavioral Research*, 1–12. https://doi.org/10.1080/00273171.2022.2155930

Hamaker, E. L., & Wichers, M. (2017). No time like the present. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, *18*, 83–90. https://doi.org/10.1016/j.cobeha.2017.07.018

Haslbeck, J. M. B., & Ryan, O. (2022). Recovering within-person dynamics from psychological time series. *Multivariate Behavioral Research*, *57*(5), 735–766. https://doi.org/10.1080/00273171.2021.1896353

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion*, *23*(8), 2117–2141. https://doi.org/10.1037/emo0001218

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2021). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*, *54*(4), 1541–1558. https://doi.org/10.3758/s13428-021-01683-6

Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method: Measuring the quality of everyday life*. Sage.

Helmich, M. A., Olthof, M., Oldehinkey, A. J., Wicheres, M., Bringmann, L. F., & Smit, A. C. (2021). Early warning signals and critical transitions in psychopathology: Challenges and recommendations. *Current Opinion in Psychology*, *41*, 51–58. https://doi.org/10.1016/j.copsyc.2021.02.008

Henry, L., Hansen, E., Chimoff, J., Pokstis, K., Kiderman, M., Naim R, K. J. B. M., Lopez-Guzman, S., Kircanski, K., Pine, D., & Brotman, M. (2024). Selecting an ecological momentary assessment platform: Tutorial for researchers. *Journal of Medical Internet Research*, *26*, Article e51125. https://doi.org/10.2196/51125

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

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality*, *34*(6), 1037–1059. https://doi.org/10.1002/per.2266

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

Ioannidis, J. P. A. (2022). Why most published research findings are false. *PLOS Medicine*, *19*(8), Article e1004085. https://doi.org/10.1371/journal.pmed.1004085

Jacobson, N. C., Bentley, K. H., Walton, A., Wang, S. B., Fortgang, R. G., Millner, A. J., Coombs, G., Rodman, A. M., & Coppersmith, D. D. L. (2020). Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bulletin of the World Health Organization*, *98*(4), 270–276. https://doi.org/10.2471/BLT.19.237107

Jacobson, N. C., Chow, S., & Newman, M. G. (2019). The differential time-varying effect model (DTVEM): A tool for diagnosing optimal measurement and modeling intervals in intensive longitudinal data. *Behavior Research Methods*, *51*, 295–315. https://doi.org/10.3758/s13428-018-1101-0

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086–1105.

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

Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling*, *25*(5), 715–736. https://doi.org/10.1080/10705511.2017.1417046

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609–619. https://doi.org/10.1111/ADD.14503

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

Kirtley, O. J., Eisele, G., Kunkels, Y. K., Hiekkaranta, A. P., Van Heck, L., Pihlajamäki, M., Kunc, B., Schoefs, S., Kemme, N. D. F., Biesemans, T., & Myin-Germeys, I. (2024, July 2). *The experience sampling method (ESM) item repository*. PsyArXiv. https://doi.org/10.17605/OSF.IO/KG376

Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920924686

Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion*, *12*(2), 256–267. https://doi.org/10.1037/a0024756

Kuppens, P., Dejonckheere, E., Kalokerinos, E. K., & Koval, P. (2022). Some recommendations on the use of daily life methods in affective science. *Affective Science*, *3*(2), 505–515. https://doi.org/10.1007/s42761-022-00101-0

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practice in Psychological Science*, *4*. https://doi.org/10.1177/2515245920978738

Lafit, G., Sels, L., Adolf, J. K., Loeys, T., & Ceulemans, E. (2022). PowerLAPIM: An application to conduct power analysis for linear and quadratic longitudinal actor–partner interdependence models in intensive longitudinal dyadic designs. *Journal of Social and Personal Relationships*, *39*(10), 3085–3115. https://doi.org/10.1177/02654075221080128

Lane, S., & Gates, K. (2017). Automated selection of robust individual-level structural equation models for time series data. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 768–782. https://doi.org/10.1080/10705511.2017.1309978

Le, H. V., Mayer, S., Bader, P., & Henze, N. (2018). Fingers' range and comfortable area for one-handed smartphone interaction beyond the touchscreen. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association of Computing Machinery. https://doi.org/10.1145/3173574.3173605

Leertouwer, Ij., Cramer, A. O. J., Vermunt, J. K., & Schuurman, N. K. (2021). A review of explicit and implicit assumptions when providing personalized feedback based on EMA data. *Frontiers in Psychology*, *12*, Article 764526. https://doi.org/10.3389/fpsyg.2021.764526

Liao, Y., Skelton, K., Dunton, G., & Bruening, M. (2016). A systematic review of methods and procedures used in ecological momentary assessments of diet and physical activity research in youth: An adapted STROBE checklist for reporting EMA studies (CREMAS). *Journal of Medical Internet Research*, *18*(6), Article e151. https://doi.org/10.2196/jmir.4954

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lugtig, P., & Smith, P. A. (2019). *The choice between a panel and cohort study design*. https://www.researchgate.net/publication/336944625_The_choice_between_a_panel_and_cohort_study_design

Mansueto, A. C., Wiers, R. W., van Weert, J. C. M., Schouten, B. C., & Epskamp, S. (2023). Investigating the feasibility of idiographic network models. *Psychological Methods*, *28*(5), 1052–1068. https://doi.org/10.1037/met0000466

Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5421–5432). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858063

Mayer, S., Le, H. V., & Henze, N. (2018). Designing finger orientation input for mobile touchscreens. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3229434.3229444

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling*, *28*(5), 807–822. https://doi.org/10.1080/10705511.2021.1915788

Medland, H., De France, K., Hollenstein, T., Mussoff, D., & Koval, P. (2020). Regulating emotion systems in everyday life. *European Journal of Psychological Assessment*, *36*(3), 437–446. https://doi.org/10.1027/1015-5759/a000595

Mehl, M. R., Eid, M., Wrzus, C., Harari, G. M., & Ebner-Priemer, U. (Eds.). (2023). *Mobile sensing in psychology: Methods and applications*. The Guilford Press.

Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Current Opinion in Psychology*, *41*, 1–8. https://doi.org/10.1016/j.copsyc.2021.01.004

Moeller, J. (2023). *The ManyMoments project*. Retrieved July 9, 2023, from https://www.erzwiss.uni-leipzig.de/en/institut-fuer-bildungswissenschaften/professuren/educational-psychology-with-focus-on-development-under-conditions-of-risk/research/manymoments

Moeller, J., Bergmann, C., Bringmann, L. F., Bastiaansen, J., Schmiedek, F., Loderer, K., Riediger, M., & Pekrun, R., & Many Moments Consortium. (2023). *ManyMoments - Improving replicability of experience sampling method studies in multi-lab collaborations* [Manuscript in preparation].

Moeller, J., Dietrich, J., Neubauer, A. B., Brose, A., Kühnel, J., Dehne, M., Jähne, M., Schmiedek, F., Bellhäuser, H., Malmberg, L.-E., Stockinger, K., Riediger, M., & Pekrun, R. (2023). *Generalizability crisis meets heterogeneity revolution: Determining under which boundary conditions findings replicate and generalize*. PsyArXiv. https://doi.org/10.31234/osf.io/5wsna

Moeller, J., Langener, A., Lafit, G., Karhulahti, V., Bastiaansen, J. A., & Bergmann, C. (2023). *The hypository: Registering hypotheses for cumulative science*. PsyArXiv. https://doi.org/10.31234/osf.io/5qgj7

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Müssig, M., Kubiak, J., & Egloff, B. (2022). The agony of choice: Acceptance, efficiency, and psychometric properties of questionnaires with different numbers of response options. *Assessment*, *29*(8), 1700–1713. https://doi.org/10.1177/10731911211029379

Muthén, B., & Muthén, L. (2017). *Mplus, version 8.0*.

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*, 123–132. https://doi.org/10.1002/wps.20513

Myin-Germeys, I., & Kuppens, P. (2021). *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies*. Center for Research on Experience Sampling and Ambulatory Methods Leuven.

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*, 1533–1547. https://doi.org/10.1017/S0033291708004947

Neubauer, A. B., Scott, S. B., Sliwinski, M. J., & Smyth, J. M. (2020). How was your day? Convergence of aggregated momentary and retrospective end-of-day affect ratings across the adult life span. *Journal of Personality and Social Psychology*, *119*(1), 185–203. https://doi.org/10.1037/pspp0000248

Neubauer, A. B., Voelkle, M. C., Voss, A., & Mertens, U. K. (2020). Estimating reliability of within-person couplings in a multilevel framework. *Journal of Personality Assessment*, *102*(1), 10–21. https://doi.org/10.1080/00223891.2018.1521418

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, *69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020

Niemeijer, K., Mestdagh, M., Verdonck, S., Meers, K., & Kuppens, P. (2023). Combining experience sampling and mobile sensing for digital phenotyping with m-Path sense: Performance study. *JMIR Formative Research*, *7*, Article e43296. https://doi.org/10.2196/43296

Palmier-Claus, J. E., Haddock, G., & Varese, F. (2019). Why the experience sampling method? In J. E. Palmier-Claus, G. Haddock, & F. Varese (Eds.), *Experience sampling in mental health research* (pp. 1–7). Routledge. https://doi.org/10.4324/9781315398341-1

Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P., Lewis, S. W., & Dunn, G. (2011). Experience sampling research in individuals with mental illness: Reflections and guidance. *Acta Psychiatrica Scandinavica*, *123*(1), 12–20. https://doi.org/10.1111/j.1600-0447.2010.01596.x

Piccirillo, M. L., Beck, E. D., & Rodebaugh, T. L. (2019). A clinician's primer for idiographic research: Considerations and recommendations. *Behavior Therapy*, *50*(5), 938–951. https://doi.org/10.1016/j.beth.2019.02.002

Piccirillo, M. L., Fritz, J., Cohen, Z. D., Frumkin, M. R., Kirtley, O. J., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., & Bringmann, L. F. (2024, March 12). *A momentary assessment on the future of ESM research*. PsyArXiv. https://doi.org/10.31234/osf.io/82bnf

Ram, N., Conroy, D. E., Pincus, A. L., Lorek, A., Rebar, A., Roche, M. J., Coccia, M., Morack, J., Feldman, J., & Gerstorf, D. (2014). Examining the interplay of processes across multiple time-scales: Illustration with the intraindividual study of affect, health, and interpersonal behavior (iSAHIB). *Research in Human Development*, *11*(2), 142–160. https://doi.org/10.1080/15427609.2014.906739

Reiter, T., & Schoedel, R. (2024, July 24). Never miss a beep – Using mobile sensing to investigate (non-)compliance in experience sampling studies. *Big Data & Research Syntheses*, *56*, 4038–4060. https://doi.org/10.3758/s13428-023-02252-9

Revelle, W. (2023). Classical test theory and the measurement of reliability. In W. Revelle (Ed.), *An introduction to psychometric theory with applications in R* (pp. 205–239). Springer.

Revelle, W., & Condon, D. (2018). *Reliability from alpha to omega: A tutorial.* PsyArXiv. https://doi.org/10.31234/osf.io/2y3w9

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184.supp

Rintala, A., Apers, S., Eisele, G., & Verhoeven, D. (2021). Briefing and debriefing in an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 119–134). Center for Research on Experience Sampling and Ambulatory Methods.

Rubin, D. B. (2018). Multiple imputation. In S. van Buuren (Ed.), *Flexible imputation of missing data* (2nd ed., pp. 29–62). Chapman and Hall/CRC.

Schlechter, P., Fritz, J., Cassels, M., Neufeld, S. A. S., & Wilkinson, P. O. (2021). The Youth and Childhood Adversity Scale: A step towards developing a new measure of adversity and its severity. *European Journal of Psychotraumatology*, *12*(1), Article 1981573. https://doi.org/10.1080/20008198.2021.1981573

Schreuder, M. J., Groen, R. N., Wigman, J. T. W., Wichers, M., Hartman, C. A., & Schreuder, J. (2022). Participation and compliance in a 6-month daily diary study among individuals at risk for mental health problems. *Psychological Assessment*, *35*(2), 115–126. https://doi.org/10.1037/pas0001197.supp

Schultzberg, M., & Muthén, B. (2017). 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. https://doi.org/10.1080/10705511.2017.1392862

Schuurman, N. K. (2023). *A "within/between problem" primer: About (not) separating within-person variance and between-person variance in psychology*. PsyArxiv. https://doi.org/10.31234/osf.io/7zgkx

Schuurman, N. K., Ferrer, E., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods*, *21*, 206–221. https://doi.org/10.1037/met0000062.supp

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, *24*(1), 70–91. https://doi.org/10.1037/met0000188

Schwartz, C. E., Sprangers, M. A., Carey, A., & Reed, G. (2004). Exploring response shift in longitudinal data. *Psychology & Health*, *19*(1), 51–69.

Shao, S., Xu, Z., Liu, Q., McClure, K., Jacobucci, R., Maxwell, S. E., & Zhang, J. (2023). *Zero inflation in intensive longitudinal data: Why is it important and how should we deal with it?* PsyArXiv. https://doi.org/10.31234/osf.io/8fscd

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences, USA*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

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

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Song, H., & Ferrer, E. (2012). Bayesian estimation of random coefficient dynamic factor models. *Multivariate Behavioral Research*, *47*, 26–60. https://doi.org/10.1080/00273171.2012.640593

Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, *30*(5), 1662–1671. https://doi.org/10.1177/10731911221113563

Soyster, P. D., & Fisher, A. J. (2019). Involving stakeholders in the design of ecological momentary assessment research: An example from smoking cessation. *PLOS ONE*, *14*(5), Article e0217150. https://doi.org/10.1371/journal.pone.0217150

Stadel, M., van Duijn, M. A. J., Wright, A. G. C., Bringmann, L. F., & Elmer, T. (2024). Considering the 'With whom': Differences between Event- and Signal-Contingent ESM data of person-specific social interactions. *Multivariate Behavioral Research*, *59*(4), 841–858. https://doi.org/10.1080/00273171.2024.2335405

Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, *24*(3), 236–243. https://doi.org/10.1207/S15324796ABM2403_09

Studer, R. (2012). Does it matter how happiness is measured? Evidence from a randomized controlled experiment. *Journal of Economic and Social Measurement*, *37*, 317–336. https://doi.org/10.3233/JEM-120364

Truijens, F. L., De Smet, M. M., Vandevoorde, M., Desmet, M., & Meganck, R. (2023). What is it like to be the object of research? On meaning making in self-report measurement and validity of data in psychotherapy research. *Methods in Psychology*, *8*, Article 100118. https://doi.org/10.1016/j.metip.2023.100118

Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: Introduction to the special section. *Psychological Assessment*, *21*, 457–462. https://doi.org/10.1037/a0017653

Trull, T. J., & Ebner-Priemer, U. W. (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. https://doi.org/10.1037/abn0000473

Vaid, S. S., & Harari, G. M. (2019). Smartphones in personal informatics: A framework for self-tracking research with mobile sensing. In H. Baumeister & C. Montag (Eds.), *Digital phenotyping and mobile sensing: New developments in psychoinformatics* (pp. 65–92). Springer.

van Berkel, N., Goncalves, J., Hosio, S., Sarsenbayeva, Z., Velloso, E., & Kostakos, V. (2020). Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, *134*, 1–12. https://doi.org/10.1016/j.ijhcs.2019.10.003

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

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, *49*, 193–213. https://doi.org/10.1080/00273171.2014.889593

Vogelsmeier, L. V. D. E., Vermunt, J. K., Bülow, A., & De Roover, K. (2023). Evaluating covariate effects on ESM measurement model changes with Latent Markov factor analysis: A three-step approach. *Multivariate Behavioral Research*, *58*(2), 262–291. https://doi.org/10.1080/00273171.2021.1967715

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment*, *6*(3), 231–243.

Weermeijer, J., Kiekens, G., & Wampers, M. (2021). Experience sampling platforms. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 103–117). KU Leuven.

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

Wichers, M., & Groot, P. C. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, *85*, 114–116. https://doi.org/10.1159/000441458

Williams, M. T., Lewthwaite, H., Fraysse, F., Gajewska, A., Ignatavicius, J., & Ferrar, K. (2021). Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *23*(3), Article e17023. https://doi.org/10.2196/17023

Wright, A. G. C., & Woods, W. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology*, *16*, 49–74. https://doi.org/10.1146/annurev-clinpsy102419-125032

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/10.31234/osf.io/6qc5x

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment*, *30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Yang, L.-Q., Wang, W., Huang, P.-H., & Nguyen, A. (2022). Optimizing measurement reliability in within-person research: Guidelines for research design and R shiny web application tools. *Journal of Business and Psychology*, *37*(6), 1141–1156. https://doi.org/10.1007/s10869-022-09803-5

Zuidersma, M., Riese, H., Snippe, E., Booij, S. H., Wichers, M., & Bos, E. H. (2020). Single-subject research in psychiatry: Facts and fictions. *Frontiers in Psychiatry*, *11*, Article 539777. https://doi.org/10.3389/fpsyt.2020.539777