Inez Myin-Germeys &
Peter Kuppens (EDS)

The Open Handbook of Experience Sampling Methodology

A Step-by-Step Guide to Designing, Conducting and Analysing ESM Studies

THIRD EDITION

Leuven University Press

The Open Handbook of Experience Sampling Methodology
A Step-by-Step Guide to Designing, Conducting and Analysing ESM Studies

The Open Handbook of Experience Sampling Methodology

A Step-by-Step Guide to Designing, Conducting and Analysing ESM Studies

THIRD EDITION

Edited by
Inez Myin-Germeys and Peter Kuppens

Leuven University Press

Published with the support of the KU Leuven Fund for Fair Open Access, and the Open Book Collective (see www.lup.be/obc)

Published in 2026 by Leuven University Press / Presses Universitaires de Louvain / Universitaire Pers Leuven. Minderbroedersstraat 4, B-3000 Leuven (Belgium).

ISBN 978 94 6270 491 6 (Paperback)
eISBN 978 94 6166 692 5 (ePDF)
eISBN 978 94 6166 693 2 (ePUB)
https://doi.org/10.11116/9789461666932
D/2026/1869/17
NUR: 770

Typesetting: Crius Group
Cover design: Andre Klijsen

Table of Contents

CHAPTER 1

Experience Sampling Methods: An Introduction

Inez Myin-Germeys and Peter Kuppens

Taking a single snapshot is usually not the best approach to understanding the whole movie.
Yet this is what we most commonly do in mental health research and practice.

Indeed, if we want to capture what people do, want, feel, experience and encounter in their normal daily life, we need to capture the movie of their normal daily behaviour and experience. Experience Sampling Methods (ESM) have been developed to track experiences in the real world and in real time, using self-reports to capture these momentary experiences as well as their context. An exponentially growing body of research is applying ESM in a diversity of fields, including behavioural science, psychology, psychiatry and psychosomatic medicine. A search for 'experience sampling' or 'ecological momentary' in the Web of Science Core Collection shows an exponential increase in articles referring to such assessment techniques over the past decade. While ESM was originally based on paper-and-pencil approaches, where a programmable watch or an alternative signalling system alerted participants that it was time to fill out the paper diary, most studies to date use digital devices like smartphones to assess the structured self-report diary. With these rapid technological developments of the last two decades, ESM has become accessible to a much wider group of researchers, and, even more importantly, it now also has clear potential for clinical implementation.

These developments, while positive and promising, have also put pressure on the ESM research community. Where are we with respect to methodological developments? When one takes a closer look at the enormous number of ESM studies that have been carried out, one thing that stands out is the enormous heterogeneity among studies. Studies differ in the number of beeps, number of days, number of questions asked, content of the questionnaires, incentives given, sampling schemes or feedback

provided. Whereas this heterogeneity may underscore the unique nature of ESM, with researchers using this technique for studying a variety of experiences in a variety of contexts, it may also point towards one of the main weaknesses of ESM research to date: There are very few guidelines substantiated by evidence of how to properly conduct an ESM study. A study by Janssens and colleagues (2018) indeed showed that most ESM researchers to date have no clear justification for the methodological choices they make for their ESM study. Whereas this is definitely problematic for research, reducing, for example, the possibility of replication, it is even more problematic when we start to develop clinical applications based on these methods.

The current book aims to help overcome this problem by providing a thorough and careful description of all the decisions one needs to make when designing an ESM study as well as the consequences of making such specific choices, thus providing an overview of the current 'state of the art'. The ESM design strongly depends on the phenomenon of interest, the expected timeframe of its occurrence and the research question. Therefore, by definition, ESM research comes in different forms. However, it is important to make informed decisions taking the consequences of specific design choices into account. It is also important to use similar research designs for similar questions to foster replicability.

1.1 What are Experience Sampling Methods?

ESM refers to structured self-report diary techniques assessing mood, symptoms, context and appraisals thereof as they occur in daily life. One crucial aspect is that participants provide data in the real world. In contrast to an experimental approach, where one zooms in on one specific aspect of experience or behaviour in a very controlled environment, real-life research focuses on the complexity of the experience in an ever-changing and uncontrollable environment. Another important difference is that experimental research typically induces change to investigate the subsequent effect on the phenomenon of interest. Real-life research usually does not modulate the real-world environment but rather observes naturally occurring changes to investigate their effects on the variable of interest. A third defining characteristic of ESM is that it assesses individuals in real time. As the focus is on the natural flow of

experiences as they occur in real life, the goal is to assess these experiences as closely in time as possible to their actual occurrence. Therefore, ESM typically includes more assessment points per day, although some authors include once-a-day assessments such as sleep diaries recorded upon awakening or end-of-day assessments.

Box 1.1. Main characteristics of ESM

Real world and real time
Prospective measurement
Self-report
Structured diary
Appraisal of context

With ESM, participants are assessed prospectively in their normal daily life with questions about their actual mental state, thus reducing retrospective biases. Retrospective recall has been defined as an active reconstruction process which is subject to cognitive biases (Stone et al., 2004), notably when subjects recall previous affective states (Levine & Safer, 2002) and in clinical samples (Safer & Keuler, 2002). More precisely, evaluations based on retrospective recall have been linked to the overestimation of positive or negative affect (Ben-Zeev et al., 2009; Shiffman et al., 1997) and disproportionate emphasis on the individual's current state and most intense experience over the assessed period, i.e., peak-and-end effect in memory retrieval (Fredrickson, 2000; Fredrickson & Kahneman, 1993). Retrospective reports also commonly require an individual to aggregate, for example, their well-being or feelings over a larger period of time, covering different contexts and situations. However, depending on the sampling scheme (number of assessments per day) and the topic of interest, ESM questionnaires could still include retrospective questions (see Chapter 4), utilized to capture as much information as possible. However, these questions would usually cover a range of a few hours (for example the time since the last beep), which is still quite different from the timescales used in general questionnaires, which include timeframes of weeks to months or even a lifetime. Chapter 4 will focus on choices in questionnaire development and the potential consequences of such choices.

The focus of ESM studies is the subjective experience of the individual. The individual is considered the privileged observer and uses self-reports to provide information about his or her mental state, mood, symptoms or context. This approach contrasts an observational one wherein one would use, for example, video analysis to observe behaviour in context (Luff & Heath, 2012). While a real-world observational approach may be more intrusive, inducing higher reactivity to the method, it also provides a different kind of information. It lacks information on the inner experience, which is exactly the focus of ESM research.

In contrast to an open diary where people write freely about their experiences when they want to, ESM is a structured diary method, usually including a limited number of open questions. It consists of a structured questionnaire assessing specific experiences, typically inquiring about experiences at the very moment of recording (see Chapter 4 for questionnaire development). ESM also requires participants to fill out the questionnaire at specific moments in time. These assessments can either be time-contingent, with sampling moments scheduled randomly or at fixed time points, or event-contingent, with sampling taking place at moments when specific events are happening (for further description, see Chapter 3). The use of different sampling schemes depends on the research question and aids in reducing reactivity to the method. Usually, ESM researchers want to capture daily life processes without altering them (unless, of course, the ESM method is used in clinical practice where altering processes is usually the aim; see Chapter 12). The former objective is not easy when a researcher asks people to self-report several times a day over several days. Chapter 3 and Chapter 4 will discuss how questionnaire development and design choices may significantly contribute to lowering reactivity to the method. Of course, asking people to self-report about very personal experiences in their day-to-day life also brings about ethical questions. What are the repercussions of such self-reporting? Can everybody do it, and what questions are appropriate to ask in that context? Chapter 5 will discuss the ethical aspects of ESM research.

Finally, ESM allows the inclusion of subjective appraisals of the context. Objective information about the context is useful but does not necessarily reflect all relevant aspects. The presence of a snake would be highly stressful to most people but not to the snake catcher who is excited by seeing a unique specimen. Appraisals of the context matter and ESM allows including that as well.

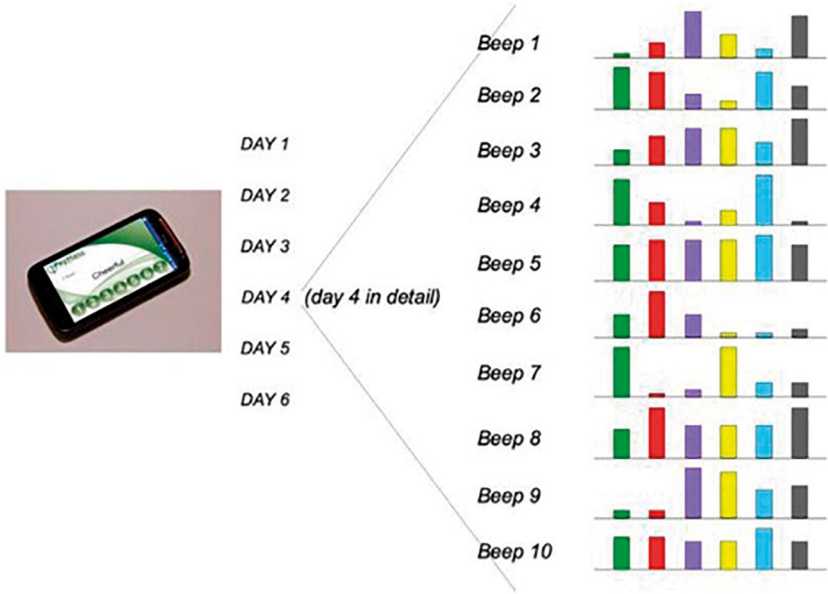


Fig 1.1 A typical ESM set-up. A typical ESM study, using a dedicated app, capturing several variables over several moments in a day during several days.

Over the years, several names have been used for ESM. ESM was first introduced by Mihyli Csikszentmihalyi and Reed Larson in their seminal late-1970s work on adolescent development (Csikszentmihalyi & Larson, 1984). Arthur Stone, another pioneer in the field, launched the term Ecological Momentary Assessment (EMA) at the beginning of the 1990s (Shiffman et al., 2008; Stone & Shiffman, 1994). Despite some people carefully describing the overlap and differences between the two, they refer to basically the same methodology. In the current book, we refer to ESM. Next to active forms of real-world monitoring, passive remote monitoring, including sensors and wearable information devices, have been developed. These measures are usually captured under the name 'ambulatory assessment', referring to both active and passive forms of real-world and real-time measurements. The Society of Ambulatory Assessment assembles experts working with all of these measurements (https://ambulatory-assessment.org).

1.2 The scientific roots of ESM

1.2.1 Ecological psychology

In the 1970s, the science of psychology was mainly focused on experimental laboratory studies, the idea being that one would get the best understanding of a specific phenomenon if one could isolate it and study it in a perfectly controlled laboratory environment. Ecological psychology, in contrast, argued that behaviour and experiences are radically situated, meaning that they can only be understood in relation to their context (Lobo et al., 2018). According to this argument, for experiences and behaviour to be fully understood, these need to be investigated under real-world circumstances outside the laboratory. Different strands of research have been developed under the umbrella of ecological psychology, the most famous being that of Gibson, who focused on the richness of environmental perceptual stimuli and the interaction between the perceiver and the world to develop a theory of direct perception (Gibson, 2015). Another famous ecological psychologist, Barker, was interested in thoroughly describing the attributes of the environment and how these attributes affected the behaviour of the occupants. He described these 'behavioural settings' through careful observation of people in their normal environment (Barker, 1975). He concluded that 'Based on these observations, the behavior of a child could be predicted more accurately from knowing the situation the child was in than from knowing the child's individual characteristics' (Barker, 1975).

ESM is rooted in the tradition of ecological psychology, with its focus on assessments in normal daily life. However, it specifically investigates the experiences and mental states of individuals and how these come about and interact with contextual factors. Mihaly Csikszentmihalyi and Reed Larson developed the method by beeping people in the real world to fill out a questionnaire, as they were interested in what teenagers think and how they feel as they live their normal lives, spending time with their friends, interacting with their parents or spending time at school (Csikszentmihalyi & Larson, 1984). Around the same time, Hurlburt developed thought sampling, a structured way of capturing streams of thought in normal daily life (Hurlburt, 1993). The focus on inner mental states, including thoughts, feelings and experiences, pushed these researchers beyond mere observation of the environment. As thoughts and feelings

such as those of loneliness or having low self-esteem are not necessarily externally observable, the ESM researchers turned towards the privileged observer – the individual who is having these experiences. Self-report, including responses to questions related both to the inner experiences and the context in which they occur, became the standard in ESM research.

So, ESM is rooted in ecological psychology with a focus on experiences as they occur in the real-world context. Researchers also claim that it is an instrument with high ecological validity. Ecological validity can be divided into *representativeness* and *generalizability* (Hermans et al., 2019). Representativeness refers to the similarity in content and experience of an experimental task with the 'real world'. ESM is high in representativeness as it measures experiences in the real world. Generalizability, on the other hand, refers to how well an experimental task predicts associated behaviour or functioning in real life. A generalizable task is not necessarily 'real world' like. For ESM, the question of generalizability pertains more to the generalizability of momentary behaviour or experiences than to overall functioning and behaviour. Do we capture the right moments, do we focus on the right experiences, and how much is this telling us about the overall picture? For example, do momentary questions of mood tell us something about overall well-being, or do snapshots of activity provide an accurate picture of functioning?

1.2.2 Quantified Self

In 2007, Gary Wolf and Kevin Kelly founded the 'Quantified Self' Movement (https://quantifiedself.com). The idea of the quantified self is that one can gain self-knowledge by self-tracking whatever variable they deem important using technology (Wolf & De Groot, 2020). With apps, sensors and wearable devices, a diversity of measures can be tracked which may be relevant including heart rate, skin conductance, breathing patterns, food consumed, number of social contacts, sleep quality, number of steps, amount of time in sedentary behavior, calories burned, kilometres travelled, goals achieved but also mental state, including mood or mental health symptoms. The goal of the Quantified Self Movement was initially to investigate what kind of self-tracking tools were available, what they could measure and how data from such measures could be analysed in ways meaningful to the individual. Since then, the Quantified Self Movement has gained enormous momentum; quantified self labs

and the Quantified Self Institute were established. Their mission has been expanded to improve quality of life by generating and sharing knowledge on the Quantified Self (Wolf & De Groot, 2020).

Although the objective of ESM seems to align with that of the Quantified Self Movement, its focus is different. First, in ESM studies, the focus is on subjective experience. Therefore, mental state, rather than health in general measured with whatever technology is available, is central to all ESM studies. Self-report is therefore indispensable in ESM. Second, the goal of ESM is to create a contextualized understanding of psychological processes and behaviour. The latter aspect is much less emphasized in the Quantified Self Movement.

Still, continuous monitoring would be an advantageous and important addition to ESM as currently, people are only assessed a limited number of times (e.g., 10 times per day). The use of wearable technology using various sensors to capture aspects of behaviour, bodily experiences and context provides a way to continuously capture relevant characteristics of the real-time and real-world interactions. Information from sensors embedded in smartphones or wearable devices can not only offer extensive information on bodily and behavioural features, tracking activity, heart rate or breathing but may also capture external context features such as geolocation, light or temperature (Arean et al., 2016; Torous et al., 2016) as well as relevant social interactions (Arean et al., 2016). Mobile or behavioural sensing, which would be the common denominator of these approaches (Mohr et al., 2020), could thus complement ESM to create a meaningful understanding of relevant person-environment interactions. It would allow the capture of additional contextual information, and it could also guide the triggering of ESM to make it more contingent on certain contextual aspects. Within the framework of this book, we will keep the focus on wearable and sensor tools as an extension or support of ESM (see Chapter 13) rather than discussing in detail how mobile sensing alone could be used in mental health research.

1.2.3 Embedded and embodied cognition and contextual science

The use of ESM to investigate experiences within and in interaction with real-world contexts is also consistent with the more recent emphasis on embodiment and embeddedness in the cognitive sciences (de Bruin et al., 2018). These 4E (embodied, embedded, enactive, extended) approaches

claim that an organism's body and the environment in which it is embedded play a fundamental role in how that organism perceives, feels, thinks and acts. According to these 4E approaches, experiences, including psychiatric symptoms, are dynamic interactive processes that occur when an individual with a certain body or brain actively engages with an ever-changing environment in particular ways (de Haan, 2020). To understand these experiences, one needs to capture that dynamic interaction with all its relevant aspects. Following this approach, the foremost research questions then become: What are the characteristics of these specific interactions? When, where and how do these experiences occur? What *are* the relevant features of these interactions, including the bodily, physiological state, thoughts and beliefs and relevant contextual factors? How can we map these patterns of dynamic interaction? Although ESM does not capture each of these aspects (e.g., bodily posture or physiological states), it does seem ideally suited to answer questions regarding thoughts, beliefs, experiences and context, making it an excellent tool for investigating mental states within a 4E-framework (Myin-Germeys et al., 2018). With ESM providing multiple assessments per person capturing different contexts, it allows temporal associations to unravel and identifies contingencies between context and experiences over time, which is also very much in line with the theoretical framework of contextual science (Zettle, 2016).

1.2.4 Within- and between-person differences and personalized approaches

As ESM assesses individuals repeatedly over time, it is an excellent tool to differentiate between-person variation from within-person variation. Between-person variation reflects how individuals differ from one another – e.g., how individuals with a depressive disorder tend to have higher negative affect and lower positive affect than healthy controls (between-person differences). Within-person variation, on the other hand, reflects how experiences within one individual can differ depending on time or context – e.g., how negative affect is higher when an individual is alone compared to when being with others (within-person differences). Although most research to date focuses on between-person differences, it has become increasingly clear that within-person variation is important as well and that the mechanisms involved in between-person differences

are not necessarily the same as those involved in within-person differences (e.g., see Fisher et al., 2018; Molenaar, 2004). Chapter 2 will further discuss the different research questions, relating to both within- and between-person variability, that can be answered with ESM. As ESM by definition includes multiple data points per individual, ESM data are hierarchical data sets (multiple assessments nested within a day nested within individuals), meaning that measurements are not independent. This has clear implications for the statistical approach, which will be extensively discussed in Chapters 8 through 11.

By examining within-person variation and how this differs from one person to the next, ESM puts the focus of research decidedly on the individual. It puts the individual at the heart of the inquiry, collecting multiple self-reports in real time in the daily life of individuals. It thus allows outlining the very specific patterns of behaviour of one individual. ESM, therefore, is also useful for single-case studies – for example, for investigating the effects of medication reduction on depression (Wichers et al., 2016) or psychosis (Bak et al., 2016). ESM data have also been studied as a series of single cases, investigating, for example, how changes in unrest serve as a warning sign for relapse in depression (Smit et al., 2019). This of course is interesting from a research perspective, and it also provides enormous opportunities for clinical application: ESM could be used to provide personalized feedback, thus enabling patients to get a better insight into their current mental states and the behaviour patterns that impact their symptoms and functioning. These personalized data could be shared with clinicians, helping them to identify personalized targets for treatment and facilitating true shared decision-making (Myin-Germeys, 2020). Furthermore, they could help identify moments when treatment is most needed, opening the way for ecological, momentary interventions and bringing the therapy out of the office and into the real life of people (Myin-Germeys et al., 2016) (see also Chapter 12). ESM could thus contribute to real needs- and patient-led personalized psychiatry.

1.3 Conclusion

ESM is an intriguing and interesting research method that has been around for more than forty years. Despite its exponential growth in many research fields, including clinical and differential psychology and mental health, it is still characterized by a lack of methodological rigor and expertise, leading to a high degree of heterogeneity and lack of replicability. This book aims to help overcome this problem by outlining the current state of the art related to all the relevant decisions that someone needs to make when designing and conducting an ESM study and analysing ESM data. It therefore aims to serve both the beginning and more advanced ESM researcher, offering more practical advice (e.g., on briefing, in Chapter 7, and on relevant digital platforms, in Chapter 6) as well as in-depth theoretical discussions of several research decisions. The ultimate goal of the book is not to direct everyone to a single, unique ESM approach but rather to provide much clearer insight into the choices at stake in ESM research and the consequences of these choices. In this way, we hope to harmonize and optimize ESM research in the years to come.

BEFORE: RESEARCH QUESTIONS AND DESIGNING AN ESM STUDY

CHAPTER 2

Research Questions That Can Be Answered with ESM Research

Peter Kuppens and Inez Myin-Germeys

Human beings are inquisitive and curious by nature (Kidd & Hayden, 2015; Loewenstein, 1994). In other words, we continuously ask ourselves questions about the world. In positive sciences, questions about the world are answered through experiments or with data. While most of this volume is concerned with how to collect data in the most reliable, valid and ethical way possible while using the experience sampling method (see Chapters 3, 4 and 5), we should not forget that the data are there in the first place to answer our questions about the world. In this chapter, we provide a structured overview of the type of research questions that can be answered using ESM data. While the overview is meant to be generic, meaning that it can be applied to any domain of study, we will give concrete examples of research to illustrate how the variety of research questions are being answered in practice.

This chapter is structured as follows. In the first part, we discuss research questions that can be asked (and answered!) through a classic observational ESM study, in which the aim is simply to measure participants throughout their normal daily activities. We review research questions that can be asked regarding one single variable measured with ESM. Next, we discuss research questions that can be asked regarding the relation between multiple variables measured with ESM. Then, we review questions that combine one or both previous kinds of questions with questions soliciting additional information at the level of the person. Finally, in the second part, we review research questions that can be asked when one goes beyond the classic paradigm and intends not only to measure daily life but also to impact daily life and examine the consequences, yielding experimental and intervention paradigms.

2.1 Research questions in observational ESM research

Most of classic ESM research can be considered observational. In observational ESM research, one is interested in observing the natural ebb and flow of phenomena in the context of daily life. These phenomena typically involve how people feel, think or behave but can range from involving lower order drives such as sex (see Impett et al., 2008) and eating (see Reichenberger et al., 2018) to involving higher faculties of the mind such as morality (see Hofmann et al., 2014), aesthetic experience (see Nusbaum et al., 2014) or just letting the mind wander (see Kane et al., 2007). The phenomena can be common and mundane (see Chin et al., 2017) or more unusual (Hillbrand & Waite, 1994) and can be studied in relation to harmless (Dickens et al., 2018) or profound contextual events (Monk et al., 2006).

Regardless of their nature, common to these phenomena is that they unfold over time and take place in connection to (within, between or around) individuals. In other words, they show both within-person and between-person variation. Following this common characterization, we introduce a systematic set of research questions that can be asked when observing phenomena in daily life: (1) questions related to the behaviour of one time-varying variable under study, (2) questions related to relations between several time-varying variables under study (these variables can come from one individual or multiple individuals as is the case in dyadic research; see Chapter 14), and (3) questions about the role of personal characteristics in time-varying variables.

In what follows, we will elaborate on and give concrete examples of these kinds of research questions. Of particular note is that this set of research questions follows the same structure as Chapter 9, regarding statistical modelling of ESM data, in the sense that it engages (1) models with one time-varying variable, (2) models with more than one time-varying variable, and (3) models with person-level variables. In other words, the reviewed research questions can be readily mapped onto the statistical models laid out in Chapter 9.

2.1.1 Research questions related to the behaviour of one time-varying variable

We start with a simplest example in which a researcher has or is planning the collection of data related to a certain phenomenon concerning how

people feel, think or behave in the context of daily life. Several basic (but important) questions can be asked related such a phenomenon:

1. How do people feel, think or behave on average?

Of course, with ESM data, one is often interested in the changes in and dynamics of phenomena over time (why otherwise do ESM research in the first place, right?). However, this does not imply that people's average feelings, thoughts or behaviour are not important to consider or study. To the contrary! One important reason that such elements are indeed important is that several core concepts in the behavioural sciences are meant to reflect typical, habitual, 'stable' forms of feeling, thinking and behaving, and average levels of variables in ESM data allow researchers to track these concepts in the context of daily life. Personality is a prime example of such a concept, and indeed, prominent theories of personality describe behaviour as a distribution of which the average or median is the most typical instantiation (e.g., Fleeson & Law, 2015). Relatedly, average levels are often assessed in standard, cross-sectional trait questionnaires on people's feelings, thoughts and behaviours. An important type of research question, therefore, examines the validity of these questionnaires by relating them to average levels in people's feelings, thoughts and behaviour, obtained in daily life (for an example of such research in the domain of emotion regulation, see Koval et al., 2023). Finally, average levels of how people feel, think and behave often hold substantial explanatory or predictive power for important outcomes and thus should not be discarded or overlooked (this is in no small part because these outcomes are often themselves intended to reflect average levels). For instance, in recent work, we showed that the personality trait of neuroticism, which is often also called emotional instability, does not so much reflect variable or unstable emotions but rather, primarily, the tendency to experience higher levels of negative emotions (Kalokerinos et al., 2020). (Examining average levels of a time-varying variable is done with random intercept models in multilevel modelling; see Chapter 9.)

2. How much do people differ from themselves compared to how much people differ from each other?

This may sound like a bit of a strange question at first, but it is actually a crucial question in ESM research. The idea of assessing people's feelings,

thoughts and behaviour (or any other phenomenon) multiple times during daily lives starts from the assumption that people's feelings, thoughts and behaviours vary across time. As previously mentioned, this is called within-person variability and should be sizeable if ESM research is to be meaningful (there is not much sense in measuring the colour of one's eyes six times a day). On the other hand, people also differ from one another. This is called between-person variability, and is studied in the psychology of individual differences.

The relative proportion of the two expresses to what extent a certain phenomenon varies within persons or between persons. An often-used fraction here is the intra-class correlation coefficient, which is calculated as the amount of between-person variance divided by the sum of the amount of (within + between-person variance) and thus reflects the relative proportion of between-person variance over the total variance. If the resulting figure is relatively large (e.g., higher than 0.5), this means that individual differences in the phenomenon under study are greater than within-person differences. If the resulting figure is relatively low (e.g., lower than 0.5), it indicates that people vary more within themselves than they differ from each other. In other words, the lower this value, the more a variable or phenomenon of interest varies over time within individuals and the more meaningful it is to examine these variations (using, for instance, ESM).

Even in domains that study purportedly stable characteristics of the mind, this proportion can be surprising. For instance, personality is considered the stable set of characteristics that determines who one is as a person and lies at the base of individual differences in how people behave, think and feel. However, work by Fleeson and colleagues (e.g., Fleeson & Law, 2015) has shown that when repeatedly assessing personality-related behaviour in daily life, the variance observed within persons is as large or even larger than the variance observed between persons, and this former variance is even comparable to that observed within persons in measures of subjectively experienced affect, which is considered to be highly context-dependent and fluctuating in nature.

3. What do moment-to-moment fluctuations look like?

Once you've established that you're indeed studying something that shows within-person variability, the next sensible question is what this

variability looks like. Identifying and trying to understand the patterns and regularities that characterize the fluctuations of your phenomenon will surely contribute greatly to understanding its nature and underlying mechanisms (for an example of this exercise in the domain of emotion or affect, see Kuppens & Verduyn, 2017).

The types of research questions that can be asked here are endless, but we will give a set of guiding examples. First, one may want to know whether the fluctuations follow a certain pattern across the day, week, year or over time more generally. For instance, Park and colleagues (2019) showed how music preferences change reliably over the course of a day (with less intense music being preferred in the mornings versus evenings). Stone and colleagues (2012) confirmed our inner suspicion that we feel better on weekend days compared to weekdays (but the good news is that once you're retired, this difference will disappear!). One may also ask whether the fluctuations are a function of progressing time itself. This function can be linear, as for instance with the personality traits of agreeableness and conscientiousness, which show a more or less steady increase in the general population throughout adulthood (Roberts et al., 2006) or non-linear, as in, for instance, the alleged U-shaped relation between happiness and age (Frijters & Beatton, 2012). A particular time-bound pattern for repeated assessment data that ESM researchers should be aware of is the initial elevation bias (Shrout et al., 2018). This bias refers to the frequently (but not always; see Arslan, et al., 2021) observed tendency for repeated subjective reports of any variable to show higher values for the first set of assessments followed by a slight decrease that then stabilizes. While the exact reason for this bias remains unknown, it is nonetheless important for researchers to be aware of the possibility of its occurrence (or any other systematic trends) in their data. This is one of the reasons that researchers sometimes shy away from using the first few assessments in the analyses of intensive longitudinal data.

Second, beyond being predicted by day, week or the progression of time itself, one may ask whether a particular phenomenon is predicted by itself over time. This kind of predictability is known as the self-predictability, time-dependency, or autocorrelation of time-varying variables. A very simple example is the weather: If you want to know what kind of weather there will be tomorrow, the weather of today is usually a pretty good predictor. In other words, the weather is self-predictable or autocorrelated

over time. The same holds for many psychological phenomena: emotion or affect, for instance (Koval & Kuppens, 2024; Kuppens, Allen, et al., 2010; Suls et al., 1998). The extent to which something is self-predictive says a lot about its nature: If something is highly autocorrelated, it means that it is generated by a system that is not very much impacted by outside influences but rather is running its own course. If, on the other hand, something is not very highly autocorrelated, it means that it is very much under the control of outside influences. If something is negatively auto-correlated, it means it follows an oscillating pattern at the frequency of measurement (this is also why autocorrelations are used to detect cyclical patterns in time-series data).

This issue of time dependency also comes into play in the question of whether the order of the fluctuations matters for understanding the within-person variability. When we say a person is unstable, we are saying that they show large variability in their feelings, thoughts or behaviour. Yet a person who feels good the first part of the week and bad the second part of the week shows the same variability and a very different pattern of fluctuations compared to a person who constantly alternates between feeling good and bad over the week. For this reason, researchers have compared indices of variability (which do not take ordering into account) with indices of instability (which look at the size of consecutive changes and therefore take the order into account) and examined which type of variability is most indicative – for instance, adaptive versus maladaptive functioning (for a more detailed discussion of this, see Jahng et al., 2008). In this realm, Thewissen and colleagues (2008) examined the importance of this distinction with respect to self-esteem and paranoia. Koval and colleagues (2015) examined this question in relation to emotions and depression.

2.1.2 Research questions related to the behaviour of multiple time-varying variables

Of course, often one is interested in observing not just one single variable but multiple variables and how they relate to or even predict one another. After all, any theory that tries to explain something has at least two elements: that which needs to be explained (the explanans) and that what explains it (the explanandum). Finding regularities in the relations between variables is the prerequisite to explaining them, so examining

relations between variables is what much research is about (especially in correlational research like most ESM research). A meaningful distinction in this respect is whether one is looking at concurrent or time-lagged relationships (see Chapter 9) between variables.

In concurrent relationships, a researcher examines how fluctuations in one variable coincide with fluctuations in another variable. The most typical examples are from research that relates variables observed in the same individual. For instance, van der Steen and colleagues (2017) examine the concurrent within-person associations between momentary stress and affective and psychotic symptoms in individuals with varying risk for psychosis. Hermans and colleagues (2020) study the relationship between the pleasantness of an activity and positive affect. When one has data from channels other than ESM, one can also examine relations between experiential, physiological and behavioural systems. As an example, Vaessen and colleagues (2018) described a curvilinear relationship between self-reported activity-related stress and cortisol fluctuations. While most research relates variables from the same individual, this does not need to be the case. When one has data from multiple individuals assessed at the same time points, one can study the relation between a variable in one person and a variable in another person over time. This is, for instance, done in research on dyads, which is discussed in Chapter 14. As an example, Sels, Cabrieto, et al. (2020) studied to what extent the emotions romantic partners experience synchronize in daily life (and found surprisingly little evidence for the existence of emotional interdependence).

In lagged relationships, one is interested in examining how one variable at one point in time is predictive of another variable at the next point in time. For instance, Thewissen and colleagues (2011) examined to what extent levels of self-esteem or anxiety were able to predict the onset of a subsequent episode of paranoia. They were interested in whether low levels of self-esteem or high levels of anxiety indicated the possibility that a paranoid episode could occur, and they therefore tested lagged relations between the former two and the latter (for these analyses, the data are lagged such that previous time values of one variable can be used to predict current time values of another variable). Another example can be found in Houben and colleagues (2017), where positive and negative mood are used to predict subsequent engagement in non-suicidal self-injury in individuals diagnosed with borderline personality disorder. A special but

often-studied type of lagged relation tracks whether the level of one variable is associated with an increase or decrease in another variable. This is often done in research on reactivity to stressors or context variables or in research on how emotions, symptoms, etc., change as a function of certain types of behaviour or of regulation efforts. For instance, Brans and colleagues (2013) examine to what extent positive or negative emotions increase or decrease in association with the use of a set of emotion regulation strategies. In the associated analyses, emotionality at one time point is predicted by emotion regulation in between that and the previous time-point, controlling for emotion at the previous time-point (note that this type of model is a simple version of the vector-autoregressive model). As a final example, one can examine to what extent a variable in one individual predicts a variable in another person over time (again an example of dyadic research questions; for more, see Chapter 14). For instance, Frérart et al. (2024) examined to what extent one's sexual desire is predicted by (and predictive of) one's partner's mood and found that not only one's own mood but also the partner's affects one's sexual desire (and vice versa).

The large benefit of this type of analyses over concurrent relations is that they indicate the temporal directionality of the relation between variables, hinting at what comes first and what comes second and therefore at patterns of sequences between variables (although causal interpretations still remain unwarranted). Both concurrent and lagged relationships can, moreover, be studied with respect to not just two but multiple variables at the same time, allowing one to chart a network of the relations between different variables, either concurrently or directionally. This network approach has become very popular in recent years in studies on, for instance, the organization of symptoms in psychopathology (see Borsboom & Cramer, 2013; Robinaugh et al., 2020) and has been applied to ESM data (see Klippel et al., 2018; Wigman et al., 2015).

All of the above is, in principle, strictly concerned with understanding the behaviour of one or more variables and how they unfold, relate and interact over time within one or more individuals (or between individuals, in some examples). Yet how this happens can vastly differ from one person to the next. To understand this variability, researchers relate individual differences in these patterns, processes and relationships to person-level variables.

2.1.3 Research questions related to person-level characteristics

Indeed, no two persons are the same, and an intriguing and important set of questions relates to how we can map and understand the individual variation in the above-reviewed patterns, processes and relationships (note that these individual differences directly correspond to the random parameters in the models reviewed in Chapter 9). As manipulation of individual differences is typically difficult or ethically objectionable (however see the next section), researchers often work with natural variation of individual differences. Roughly two types of questions are possible here.

The first type of question charts the observed variation and investigates the underlying structure and organization. As individuals may differ in so many respects related to the processes and patterns reviewed above, it may be wise to sometimes try to structure these individual differences and see where there is overlap and distinctiveness (a process undertaken by researchers trying to understand the fundamental dimensions involved in say personality, values, intelligence and so on). For example, Dejonckheere and colleagues (2019) examined the underlying dimensionality of commonly studied affect dynamic indices, showing distinct groupings of ways people differ in the patterns and regularities underlying the dynamics of their emotional lives.

The second type of question pertains to trying to understand observed variation by directly relating it to third (person-level) variables. The list of variables pertaining to individual differences that typify and/or shape people's feelings, thoughts and behaviour is endless, but categories or variables that are often studied are, for instance, age, gender personality, psychiatric illness (severity), early childhood experiences, trauma and stress exposure, cognitive functioning, genetic constitution and expression, biological substrates or indicators such as peripheral physiology, brain structure, activity and connectivity, and interactions between all of the above such as, for instance, gene x environment interaction studies. In fact, many of the examples of concrete research reviewed above often include person-level variables to understand variation in the within-person processes they observe. For instance, Vaessen and colleagues (2018) examined the relation between perceived stress and cortisol as a function of risk for psychotic disorder. Sels and colleagues (2020) examined the degree of emotional synchrony between partners as a function of relationship duration and satisfaction. In the realm of more

biologically oriented individual differences, Collip and colleagues (2011) examined how variation in the catechol-O-methyltransferase (COMT) Val158Met polymorphism is related to stress reactivity in healthy and psychotic individuals. Provenzano and colleagues (2018) examined how brain activity in response to social inclusion or exclusion is related to the dynamics of affective experiences in daily life (note that including person-level variables as moderators is also covered in Chapter 9 of this volume).

Of course, here it is also important to stress that the design or data remains correlational and that causal interpretations of the role of these individual difference variables are unwarranted. Experiments which deliberately manipulate factors that change people's feelings, thoughts and behaviours in daily life are needed for causal interpretations.

2.2 Research questions involving non-natural variation

As ESM is quintessentially about studying people's normal daily lives, most research is concerned with recording and observing the object of study during the normal daily routines that make up our lives. Despite that much can be learned from such methods, they are limited in the sense that while sometimes researchers may be interested in the effects of particular types of events on people's feelings, thoughts and behaviours, these may be difficult to predict, let alone control, in daily life. They are also limited in the sense that normal daily life may more often than not be rather ordinary and may not always elicit the feelings, thoughts or behaviours we are interested in. Moreover, there is an increasing interest in studying the effects of interventions not only in the lab or in the clinical centre but also in the realm of human functioning in daily life. For these reasons, ESM is being used to study forms of non-natural variation in the context of people's lives.

A first step in this type of research is when researchers decide to observe people using ESM during specific periods or surrounding particular significant events. There are several interesting examples in scientific literature where ESM has been used to study people's feelings, thoughts and behaviour around specific types of events. For instance, Kalokerinos and colleagues (2019) studied students' emotions from several days before to a week after receiving their exam grades in first year of higher

education, a milestone in the context of Belgium education. In a similar vein, Belisario and colleagues (2017) studied the course of pregnant women's mood and mood disorder symptoms over the course of pregnancy. In a famous n=1 example of researchers examining changes in the wake of an important event, Wichers and colleagues (2016) examined symptom fluctuations before and after decreasing anti-depressant medication in a single patient. These kinds of designs are very meaningful in studying anticipatory and regulatory processes in the context of important life events. An important requirement, however, is that the event is predictable so a study can be set up around it. And, as we all know, not all that happens to us in life is predictable.

Therefore, researchers have tried to introduce manipulation into people's lives to study its effects. Rather than using predictable events, here researchers create the events themselves. For example, Koval & Kuppens (2012) sampled participants' emotions in daily life before and after they were subjected to a classic stress manipulation (having to give a public speech). While manipulation is classically utilized in laboratory research, the interesting element in this type of ESM research is that researchers can examine effects of manipulations beyond the lab, tracking how they are anticipated and live on in daily life.

Finally, researchers may want to test the effects of interventions meant to benefit people's everyday feelings, thoughts and behaviours. In a way, whether or not people's real lives are affected by certain types of intervention or treatment should be the ultimate evaluative criterion by which to evaluate them (see Chapter 12). Researchers are therefore increasingly incorporating ESM in outcome assessment of new or existing treatments. For instance, Van der Gucht and colleagues (2019) examined the effect of a mindfulness-based intervention on individuals' ability to differentiate negative emotions in the context of daily life. A special example of intervention is when ESM data is used as an element of the intervention as in when, for instance, idiographic data is used to provide feedback and lifestyle advice (see Kramer et al., 2014). Finally, the use of smartphones in contemporary ESM research enables the possibility of delivery of interventions and their evaluation at the same time using the same smartphone. This is the domain of so-called ecological momentary intervention (EMI), which lies at the base of the current spread of mobile health and mental health applications (for reviews, see Heron & Smyth, 2010; Myin-Germeys et al., 2016; see also Chapter 12).

CHAPTER 3

Designing an Experience Sampling Study

Leonie Cloos, Gudrun Eisele, Egon Dejonckheere, and Yasemin Erbas

The design of an experience sampling study involves many decisions that can have a profound impact on the data quality and the ability to answer substantive research questions. This chapter aims to provide an overview of the different parameters that need to be considered when designing an ESM study as well as to promote careful deliberation about the optimal value for study settings as a function of the research question researchers wish to answer. This chapter provides researchers with the necessary tools and guidance to design and set up their own studies.

Before we start, however, we would like to make two general disclaimers. First, while ESM is certainly not a new data collection method (its first use dates back to the eighties; see Chapter 1), methodological research on the design characteristics has lagged. In the last 10 years, however, the field has started to evaluate the methodological standards and conduct empirical research to determine systematic guidelines for the optimal parameter specification for ESM protocols (Trull & Ebner-Priemer, 2020). In the revision of this book, we incorporate the latest evidence on ESM design. Yet it is important to realize that some of our recommendations are still based on personal experience rather than robust scientific evidence. Second, it is important to realize that when designing an ESM study, there is no one-fits-all solution. The decisions you have to make and the optimal design for your study are highly dependent on the specific research question, study population, research infrastructure and so on. Therefore, this chapter cannot advise what the best design is for your particular ESM study. Nevertheless, we provide a general framework to carefully design an ESM study that is optimal for your research question.

We will describe this framework using illustrative examples of five ESM studies that differ in study design. These studies will give a first impression of what types of study designs are possible and what design parameters must be considered. We will provide a descriptive overview of

a number of design parameters and discuss several different options for each. The list of parameters discussed here is not meant to be exhaustive; it is merely a selection of what we think are some of the most important parameters to consider.

3.1 A selection of five ESM studies as examples

Here, we provide a short discussion of five different ESM studies, differing drastically from each other with respect to their design. We selected these examples to outline different types of research questions and what sort of study design they might require. In Table 1, you will find an overview of the most important characteristics of the five studies.

3.1.1 Example 1: The 'traditional' ESM study

The first example that we discuss is a study by Myin-Germeys and colleagues (2001). In this study, the researchers assessed whether the way that individuals emotionally respond to daily life stress is a vulnerability marker for psychotic illness; the study thus examines between-person differences in within-person associations. The **sample** consisted of 150 participants that were divided into three groups consisting of 50 participants each: 1) patients with psychotic illness (high vulnerability), 2) their first-degree relatives (intermediate vulnerability), and 3) control subjects (low vulnerability).

The **study duration** of the ESM part of this study was six consecutive days. During an initial briefing session, participants received a digital wristwatch and a set of ESM self-assessment forms collated in a booklet for each day. The watch served as a **device** to signal the moment of assessment, prompting participants to complete an ESM questionnaire with an **assessment frequency** of 10 times a day. Assessments were scheduled with a random **sampling scheme**, with signals occurring at random moments during the waking hours (from 7:30 a.m. until 10:30 p.m.). The **questionnaire length** was 46 items. It included measures of positive and negative emotions as well as of stress. Positive emotions were assessed with four items (happy, relaxed, satisfied, cheerful), and negative emotions were assessed with five items (down, guilty, insecure, lonely, anxious). All emotion items were assessed on 7-point Likert

scales, ranging from 'not at all' (1) to 'very' (7). For stress, four different items were assessed: event-related, activity-related, thought-related and social stress. To measure event-related stress, participants were asked to report and rate the most important event that occurred between the current and the previous measurement occasion on a 7-point bipolar Likert scale ranging from 'very unpleasant' (-3) to 'very pleasant' (3). Activity-related stress was measured with three items (all on 7-point Likert scales) assessing participants' current activity (e.g., 'I do not have the skills to do this activity'). Thought-related stress was measured with a single item (i.e., 'my current thought is unpleasant'), rated on a 7-point Likert scale. Finally, if others were present, social stress was assessed with two items (e.g., 'I don't like the company'). At each measurement occasion, participants also indicated the time at which they completed the ESM questions.

During the sampling period, participants were called once by the researchers to assess whether they were complying with the instructions. To know whether participants completed the ESM questions within a 15-minute timeframe, the time at which participants indicated they completed the ESM questions was compared with the actual time the digital watch signalled that the participant should complete the questionnaire. All ESM questionnaires that were completed more than 15 minutes after the signal were excluded from analyses. Out of the 150 participants, two relatives stopped collaboration, two patients did not return their booklets and one relative and six patients completed less than 20 ESM questionnaires and were therefore excluded from analyses.

3.1.2 Example 2: Three-wave ESM study

The second study discussed in this chapter is by Dejonckheere and colleagues (2018). In this study, the researchers investigate whether the relation between positive and negative affect (i.e., affective bipolarity) in daily life is altered in people who experience depressive symptoms (*between-person* differences) and whether the degree of affective bipolarity is predictive for future depressive symptoms (*within-person* change). The study consists of three individual ESM periods separated by larger time intervals, implementing the same experience sampling protocol (also referred to as measurement bursts; Stawski et al., 2015). A three-wave ESM design is ideal for the study of short-term variability, long-term change

and the individual differences therein. It is therefore an ideal design for the current study given its interest in both between-person and within-person effects. Each wave consisted of one baseline session in the lab followed by the start of the experience sampling period the next day. The **study duration** of each individual ESM wave was seven consecutive days. The waves were spread over one year. The second wave took place four months after the first wave, and the third wave took place 10 months after the first wave.

At the start, the study had a **sample size** of 202 participants, which is larger than the first study's sample size. However, there was attrition over time: In the second wave, 10 people dropped out followed by a dropout of 13 participants in the third wave. At the beginning of each ESM wave, participants attended a baseline session where they completed a battery of self-report questionnaires, lab tasks and a structured clinical interview. At the end of the baseline session, participants received the assessment **device** (a Motorola Defy Plus smartphone) along with instructions for its use and started the ESM phase.

The ESM **sampling frequency** was again 10 assessments a day that were administered in a semi-random **sampling scheme**. Similar to the previous example, the waking hours of each day (from 10 a.m. until 10 p.m.) were divided into 10 equal time intervals, and at a random time within each interval, participants received a momentary assessment. The **questionnaire length** was 24 items, including eight items measuring their current positive (happy, relaxed, cheerful) and negative (sad, anxious, depressed, angry, stressed) emotions presented in a randomized order at each assessment. All emotions were measured on a continuous slider scale from 1 ('not at all') to 100 ('very much'). The positive and negative emotion items were averaged into positive and negative affect scores at each occasion and correlated with each other across measurement occasions within participants, serving as an indicator of their level of affective bipolarity.

As an incentive, participants were paid up to €60 per wave, depending on how many of the assessments they completed. To motivate them to stay in the study, they were paid an additional €60 for completing all three waves. On average, participants completed 87.27%, 87.87%, and 88.35% of the momentary ESM surveys for the first, second and third waves, respectively.

3.1.3 Example 3: High-intensity sampling design

The third example is a study by Kuppens and colleagues (Kuppens, Oravecz, et al., 2010). In this study, the researchers aimed to investigate individual differences in short-term affective fluctuations using the DynAffect model. According to this model, three major processes underlie individual differences in how our emotional experiences change over time: the affective home base, variability and attractor state (Kuppens, Oravecz, et al., 2010). To evaluate this model, the researchers analysed data from time points sampled at very close and fine-grained intervals. This study differs from the previous examples by focusing on micro-level, within-person changes – specifically, changes occurring within individuals over extremely short timeframes.

The study started with a lab session where participants completed self-report questionnaires and received a measurement **device** (Tungsten E2 palmtop computer) along with instructions for its use. After this initial session, the ESM protocol started. The total **study duration** was only four consecutive days. The **sampling scheme** divided the waking hours of each day into 50 equal time intervals, and the palmtop was programmed to prompt a questionnaire once within each interval. As a result, the average time between two momentary surveys was 17 minutes. Since this study had a very intense **sampling frequency** of 50 momentary assessments a day, the **questionnaire length** was shorter than the one used in the studies above. At each sampling occasion, the participant received only seven items. One of these items was an affect grid, a single-item measure designed to simultaneously assess subjectively felt valence and arousal (Russell et al., 1989) used as an indication of participants' core affect.

The **sample size** was 60 university students. Participants were paid €50 for their participation, and they completed 87% of the momentary assessments on average.

3.1.4 Example 4: Single-case design

The fourth study is by Wichers and colleagues (2016). In this study, the researchers investigated whether change in emotional inertia (i.e., the degree to which an emotional state is self-predictive) can be considered an early warning symptom for the onset of a Major Depressive Episode (MDE). The **sample size** of this study consisted of only one participant who

was a clinical patient: a 57-year-old male patient with a history of multiple MDEs who had been using antidepressants for the previous 8.5 years (see Groot, 2010 for details). Throughout the study, his antidepressant was gradually discontinued. The participant and the researchers were blind to the dose reduction scheme.

While the **sampling frequency** and **sampling scheme** were again similar to the first two examples – 10 measurements a day administered semi-randomly – the **study duration** was far longer. The participant completed 239 days of assessment, during which he received 1,474 momentary questionnaires. The measurement **device** used in this study was a palmtop computer. The **questionnaire length** was 13 items, including 11 emotion items (*irritated, content, lonely, anxious, enthusiastic, cheerful, guilty, indecisive, strong, restless* and *agitated*) that were used to compute indices for emotional inertia. All items were rated on a 7-point Likert scale. Once every two weeks there was an additional trait questionnaire assessing his psychological well-being (i.e., the severity of his depressive symptoms).

Since the participant was one of the initiators of this study, he was highly intrinsically motivated to complete the momentary questionnaires daily for an extended period of time. As such, there was no additional monetary incentive, and the participant complied with 62% of the momentary surveys.

3.1.5 Example 5: The weekend design

The final example we discuss is a longer-term study by Howard and Lamb (2024) focusing on undergraduate alcohol drinkers and differences in compliance over time. This study's **sample size** was 194 university students under the age of 25 who reported drinking alcohol at least twice per month. The data collection took place from Thursday through Sunday each week. The study contained a total of 13 assessments each week delivered over four days with **sampling frequency** varying by day. On Thursday, Friday and Saturday, participants received four assessments per day at fixed time points (10 a.m., 4 p.m., 7 p.m. and 11 p.m.) while on Sunday they received only one assessment at 10 a.m.. Participants had two to four hours to respond to the questionnaires. The **questionnaire length** could be up to 45 questions due to the use of branching or follow-up items, where a follow-up question would be asked depending on

the participant's response to the previous one. The total **duration** of the study was 14 weeks.

Participants used their smartphones as an assessment **device**, receiving text messages with links to the online surveys at the scheduled assessment times. Alongside these momentary assessments, participants completed an extensive intake battery that gathered demographic information and data on personality traits and other mental health indicators. The study specifically examined the relationship between these demographic, personality, and mental health characteristics and participants' compliance with the study protocol.

The overall compliance rate across the study was 76.5%, with notable trends over time. Compliance was highest in Week 1, at 88.9% but gradually declined over the 14-week period, reaching 70% by Week 14. This decline underscores the challenges of maintaining participant engagement in long-term studies.

We have now briefly discussed five different daily life studies with four different ESM protocols. As you can see from these examples, there is large heterogeneity in the adopted ESM protocols. This is because the study design was tailored to the specific research question the researchers aimed to answer, the study population and so on. As the goal of this chapter is to provide tools for designing an ESM study, we delve deeper into the parameters given in the examples above to help understand these parameters and how they interact. In the remainder of the chapter, we will use these studies to explain the various ESM parameters and their options.

Table 3.1. An overview of the most important characteristics of the five example ESM studies.

ESM Study	Sample	Study duration	Assessment frequency	Sampling scheme	Questionnaire length	Device
Traditional	150 (patients, relatives, controls)	6 days	10/day	Semi-random	46 items	Wristwatch
3-Waves	200 students	3 x 7 days	10/day	Semi-random	24 items	Motorola Defy Plus
High-intensity sampling	50 students	4 days	50/day	Semi-random	7 items	Tungsten E2 palmtop
Single-case	1 MDD patient	239 days	10/day	Semi-random	13 items	Palmtop
Weekend	196 college students	14 weeks (4 days/week)	max. 4/day	Interval	max. 45 items	Personal Smartphone + survey link

3.2 The most important design parameters of an ESM protocol

In the second part of this chapter, we will introduce several ESM parameters. We will do this by first giving a descriptive summary of the parameters. Next, we will use the examples of the ESM studies that we discussed earlier to illustrate the range of possibilities. As highlighted in Table 3.1, we will discuss the following parameters: study duration, assessment frequency, sampling scheme, questionnaire length and study device. As we already mentioned, this is not a conclusive list. Rather, these are the most basic and some of the most important parameters that structure an ESM protocol. Consequently, these are the factors that must be considered first when designing your ESM study, regardless of what type of study you are planning.

3.2.1 Study duration

Study duration refers to the number of days that a given ESM protocol lasts. As we have seen from the examples, there is considerable heterogeneity in study duration (ranging from four days to 239 days). As such, there are many possibilities, and the specification of this parameter mainly depends on the research question. Are you interested in the relationship between certain variables across individuals (for instance, in the relationship between stress and negative affect across people with a different vulnerability for psychosis; see the 'traditional' ESM study)? Or in processes within individuals (for instance, whether changes in the level of emotional inertia can predict the onset of a depressive episode, as seen in the 'single-case' study)? Whether you are looking into relations on the between-person or within-person level, you must know on what **timescale** these processes unfold. For example, are these changes expected to unfold over longer periods of time (as was the case in the 'single-case' ESM study) or from minute to minute (as was the case in the 'high-intensity' ESM study)?

As we have seen, the two 'typical' ESM studies (example 1 and example 2) lasted six and seven days. But this is not the gold standard; several meta-analyses show considerable variability in study durations. For example, Vachon et al. (2019) analysed 79 datasets and found that study durations ranged from one to 150 days with an average of 11.2 days and 68% of studies lasting between two and 10 days. Similarly, research

on well-being reported a mean duration of 12.8 days (range: 1–56 days; median: 7 days; de Vries et al., 2020) while another meta-analysis of 477 ESM articles found a very similar result: a mean duration of 12.4 days (range: 2–180 days; median: 7 days; Wrzus & Neubauer, 2023). In the realm of mood and anxiety assessments, Hall et al. (2021) noted that most studies lasted either seven or 14 days, with an overall mean of 22 days and a range from 1–240 days. Often, this information is not easily identified in studies. For instance, a review of 36 smartphone-based mood assessment studies found that only 14 reported the duration of the self-reporting period, which ranged from four days to 18 weeks (Sarsenbayeva et al., 2023).

Importantly, typically, when running an ESM study, not all participants start on the same day (e.g., some may start on a Monday, others on other days of the week). Research using the day reconstruction method (e.g., Kahneman et al., 2004) shows that people engage in different categories of activities on different days, and these categories can impact psychological phenomena (e.g., emotions) differently. Therefore, a duration that will lead to sampling both weekdays and weekend days in all participants, eliminating potential differences due to weekdays (when most people work or study) versus weekend days (when most people have more time for leisure activities and social contact), is recommended. This applies to the last example where researchers only looked at the weekend as that is when it is socially acceptable to drink alcohol. But perhaps you are only interested in weekdays (e.g., when studying people's behaviour in the office) or only on weekend days (e.g., when studying how people divide their free time or studying how working parents interact as a family with their children). In these cases, a protocol of seven consecutive days would not be very useful, and you could decide to sample only during office hours or during the weekends (depending on the focus of the study) for a number of weeks.

The ESM measurement periods can also be spread out when changes in affect dynamics over time are anticipated. Likewise, when studying specific behaviours that are relatively infrequent (such as for instance, sexual activity), it may be necessary to extend the sampling period to increase the likelihood of capturing these occurrences (Serre et al., 2012). For example, in the 3-waves example study, the participants were repeatedly sampled for seven days. A recent review found that 81.1% of ESM studies use one wave of data collection (de Vries et al., 2020). Some ESM

studies have left the number of days up to the participants and some have used multiple waves.

An ESM study can also be interrupted by measurement-free phases. The 'three-wave' study also had an ESM protocol of seven days, but the protocol was repeated three times (i.e., three waves over the course of one year). Here, the researchers were interested in studying between-person differences in within-person change over a longer period. In contrast, the 'single-case' study, where the focus was on within-person change only, lasted 239 days, which is quite a long duration for an ESM study. Finally, the 'high-intensity' study was quite a bit shorter than the other studies, lasting only four days, because the researchers were interested in micro-level changes (and when we discuss the parameter assessment frequency, it will be clear that study durations become shorter when the assessment frequency is very high and the study protocol is very demanding for the participants (see also the section in this chapter on interdependencies between ESM parameters).

3.2.2 Assessment frequency

This design parameter refers to the number of momentary assessments (also referred to as time points, sampling occasions, notifications, beeps or prompts) per day, which mainly depends on what a meaningful frequency is for the topic of the study (see section 3.3 of this chapter for more information). For instance, if the topic of study is sleep quality, it is enough to assess this only once every morning. It would not make sense to ask this question multiple times throughout the day because the answer will not change. However, if the topic of study is people's current emotional states, a highly variable construct, a higher sampling frequency is likely more informative (Kuppens, Oravecz, et al., 2010).

On average, in ESM studies self-report measures are typically administered around five to six times per day. For example, the review by Hall et al. (2021) reported an average of 5.63 assessments per day, with most studies applying between one and 10 assessments daily (range: 1–64). Similarly, a meta-analysis by Wrzus and Neubauer (2023) found a mean of 6.53 assessments per day (range: 1.7–81), noting that the number of assessments was unrelated to compliance and dropout rates. Notably, a meta-analysis of studies in patients with severe mental disorders by Vachon et al. (2019) reported a mean of 6.9 assessments per day but found

that a higher number of assessments was associated with a slight decrease in compliance in this population. In studies related to well-being, 42 studies employing random sampling reported an average of five prompts per day (de Vries et al., 2020). Lastly, a review of smartphone-based mood assessments by Sarsenbayeva et al. (2023) found that among the studies, 24 utilized daily mood assessments ranging from one to 12 assessments per day.

An important question concerning assessment frequency is whether micro-level or macro-level changes are being investigated. If we look at our examples, we see that the 'high-intensity' ESM study, which investigated micro-level changes in affect, had an assessment frequency of 50 momentary assessments per day. There should ideally be some theoretical reason behind the number of assessments. There is always the option to opt for the minimum number of assessments based on the planned analyses and practical constraints. However, this approach carries the risk of too few assessments being taken. For example, if we examine the link between a social interaction at one time point and emotions at the next, but the assessments are too far apart, there may have been other social encounters in the meantime, and we may attribute a direct association that does not exist. Furthermore, if there is not enough data and changes in the data are missed, complex and nonlinear patterns could remain undetected. In general, fewer assessments leads to an oversimplification of the data, potentially missing critical insights that arise from more frequent or nuanced measurements (Cloos et al., 2024; Hopwood et al., 2022). Moreover, the reliability of model-based estimates such as the autoregressive effect, typically estimated with time series data, may be lower when there are not enough assessments (Adolf et al., 2021). If there are restrictions on the assessment frequency, autoregressive estimates can be improved by increasing the number of participants to compensate for a shorter time series (Hecht & Zitzmann, 2020).

Two additional parameters related to variation in assessment frequency are the start point and endpoint of the day and the between-prompt interval. The start point and endpoint of the day refer to the time of the first assessment and the time of the last assessment, respectively. For these parameters, it is important to find out what the typical waking hours of your population are and make sure that in normal circumstances, all assessments fall within these hours. It seems that participants get more responsive to assessments over the course of a day. A recent re-analysis of

10 ESM datasets, including 1,717 individuals with different mental health conditions, showed that the highest compliance was between 12 p.m. and 1:30 p.m. (83%) while the lowest compliance was between 7:30 a.m. and 9 a.m. (56%; Rintala et al., 2019). However, this analysis may be driven by participants' preferred assessment times as later times are generally favoured (Dejonckheere et al., 2024). Moreover, the interval between notifications is of course highly dependent on the number of assessment points and the start points and endpoints. For example, if you only sample students during school hours, the timeframe to space the assessments will be smaller, and the more momentary assessments you impose, the smaller the interval between two consecutive assessments will generally be. However, the between-prompt interval also depends on the sampling scheme applied in your ESM protocol, which we will discuss next.

3.2.3 Sampling scheme

The sampling scheme refers to the way in which assessments are generated over the duration of a study. Here, we will discuss the fixed scheme, the random scheme and the semi-random sampling schemes, which are all signal-contingent, as well as the event-contingent sampling scheme (see Figure 3.1 for a graphical comparison).

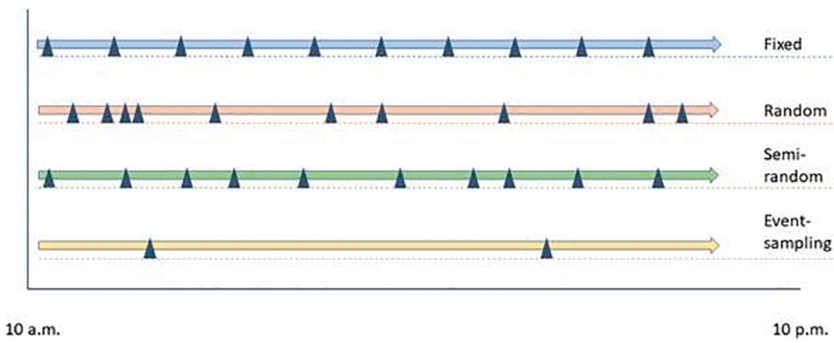


Fig. 3.1 A graphical comparison of the four different sampling schemes, where the triangles indicate a sampling occasion.

In a **fixed sampling scheme**, assessments are generated at pre-determined and equally distributed time points. For instance, they can be generated

at the start of every hour or between the start point and endpoint of the ESM protocol. There are significant advantages to a fixed scheme. When the assessments are scheduled at a fixed time, they are predictable. This will likely result in higher compliance with the ESM protocol (Vachon et al., 2019). Moreover, many statistical models rely on the assumption that the distance between time points is equal (Bringmann et al., 2013). If the assessments are fixed at equal time intervals (e.g., once every hour), which is only possible when sampling occurs following a fixed scheme, this assumption is correct. However, if a fixed sampling scheme incorporates unequal intervals or allows participants to delay their responses – as seen in the weekend study by Howard and Lamb (2024) – this assumption is violated.

One disadvantage of such a fixed scheme is the fact that the prompts are predictable, which decreases the ecological validity of the study and increases reactivity to the method as people may change their behaviour in anticipation of the prompts (e.g., not starting an argument with your spouse because you anticipate an assessment in two minutes). This would be unfortunate because high ecological validity is one of the main assets of ESM. Moreover, because assessments always take place at the same time, selection bias may occur, meaning that certain moments can be overrepresented or underrepresented. This could, for instance, be the case when one of the assessments is always exactly at noon when participants are having lunch with their colleagues. Conversely, a fixed scheme can also result in important periods of the day being missed, which, again, is problematic for the representativeness of the data. For instance, if a participant only has social interactions in the evenings but the ESM protocol does not include evening assessments, data may not be missing at random but rather in an explainable pattern. The structural absence of information will affect the generalizability of the conclusions.

In a **random sampling scheme**, assessments are randomly generated within a single time interval over the day. The random sampling scheme has exactly the opposite advantages and disadvantages of the fixed sampling scheme. The main advantage of a completely random sampling scheme is that the measurement prompts are not predictable, and this unpredictability increases the ecological validity and reduces reactivity to the method. However, the unpredictability of the assessments can make it more burdensome for participants to keep up with the ESM protocol. Moreover, this type of sampling can, in unlucky cases, result in a very

unequal distribution of the assessments. For instance, if most assessments are assigned to take place in the morning, information about the rest of the day will be missing. In such a case, the data will not be representative of the entire day. Furthermore, an uneven distribution of assessments will also lead to unequal distances between the time points. In this case, many statistical models that rely on the assumption of equal intervals cannot be used, and alternative continuous time models must be used (de Haan-Rietdijk et al., 2017). To our knowledge, this type of sampling is not used very often because the advantages generally do not outweigh the disadvantages.

In a **semi-random sampling scheme**, assessments are randomly generated within multiple pre-defined time intervals. For instance, the time between the start point and endpoint of the day can be divided into equal time intervals (and the number of intervals corresponds with the number of assessments per day), with a measurement prompt being generated randomly within each interval. There can also be additional rules imposed on this type of scheme – for instance, that there has to be a minimum window of at least 10 minutes between two consecutive assessments.

The semi-random interval scheme is the most commonly used sampling scheme as it is a trade-off between the fixed and random interval schemes. Therefore, the advantages and disadvantages are also a combination of the advantages and disadvantages of the fixed and random interval schemes. For instance, there is some predictability but also some unpredictability. This results in a relatively high ecological validity compared to the fixed sampling scheme as the variation in assessment timing captures a broader range of daily experiences. It reduces participant burden and retains compliance compared to the random interval scheme. Finally, while the time in between consecutive assessments is not exactly equal, in a semi-random sampling scheme it is relatively safe to assume equality because the differences in the intervals will generally equal out.

In an **event-contingent sampling scheme**, participants fill in a momentary survey when a specific event occurs (e.g., a panic attack or a binge-eating episode). This type of sampling scheme is used when the research investigates rare events or very specific situations or behaviours because the other sampling schemes may miss them. Explicitly asking participants to respond to a questionnaire after such an event ensures that the event is not missed. The main advantage of this event-based

sampling scheme is that the assessments are tailored to study specific, potentially infrequent events. However, a disadvantage is that it requires an active participant: While in the other sampling schemes participants are reminded by a device to answer the questionnaires, here participants need to initiate the questionnaire themselves. Depending on the events of interest, this could affect participant burden and compliance. For instance, if you are interested in the effects of physical activity on mood, this will not be too complicated for the participant. In fact, a recent study investigating interpersonal behaviour and affect in social situations showed that the quality of data collected through a random sampling scheme and an event-sampling scheme resulted in a similar data quality, but the event-sampling scheme resulted in a higher number of reported social interactions (Himmelstein et al., 2019). However, it becomes more complex if, for example, you are interested in the consequences of panic attacks or binge-eating and you require the participant, who is already burdened by the event itself, to initiate and respond to questions immediately after such an event. Furthermore, it is possible that a selection bias can occur when participants must decide which events to include. For instance, when participants are asked to report every stressful event, the type of events that are selected may differ between participants. Additionally, because the assessment will only be initiated *after* the event has taken place, the questions concerning experiences during the event will be answered retrospectively. The known problem with retrospective questionnaires is that recall biases can influence the accuracy of the memory of the event (Van den Bergh & Walentynowicz, 2016) and therefore the quality of the data. Finally, an important disadvantage of the event-sampling scheme is that it provides no information about what happens in the time between events, and it will not be possible to assess the predictors or the consequences of these events. Event-sampling schemes are therefore often used in combination with the other sampling schemes.

Most studies use a time-contingent approach (de Vries et al., 2020; Vachon et al., 2019; Wrzus & Neubauer, 2023), sometimes combining both fixed and semi-random sampling approaches (Hall et al., 2021). *Event-contingent* sampling, while theoretically beneficial for capturing specific events, remains rarely used due to the increased burden it places on participants (Dawood et al., 2020; Wrzus & Neubauer, 2023; Hall et al., 2021).

3.2.4 Questionnaire length

Questionnaire length refers to the number of questions assessed at each measurement occasion. Again, there is no gold standard for the optimal number of items. Different review papers report an average of 13 to 22 items in an ESM questionnaire with a range between one and 60 items (de Vries et al., 2020; Vachon et al., 2019). Like most design parameters, the number of items depends on important study factors (see 3.3 for more information). As discussed in Chapter 4, on questionnaire development, there is research showing that generally, a longer momentary questionnaire predicts lower compliance. In a recent study, Eisele and colleagues (2020) gave students either a 30- or a 60-item questionnaire three, six, or nine times per day for 14 days and found that a longer questionnaire negatively affected both data quality and compliance. In a study by Hasselhorn and colleagues (2022), students received 33 or 82 items three times a day. The questionnaire length did not affect the perceived burden or compliance, but longer questionnaires had a lower data quality, providing smaller within-person variability of mood and lower relationships with extraversion. While several meta-analyses have reported links between questionnaire length and compliance (Morren et al., 2009; Vachon et al., 2019), others have not confirmed this relationship (Jones et al., 2019; Ono et al., 2019; Ottenstein & Werner, 2022; Rintala et al., 2019; Soyster et al., 2019). These discrepancies may be explained by the adaptation of other study factors – such as longer questionnaires typically being paired with lower sampling frequencies (Dejonckheere et al., 2024) – or by differences in whether studies examined completion times rather than questionnaire length. In our example studies, the questionnaire in the 'traditional' ESM study consisted of 46 items, the 'three-wave' ESM study had 24 items, and the 'high-intensity' ESM study featured only seven items but implemented a high sampling frequency of 50 questionnaires per day. The 'single-case' ESM study included only 13 items (though it lasted much longer), and in the 'weekend' ESM study, the number of items differed across surveys (9–45). An interesting observation from the weekend study is that the morning survey, despite having the longest questionnaire, achieved the highest compliance, likely because it was available for four hours and conducted in the morning when students are less occupied (Howard & Lamb, 2024). These examples illustrate how decisions regarding questionnaire length interact with other study parameters.

3.2.5 Response times

For the questionnaire administration, two parameters are relevant to study design. Firstly, the amount of time participants have to initiate an ESM questionnaire, also referred to as the response delay, and secondly, the amount of time it takes a participant to complete a questionnaire once it is initiated, also referred to as the completion time. While both parameters are often not reported in publications, every ESM researcher still needs to decide on these parameters. A recent study (Eisele, Vachon, et al., 2021), showed that studies differ significantly regarding the allowed **response delay**: while some studies considered responses only to be valid if they were initiated within seconds of the prompt, others adopted a response delay of hours, and most studies allowed response delays up to 30 minutes (Scollon et al., 2009). Importantly, Eisele, Vachon, et al. (2021) show that longer response delays can have a negative impact on the reliability of the data. While it is currently not clear why this is the case, according to the authors, there are two possible explanations: With regard to momentary items, it is possible that larger response delays can increase sampling or self-selection bias, meaning that not all situations have an equal chance of being measured (e.g., participants wait for a calmer moment, resulting in a mood-related bias). Concerning retrospective questions, larger response delays may increase recall bias. In our example studies, the allowed response delay was 15 minutes in the 'traditional' ESM study, 90 seconds in the second 'three-wave' ESM study, 90 seconds in the 'high-intensity' ESM study, 15 minutes in the 'single-case' ESM study and up to four hours in the 'weekend' ESM study.

While the relation between **completion time** and data quality has also not been extensively studied, there are indications that the principles that apply to response delay may also apply to completion time. For instance, a recent study showed that longer completion times related to lower data accuracy (van Berkel et al., 2019). Indeed, some researchers therefore limit the completion time. For instance, the 'three-wave' study referenced above used the ESM software MobileQ (Meers et al., 2020), which applies a time limit of 90 seconds per item. Importantly, while a large completion time can indicate inaccurate or unreliable answers, so can a very short completion time. Participants need time to properly process and answer the questions. An extremely short completion time can therefore be an indication of careless responses, and this may need to be accounted for.

For this reason, setting a limit for the minimum completion time may be necessary to monitor data quality (e.g., McCabe et al., 2012, Jaso et al., 2021).

3.2.6 Study device

A device refers to the instruments that are used for data collection. ESM data can be collected in many ways. While ESM studies originally utilized the paper-and-pencil method (in combination with a programmed wristwatch that would prompt participants to fill in a momentary survey), nowadays, **electronic devices**, predominantly **smartphones**, have become the preferred method for data collection (Trull & Ebner-Priemer, 2013). Indeed, the rise in experience sampling research is likely tied to the widespread availability and use of electronic devices (Kuppens et al., 2022). Many people already own a mobile phone and have the necessary technology to participate in an ESM study. The ESM study can be easily implemented on the participants' smartphones using ESM applications. However, there may be contexts where a research-dedicated device is preferred. For instance, when conducting a study with children, teachers may not want pupils to have access to their smartphones in class but could allow research-dedicated devices. Other populations (prisoners, for instance) may not always have access to smartphones, and using research-dedicated devices may be necessary.

Apart from the paper-and-pencil method and the use of electronic devices, there are other methods to collect ESM data. One possibility is data collection through **phone calls**, where the researcher calls the participant and the participant answers the questions through the phone (see the Midlife in the United States (MIDUS 2) study for an example in which researchers called the participants once a day for seven days to collect their answers; Ryff et al., 2017). Another option is using **online interfaces** (such as Qualtrics or Survey Monkey) where the participant answers the questions online and the links to the questionnaires can be sent to the participant in various ways (e.g., by email or text message or through platforms such as www.surveysignal.com). The 'weekend' study from the above example studies sent a link to participants' smartphones where they could open and answer the question. Data from the above-discussed 'three-wave' ESM study was collected with research-dedicated smartphones while the 'high-intensity' study used a palmtop and the 'single-case' study used a dedicated research device.

All these sampling methods have their advantages and disadvantages. The paper-and-pencil method has the important advantage that it is free to use. However, there is little control over when the participant fills in the questionnaire (e.g., Stone, Broderick, et al., 2003), and if a paper diary gets lost (e.g., it is easy to 'forget' one on the bus), all data is lost. Additionally, the data preparation process is long, and the data needs to be entered manually, which makes it susceptible to errors. A final potential drawback is the occurrence of *backfilling*, when participants fail to complete the questionnaires at the required times and quickly fill in the questionnaires before returning them to the researcher (e.g., Stone, Broderick, et al., 2003). This phenomenon may be especially problematic when the reward for study participation is dependent on compliance. Naturally, this could have detrimental consequences for data validity. It must be noted, though, that another study by Jacobs and colleagues (2005) reported that most people comply with the ESM protocol when using paper-and-pencil approaches. In addition, a study comparing the paper-and-pencil approach with one using digital devices found similar results (Green et al., 2006).

Using electronic devices has a lot of advantages and solves most issues associated with paper-and-pencil techniques. The data preparation process is much shorter compared to a paper-and-pencil method. It is also automatic, so data will not need to be entered manually, and therefore there is a lower risk of error. Electronic devices are also small and light and therefore very easy to carry along in daily life. Moreover, the researcher can choose to use an ESM application that is installed on the participants' phones, meaning no extra devices need to be carried at all. Another advantage is that with electronic devices, each questionnaire entry will receive a timestamp. This way, the researchers will know exactly when a questionnaire was filled. Furthermore, the researcher can program the questionnaires in such a way that the participant is required to answer them within a given timeframe (for instance, within 10 minutes of the actual prompt) and make it impossible to access the questionnaire after this period. Depending on the type of software and hardware used in the ESM study, it is also possible that the data on the electronic device backs up automatically – for instance, when connected to a mobile network. This way, there is a smaller chance of losing the data, and if the electronic device does get lost, usually only a small amount of data will be lost with it. Finally, given vast technological developments, there are many different types of applications available to suit the different needs

of researchers, and many of these applications are open-source and free to use for non-commercial activities (see Chapter 6 for more information).

With all these advantages, it is no surprise that electronic devices are now the most commonly used devices to collect ESM data. Nevertheless, there are also some important disadvantages. For starters, there may be issues with the usability/accessibility of electronic devices. There may also be technical issues with electronic devices: there can be hardware or software bugs, incompatibility of the ESM application with certain devices or battery problems. Also, similar to the paper-and-pencil method, participants may lose their devices. Because the data can be backed up regularly, data loss will be less of a problem here. However, these devices are often very expensive, and losing them can have consequences for the ESM study. Moreover, for most people, smartphones are mundane devices that they use routinely without much thought. As a result, when participants fill out questionnaires on their phones, the task may feel automatic and relatively insignificant. This sense of familiarity can reduce the cognitive load and make the data collection process smoother, but it also raises the risk of participants not fully engaging with the questionnaires. Moreover, when the ESM software runs on participants' own device, their own applications may interfere with the ESM application (e.g., receiving a phone call or text message while completing a momentary survey). Finally, while it is a big advantage that there are many ESM applications available to researchers, some of these applications are rather inflexible and/or expensive. Therefore, researchers need to consider very carefully which application they should use to collect their ESM data (Walsh & Brinker, 2016).

Finally, as mentioned before, there are other methods that can be used to collect ESM data, such as phone interviews or online surveys (e.g., with links to the surveys sent through text messages or emails). However, research shows that these methods often result in lower compliance than the use of electronic devices or the paper-and-pencil method (Stone, Shiffman, et al., 2003). Therefore, using these alternative methods may not be indicated.

3.3 Fine-tuning the ESM parameters of your study: A guiding framework

In the previous paragraphs, our discussion of the different parameters that shape an ESM study was confined to a purely descriptive level,

covering each of the criteria in an isolated way as if they exist entirely independently. However, now that we know which parameters to consider, the next logical steps are to explore potential interdependencies between these study factors and to evaluate how a particular parameter blend may give rise to a unique protocol that is suited to answer the specific research questions you have in mind. Indeed, as we mentioned in the introduction of this chapter, each research hypothesis calls for its own specific study design, and there are no clear-cut rules or fixed regulations that should be followed. Instead, this section aims to provide a framework that may guide you in your decision-making and whose ultimate goal is promoting careful consideration about the optimal set-up for your ESM study's parameters in light of the study's research question.

Contrasting the five different ESM studies we reviewed earlier, it becomes clearly evident that there is considerable heterogeneity in the protocols the researchers adopted. This heterogeneity is an inherent consequence of the fact that each ESM study was designed to answer a different research question. Indeed, a qualitative study with different interviews from a renowned group of ESM experts (n = 74) revealed that the research question was the most important factor in determining the design choices for a daily life study (Janssens et al., 2018).

You could generally argue that the configuration of your study settings is the result of inevitable tension between the practical constraints of reality and the ideal scenario in which you would like to test your hypotheses. Although this dichotomy is applicable to any empirical test, it may be particularly valid in the case of ESM research, where scientists ultimately want to have a detailed and accurate understanding of what is happening in participants' lives at every given moment but are limited by the feasibility of their protocol. Put differently, ESM researchers are required to make a trade-off between answering their research question in the best possible way (i.e., ideal scenario) and the study *feasibility* (i.e., reality; Hektner et al., 2007). In the next paragraphs, we will discuss these two opposing forces in more detail.

3.3.1 Force 1: Answering your ESM research question in an ideal world

How would you answer your research question if you had all the resources (i.e., time, money, participants, etc.) in the world? In an ideal ESM study, you could test your hypotheses by (a) *continuously* monitoring all the

people in the world, preferably with a (b) *non-obtrusive*, (c) *implicit* and (d) *unlimited* set of measures. Rather than a linked sequence of momentary snapshots, (a) continuously tracking participants refers to receiving an uninterrupted stream of information about people's daily lives with a level of detail similar to a movie (Cloos et al., 2024; Hollenstein, 2021). Ideally, the measurement would (b) not interfere with participants daily routine, alter subjects' behaviour (e.g., participants may still want to take a bath to relax even though they anticipate a measurement occasion in the near future; Myin-Germeys et al., 2009); furthermore, the responses would not be affected by important life events (e.g., even if participants were scared to death, they would still complete a momentary survey about their feelings; Sun et al., 2021). In fact, participants would (c) not even have to actively reflect on the questions they are presented with. Researchers would derive internal experiences such as mood, thoughts, symptoms or attitudes from indirect measures. This practice would get rid of any response biases (e.g., prompting participants about momentary moral behaviour would always result in honest responses; Fisher, 1993) and eliminate reactivity issues (e.g., repeatedly asking patients with binge-eating disorder about their unstoppable urge to eat would not increase the frequency of binge episodes; Boon et al., 2002). Finally, researchers would be able to monitor participants' lives in all its facets, (d) not having to make an a priori selection of the constructs they wish to track nor having to specify a predefined study period. Similarly, regarding number of participants, the researchers would not have to draw a particular subsample from the population about which they wish to make inferences. Taken together, in an ideal world, maximizing the parameters of the ESM protocol would allow you to obtain the most valid answer to your research question.

3.3.2 Force 2: Answering your ESM research question in a world with practical constraints

Unfortunately, an effective implementation of this ideal protocol is impossible due to laws and practical constraints. To gain insight into the dynamics of participants' internal world, i.e., to capture a person's subjective experience, self-report is the most valid method (Coan, 2010; Kuppens, 2019). Although attempts to infer people's internal states via passive sensing are well underway (Mehl et al., 2021), these will not be

able to replace self-report but rather will merely support it. However, having one's daily routine repeatedly interrupted to reflect and report on mood, thoughts or symptoms is quite challenging and burdensome for participants to carry out on a regular basis (e.g., Fuller-Tyszkiewicz et al., 2013). Consequently, researchers need to carefully consider the duration of their study. Similarly, researchers cannot harass participants constantly, imposing measurement occasions at whatever times of day and however often they want (Silvia et al., 2014). Moreover, the number of items and measures participants receive at each assessment need to be limited (Eisele et al., 2020). Finally, in terms of participants, limited monetary compensations or research devices prohibit the inclusion of an infinite number of participants in a study. Instead, anticipating a predefined effect size, researchers should perform rigorous planning of the sample size (Lafit et al., 2021; Lafit, Revol, et al., 2024; Lafit, Artner, et al., 2024). In sum, reducing the parameters of an ESM study has the advantage that it is less burdensome for participants and hence more feasible to carry out.

The ideal and realistic scenarios work in opposing directions with respect to determining the optimal value of ESM parameters. Theoretically, maximizing the intensity of your ESM parameters provides you with more information about people's life (*data quantity*) and hence would allow you to answer the research question at hand. However, in reality, this maximization comes at the cost of additional burden for participants, likely leading to fatigue, careless and unreliable responding, and maybe even drop-out from the protocol, undermining the reliability and validity of obtained information (*data quality*). Put differently, under real circumstances, there is an inherent trade-off between the quantity and quality of obtained information, dissolving a strict linear relation between the quantity and the quality of collected data (see Figure 3.2). Although more information is better to a certain degree, further increasing the intensity of your ESM protocol or increasing the data quantity will not automatically lead to an increase in the quality of the data. On the contrary, at a certain point, intensifying the ESM protocol to increase the quantity of the data is associated with a decrease in the quality of the data. Consequently, when tailoring your ESM study around a specific research question, the implicit aim should be to pursue the saddle point in this curve-linear relation (i.e., where the relation between quantity and quality flips from positive to negative). Around this value, the ESM protocol allows you to collect a maximum amount of information with minimal burden or

resources. This sweet spot is where the ideal values of the assessment frequency, study duration, and questionnaire length come together to yield the perfect trade-off between data quantity and quality. Indeed, a meta-analytic summary of ESM studies shows that this sweet spot can be reached if researchers adjust the number of assessments per day to the study duration (Wrzus & Neubauer, 2023). Importantly – as we will reiterate – the position of this optimal value will be different for different research questions (see Figure 3.2).

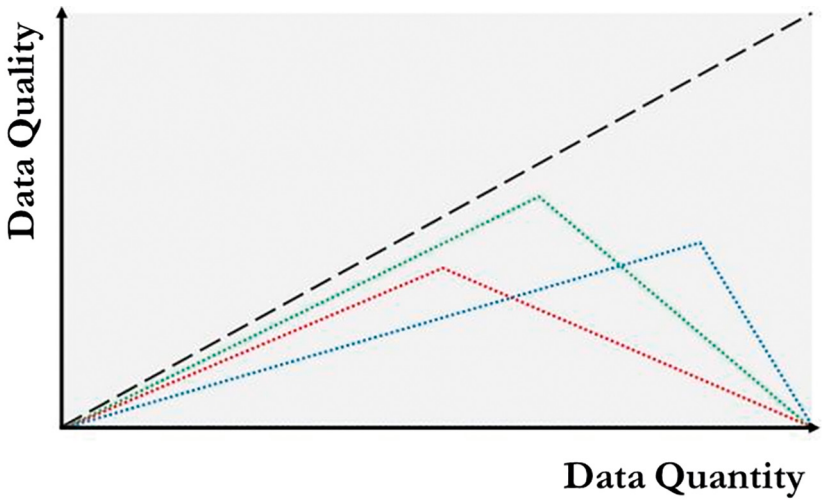


Fig. 3.2 A graphical representation of the inherent tension between the data quality and the data quantity. The black long-dashed line represents the ideal scenario, where intensifying the ESM protocol leads to more data and information to answer your research question. The small-dashed lines refer to ESM studies carried out under real circumstances, where a further increase in the intensity of an ESM protocol compromises the quality of the data. The implicit aim is to find the saddle point in this relation, where quantity and quality are maximized, while the participant burden is minimized. This theoretical point may differ for different studies (compare red, green and blue lines), and is a function of participant characteristics, construct features and statistical analyses (see below).

3.3.2.1 Quantitative indicators of data quality

The most well-known quantitative indicators of data quality are compliance and attrition. **Compliance** can be defined as the ratio of the number of measurement occasions that participants actually completed over the

theoretical maximum number of measurement occasions allowed by the protocol (Vachon et al., 2019). As such, this percentage is inversely related to the number of measurement occasions that were missed by a participant. Missing measurement occasions is inevitable in daily life studies, and a few missed measurement occasions are generally not problematic. Indeed, in naturalistic settings, answering a measurement prompt may not always be feasible either because the participant does not hear or see the notification (e.g., when taking a nap; McLean et al., 2017) or because the completion of the momentary survey is considered inconvenient (e.g., when taking a shower), unsafe (e.g., while driving) or inappropriate (e.g., while having sex). In some cases, however, missing measurement occasions can influence the representativeness of the data and undermine the interpretation of the results – specifically when the pattern of missed occasions is no longer completely random but consistently dependent on particular moments or situations in the life of the participant (Silvia et al., 2014). For example, when a participant consistently fails to complete the first survey of each day, the ESM data of that participant may not accurately represent his or her morning routine (Rintala et al., 2019). Scanning for temporal regularities in missing data within study days and across the entire study period is therefore crucial. Similarly, missing data is problematic when it depends on the phenomenon being investigated (e.g., when a researcher is interested in panic attacks, but the patient consistently fails to rate his anxiety symptoms when experiencing extreme levels of panic). Here, completed measurement occasions may not provide a full and representative answer to the research question at hand. However, due to the fact that the data is missing, it is very difficult to determine what people were actually doing at the time a measurement occasion was missed. The recent development of unobtrusive measurement devices allows researchers to bypass this obstacle. In an inventive ESM study, for example, Sun and colleagues (2021) used Electronically Activated Recorders (Mehl, 2017) that logged short audio snippets of participants' real-world behaviour and surroundings at the time of a missed prompt (later category-coded by the researchers). They found very little evidence that missing an ESM survey was related to psychological constructs that are typically of interest to ESM researchers (e.g., positive or negative emotion, social interactions, etc.), providing reassuring evidence that missing data is not entirely dependent on specific instances that happen in the lives of participants.

Two meta-analyses of ESM compliance found an average compliance rate of 79% across different ESM studies (Wrzus & Neubauer, 2023; Vachon et al., 2019). This benchmark can serve researchers as they design their studies and set expectations for participant compliance.

Attrition (also called retention or drop-out) refers to the proportion of participants that does not reach the end of the ESM protocol but rather prematurely wishes to abort the study (Vachon et al., 2019). Usually, attrition is strongly related to compliance because the participants who quit the study early are typically those who find the protocol too burdensome and too much of a disturbance to their everyday lives (Delespaul, 1995). For this reason, it is common practice that these cases are not included in the final sample for analyses because researchers implicitly assume that their unreliable answers may invalidate their conclusions. Again, some attrition is to be expected, but the question remains to what extent instances of drop-out relate to particular participant characteristics (Ji et al., 2018; Rintala et al., 2019; Vachon et al., 2019). If drop-out is not completely random, the danger exists that the study sample may not accurately represent the population you are trying to make inferences about (e.g., when the most depressed patients drop out from an ESM study because they feel too tired or down to complete surveys on an hour-to-hour basis, drawing valid and representative conclusions about symptom fluctuations in major depressive disorder may be difficult; Houben et al., 2021). Relatedly, sampling bias can undermine the data quality, for example, students who have more free time between classes are more likely to sign up for studies, leading to a sample that may not be representative of the broader population. This type of bias limits the generalizability of the findings as the experiences of these participants may not reflect those of individuals with different schedules or responsibilities. In the problematic case that a substantial group of subjects does not reach the end of the protocol, the study was probably infeasible in its totality. In this regard, qualitative feedback after the study period may help you understand why participants found the study too demanding. For example, in a study in which the prompts were sent too early in the day, taking part in the study was incompatible with participants' jobs, carrying around another mobile device next to their personal smartphone was irritating, etc. (Eisele et al., 2020). If person-level information was collected during a baseline session prior to the ESM study, it may be worthwhile to statistically check whether dropouts differ in meaningful ways from participants who successfully

completed the entire protocol (e.g., clinical status, age, gender, etc.; Dejonckheere et al., 2021; Scollon et al., 2009).

Finally, sometimes participants do not pay attention to the question and just answer without really considering or taking it seriously, which undermines the quality and interpretability of the data. This is referred to as 'careless responding.' To date, careless response methods have largely been identified post-hoc, meaning that careless responding is identified once the data have been obtained by evaluating them on quantitative parameters related to response times (e.g., very short completion times) and response content (e.g., inconsistent responses to similar or opposite items, strings of same responses to different items). However, although these methods based on response time and response content parameters have been used in cross-sectional survey data (Meade & Craig, 2012, Jaso et al., 2021) and implemented in the experience sampling context, there has been little promise that they actually improve the data quality (Welling et al., 2021). Careless responses are the result of participants' lack of effort and intention. Effort and intention can be checked with so-called 'careless response items' in which an instructed response is required (e.g., 'please choose disagree for this item'). A disadvantage of such items is that they are easily identified (Ward & Meade, 2023) and thus participants may learn to watch out for them, especially in the context of the repeated measures. Recently, Jaso and colleagues (2021) introduced a promising online detection method for identifying careless responses, but its adoption in studies has been limited so far. While this method shows potential, it still requires further refinement and broader implementation. Other studies have been more critical of the overall concept and definition of careless responding. Using simulations and empirical data, Schroeders (2021) and colleagues have shown that real-world data can be far more erratic than we may believe, potentially leading to misidentification of genuine data variability as careless responding.

3.3.2.2 Qualitative indicators of data quality

Interference and reactivity are two of the most important qualitative measures of data quality. We define them as 'qualitative' because these indicators can typically not be condensed into a single number and are therefore harder to evaluate directly compared to the quantitative indicators. **Interference** refers to the degree to which taking part in an ESM

study hinders the occurrence of naturalistic or authentic behaviour (e.g., playing sports, driving a car, taking a nap, etc.; Scollon et al., 2009). Again, some interference in ESM is inevitable because repeatedly being asked to complete short momentary surveys while participating in ongoing activities can be experienced as a hindrance, potentially altering participants' behaviour (Hormuth, 1986). However, whether interference becomes problematic by substantially undermining the data collected depends on both the participants' characteristics and the ESM protocol adopted. Interference may be more of an issue when participants already have a lot on their plate in addition to taking part in a study (e.g., parents with a newborn, students taking finals, couples organizing their wedding day, etc.). With respect to features of the study itself, the assessment frequency and the exact timing (sampling scheme) of the measurement prompts may affect the overall study interference. Specifically, having to complete more surveys daily in combination with the fact that participants can predict upcoming measurement prompts (e.g., in fixed-time sampling schemes; Verhagen et al., 2016), increases the risk that they will organize their lives around the study instead of vice versa.

Relatedly, **reactivity** refers to the degree to which measuring operations directly affect participants' responses provided in the momentary questionnaires (Barta, 2012). Although reactivity may be an issue for all researchers relying on self-report methods, it can be particularly problematic in ESM studies because the repeated nature of the surveys may lead participants to pay unusual attention to their internal states or own behaviour (Scollon et al., 2009). The mechanisms by which repeated self-assessments may initiate change in responses can vary from active reflection (e.g., 'I seem to be a person who is feeling down regularly'), to introducing different reference values over time (e.g., 'at least I'm not feeling as down as yesterday'), to social desirability (e.g., 'I must never feel sad') to installing feedback processes (e.g., 'I typically feel sad after talking to my mother; I may want to cut down on these interactions'), yet the degree of reactivity likely differs following a number of factors (Ram et al., 2017). For example, reactivity may be a function of the specific study domain, leading some phenomena under investigation to trigger more reactivity than others. Although to our knowledge systematic reviews in the ESM literature do not exist in this regard, reactivity seems to be more central in the context of, for example, rating substance use (e.g., drinking behaviour; Buu et al., 2020) or depressive symptoms (e.g., alleviation of

depressed feelings; Broderick & Vikingstad, 2008; Kramer et al., 2014) than rating pain symptoms (Cruise et al., 1996; Stone, Broderick, et al., 2003; von Baeyer, 1994), body image or self-esteem (Heron & Smyth, 2013; Leahey et al., 2007) or eating behaviour (Munsch et al., 2009). Subject susceptibility may evidently also play a role in reactivity. Conner and Reid, for instance, found that reactivity in happiness was particularly evident in participants who showed low levels of trait negative affect (2012). In terms of study characteristics, the specific phrasing of the survey items likely introduces more or less reactivity (see Chapter 4 on how to avoid reactive questions in an ESM questionnaire and Chapter 5 on the associated ethical considerations). Similarly, the regularity with which participants need to complete assessments may affect the observed reactivity (e.g., evaluating emotions every five minutes versus once per day; Ram et al., 2017). Finally, differences exist in the way reactivity is typically operationalized (e.g., evaluating the temporal slope of a construct during an ESM protocol, contrasting pre- versus post-baseline ratings, exploring changes in response variability in function of time, etc.), which may also explain why different ESM studies come to different results in terms of reactivity (De Vuyst et al., 2019; Vachon et al., 2016). Consequently, drawing unequivocal conclusions regarding the problematic role of reactivity in ESM is quite difficult, but acknowledging its existence is critical (Barta, 2012).

3.3.3 Study factors that determine the optimal value of your ESM parameters

3.3.3.1 Sample characteristics

A first important point of attention pertains to whether your sample of interest has a clinical diagnosis or whether you are studying vulnerable populations. Using ESM as a data collection method in clinical groups is certainly possible (Myin-Germeys et al., 2018; Trull & Ebner-Priemer, 2009) as illustrated by the plethora of daily life research with patients suffering from depression (e.g., Heininga et al., 2019), borderline personality disorder (e.g., Houben et al., 2021), psychosis (e.g., Myin-Germeys et al., 2009) and so on. Nevertheless, their clinical status may call for an ESM protocol that is tailored around their specific needs and vulnerabilities. First, it may be important to evaluate the start point and endpoint of the day depending on their diagnosis. For example, depressed individuals

typically suffer from disturbed sleep patterns (Nutt et al., 2008). Prompting them too early in the day may interfere with their natural sleeping routine, leading to interference or reduced compliance. Conversely, not prompting them at night (when they are awake) may cause researchers to miss important information about their sleeping behaviour, invalidating the representativeness of the data for subjects' daily life. Individually adapting the start point and endpoint of an ESM day to the sleep-wake cycle of a participant or the assistance of passive sleep sensors that allow for conditional prompts only when the participant is not asleep (e.g., McLean et al., 2017) may overcome this difficulty. Relatedly, clinical status may also have implications for the study duration, assessment frequency and questionnaire length. Overall, compliance and retention rates are still acceptable in patient groups (Soyster et al., 2019; Ono et al., 2019). However, some meta- or pooled data analyses indicate that clinical populations tend to miss more measurement prompts and drop out from an ESM study more frequently, and this is particularly true of participants with psychosis (Rintala et al., 2019; Vachon et al., 2019). The fact that their psychiatric illness already introduces substantial burden into their everyday lives and the experience of psychopathological symptoms could considerably interfere with taking part in an ESM study (e.g., because of concentration difficulties, fatigue, paranoia, etc.) likely explains this finding. Nevertheless, ESM in clinical populations is certainly possible.

Next, it is critical to evaluate the developmental phase of your population of interest or the age group to which they belong (e.g., children versus university students versus elderly). Again, ESM research can be carried out with participant groups from across the lifespan (see Cordier et al., 2016; Rah et al., 2006; Schmiedek et al., 2010), but age-related characteristics potentially call for specific parameter settings. First, research shows that younger participants are typically less compliant (Rintala et al., 2019), which you may wish to account for in ESM protocols when studying younger age groups. Second, age may have implications for the device you use to collect your ESM data (Gould et al., 2020). While very young children may not be allowed to use mobile phones or other digital devices in some of the contexts they typically encounter (e.g., during classes, before they have finished homework, etc.), elderly people may not be very familiar with recent technologies, making a smartphone-based ESM study impractical. Relatedly, older participants' vision abilities may be impaired, leading them to favour old-school paper-and-pencil techniques

over answering survey items on small, contemporary smartphone touch screens. In extreme cases, parents or caregivers may even have to assist participants in completing the questionnaires (e.g., interviewing or observing; Bartels et al., 2020; Bouisson & Swendsen, 2003; Lamont, 2008; van Knippenberg et al., 2018). Attenuating the duration, assessment frequency and questionnaire length of your ESM protocol may be justified.

3.3.3.2 Construct features

Researchers should think carefully about the prevalence of the specific feelings, symptoms or behaviours being investigated. Among other parameters, these will have important consequences for the duration of the ESM study because the regularity with which the phenomenon of interest occurs determines how representative a sample of momentary assessments will be for people's daily life as a whole. While instances that occur frequently (e.g., eating and drinking) may only need a handful of assessments to adequately cover a typical day in participants' everyday lives, rare behaviours or experiences (e.g., binge-eating or drinking alcohol) will generally require longer study periods.

Here, it is important to note that the phrasing of survey items can affect the degree to which participants will endorse statements about momentary states and that altering the phrasing may change a construct's measured prevalence (see also Chapter 4 on how to avoid the formulation of extreme survey questions). For example, when interested in the determinants of non-suicidal self-injury in patients with borderline features, the urge to self-injure is known to be more prevalent than the actual act (Snir et al., 2015). While urge and act may certainly not be entirely equivalent, it is justifiable to monitor the urge to auto-mutilate instead of the specific act for appropriate study duration purposes. Similarly, in groups of depressed patients, momentary endorsements of experienced euphoria versus happiness differ in terms of frequency due to differences in item intensity (Heininga et al., 2019).

If you choose to use multiple items to measure the same construct in ESM, it's important to be aware that the relationship between the items can fluctuate over time. For instance, when using various emotions to assess affect, the emotions may not always co-occur or change in perfect synchrony (Zelenski & Larsen, 2000). For example, emotions like anger and sadness both represent negative affect, but the likelihood that they

will occur together in the same moment is often low. This variability means that the underlying construct, which is measured by combining these items, may shift across different time points (Vogelsmeier et al., 2023). This should be considered when designing a questionnaire with multiple items for a construct (Cloos et al., 2023).

The prevalence of the constructs will also have implications for the sampling scheme that is adopted in the ESM study (Ram et al., 2017). With time-contingent ESM designs, the real-time occurrence of very rare feelings, symptoms or behaviours may be underreported (e.g., 'Are you having a panic attack right now?'; Verhagen et al., 2016), making this type of sampling scheme generally less suitable when studying exceptional events (Himmelstein et al., 2019). In contrast, event-contingent sampling may capture these infrequent phenomena shortly after they occur, but this sampling scheme has the drawback that it typically cannot give real-time information about antecedents as participants only initiate the questionnaire after the phenomenon of interest takes place. Alternatively, modifying the phrasing of an item to capture experiences in between two measurement occasions (e.g., 'Did you have a panic attack since the last prompt?') may pick up on infrequent behaviour or experiences using time-contingent sampling schemes, but it has the disadvantage that recall biases may distort accurate memory retrieval (e.g., knowing the outcome of an event is known to change one's perspective on it; Colombo et al., 2020).

Closely related to the prevalence of psychological constructs is the timescale on which your phenomenon of interest operates (Hopwood et al., 2022). 'Timescale' here refers to the question of how quickly a particular emotion, symptom or behaviour is subject to change (Boker et al., 2009), and it is critical to evaluate in light of the assessment frequency of the ESM design. On the one hand, oversampling the construct to the extent that temporal changes can hardly take place between measurements runs the risk that the protocol is perceived as too burdensome for participants to engage with. For example, when interested in the dynamics of people's sleep quality, it will suffice to assess this construct on a daily basis (e.g., first prompt of every morning; Kasanova et al., 2020) because the rate of change of being awake versus asleep typically follows a *diurnal cycle*. In contrast, mood fluctuations are known to be more volatile (Kuppens, Oravecz, et al., 2010) and therefore likely require multiple assessments per day. On the other hand, under-sampling the construct jeopardizes

the tracking of relevant fluctuations between measurement occasions (Dejonckheere & Mestdagh, 2021) and generates data that do not accurately reflect the deterministic properties of the symptoms or emotions under study (Schiepek et al., 2016). For example, when studying emotional recovery in response to real-life stressors, a low temporal resolution runs the risk that full recovery took place in between assessments, making it impossible to accurately describe individual differences in emotional recovery (Mestdagh & Dejonckheere, 2021). Determining the appropriate timescale on which psychological constructs change is far from trivial, and relying on different timeframes will produce different conclusions (Neubauer & Schmiedek, 2020). As a guiding principle, to adequately pick up on the serial dependency between discrete assessments, the time interval between consecutive measurement occasions should be shorter than the rate of change of the phenomenon under study (Ram et al., 2017).

3.3.3.3 Statistical analyses

Finally, it is vital to tailor the ESM protocol to the statistical analyses that will be performed to obtain the answer to the research question in mind. As with data collection methods, it is advisable to think about the statistical models or techniques you will rely on to test your hypotheses before collecting the data. Modelling a research question involves making an abstraction of reality (Box, 1976), and important assumptions often underlie the translation of a research question into a statistical model.

A first query involves the level of analysis. Are you mainly interested in between-person associations, or is your research question concerned with within-person associations? (Or both; see also Chapter 4 on assessing psychological constructs with ESM.) A focus on between-person differences has implications for your sample size (Lafit et al., 2021). If interested in the effects of a person-level trait, characteristic or ability, it may be worthwhile to stratify your sample on that variable of interest to ensure the full range of levels is represented in your study (Ingram & Siegle, 2009). For example, when evaluating individual differences in the structure of everyday emotion in function of depressive complaints, a depression pre-screening instrument can be used to recruit a final sample that experiences a wide and balanced range of depressive symptoms (e.g., Dejonckheere et al., 2018). In addition, for burst design studies where an ESM study is composed of multiple sampling waves, it is desirable to enrol

slightly more participants into your study than required because drop-out rates likely increase as the interval between measurement periods becomes larger (Dejonckheere et al., 2018). However, when the focus of your research question is related to within-person relations, the most important parameters to consider are study duration and assessment frequency. Here, the number of completed surveys is pivotal to making generally and ecologically valid inferences. Because interference makes it rather unlikely that participants will complete all measurement occasions (Vachon et al., 2019), it is advisable to anticipate a certain percentage of absent assessments by introducing slightly more assessment occasions than those that are strictly needed.

A specific type of research question related to the within-person level involves a temporal component (e.g., studying how emotions or symptoms change over time; Dejonckheere et al., 2017; Houben et al., 2015; Koval et al., 2015; see also Chapter 4 for more information). To adequately study temporal relations, it is important to make sure that you capture sufficient autocorrelation in the constructs of interest (Dejonckheere & Mestdagh, 2021). Autocorrelation has to do with the question of how well you can predict the value of a construct from its previous assessment (Koval et al., 2013) and is partly determined by the assessment frequency of your ESM protocol (Bulteel et al., 2018). That is, when the temporal resolution of repeated assessments is too low (i.e., the interval between two assessments is too large), you will not be able to capture serial dependencies between assessments. This should encourage an increase in the measurement frequency of the protocol.

Lastly, it is important to consider the assumptions underlying your model of ESM study because violations may lead to biased model parameters. Some modelling techniques, for example, assume that the time window between consecutive measurement occasions is fixed (i.e., equidistant assessments; Bringmann et al., 2013), affecting the selection of the sampling scheme. Similarly, other models can only be applied to stationary time series, meaning that the mean and variance of a construct remain relatively stable over time (Box et al., 2015).

3.3.4 Interdependencies between ESM parameters

In the previous paragraphs, we highlighted how researchers should select the optimal value of an ESM parameter in function of a single participant

characteristic, construct feature or statistical analysis. However, to further maximize the validity of your ESM protocol for a given study, it is preferable to consider all these factors simultaneously while also acknowledging the relative interdependencies between ESM parameters. Indeed, changes in one specific parameter may have consequences for other parameters when researchers want to maintain similar levels of validity (Eisele et al., 2020). Although this observation seems to make things even more complex, the goal remains unchanged: Designing an ESM study that is as unobtrusive as possible but allows for answering the research question in the best possible way.

To illustrate the importance of considering the relative interdependencies between ESM parameters, let's consider and compare the protocols of two real-life ESM studies: the 'single-case' ESM study and the 'high-intensity' ESM study (see Figure 3.3). In both examples, the researchers aimed to answer a specific research question in the most valid way, which led to a unique combination of the parameters related to study duration, questionnaire length and assessment frequency. In the 'singe-case' ESM study, the researchers were interested in predicting the onset of a Major Depressive Episode based on changes in emotion dynamics. Because depressive relapse does not take place very often, this construct feature (i.e., prevalence of the phenomenon of interest) had major implications for the duration of the ESM study. That is, the study duration had to be considerable (i.e., 239 days) to observe a phase transition from healthy to clinical. However, keeping participant burden in mind, this specification also has consequences for the assessment frequency and questionnaire length of the protocol. In the interest of not undermining the data quality, it is likely that the optimal value of these parameters was kept lower than that in the prototypical ESM study. Indeed, to minimize interference, poor compliance or drop-out, assessment frequency and questionnaire length were lower than usual. In contrast, in the 'high-intensity' ESM study, the researchers were interested in capturing micro-dynamical changes in people's affective life. Here, the crucial factor that determined the study settings was the timescale on which the phenomenon of interest operates. To effectively capture minute-to-minute changes in affect, the assessment frequency of the ESM protocol needed to be substantial, requiring a value that was higher than that in the average ESM study (i.e., 50 times per day). Once more, the selection of this value also had implications for the other aspects of study settings. To ensure the protocol would not be too

burdensome, the researchers decreased the duration of the study and the questionnaire length.

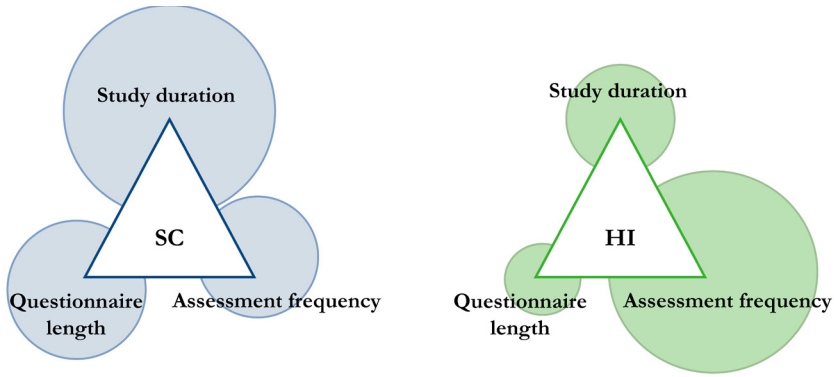


Fig. 3.3 A graphical representation of the inherent interdependencies between ESM parameters. Larger circles indicate higher parameter values. SC = Single-Case ESM study (see Example 4); HI = High-Intensity ESM study (see Example 3).

3.3.5 Conclusion: Pilot your study

As emphasized throughout this chapter, the presented guidelines do not have an imperative or exhaustive status. Rather, an overarching and general framework was designed to guide you in your decision-making about the specification of ESM parameters for a given study and to point to specific considerations that should be accounted for in pursuit of optimal study settings. This is in part because systematic inquiry about the implications of ESM design choices is still lagging (Himmelstein et al., 2019). Answering these types of questions remains difficult because (a) evaluating the effect of isolated ESM parameters is challenging due to the interdependent nature of the different parameters, and (b) every research question, participant sample, construct of interest and statistical analysis may require slightly different design settings. Clearly, all this precludes a one-fits-all solution for the optimal ESM design. But does this mean we have to set up our ESM study in a vacuum? Certainly not. As mentioned earlier, these guidelines may help you to consider certain factors – perhaps not previously thought of – that shape your ESM study. Furthermore, with the rising frequency of ESM studies, there are more protocols available

from other studies that can be used to set up a study. This rising frequency will also aid replicability and reproducibility. Above all, it is crucial to design an initial plan for a given ESM protocol and to try it out on a pilot sample. When severe data quality issues arise, it may be worthwhile to redesign and adapt the protocol in function of the quantitative and qualitative feedback received. Quantitatively, problematic compliance or severe drop-out rates may inform the researcher that the protocol was too burdensome. Qualitatively, incorporating a user-centred approach to enhance participant engagement can improve data quality (Bülow et al., 2025). Interviews with participants may reveal specific aspects of the design that undermine the data quality of the ESM protocol (e.g., reactivity due to phrasing of certain questions, sampling schemes that start too early, interference with normal day to day activities, etc.). In sum, designing a good ESM study is a process of trial and error like any other data collection method in psychological science.

CHAPTER 4

Questionnaire Design and Evaluation

Gudrun Eisele, Leonie Cloos and Marlies Houben

The validity of ESM research findings stands or falls on the use of appropriate questionnaires. The development of such questionnaires is arguably one of the most challenging parts of ESM research since it requires both conceptual clarity and methodological rigor as well as some foresight as to what to expect during the course of the data collection period. In this chapter, we aim to give an introduction on how to design a questionnaire while paying attention to the intricacies inherent to ESM. While there is currently no standardized procedure, we provide an overview of issues and considerations that need to be addressed at different stages of questionnaire design. First, we discuss how to identify the need for a new measure and define what the study seeks to measure. Second, we list some guidelines for the construction of optimal ESM items that capture dynamic daily-life experiences and address considerations regarding the use of multi-item scales versus single-item measures. Third, we discuss principles and considerations that are essential to constructing an ESM questionnaire consisting of several items and scales. Fourth, we give an overview of tools to assess the quality of individual items and scales. Finally, we discuss assessments that can complement self-reports.

4.1 Defining the target construct and identifying the need for a new measure

Before engaging in any questionnaire development work, formulating a clear definition of the target construct is crucial. What is it that you want to measure? How is it similar to or different from other constructs? ESM can be used to assess a broad range of different concepts in daily life (Wrzus & Neubauer, 2023); for instance, it has been employed to capture the current context (‘Who are you with at the moment?’), appraisals thereof (“This is a pleasant company), affective experiences (‘How happy

do you feel right now?'), the occurrence of behaviours ('Since the last beep, have you consumed alcohol?'), physical experience ('How hungry do you feel right now?'), cognitive states ('Since the last beep, have you been ruminating?'), etc. A clear definition of the construct you want to measure can reduce ambiguity and confusion and help other researchers synthesize results later. Conversely, a lack of clarity at this stage can contribute to fragmentation of the literature (Weidman et al., 2017). Recent reviews of ESM studies indeed document large variations in how constructs are operationalized across studies (e.g., Janssens et al., 2024; Beames et al., 2025), underlining the importance of properly defining measures and checking existing instruments before developing new ones.

ESM items are intended to capture dynamic phenomena that are expected to evolve and change over time on a relatively small timescale (i.e., across hours or days rather than months or years). This contrasts with static phenomena that are expected to remain stable for a relatively longer period, for which ESM is less well suited. To illustrate, 'How hungry are you right now?' is a meaningful ESM item since for most people appetite tends to fluctuate during the day, depending on food intake. In contrast, a person's response to 'What is your favourite food?' is expected to be more stable. Food preferences can change, but it is unlikely they would on a timescale of hours or days. Therefore, repeatedly assessing this throughout the day would be redundant and not meaningful. A meaningful ESM item should capture state-like features that assess something specific to the present moment and thereby show sufficient within-person variability. A very general formulation tends to measure trait-like features rather than momentary states. To illustrate, 'Are you an impulsive person?' captures impulsivity as a more stable trait-like feature while 'Do you feel the urge to do things impulsively right now?' tends to capture momentary impulsivity. Note that simply adding 'at the moment' to a question is not sufficient to make it sufficiently variable. Questions such as 'At the moment, are you an impulsive person?' can be confusing as they simply mix trait-like assessments with momentary time references. Indeed, the need to capture a time-varying construct emerged as one of the central criteria for evaluating ESM items in a Delphi study among 43 ESM experts. This exercise led to the development of ESM-Q, a quality evaluation tool for ESM items consisting of 25 quality criteria (Eisele et al., 2024).

Note that not all phenomena are strictly momentary states or stable traits. For example, from a clinical perspective, a depressive episode typically refers to a period of at least two weeks during which certain specific depressive features are present, implying (limited) stability over time. However, it is also thought to express itself in changes at the momentary level and fluctuations of depressive features over time. While more subtle fluctuations at the momentary level can be captured through ESM ('I feel down right now'), it would not be appropriate to employ ESM to inquire about the supposedly stable intensity of a depressive episode over a long timeframe ('I have been down the past week'). Generally, the reference period for ESM items should focus on phenomena that range from minutes to hours.

Note that sometimes concepts are assumed to be static from a theoretical perspective while realistically they can fluctuate across time within individuals; some examples are self-esteem or personality (Oosterwegel et al., 2001; Santangelo et al., 2017; Thewissen et al., 2008). Whether a phenomenon is state-like rather than trait-like can be psychometrically evaluated using intra-class correlation coefficients (for more information, see section 4.3.2, describing the assessment of the quality of one or more items, and Chapter 9).

Even in the relatively new field of ESM research, many phenomena have already been assessed. Therefore, it is important to first establish whether a new measure is needed or whether there are already high-quality items that can be used to assess the target construct. In addition to scanning relevant ESM literature, searching the ESM item repository may help researchers find existing items of their target construct. The ESM item repository (https://www.esmitemrepository.com/; Kirtley, Eisele, et al., 2025) is a publicly available item bank already containing more than 3,000 ESM items at present. We recommend consulting the item repository before setting up a new ESM study and contributing new items if you have already collected relevant data. It is important to note that the repository does not currently contain information on the reliability or validity of the included items. Rather, its current primary function is to give researchers an overview of existing measures and their references. Therefore, the repository represents a useful starting point, but it is important to consult existing evidence of validity or evaluate items before using them in a new study.

4.2 Constructing individual ESM items

Once the target construct is defined and the need for a new measure is established, you can start constructing new ESM items. Some basic principles and considerations are essential to pose a meaningful ESM question that can easily be answered by all participants under any circumstances.

4.2.1 Different timeframes

When we formulate individual ESM items, a choice must be made concerning the timeframe or time reference in the question. Overall, two options are customary: a question can refer to the present moment (for example, 'How sad are you right now?') or to a specific time interval (e.g., 'Since the last beep, have you had conversations with someone?', 'Have you taken your medication today?').

One of the main goals of ESM is to assess dynamic and momentary states in daily life. Therefore, in many instances, researchers want to assess phenomena in the present moment. This can be done using different terms, such as, 'How sad are you right now?', 'Are you currently experiencing anger?', 'How happy do you feel at this moment?' Time references to the present reduce recall bias to a minimum and are therefore preferred in many instances. Keep in mind also that in the case of momentary assessments, explicitly adding the time reference will improve the clarity of the question. To illustrate, 'How happy are you?' could refer to momentary happiness but also overall satisfaction with life, which is likely less changeable. 'How happy are you right now?' is clearer as it makes the momentary nature of the assessment explicit.

Researchers can also assess a construct within a specific time interval rather than in a specific moment. The most commonly used time interval is that between the moment the question is being asked and the previous ESM questionnaire, referred to using terms such as 'since the last beep' or 'since the previous questionnaire'. Repeatedly using this time reference allows us to capture the entire day. However, any other time interval can be used to assess concepts for practical or theoretical reasons, such as 'in the last hour', 'since this morning' etc. Questions can also be asked daily at the start or the end of the day – for example, 'Did you sleep well, last night?' or 'How pleasant was your day today?'

Referring to a specific time interval, such as in between assessments, instead of asking subjects to report on the moment itself can be useful for several reasons. First, time intervals are suited to capture behaviour, thoughts, events etc. that occur less frequently. If such concepts are assessed momentarily ('right now'), the risk exists that many, if not all, occurrences will be missed. Asking about the occurrence using a longer time interval will likely give a better picture of the overall occurrence of such phenomena. To illustrate, researchers could be interested in assessing the occurrence of binge eating in daily life. If they repeatedly ask participants whether they are currently engaged in binge eating, they will miss many or most occurrences of such behaviour unless the assessment schedule and the occurrence of the behaviour line up perfectly, which is very unlikely, so assessing whether one or more binge-eating episodes occurred since the previous assessment results in a better overview of the overall occurrence. Second, time intervals might be more suited to assess phenomena that do not happen in a moment but rather over time, requiring assessment over a longer period. For example, the use of emotional regulation skills typically occurs over time (i.e., over the course of minutes to hours) rather than at one particular moment. It is easier to assess such phenomena retrospectively by reflecting on what a person did in the previous minutes/hours. Third, some states are difficult to assess in the moment because the participant might be unable to respond to questions while being in the state of interest. Examples of such states could be dissociative states or epileptic seizures. Fourth, sometimes one may be able to choose a specific timescale that matches the occurrence of the construct perfectly. To illustrate, if sleep or medication use are only expected to occur once per day or once per night, it would be most appropriate to assess sleep quality or the use of medication using the time reference of 'today' or 'last night'.

Assessing phenomena in specific time intervals rather than momentarily also has some disadvantages or implications that should be kept in mind. First, the larger the time intervals, the higher the likelihood of retrospective biases and undue influence of the current mental state on the appraisals of past moments, especially in populations that might have inherent memory problems. Second, if it is desirable to model associations between variables, the time reference used to assess each variable has implications for the nature of the relationship being modelled. For example, if the association between two variables – A and B – is modelled, with

variable A being assessed momentarily and variable B since the previous assessment, the association will be of a mostly prospective nature despite both variables being assessed at the same time. An association between the two will show whether and how variable A at time t is related to variable B assessed in the preceding time interval between time t-1 and time t. Third, if the occurrence of behaviour or thoughts is assessed in a specific time interval, it is unknown when exactly they happened within that time interval unless it is specifically assessed. For example, an event reported in the previous time interval could have occurred immediately before the question was asked or at the start of the time interval. Especially for larger time intervals, this can lead to large variability in time of occurrence and can have implications for prospective relationships that researchers may want to model. Moreover, there is the possibility that more than one phenomenon of interest occurred during the time interval. If the participant is then asked to appraise the phenomenon or the impact it had, it would be unclear to the participants as well as the researcher to which occurrence of this phenomenon this appraisal pertains. To illustrate, a researcher may want to assess whether the participants experienced any stressful situations in the past hours and require them to rate how unpleasant they were. If a participant can recall two distinct stressful situations in the past hours, however, it is unclear which event should be appraised. To summarize, many different arguments can be made for and against the use of different time references. The choice of the time reference of items should be driven by the construct that researchers aim to measure and the type of research question they aim to answer. If a person is interested in examining the concurrent relationship between two variables, it is important to make sure that both variables are assessed using the same time reference. To illustrate, if a person is interested in whether people feel cheerful while exercising, both current-moment exercise and positive affect should be assessed momentarily. However, if a person wants to obtain information regarding the extent to which a person exercises in daily life, one could consider assessing the amount of exercise since the previous assessment. This would provide a better estimate of overall exercise than repeated assessment of exercise in the moment. If the research question is whether exercise is followed by positive affect, one could associate exercising since the last assessment with momentary positive affect and possibly also with positive affect at subsequent moments in time to capture the duration of the effects of exercising on mood.

The importance of clearly defining and matching the timeframe of an ESM item to the expected variability in the construct also emerged in ESM-Q, indicating that there is consensus among ESM experts that this is an important consideration (Eisele et al., 2024). As discussed above, both momentary and the retrospective framing have advantages and disadvantages. When items are assessed in the moment, the assessments occur intermittently, leaving gaps between them in which important experiences are missed (Hollenstein, 2021). When items are asked retrospectively, the score represents a summary over a longer period, which may be influenced by the most intense or the most recent memory of that period (Fredrickson, 2000). To capture changes and important peaks of intensity in between assessments, research can apply Intensity Profile Drawings, in which participants retrospectively draw the changes in a construct over the period in between two assessments (Cloos et. al., 2024). For instance, a participant may have indicated feeling very positive during the last assessment and may again feel very positive at the current assessment moment. However, it is possible that in the time between the two assessments, important fluctuations in positive affect were experienced. In a study using Intensity Profile Drawings, the participant can indicate these fluctuations by drawing the level of positive affect over time.

4.2.2 Wording

Some basic principles should be kept in mind regarding the formulation of ESM items. First, keep each question short and to the point. This will ensure that each question is easy to read, fits the screen of the mobile device well and can be quickly answered. It is important to minimize the burden and interference with daily activities that ESM may pose, as is apparent in ESM-Q (Eisele et al., 2024). Second, avoid extreme wording in ESM questions. If extreme terms are used to describe a concept, it will apply to only a limited number of time points and persons. This will result in limited within-person variability and potentially skewed data. Note that what is considered extreme depends on the type of population a researcher is investigating. To illustrate, while rage, ecstasy and anxiety can all occur in the daily life of some people and can be captured through ESM, these items are likely to show limited variability in the majority of the participants. Most people likely score on the lowest range of the response scale, creating a floor effect (note, however, that such items

may show very high variability in specific populations, such as patients with bipolar disorder or psychotic conditions; Myin-Germeys et. al., 2001). In contrast, irritation, enthusiasm and stress are experiences that most people experience in mild to high levels. Third, explicit assessment of a concept is often, but not always, preferred. Sometimes, explicit assessment can actually induce thoughts or behaviour. To illustrate, imagine a researcher wanting to assess the presence of intrusive thoughts in daily life. Explicitly assessing these will likely lead to more intrusions. In such cases, an implicit assessment would be preferred (e.g., 'What was the most prominent thought in the past hour?'). Additionally, for more complex concepts, explicit assessment does not always provide the most valid assessment as people might lack explicit knowledge themselves. To illustrate, concepts such as emotional instability can be explicitly addressed by asking people to indicate the extent to which their emotions changed abruptly since the last assessment. However, people do not always have the insight to correctly estimate the volatility of their emotions. Therefore, repeatedly asking persons to rate their current emotional states and then modelling the degree of instability using statistical indices would provide a more valid assessment of their emotional instability. Fourth, make sure to avoid excessive use of negative wording and to use neutral wording as much as possible to avoid response bias. If a researcher wants to assess depressive features such as loss of appetite, weight loss and anhedonia, one could consider asking about appetite, weight changes and experience of pleasure to avoid having a set of negatively formulated items that can put the participant in a negative mindset. Fifth, although a researcher is typically interested in assessing specific theoretical concepts, it is very important to avoid jargon in the actual ESM items, a consideration that, again, also featured among the agreed-on quality criteria of ESM-Q (Eisele et al., 2024). Instead, use terms and words that participants themselves would use to describe a concept. When in doubt, ask non-academic relatives to proofread the items, organize discussion groups with your target population to make sure that questionnaires are clear and relevant, and pilot test all items. Sixth, avoid questions in which participants are asked to reflect on behaviour, thoughts or feelings that have occurred and then make judgments about why or under which circumstances they occurred. Instead, each relevant component can be assessed separately, and statistical models can be applied to test the association between the relevant components.

To illustrate, if a researcher wants to know whether someone feels sad when they are alone, one could consider asking participants to indicate the degree to which they felt sad while being alone since the previous assessment. However, responding to this question requires participants to rate several components (feeling sad and being alone) separately and to connect them, which is complex and difficult. Moreover, such questions more likely trigger socially desirable answers or answers that reflect global self-views rather than actual momentary states. Therefore, a more optimal course of action would be to ask participants to rate their current sadness and then indicate whether they are currently alone or not. Afterward, one can easily model the association between sadness and being alone. Seventh, it is important to formulate questions in such a way that they are relevant and can be answered in any type of situation. For example, if the research question centres on social contacts, asking participants 'Did you initiate the contact?' can be very meaningful in some, but definitely not all, social contexts. Instead, questions should be formulated in such a way that they are relevant in a large variety of contexts. Finally, standard recommendations on questionnaire wording such as using unambiguous wording and avoiding double negations should be followed (see ESM-Q; Eisele et al., 2024).

Box 4.1 Checklist for optimally formulated ESM questions

1. Make them short and to the point
2. Avoid extreme wording
3. Consider implicit rather than explicit assessments
4. Avoid excessive use of negative wording
5. Avoid jargon
6. Avoid reflective questions
7. Questions should be relevant in all or most contexts
8. Unambiguous wording
9. Avoid double negations

4.2.3. Response scale options

Different response options exist depending on the type of question. For continuous variables, such as affective intensity or sleep quality, broadly

speaking, two different types of scales (and combinations) can be used: continuous scales (visual analogue scales; VAS) and discrete scale options such as Likert scales. These scales can be unipolar or bipolar if two concepts are assumed to be negatively related. Note that 7-point Likert scales are relatively common in ESM research (Hall et al., 2021). For categorical variables such as current activity type or the binary assessment of the occurrence of behaviour, multiple or single-choice response options are most appropriate. Bivariate data can be collected using a 2-D grid in which two theoretically related variables are assessed simultaneously. Examples include dyadic behavioural or emotional data and the valence and arousal dimension of core affect. In addition, intensity drawings on a 2-grid have been suggested recently to assess temporal fluctuations in affective experiences (Cloos et al., 2024). Finally, text data or voice recordings can be collected, allowing open-ended responses (Hoemann et al., 2023). This can be used to collect certain thoughts, remarks or specifications of previously reported answers.

4.2.3.1 Effects on responses

It has been suspected that the choice of response scales influences responses (e.g., distribution, precision) and their psychometric properties. Continuous scales have an intuitive advantage as they theoretically allow higher precision than discrete scales. Sometimes participants might, for instance, want to assign a score between two answer options, which is not possible when discrete scales are used. However, findings from cross-sectional questionnaire research suggest that discrete scales may be more user-friendly: in one study, they were found to lead to shorter response times and lower dropout rates than continuous scales (Couper et al., 2006). In ESM research, findings on differences between these two scale types have been mixed. In an analysis of seven ESM datasets, researchers detected higher rates of multimodality associated with VAS compared to Likert scales, providing tentative evidence in favour of Likert scales (Haslbeck et al., 2023). However, a first experimental ESM study into the differences between VAS and Likert scales found largely similar characteristics between the two scales but some first evidence in favour of VAS scales as they were found to result in higher correlations with external measures and reduced floor effects (Haslbeck et al., 2025). The mixed previous findings underscore that more research is needed to

establish differences between the response scales. Even within one scale category, small variations in the scale can have effects on the collected data. The use of ticks on visual analogue scales has, for instance, been found to influence the distribution of responses in cross-sectional questionnaires (Matejka et al., 2016). With continuous scales, the use of a slider leads to different responses than with scales where participants click to select a response (Funke, 2016). However, most differences detected seem to be subtle, and previous research suggests that findings obtained with different scale types are largely comparable.

In any case, if possible, the variety of response scale used should be kept consistent throughout the questionnaire to avoid confusion for participants. In addition, as highlighted in ESM-Q, care should be given to formulate a clear response scale with anchors that match the question (Eisele et al., 2024).

4.2.3.2 Using single-item measures or combining multiple items into a scale

When developing a new ESM measure, researchers need to decide whether their construct can be captured with a single item or whether they want to combine multiple items into a scale. While the use of single-item measures was long discouraged, they have experienced a recent revival in psychometric literature (Allen et al., 2022). The most salient advantage of single-item measures is their potential for reducing the questionnaire length and participant burden. In cross-sectional questionnaires, research has documented that participants can react negatively to repetitive scale measures (Allen et al., 2022). Given research indicating a link between shorter ESM questionnaire length and data quality and quantity (Eisele et al., 2020; Hasselhorn et al., 2022), limiting participant burden by reducing the number of items is an important consideration in ESM studies. In addition, recent ESM research indicates that in some cases the predictive power of single-item measures may outperform that of their multi-item counterparts (e.g., Cloos et al., 2023). However, depending on the target construct, single items may not be able to cover its full breadth. For instance, momentary expressions of personality may be characterized by numerous different behaviours that cannot be summarized in a single item. In such a case, combining multiple items into a scale may be more appropriate. When doing so, the uni-dimensionality of the scale needs to be considered, i.e., whether the items indeed measure a

single construct (McNeish, 2022). Finally, the choice of using a single-item measure or a scale has consequences for how the reliability of the measure can be evaluated (approaches based on internal consistency do not apply to single-item measures; see paragraph on reliability). Studies that compare both kinds of measures can help determine the added value of using multiple items to assess a given construct (see Cloos et al., 2023).

4.3 Constructing a questionnaire

Once the different (sets of) items that we want to include are identified, they then need to be assembled into a coherent questionnaire. We will discuss different considerations to make regarding the order of the questions, the length of the final questionnaire and the inclusion of control questions.

4.3.1 Order of questions

Regarding the order of the questions, it has been suggested to start with the most transitory constructs (thoughts and feelings), and finish with questions about context that are not likely influenced by what was asked before (e.g., location, company; Palmier-Claus et al., 2011). Since participants may sometimes only fill in the beginning of the questionnaire, a possible solution can be to ask the most important questions first. However, generally, participants seem to either not respond altogether or fill the whole questionnaire; partial responses are unusual (Silvia et al., 2013).

Another decision to make is whether participants should always be asked the questions in the same order or whether the order of questions should be randomized. While keeping the order constant can make the assessments feel repetitive and possibly boring, it could also allow participants to answer faster, which might reduce burden. However, it is also possible that questions influence answers to subsequent questions. For instance, a participant perceiving questions about symptoms as confronting could induce negative mood and influence how they answer the remaining questions. Randomizing the order of questions can prevent these systematic sequence effects from introducing bias to the data. If a researcher chooses to randomize the order of items, it is also important to consider whether groups of items (e.g., all items measuring affect) should

be randomized separately or whether all items can be mixed irrespective of their category. An argument for keeping items of the same group together is that participants might find it easier not to change between topics too often. Besides changes between topics, it might also be preferable to list questions with the same timeframe subsequently in the questionnaire. For instance, if a questionnaire contains both momentary questions and questions about the time since the last assessment, it might be easier not to jump between timeframes but to keep items using the same timeframe together (e.g., ask all the momentary items first and then questions about the time since the last beep). Similarly, it might be preferable to group items scored using the same scale to avoid participants having to switch between scales and potentially make scoring errors. For instance, while switching from a unipolar Likert scale (e.g., 1 to 7) to a multiple-choice list should take minimal effort, frequent switching between a unipolar and bipolar (e.g., -3 to +3) Likert scale might slow participants down or confuse them.

4.3.2 Length of questionnaires

It is important that the questionnaire is as short as possible so that interruptions to the participants' daily lives are minimized (see also Chapter 3). There are currently no clear-cut rules for how many items an ESM questionnaire can consist of. In ESM literature, questionnaire lengths range from 1 to more than 100 items (Morren et al., 2009; Ono et al., 2019; Vachon et al., 2019). A general guideline is that questionnaires should take no more than three minutes to fill and preferably less (Kimhy et al., 2012). When designing the questionnaire, a balance needs to be found between gathering sufficient information and not overburdening participants with an overly lengthy questionnaire. It is expected that data quality and quantity diminish with longer questionnaires. This is supported by a recent experimental study in which a 60-item ESM questionnaire was associated with higher perceived burden, lower compliance, and increased self-reported careless responding than a 30-item version of the same questionnaire (Eisele et al., 2020; Ulitzsch et al., 2024). Two analyses of a second experimental study discovered signs of lower data quality and increased use of response styles associated with longer ESM questionnaires (Hasselhorn et al., 2022; Hasselhorn et al., 2024).

Likewise, a meta-analysis also found that shorter diaries were associated with higher compliance rates (Morren et al., 2009). Other

meta-analyses have not found a relationship between questionnaire length and measures of data quality and quantity (e.g., Ono et al., 2019; Vachon et al., 2019). However, it is important to note that meta-analyses on this question are limited by the variation of questionnaire lengths in included studies. In addition, design choices are often adapted to each other in published studies, making it hard to make firm conclusions about the effect of a single design choice. Ways to reduce questionnaire length include branching questions that are relevant only in certain situations. For example, the question 'Do you feel comfortable in this company' can only be shown when the participant answered 'No' to the question 'Are you alone?' An important consideration, however, is that branches of questions should, if possible, be equally long. If answering yes to one question leads to 10 additional follow-up questions while responding with no is only followed by one question, participants might quickly learn to avoid the first option. Other forms of adaptive testing, such as adjusting content depending on previous input and adjusting frequency testing, can also be useful. In the former, if a participant indicates strong negative feelings, extra follow-up questions further explore the nature of these feelings. This reduces the length of the questionnaire if the process of interest is absent. To illustrate the latter, if a researcher is particularly interested in the response to stressful events, more frequent assessments could be triggered after the participant has indicated experiencing a stressful event. This allows investigating moments of interest in higher temporal resolution while reducing burden at other times. Another alternative is the use of a planned missing design, where only some items out of a larger item pool are selected per assessment moment (Silvia et al., 2014). To illustrate, we could have a pool of four items measuring positive affect. Instead of assessing all these items every time, we could measure a randomly chosen subset of three items per assessment moment. This reduces the burden for participants while still assessing all the items at the price of a slight increase in the standard errors of estimates.

The goal of the study also has consequences for the construction of the questionnaire. If the goal is to assess without intervening, it can be good to 'hide' items in longer questionnaires. For instance, it might be preferable to include items about positive states such as happiness in a questionnaire even if one is interested solely in symptoms such as depressed mood. This might help to prevent participants from focusing more on their symptoms as a result of the ESM questionnaire and becoming more reactive to the

method. If, on the other hand, one wants to intervene using the questionnaire, it can be better to emphasize items and to have few distracting items in the questionnaire.

4.3.3 Control questions

Even if the construct of interest is measurable with relatively few items, it can be beneficial to include other items for several reasons. Firstly, it is important to assess sufficient information since lurking variables might otherwise distort results. Control questions such as 'Are you in pain?' or direct reactivity questions (e.g., 'Did this question bother you?') can help to rule out alternative explanations for findings. Additionally, accuracy checks can be included in the questionnaire to detect careless responding. These checks can, for instance, consist of directed response items ('Please select answer option 2'), an item with a verifiable answer ('2 + 2 = ?') or an exact repetition of an item in the same questionnaire ('How happy are you?' – 'How happy are you?'). If careless responding is detected, different approaches can be taken. A conservative approach is to exclude data of the participant from all analyses. A researcher can also choose to exclude only data from the specific assessment moment where inattentiveness was detected. Alternatively, sensitivity analyses can be conducted to see to what extent including the careless responders changes results (for a detailed discussion, see Edwards, 2019).

4.4 Assessing measurement quality

Once the items and questionnaires are developed in line with the guidelines above and have passed a first face-validity check by yourself and colleagues, it is important to evaluate them systematically. Broadly speaking, measurement quality refers to the ability of the item(s) to either predict a relevant outcome or serve as proxies for the underlying construct that the researcher is aiming to measure. In this chapter, we focus on evaluating how well items perform in measuring an underlying construct. For instance, if a researcher wants to measure anger, this means that we want to know to what extent the latent construct ('momentary anger') is captured by the items that we designed (e.g., 'I feel irritated', 'I feel annoyed', 'I feel upset', etc.). The brevity of ESM questionnaires makes

it even more critical that the items are carefully designed and selected. In the following, we give an introduction to how items can be tested thoroughly, first by expert review and in small pilot studies then in larger samples, to gather further evidence for the validity and reliability of the measure.

4.4.1 Expert review

The evaluation of newly developed items or questionnaires by expert judges is a crucial part of the measure development process (Boateng et al., 2018; Simms, 2008). Ideally, the reviewing experts were not involved in the development of the items and are familiar with the target construct (Boateng et al., 2018). Experts can evaluate the validity of items and help to identify problems in wording or instructions. The review process can be guided by a formal evaluation tool such as ESM-Q. As briefly introduced above, this set of 25 quality criteria for ESM items was recently developed in a Delphi study with ESM experts and can be used to assist the review of newly developed items (see Eisele et al., 2024).

4.4.2 Pilot testing

An equally important part of the measurement development process is the pilot testing phase (see also Chapter 3). Many errors can be identified by testing a measure in the field. We believe that it is good practice to first pilot test the questionnaire extensively and then ask others to use it. Ideally, pilot tests should also be conducted with subjects from the target population. Questions that need to be addressed are: Is the wording ambiguous? Are the categories clear? Do some questions apply less to some situations? Do the questions and response categories cover all important aspects of the participants' real-life experience? Below, we illustrate the value of piloting using a real pilot study example.

Example 1: The item 'Are you alone?' was identified in pilot tests as being highly ambiguous. While some researchers identified sitting in their office with colleagues (while not interacting) as being alone, others would classify this same situation as not being alone. Being at home and alone in a room while others were present in the same house was another situation that led to different responses. The presence of pets can also be classified as company by some people while others would indicate

that they are alone despite their pet being present. Our solution was to explicitly brief participants to be able to classify alone and in company situations ('With being alone we mean that no other people are present in the same room or space. This means that if you walk down the street and are surrounded by strangers that you don't interact with, you are not alone. If you are at home in your own room, but someone else is in the house, but not in the same room, this would mean that you are alone').

To dig deeper into the meaning of questions, cognitive interviewing techniques can be useful (see Schorrlepp et al., 2025 for a discussion of cognitive interviewing in the ESM context). These approaches can, for instance, help to understand ways that participants approach more complex questions (e.g., 'Since the last beep, have I tried to distract myself from my feelings?') and what meanings participants give to response options. Even during data collection, interviews can help to identify problems that did not appear during pilots or facilitate the interpretation of results. An example of such a problem is given below:

Example 2: We used the item 'I can do this well' in combination with an item assessing the current activity. This item aims to measure a form of activity-related stress. However, in combination with certain activities (such as showering or watching television) it can turn out strange, as was remarked by some participants in a recent study. A 'not applicable' answer option might help to avoid this problem in the future.

4.4.3 Intra-class correlation

As mentioned above, it is important to check if the construct of interest changes within persons in a way that justifies assessing it frequently. The intra-class correlation (ICC) can be helpful in answering this question since it is a measure of how much variance of a variable is due to stable between-person differences (i.e., differences in a variable between people) as opposed to within-person fluctuations (i.e., differences in scores on the same variable within a participant). It can be calculated by dividing the between-person variance by the overall variance of a variable (between plus within-person variance). Very high ICCs indicate that the variable does not vary much within a person, which would mean that frequent measurements are redundant. Bolger and Laurenceau (2013) state that ICCs ranging from 0.2 to 0.4 are typical in ESM studies. An ICC of 0.4 for the item 'I feel irritated' would mean that 40% of the variance of this item

is due to stable differences between persons while 60% of the variance is due to momentary fluctuations within persons. See Chapter 9 for details on the calculation of the ICC.

4.4.4 Different forms of validity

When both the initial review by researchers and the pilot testing have been favourable, additional evidence for the validity of our measure can be gathered. While an expert review and short pilot test can be implemented in most research projects, evaluating the validity of a measure comprehensively will take more time and effort and may therefore require dedicated studies. Because we usually have no way of directly assessing the latent construct, we need to rely on other indices that can provide evidence for the validity of our measure. Such evidence can be based on the internal structure of our measure or on the relationships between our measure and other variables (a distinction based on the Standards for Educational and Psychological Testing; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Evaluating validity based on the internal structure of a measure typically entails testing the claim that items from a scale indeed measure a single underlying construct. Multilevel applications of factor analysis can be used to examine this structural validity or uni-dimensionality of a construct. If we, for instance, include 10 items that are all supposed to assess the underlying construct 'momentary anxiety', a one-factor solution should provide the best fit to these items, and all our items should load strongly on this single factor. If, however, we detect that our unidimensional construct is in reality made up of two sub-constructs, such as, for instance, feelings versus behavioural tendencies, we need to reconsider our conceptualization of 'momentary anxiety'. Additionally, if we find that our construct is not unidimensional and/or the loadings vary a lot across all items, the calculation of a sum score of the items might not be sensible. In ESM data, the uni-dimensionality of a measure needs to be evaluated at both the within- and between-person level. Importantly, the internal structure of a measure can further vary both across persons and within persons over time. While multigroup, multilevel factor analysis allows comparing static measurement invariance across groups of people, newly developed methods such as latent Markov factor analysis

(see Vogelsmeier et al., 2024) and Dynamical Structural Equation Modeling (McNeish & Hamaker, 2019) can be employed to evaluate measurement invariance over time in an ESM setting. With these methods, the factor structure is allowed to vary across moments and between individuals, providing a way to deal with measurement non-invariance in panel data (Adolf et. al., 2014).

When evaluating validity based on the relationships between a measure and other variables, different types of relations can be considered. Constructs do not exist in isolation but are usually embedded in a theoretical net (also termed nomological network), which defines one construct's relationships with the other constructs in the net (Raykov & Marcoulides, 2011). Assessing the relationship between one measure and other measures in the nomological network is a form of evidencing external validity. A good measure needs to relate to other constructs in predictable ways. Like to the internal structure of a measure, the relationships between different constructs can be different at the within- and between-person levels (Schuurman, 2023). For example, individuals might rarely feel fear and anger at the same time, meaning that these two emotions might be relatively unrelated or even negatively related at the within-person/state level. It is, however, possible that individuals who frequently experience fear also frequently experience anger (e.g., because of a relatively high level of the personality trait neuroticism), meaning that the two emotions are positively related at the between-person/trait level. Such a discrepancy between within- and between-person relationships was observed by, for instance, Borah and colleagues (2018) for items measuring subdomains of aggression. In this example, a two-factor structure led to the best fit at the within-person level while a single factor was sufficient to explain variance at the between-person level. Eisele, Lafit, et al. (2021) give another example of different factor structures for measures of affect depending on the level of analysis. They identified a two-factor structure of affect at the between-person level while five factors were needed to explain variation within individuals. These examples highlight that a clear conceptualization of the construct at both within- and between-person level is necessary.

When a construct is expected to relate in certain ways to other ESM measures, correlations between the measure being validated and these measures can also be used as indices of convergent and discriminant validity (Shrout & Lane, 2012). Again, correlations can be different at the

within- and between-subject level. Finally, validity can also be assessed by investigating a construct's relationships with other, non-ESM measures. We can, for instance, investigate how ESM measures relate to non-self-report (e.g., Maher et al., 2018) or retrospective measures of the same construct (e.g., Forkmann et al., 2018). However, it is important to keep in mind that discrepancies between different assessment methods do not necessarily imply that the ESM measure is not valid as the different measures might simply tap into different concepts (Dubad et al., 2018; Robinson & Clore, 2002). Again, only a careful conceptualization of the construct and its expected relationships with other constructs can help researchers decide on validity and on what information one is most interested in. Furthermore, validity of a measure is dependent on the sample and therefore needs to be considered anew for every conducted study (Shrout & Lane, 2012).

4.4.5 Reliability

Another aspect of measurement quality is the reliability of the instrument, which can be operationalized as the proportion of variance in responses that is due to true variance as opposed to random error. We usually do not know how much variance of our measure is real and how much is due to random error. However, if we have multiple repeated measures of the same construct, we can estimate these proportions.

In cross-sectional research, when the focus lies on determining the reliability of assessments of between-person differences, there are different ways of obtaining these repeated assessments which lead to different types of reliability. When the whole instrument is administered twice, the test-retest reliability can be calculated. When the same construct is rated by two raters, the interrater reliability can be assessed. When we have multiple items that measure the same underlying construct, internal consistency can be used as a measure of reliability. Of course, a particular feature of ESM data is that even though we have repeated measures of a construct, the construct is expected to change. Therefore, reliability is typically operationalized as internal consistency, and it is necessary to distinguish the reliability of assessing between-person differences and the reliability of assessing within-person change in the construct. Different ways of calculating the latter type of reliability have been proposed in ESM literature (see for an overview Castro-Alvarez et al., 2025). These proposed

methods require multiple items measuring the same construct at every assessment. Nezlek (2017) describes how estimates of between-person and within-person level reliability can be obtained. In the approach Nezlek discusses, a three-level model is fitted with individual items nested in measurement occasions, and occasions are nested in persons. Within-person reliability can then be calculated based on the random effect of occasion and of item (i.e., the residual variance) that are in the output of standard multilevel software. Along similar lines, Cranford and colleagues (2006), suggested using generalizability theory to obtain different measures of reliability utilising standard multilevel software. However, both approaches assume that all items are equally predictive of a construct, which might not be a realistic assumption in many cases. Therefore, numerous alternative approaches have been introduced that make use of confirmatory factor analysis (CFA) to calculate omega as a reliability measure and take into account that items might show different relations with an underlying construct. A confirmatory factor analysis is a method that can be used to investigate whether items that are expected to measure the same underlying factor are indeed related in expected ways. It allows distinguishing the amount of variance in each item that is due to the common factor and the amount due to remaining error. Based on these different types of variances, a measure of reliability can be calculated. The approaches to calculate reliability with help of a CFA differ in the details of the computation of the factor structure. The CFA, can, for instance, be applied in a two-step procedure by first fitting a multilevel model to obtain within- and between-person variance-covariance matrices (Goldstein, 2011; Viechtbauer, 2017) (for an application, see Forkmann et al., 2018). Separate CFAs for each level can then be performed with the obtained variance-covariance matrices, and results are then used for the calculation of the two omegas at between- and within-person levels. In another approach, the omegas for each level are estimated with help of structural equation modelling, which allows fitting a multilevel CFA (Bolger & Laurenceau, 2013; Muthén & Asparouhov, 2011). Alternatively, when sufficient time points per person are available, reliability measures can also be calculated per individual. This means that, analogous to the approaches above, CFAs and omegas can be calculated separately per person, and the distribution of reliabilities can be inspected (Fuller-Tyszkiewicz et al., 2017; Shrout & Lane, 2012). In recent years, alternative methods were suggested that allow calculating person- (and item-) specific

reliabilities, namely mixed-effects trait-state-occasion modelling (Castro-Alvarez et al., 2022) and multilevel dynamic factor modelling (Xiao et al., 2023). Methods also differ in the extent to which they take the auto-correlated nature of ESM data into account.

All methods above focus on internal consistency measures of reliability and are therefore limited to the use of multi-item measures. Given the popularity of single item measures in ESM studies, researchers have also suggested ways to assess test-retest reliability of a measure in an ESM setting. For instance, Schuurman and Hamaker (2019) suggested measures of reliability based on different components of multilevel vector-autoregressive models, and Dejonckheere et al. (2022) suggested calculating the reliability of single items using repeated assessments of the same item at the same or a close time point.

While the calculation of reliability in ESM studies is an active research area, we hope to have given the reader a first glance at currently available methods. Since a detailed description of the individual methods is beyond the scope of this chapter, we refer the interested reader to the references above and an overview paper by Castro-Alvarez et al. (2025).

4.5 Beyond self-report

Self-report measures will always remain susceptible to certain errors. While compared to other methods ESM greatly reduces errors due to recall bias, there are other forms of biases that cannot be eliminated. One example are biases due to social desirability. Some states or behaviours might be more sensitive to such distortions than others. Dietary intake, for instance, is thought to be prone to socially desirable responding (Schembre et al., 2018). Measures can be taken to reduce the influences of these errors (such as stressing anonymity during participants' training). However, in some cases, it might be a better option to use objective measures instead of self-report. Additionally, even when participants are honest, self-report remains subjective, and when one is interested in an objective parameter, objective measures are preferable. Many alternatives to self-report have appeared in recent years. Examples include the use of actigraphy to assess sleep and movement, sound snippets to assess social interaction, pictures to assess dietary intake, GPS to assess location, and sensors to assess heart rate, blood pressure or body temperature (Conner & Mehl, 2015). Most

smartphones have built-in sensors that can be used to passively gather a wide range of objective measures. These types of passive assessments can also reduce the burden placed on participants. However, the quality of data from these sensors can be low. A combination of self-report and objective assessment in daily life can be especially powerful in creating a comprehensive conceptualization of a construct. This topic is discussed in more detail in Chapter 13.

CHAPTER 5

Ethical Issues in Experience Sampling Method Research

Olivia J. Kirtley

Along with the myriad possibilities that ESM brings, there also come additional ethical considerations and responsibilities. Just as there are currently few methodological 'gold standards' for ESM research, there are also no ethical gold standards specifically for ESM research. However, a consensus statement on ethical and safety practices for digital monitoring studies of suicidal behaviour has been published (Nock et al., 2021), which may also provide some pointers for ESM studies more broadly. Mostly, current ethical practices have evolved organically from those of previous ESM studies as well as the changing requirements of data protection legislation such as the European Union's General Data Protection Regulations (GDPR). A review by Capon and colleagues (2016) highlighted several relevant ethical considerations for ESM research, including data storage and transfer, data ownership, user anonymity, access to technology and communication of clinically relevant results. Many of these ethical considerations are not unique to ESM research, for example, obtaining informed consent and maintaining participants' privacy, however, the real-time and remote nature of ESM data collection and the high level of detail within ESM data can present additional ethical challenges. Moreover, advances in technology that allow the collection of passive data mean that even seemingly routine aspects of ethically conducting research, such as informed consent, require additional thought. For a discussion of privacy issues in mobile sensing research, see Hong (2023).

In this chapter, I focus on ethical considerations of specific relevance to ESM studies in the broad area of mental health research including clinical and non-clinical samples. I cover five key aspects of ESM research ethics: inclusivity, privacy and consent, responsibility for intervening when the researcher is concerned about a participant, participant burden, and reactivity.

5.1 Inclusivity in research

While the early days of ESM research saw participants recording their responses to momentary questionnaires using a pen, paper and a digital watch, the smartphone is the contemporary ESM researcher's method of choice for gathering data (Myin-Germeys et al., 2018; van Berkel et al., 2018). Smartphone use is now almost ubiquitous, with almost seven billion smartphone-users worldwide (Statista, 2025). Furthermore, people who are frequently underrepresented in research – including Black and Hispanic individuals, people who have low incomes and a lower level of education, and those living in rural areas – are among some of the most 'smartphone-dependent' individuals, meaning that they use their smartphones as their primary means of contact and internet access (Pew Research Center, 2025). Smartphone ownership is also high among adolescents (van Roekel et al., 2019), and it is increasing among individuals with mental health problems (Torous, Wisniewski, et al., 2018). Access to smartphones and the internet is, however, not the case everywhere or for everyone. The percentage of individuals owning a mobile phone is lower in low-income countries (47%), lower-middle-income countries (70%) and upper-middle-income countries (84%) relative to that in high-income countries (94%) (ITU, 2025). Furthermore 2023 statistics show that women worldwide and especially women in low- and lower-middle-income countries were less likely to own a smartphone (ITU, 2025). This creates a steep 'digital divide' – the term given to the gap between individuals with access to digital technology, e.g., the internet, smartphones, computers, etc., and those without (Steele, 2019). This means that for some individuals, especially those in low- and middle-income countries, the digital divide in terms of smartphone ownership and internet access may be a barrier to participation in ESM research, and consequently to researchers from low- and middle-income countries conducting ESM research.

Even where smartphone ownership per se is not a barrier to participation, usability and compatibility issues with ESM apps may still hamper inclusivity efforts. When adapting our large-scale adolescent mental health study (SIGMA; Kirtley, Achterhof, et al., 2021; Achterhof et al., 2025) to enable completely remote ESM data collection during the COVID-19 lockdown, we found that a small proportion of adolescents had 'hand-me-down' smartphones from parents or older siblings, which caused compatibility issues with the ESM app we were using. An additional

issue is that some ESM apps do not work on all makes and models of smartphones. ESM apps can also be battery and data 'hungry', which, for individuals who are smartphone-dependent, may create a barrier to these apps' download and use. For further discussion of this and other software and hardware issues, see Chapter 6.

The bottom line is that researchers should bear in mind whether their target population is likely to experience issues with smartphone ownership, up-to-date functionality or internet access. One way of addressing this is to loan participants a smartphone for the duration of the study. This ensures that potential participants are not deterred from taking part due to lack of a smartphone and also circumvents possible compatibility issues between the ESM app and the participants' smartphone. Lending participants smartphones does of course increase the cost of conducting the research, so it should be considered as early as possible when planning an ESM study, i.e., at the point of applying for funding.

5.2 Privacy and consent

In some ESM studies, participants are asked to report on behaviours that are heavily stigmatized, e.g., suicidal behaviour, or even illegal, such as use of illegal substances or underage drinking. Anonymity and confidentiality are often key in studies aiming to collect information on potentially sensitive topics (Tourangeau & Yan, 2007). Management of expectations is crucial, and this should be addressed in the informed consent and participant information materials. It should be made clear to participants whether or not their responses will be monitored and if so, how often and by whom. When ESM research involves adolescents, expectations regarding privacy and confidentiality should also be made clear to parents and adolescents themselves. It is important that researchers are aware of their legal duty of care regarding confidentiality of information disclosed to them during the study – especially by children and adolescents, and that this duty of care may also differ according to country or state.

In some instances of passive data collection, issues of privacy and consent also extend to those around the participant, especially when ESM involves collecting audio or photo/video samples. For example, these issues may arise if data collection utilizes an Electronically Activated Recorder, which captures short snippets of audio during participants' everyday

life (Mehl, 2017). Whilst informed consent is necessarily obtained from study participants, individuals around the participant may not have had an opportunity to consent yet may inadvertently be sampled during the study. Robbins (2017) tackles these legal and ethical issues in an excellent paper and makes several practical suggestions to ensure bystanders are informed they may be recorded, e.g., participants can wear a pin badge saying, 'this conversation may be recorded!' Mehl (2017) also discusses a multi-step process for protecting participants' privacy in research utilizing a Electronically Activated Recorder, including allowing participants to review their collected data before making it available to the research team. A 'privacy by design' approach to passively collected data is strongly encouraged (Hong, 2023).

The previous two considerations concern issues of consent and privacy prior to or during data collection, but what about after data collection has occurred and a researcher would like to share those data? Open data is an excellent way of increasing transparency and enabling analytic reproducibility (Munafò et al., 2017), yet it remains an especially thorny issue in mental health and medical research due to concerns about privacy and participant identification (see Walsh et al., 2018 for further discussion). If a researcher would like to share ESM data, this must of course be de-identified, but importantly participants' consent must be obtained a priori. Fortunately, Soderberg and colleagues (2019) have created resources and example texts for informed consent forms and institutional ethics boards, which can be used to facilitate obtaining consent for data sharing. These are freely available on the Open Science Framework (osf.io/g4jfv). Even if permission for data sharing is explicitly provided in the informed consent form, other guidelines regarding country- or region-specific privacy legislation will also need to be followed, e.g., GDPR in Europe, and potentially also those of funders and institutions. We advise any researchers planning on sharing data to seek advice from the relevant privacy, legal and ethical departments at their institution as early as possible – ideally, at the study design phase – including for reuse of existing data.

Although many people equate open science with open data, there are myriad ways in which to increase the transparency, reproducibility and replicability of your ESM research. One way of achieving this is pre-registration (or post-registration in the case of existing data; Benning et al., 2019), whereby researchers create a locked, time-stamped, un-editable plan of their research questions, hypotheses and analysis plan before data

collection (or before data access and analysis in the case of post-registration). To aid ESM researchers in pre- or post-registering their research, we have developed a specialised registration template for ESM research, which is accompanied by an open-access tutorial paper (Kirtley, Lafit, et al., 2021). More recently, a template for registration of studies using passive sensing data has been developed (Langener, Siepe, et al., 2024).

Registered Reports are an enhanced form of registration whereby introduction, methods and a detailed analysis plan for a study – essentially, the proposal for a study – are submitted to a journal for peer review prior to data collection (or access in the case of existing data) (Chambers & Tzavella, 2022). The plan for the study is then peer-reviewed, and once any peer-reviewer comments have been implemented, the paper receives Stage 1 In Principle Acceptance. This means that, providing the researchers now conduct the study exactly as they said they would in their Stage 1 submission, the journal commits to publish the full article irrespective of the direction and statistical significance of the results. Registered Reports are designed to reduce publication bias and wastages of resources (on flawed ideas) but also to enable researchers to receive reviewer feedback at the point when it is most useful: prior to data collection and analysis. Following data collection and analysis, researchers write up the results and discussion sections and combine these with the introduction, methods and results sections, as per a typical journal article, and submit the manuscript for further peer review. After successful peer review of the now complete manuscript, the paper is accepted and published. For an example of a Registered Report from our own research group, using existing ESM data on self-harm thoughts and behaviours in daily life, see Janssens et al. (2023). For examples of other Registered Reports using ESM/EMA data, see King et al. (2025) and Dora et al. (2024).

As ESM research often entails a large number of researchers working with large ESM datasets to answer many different primary research questions over time; this raises challenges for registration and Registered Reports using ESM data due to the need to limit data access and reduce data-dependent decision-making. To confront these challenges, following Scott & Kline (2019), the Center for Contextual Psychiatry at KU Leuven developed a data checkout system: Data cuRation for OPen Science (DROPS; Kirtley et al., 2020; Kirtley, 2022), which enables researchers to check out only the specific variables they will use for a particular, post-registered

project and to receive a time- and date-stamped receipt showing who had access to what and when.

Aside from pre- and post-registration, Registered Reports, and controlling and documenting data access, ESM researchers wishing to enhance the transparency and replicability of their work can also share their study materials, including participant forms and briefing documents, analytic code, and also measures by, for example, contributing ESM items to the ESM Item Repository (Kirtley, Eisele, et al., 2025; www.esmitemrepository.com). For a general discussion of open science in ESM research see Löchner et al. (2025). See Wrzus and Schoedel (2023) for a review of considerations around transparency in mobile sensing research.

5.3 Real-time data, real-time responsibility?

While the existing body of literature suggests that taking part in ESM research is unlikely to result in participants experiencing adverse effects, in some cases, there may still be an increased risk that an adverse event could occur during an ESM study. Here, we are primarily referring to studies where participants are selected on the basis of engaging in behaviours that may put them at risk – for example, suicidal behaviour, eating disorders or substance use. In these cases, the purpose of the study is likely to attempt to capture these behaviours. Prior to commencing data collection, researchers should carefully consider three questions: 1) Will they intervene? 2) How will they intervene? 3) Who will intervene? These questions have been considered in an excellent review by Jacobson and colleagues (2020), and we discuss each of them below. Based on my personal expertise, in this section, I mostly provide examples from research on self-harm and suicide. See Hoelscher et al. (2025) and Kiekens and colleagues (2021) for a detailed discussion of ethical issues involved in ESM research on Non-Suicidal Self-Injury and Kirtley, Sohier, et al., (2025) for considerations regarding ethical issues in ESM research on self-harm.

First, the researcher must decide whether they will intervene and if so, under what circumstances. Perhaps, 'hot questions' will be used. These are items within the ESM battery whereby if a participant endorses a particular response to a question, the researcher is immediately alerted and, if necessary, can take action – for example, they might do so if a participant indicates a response of >5 on a 1–10 scale for intensity of

suicidal ideation. It is worth noting, however, that asking questions on potentially sensitive topics does not in and of itself confer a necessity to intervene. Deciding upon a practically meaningful and sensible threshold for intervention is a critical question, to which extant literature can provide few concrete answers, especially given the widely acknowledged inability to accurately predict suicide risk using risk assessment tools (e.g., Franklin et al., 2017; Quinlivan et al., 2017; Steeg et al., 2018). Research by Kleiman and colleagues (2018) demonstrating multiple distinct 'phenotypes' of suicidal ideation underscores this challenge; some individuals report consistently high levels of suicidal ideation whereas some report consistently low levels and others report highly variable levels of suicidal ideation. A study sample may include individuals with the full range of these behavioural phenotypes, meaning that a 'one size fits all' threshold is likely to miss some individuals in distress whilst 'over-monitoring' others. Consequently, it is also highly unlikely that such thresholds are portable across studies, especially with different populations. Co-designing thresholds for intervention may be the optimal method for ensuring that safeguarding works for participants as well as researchers. Research has indicated that monitoring of ESM responses is desirable both for researchers (Nock et al., 2021) and individuals with lived experience of mental health problems (Dewa et al., 2019). If researchers intend to monitor responses as part of their safety protocol, the frequency with which responses will be monitored, the circumstances under which intervention will be initiated and the nature of the resulting intervention should be clearly explained to participants in the participant information documents as well as during briefing to align with principles of informed consent (Hoelscher et al., 2025). Equally, if responses will not be monitored, this should be clearly communicated to participants. Expectation management is critical.

Second, having decided an intervention will be initiated, the researcher must consider what form this intervention will take. In some studies, endorsing a particular response to a 'hot question' causes instant, indirect intervention via a pop-up window displayed on the smartphone screen, providing the participant with details of local and national support services. Kleiman and colleagues (2017) used this method in their anonymous study of adults with suicidal ideation, and we have also used this in our own work within the SIGMA study (Kirtley, Achterhof, et al., 2021) where ESM responses were provided pseudonymously. This is a relatively

low-threshold intervention, requiring little in the way of additional staffing or resources, and provides instant support information to participants who may be in distress. Another more intensive method of intervention was used in Glenn and colleagues' (2020) study of suicidal ideation in adolescents; a member of the research team checked participants' ESM responses twice per day and made telephone contact with participants within 24 hours if responses gave cause for concern. In the DAILY study on NSSI (Kiekens et al., 2024) and the SCOUT study on suicidal thoughts and behaviours (Kirtley, Claes, et al., 2025), carried out by members of our research group, an intensive real-time safety alert protocol was used whereby the research team contacted participants by phone upon receiving a safety alert triggered by a combination of responses to several ESM items. Concerns have been raised regarding whether responding to safety alerts may affect the data collected by, for example, causing participants to alter their responses to avoid being contacted by researchers or by the researcher contact acting as an intervention to reduce suicidal ideation (Bentley et al., 2024). Changing responses was more common among adolescents than adults, but evidence regarding intervention effects of the safety protocol was less clear-cut (Bentley et al., 2024). As such, empirical investigation of the effects of safety procedures on data quality, quantity and participants' experiences is an important area for future research. For an excellent discussion of different intervention and safety monitoring protocols for ESM studies of suicidal behaviour, see Bai and colleagues (2020), and for the same on self-injurious thoughts and behaviours, see Bentley et al. (2021).

Third, once a researcher has decided upon a more intensive intervention, who will be responsible for delivering it? These more 'hands-on' monitoring and intervention procedures require adequate staffing to monitor participants' responses, which becomes challenging as sample size increases. It is also essential that research staff are adequately trained in how to support someone experiencing an acute mental health crisis (Hoelscher et al., 2025; Kiekens et al., 2021). Not all ESM researchers studying mental health are clinicians, so it is important that appropriate specialist training is provided. Another strategy is to collaborate with a clinician who agrees to be the person to contact the participant if a participant's responses give cause for concern. It is also advisable to collect the contact details of participants' clinicians at study onboarding so that their own trusted clinician can be reached by the research team should

they become distressed. This may also avoid interrupting participants' continuity of care, which could occur if researchers were to intervene in a crisis; the participants' clinician would be better placed to treat but may be unaware of such crises. Participant anonymity does limit the possibilities for researchers to intervene should participants indicate that they are about to or have already engaged in potentially risky behaviours. Sometimes individuals may only be willing to participate in ESM studies where anonymity and thus non-intervention are guaranteed, especially regarding sensitive topics. In these cases, the provision of extensive support information and researcher contact details is even more important, and supportive pop-ups during or after ESM completion may act as a remote intervention.

5.4 Participant burden

The feasibility and acceptability of ESM research has been demonstrated time and time again, across a wide range of different populations, including individuals experiencing psychosis (Kasanova et al., 2018), Borderline Personality Disorder (Houben & Kuppens, 2019), depression and anxiety (Schoevers et al., 2020), bipolar disorder (Schwartz et al., 2016), suicidal ideation (Glenn et al., 2020; Kleiman et al., 2017), non-suicidal self-injury (Burke et al., 2021; Kiekens et al., 2020; Victor et al., 2019), eating disorders (Stein & Corte, 2003), alcohol misuse (Poulton et al., 2019), high-risk poly drug-use (Roth et al., 2017), and chronic pain (Kratz et al., 2017). Yet, ESM research is an *intensive* longitudinal method due to participants' provision of multiple responses per day over a period of days or even weeks, and this can also make for an intense experience for participants. The 'cost', i.e., burden, to participants is an essential ethical consideration in ESM research and must be balanced with the benefits to participants. Participants may experience burden if questionnaires are too long (Eisele et al., 2020), if beeps interfere with their normal daily activities or if they are also completing other measures alongside ESM (Bos et al., 2019).

Researchers and ethical committees are also frequently concerned about ESM research being too burdensome for so-called 'vulnerable groups', e.g., groups considered vulnerable based on mental health status, age or other factors. In our recent research with adolescents from the

SIGMA study, we found that young people who reported thinking about or engaging in self-harm in daily life reported slightly higher levels of beep disturbance – indicated by their response to the item 'this beep disturbed me' – than adolescents who reported no self-harm thoughts or behaviours in daily life (Kirtley, Sohier, et al., 2025). Moreover, at moments when adolescents' self-harm thoughts were more intense, they were more likely to report higher levels of beep disturbance. However, there were no differences in beep disturbance between adolescents with and without a lifetime history of self-harm thoughts and behaviours. Consequently, excluding individuals with a lifetime history of self-harm thoughts and behaviours from ESM research due to fears about studies being too burdensome appear misplaced, and a dynamic approach to considering participant burden and vulnerability in ESM research is preferred. For a full discussion of this issue, see Kirtley, Sohier, et al. (2025).

How, then, can researchers minimize participant burden in ESM research? Research by Eisele and colleagues (2020) demonstrated that participants perceived longer ESM questionnaires as more burdensome than shorter questionnaires, but there was no significant difference in participant burden as a function of sampling frequency, i.e., the number of questionnaires. Researchers looking to minimize participant burden while maintaining a sufficient sampling frequency should therefore ensure that questionnaires are kept short. See Chapter 4 for more information about ESM questionnaire development.

5.5 Reactivity

Numerous researchers have raised the question of whether being repeatedly asked about particular thoughts and behaviours may in fact induce those thoughts and behaviours (e.g., Myin-Germeys et al., 2009) or may cause participants to alter their behaviour (Palmier-Claus et al., 2011). This is also a frequently recurring concern of ethical committees. Low compliance rates may indicate reactivity in the form of elevated distress (Kleiman et al., 2017); however, this is difficult to ascertain without substantive examination of participants' reasons for missing prompts. Individuals who recently attempted suicide (Husky et al., 2014) or engaged in self-harm (Law et al., 2015) had lower compliance rates during seven- and fourteen-day ESM protocols, respectively, than individuals without a

history of suicide attempts. For individuals with recent suicide attempts, compliance was unrelated to study duration (Husky et al., 2014). No differences in compliance as a function of lifetime history of self-harm thoughts and behaviours were found in Kirtley, Sohier, et al.'s (2025) study with adolescents. Law and colleagues (2015) also found no significant effect of repeated daily questioning about suicide on individuals' suicidal thoughts and behaviours. An elegant study by Coppersmith and colleagues (2022) tested whether suicidal ideation was increased by being repeatedly assessed during ESM in addition to whether frequency of assessments was associated with changes in suicidal ideation. In this study, ESM assessment of suicidal ideation was not associated with changes in suicidal ideation intensity. Furthermore, assessment frequency was not associated with increased severity of suicidal ideation, and when suicidal ideation intensity increased, this was not associated with decreased responding to ESM questionnaires (Coppersmith et al., 2022). This study provides the strongest evidence to date that repeatedly asking about suicidal ideation has no significant iatrogenic effects. Yet a small proportion of individuals experiencing suicidal ideation may experience negative effects of ESM participation (Kivela et al., 2024). Future research should investigate who may experience negative effects from participation in ESM research on suicidal ideation and why. Beep disturbance, however, does not predict the presence or intensity of self-harm ideation in adolescents (Kirtley, Sohier, et al., 2025).

Other studies from the substance abuse field have even found that participants report a positive reaction associated with repeatedly being asked about their substance use and other risk behaviours, including increased introspection and awareness about both positive and negative behaviours (Roth et al., 2017). It may be that certain aspects of reactivity are highly specific to particular populations or question topics – an empirical question which emerging research is beginning to address (e.g., van Ballegooijen et al., 2016). In sum, while researchers and ethics committees are often concerned about reactivity to ESM in terms of iatrogenic effects, the existing literature suggests that these concerns are unfounded and that dynamic rather than static approaches to iatrogenic effects and participant vulnerability should be adopted (Kirtley, Sohier, et al., 2025). Further discussion of measurement reactivity in ESM research can be found in Chapters 3 and 4.

5.6 Conclusion

In this chapter, I have discussed five key ethical issues of specific relevance to ESM research and, where possible, provided potential options for addressing these challenges. Little research has substantively considered ethical issues within ESM research although there are some notable exceptions that demonstrate encouraging findings. The digital divide may represent a growing ethical issue for ESM research as more sophisticated apps and technologies are developed for passive monitoring. These inevitably lead to new challenges regarding privacy and ensuring participants sufficiently understand the ESM study in question. Some participants may incidentally experience heightened distress during an ESM study, and in these cases, it is imperative that researchers have a thorough plan in place for whether they will intervene, how, and who will be responsible for intervening. Managing participants' expectations, especially regarding intervention, is crucial and may also help to reduce the burden of taking part in ESM research. Recent research has also shown that burden is optimally reduced by keeping questionnaires brief. Finally, even though reactivity to ESM is a perennial concern of institutional ethics boards, especially when studying suicidal behaviours, existing research suggests this worry is often overestimated. Of course, this does not remove the need for a thorough safety protocol to ensure participants' well-being during ESM research participation. However, one-size-fits-all and static approaches to ethical issues in ESM research are inappropriate. Dynamic challenges require dynamic solutions, even in relation to ethical issues.

DURING: CONDUCTING AN ESM STUDY

CHAPTER 6

Experience Sampling Platforms

Jeroen Dennis Merlijn Weermeijer, Glenn Kiekens and Martien Wampers

Thus far, we have explored the types of research questions that can be addressed using the experience sampling method (ESM) and learned how to design a methodologically rigorous and ethically sound study. The next important step that needs to be taken is the programming, scheduling and delivery of the study's content. This is typically done using an Experience Sampling Platform that integrates and allows complex communication between the hardware (e.g., smartphone, wearable) and software components (e.g., dashboard, app). Numerous platforms already exist that allow researchers to operationalize their protocol in everyday life, making it difficult for researchers to decide which platform to use. Instead of attempting to give an exhaustive overview (which would quickly be outdated), this chapter (1) provides considerations on essential software and hardware components of ESM platforms, (2) discusses legal and practical challenges that may help guide the choice for a particular platform and (3) ends with providing a comparison of five excellent ESM platforms currently available. This does not mean that other platforms not included here should not be considered when deciding the right platform for your study. Instead, it aims to provide researchers with a starting point by highlighting commonalities and meaningful differences between platforms.

6.1 The online dashboard

Most central to each ESM platform is its online dashboard. This website can be accessed through a web browser and consists of multiple web pages for implementing ESM questionnaires, sampling schedules, enrolment of participants, data analytics and downloading of data. In what follows, each of these modalities is considered.

6.1.1 ESM questionnaires

ESM questionnaires measure thoughts, experiences and behaviour in real-time (see Chapter 4). While the design of an ESM questionnaire is extensively covered in Chapter 4, it is important to note that the questionnaire itself also needs to be programmable. For example, when some questions require a particular response scale option (e.g., multiple choice or slider) this scale must be available for use on the online dashboard that is used for the study. Across ESM-platforms, four basic question types are frequently used: modifiable slider questions, checkbox questions, radio questions and open questions. For modifiable slider questions, the dashboard ideally allows for the anchors and range to be freely adjusted. The slider can function as either a Likert scale (e.g., ranging from 1 to 7) or a continuous scale (e.g., ranging from 0 to 100). Checkbox and radio question types are used for multiple-choice questions. Checkbox questions allow participants to select multiple answer options whereas radio questions restrict the selection to only one response option. Finally, open questions allow for the possibility of receiving written qualitative feedback from participants.

Two important considerations are to be noted when considering more complex questionnaires that use branching (see Chapter 4) or audiovisual stimuli. First, when you plan to utilize branching, verifying whether this feature can be applied is crucial as dashboards do not always offer this feature for all question types (e.g., it might be possible to branch a radio button question type but not a slider question type). Second, when using audio-visual stimuli (e.g., pictures, videos or sound clips) it is important to be aware of the potential mobile data costs of these types of questions to participants.

6.1.2 Sampling schedules

In Chapter 3, four different types of sampling schemes were considered: fixed, random, semi-random and event-contingent sampling. The first three are signal-contingent, meaning that participants are requested to fill out a questionnaire on a smartphone or wearable each time they receive a push notification (‘beep’). Event-contingent sampling, as the word itself explains, is not contingent upon a random beep but asks participants to initiate a questionnaire each time a predetermined event has occurred (e.g., after smoking a cigarette).

When designing a study that follows signal-contingent sampling (i.e., fixed, random or semi-random), a survey schedule must often be created on the dashboard from scratch. For example, if a researcher wants to use a semi-random sampling scheme, each time point needs to be specified. In a larger-scale study, it may quickly become too time-consuming to do this for each participant individually. Hence, a dashboard that allows one to use a prespecified schedule template or the ability to copy a created schedule would be preferred. This not only saves time but also ensures that all participants receive notifications at the same time. This is advantageous within group settings, where it would be troublesome if smartphones continuously kept ringing asynchronously (e.g., within group therapy, classrooms and shared office spaces). Other settings may benefit from (semi-)random sampling schedules that differ for each participant. In this case, it is advised to use a dashboard that offers individualized (semi-) random sampling. What this entails is a feature in which a researcher needs to indicate – instead of individual time points – the length of the time interval in addition to how many notifications should be presented randomly and differently for each participant.

When using event-contingent sampling, two types of questionnaire initiation are to be considered: self-initiated and device-initiated based on passive data. The implementation of self-initiated questionnaires is straightforward. The dashboard needs to make the questionnaire permanently available and accessible (e.g., on the home screen of an ESM app), after which a participant can self-initiate a response when needed. Initiating questionnaires based on a device data related to bodily conditions (e.g., increased heart rate) or environmental conditions (e.g., sound or GPS location) is more complicated. The challenge is for the ESM platform to integrate the passively collected sensor data, analyse it and trigger a questionnaire when a particular condition is met. However, as technology advances, ESM platforms are starting to offer this type of data collection for their sampling schemes (for a recent example, see Hoemann et al., 2020). Researchers should, however, be mindful that this requires accurate wearable technology and good decision rules for when to trigger a questionnaire, which may be tedious to develop. For example, considering heart rate alone would not be sufficient to trigger a questionnaire investigating the moments that lead up to a panic attack as heart rate may increase by intensive movement. Similarly, battery constraints can render real-time interventions

based on continuous passive data impractical as analysing sensor data or streaming it to another device for data integration consumes significantly more battery life than merely storing it locally. For instance, to integrate passive and active data, Bögemann and colleagues (2024) had to limit their analysis to only 10 minutes of passive data before an ESM beep + ESM questionnaire data to prevent rapid battery depletion of their wearable device.

6.1.3 Enrolment of participants

Ideally, the ESM dashboard will generate a single study code, scannable QR code or web link that allows participants to enrol in the ESM protocol through the platform's smartphone app (discussed later). This is important as some dashboards do not allow this and instead require the researcher to generate an individual code for each participant, which quickly becomes time-consuming. Equally important with respect to the enrolment of participants is whether the dashboard allows for a flexible rather than fixed start of the ESM schedule. A fixed start specifies a single starting date for all participants regardless of when they enrol (e.g., first beep on Monday, June 15). A flexible start sets a starting point relative to the enrolment date (e.g., first beep on the first Monday following enrolment or the morning after enrolling in the study).

6.1.4 Data analytics

Once people have started their ESM monitoring period, it may be advantageous to keep an overview of participants' involvement. A dashboard that allows for checking compliance makes it possible to quickly identify data collection problems or risk of drop-out. While a participant can then be contacted by the research team if needed, some dashboards also allow researchers to immediately follow up with participants via the app itself. Additionally, dashboards that allow for data visualization make it possible to stimulate compliance by facilitating the provision of visual feedback. Similarly, these dashboards allow the software to be used in more clinically oriented (study) settings. However, it is important to note that the analysis options and visualizations available on a dashboard or app are often limited, and requiring additional development of new onboard analysis tools or visualizations unique to a clinical study, for instance, is

expensive. For these types of studies, we hence recommend discussing additional development costs with developers well in advance.

6.1.5 Data download

Data collected with an ESM platform is typically stored on a secure database that is managed by the platform provider. However, some platforms offer the opportunity to use your own database for storing data (e.g., RADAR). However, this requires significant technical skills to set up and maintain. Platforms that do not allow you to set up your own database instead have a function to export data, which is typically done via the dashboard. This export process concerns aggregating all data into a single data file that can be used for statistical analysis. To ensure the exportation in the required (long-data) format (e.g., .csv file), we highly recommend researchers to test this process out before the start of the study.

6.2 ESM apps

So far, this chapter has discussed important features to consider when selecting an ESM platform. An ESM platform is, however, much more than just the dashboard. We now turn our attention to the ESM app itself. In this section, we highlight two important considerations related to ESM apps. Afterwards, we consider three advanced app features that may impact compliance.

6.2.1 Native or hybrid

There are two main types of ESM apps: native and hybrid. While the former is developed to function and work with only one type of system software (e.g., Android or iOS), the latter is based on web technologies and works cross-platform (i.e., apps that works on both Android and iOS). Native applications are typically faster and can take full advantage of special features unique to the system software for which they are developed (Ajayi et al., 2018). This may be relevant to consider with mobile sensing (see Chapter 13). However, development and maintenance costs, and hence also subscription fees, are often higher for native apps compared to hybrid apps. This is because native apps require a unique and more

difficult codebase for each system software they are run on whereas hybrid apps share a single codebase that is generally easier to implement. While performance may favour native apps for those interested in mobile sensing, experts suggest that hybrid apps may eventually equal and possibly even outperform native applications as technology advances (Huynh et al., 2017).

6.2.2 Push notifications: a warning

An ESM app uses push notifications to signal participants to fill out a questionnaire, but these may not work seamlessly on all smartphone models and operating systems due to hardware- and software-based fragmentation or, in other words, inconsistency in how different operating systems handle app features, which causes varied performance across devices (Han et al., 2012). Smartphones that run on iOS and Android cover about 99% of the current market share (Karthick & Binu, 2017), and for both, there are different operating system versions in circulation. Similarly, phones run on different hardware (i.e., processors, sensors, graphic cards, etc.). This issue means that not all smartphones may be compatible with the selected ESM app. It is therefore crucial to test whether the app is compatible with the smartphone of the participant. This is especially true for Android smartphones, for which manufacturers often develop unique 'skins' for different smartphones. These skins give each smartphone its own unique user interface. This is why phones running on the same version of Android can look and behave differently. These skins can furthermore effect whether push notifications related to an ESM app are allowed by default. When working with Android phones, it is therefore even more important to check whether manual adjustment to app privacy settings is possible and required (i.e., overriding default settings to allow an app to send push notifications). Additionally, updates to the operating system may include changes to default settings. Hence, it is recommended to check whether notifications are still coming through after such updates.

6.2.3 Helpful app features

Three app features can benefit ESM research: sound intensity and duration, font size, and offline notification. First, the sound intensity and duration of push notifications can be increased on some apps, which makes it

easier for participants to notice them in noisy environments. Second, it may be helpful to check whether the font size of the text displayed within the app is adjustable so that every participant can comfortably read the questions or information provided on the app. Third, when sampling in remote areas or at moments when people may have low connectivity (e.g., commuting to work on the train), it may be helpful to use an app that is capable of functioning offline.

6.3 Wearables

Wearables concern technologies that can be worn. In the context of ESM, these technologies are most commonly used for passively collecting physiological data (e.g., heart rate and galvanic skin response) and movement (e.g., accelerometer data and relative geographical position). Wearables can come in many different forms and shapes. For the measurement of heart rate, there are, for instance, smartwatches (Tison et al., 2018), rings (Magno et al., 2019), chest patches (Liu et al., 2018) and even earpieces (Passler et al., 2019). The scope of what is possible concerning measurement with wearables is covered in detail in Chapter 13.

While the dashboard and ESM app are prototypical for each ESM platform, the inclusion of one or more wearables is not. Currently, only a limited number of providers are capable of integrating data from a wearable. This point is mainly of importance when one wants to trigger questionnaires based on data collected from a wearable (i.e., a particular type of event-contingent sampling). When this is not the case, one may use commercially available wearables and aggregate the data of both ESM questionnaires and passively collected data using external software such as R or Python. However, when using commercially made wearables, one should be wary of how certain measures are calculated and of the limited access to raw physiological data. For instance, a wearable may claim to measure or indicate a level of stress without users being able to see how this ‘stress’ is calculated or access the underlying physiological signals (e.g., heart rate variability or skin conductance). This is because the computation of such measures may be hidden and protected by the intellectual property rights of the developer. Thus, raw physiological data may not necessarily be available from commercially developed devices.

6.4 Legal considerations

There are laws related to ESM software and hardware that need to be taken into consideration, including laws on data privacy and the use of electronic devices as well as laws on the use of ESM in clinical care.

6.4.1 Data privacy and electronic devices

Chapter 5 addressed the highly personal and often sensitive nature of ESM data, calling for responsibility regarding data privacy and protection. ESM platforms based in the EU will by default be required to meet the European Union's General Data Protection Regulations (GDPR, https://gdpr- info.eu/) which came into effect in 2018. While researchers outside of the EU do not need to adhere to GDPR, they will often also have country-specific laws to which they need to adhere (Greenleaf, 2017). This means that the choice of platform may also be determined by the data privacy laws in the country in which an ESM study is conducted. For example, GDPR demands that any data collected on EU-citizens needs to be stored on a database that meets GDPR regulation.

6.4.2 Clinical use of ESM software: a medical device?

Both the EU and US have strict laws surrounding medical devices with different definitions of what medical devices are. Regulation (EU) 2017/745 on Medical Devices (MDR), article 2(1) (OJ L 117) states: '[M]edical device means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease, diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability, investigation, replacement or modification of the anatomy or of a physiological or pathological process or state, providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations, and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means [. . .]'.

US Federal Food, Drug, and Cosmetic Act, section 201(h) states that a medical device is 'An instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is: recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them, intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or intended to affect the structure or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes [. . .]'.

At first sight, the use of ESM in clinical practice matches both EU and US definitions. However, as with many laws, exemptions are possible. For example, the US Food and Drug Administration states that while some software may meet the definition of a medical device, it intends to exercise enforcement discretion when the software poses a low risk to the public (US Food and Drug Administration, 2019). This in itself is a vague statement as the words 'intend' and 'low risk' are undefined. However, they do provide examples under which we also find diagnoses and treatment of psychiatric conditions (US Food and Drug Administration, 2019, p. 22). It is currently unclear whether the same exemption intention applies to the clinical use of ESM software within the EU. But app stores are full of mental health apps (unrelated to ESM) that fit the definition of medical devices. To the best of our knowledge, none of these are classified as medical device software. Therefore, in the EU and the US, there is an equal amount of vagueness surrounding the applicability of medical device regulation on clinically used mental health apps (which may include ESM software).

6.5 Sustainability of ESM software and hardware

The different elements of an ESM platform consist of various types of hardware. This includes the collection of elements (i.e., physical objects) that make up smartphones, wearables, laptops, databases and servers. Each of these devices in turn runs on its own system software (e.g., Windows, Mac, Android and iOS). System software provides a platform for the use

of other types of software such as, for instance, application software (i.e., apps, database management software, etc.). System software therefore acts as an interface between hardware and apps. This implies that when system software updates, application software may need to be updated as well.

System software is typically backed by major multinationals (e.g., Microsoft/Apple). These multinationals employ a solid workforce whose task it is to continuously improve the system software. In contrast, application software is often not backed by a multinational or even a company (e.g., an app could be developed by a singular person). The updating of application software to remain compatible with updated system software is a vital element of sustainability. The management of application software by a single person requires considerable investment, which may threaten the sustainability of the platform. Similarly, without support for new developments, the application software may quickly become outdated. An ESM platform should therefore preferably involve a multidisciplinary team of programmers, researchers and medical health professionals to stay operational as well as innovative.

6.6 Recommended ESM platforms

Up until this point, we have described different relevant elements and considerations when deciding upon the right ESM platform for a study. In this section, we compare five excellent ESM platforms: m-Path (https://m-path.io/landing/), Movisens (https://www.movisens.com/en/), RADAR (https://radar-base.org/), SEMA3 (https://sema3.com/) and Expiwell (https://www.expiwell.com/). These platforms have been selected based on merit, geographical location in light of legislation and perceived sustainability.

6.6.1 Overview of ESM platform features

In the table below, we provide an overview of the selected platforms. The content within the table is based on personal correspondence with representatives of each of the platforms (November 2020 through January 2021; updated Augustus 2024). This table can be used to see how a given platform would compare against the five platforms considered here.

Table 6.1. Overview of presented ESM platforms [a]

	m-Path	Movisens	RADAR	SEMA3	Expiwell
Online dashboard					
Slider questions	Yes	Yes	Yes	Yes	Yes
Checkbox	Yes	Yes	Yes	Yes	Yes
Radio buttons	Yes	Yes	Yes	Yes	Yes
Open questions	Yes	Yes	Yes	Yes	Yes
Picture stimuli	Yes	Yes	I.D.	Yes	Yes
Video stimuli	Yes	Yes	No	No	Yes
Audio stimuli	Yes	Yes	Yes	No	Yes
Branching	Yes	P.	Yes	Yes	Yes
Signal-contingent: fixed and (semi)random	Yes	Yes	Yes	Yes	Yes
Signal-contingent: individualized (semi) random	Yes	Yes	I.D.	Yes	Yes
Event-contingent: initiated by passively collected data	Yes	Yes	I.D.	No	Yes
Event-contingent: self-initiated	Yes	Yes	Yes	Yes	Yes
Templates	Yes	Yes	Yes	Yes	Yes
Data visualization	Yes	P.	No	Yes	Yes
Compliance check	Yes	Yes	Yes	Yes	Yes
Data download	Yes	Yes	Yes	Yes	Yes
ESM app					
Native/Hybrid	Native	Native	Native	Hybrid	Native
Operating system compatibility	Android/ iOS	Android	Android/ iOS	Android/ iOS	Android/ iOS
Adjustable notification sound & durations	Yes	Yes	No	P.	Yes
Adjustable text size and font	Yes	Yes	No	No	Yes
Offline	P.	Yes	P.	Yes	Yes
Data communication	Yes	Yes	No	Yes	Yes
Mobile sensing	Yes	Yes	Yes	No	Yes

	m-Path	Movisens	RADAR	SEMA3	Expiwell
Wearable					
Integrated data	Yes	Yes	Yes	No	Yes
Legal					
GDPR compliant	Yes	Yes	Yes	P.	Yes
MDR compliant	No	No	No	No	No
510(k)	No	No	No	No	No
Other	No	No	No	Legal framework Australia	No
Profile					
Founding date	2019	2009	2016	2013	2015
Country	Belgium	Germany	UK	AUS	US
Number of paid employees	6	16	U.	2	U.
Number of active users	175+	850+	U.	500+	5400+
Cost[b]					
Free	Yes	Yes	Yes	Yes	Yes
Premium	Yes	Yes	No	No	Yes

Note: U. = undefined, P. = partial, I.D. = in development.

[a] *For additional ESM platforms, please see: https://docs.google.com/spreadsheets/d/18R9x9Qbl9t-ADJGpJBJID_T9EWZeQ_4W3OFdn3i KRU7U/edit#gid=204277638*

[b] *Free versions may be limited. Similarly, premium prices may vary depending on study design and are furthermore subject to change. Hence, they are not specified in this table. For free version restrictions as well as official prices of premium versions, please consult the original platform websites.*

6.6.2 Practical advice

Each platform outlined above has unique characteristics that go beyond the scope of basic ESM. For example, with m-Path one can create questionnaires that change dynamically based on user input as well as provide psychoeducation and exercises on a separate window inside the app. It is furthermore the only platform that allows users to create, save and share content (e.g., questionnaires and EMIs) with one another. Comparably, Movisens is currently the only platform to offer wearables that are developed in-house, and RADAR is the only platform that allows users

to set up their own database for data collection. While these additional features make each platform different, it is important to stay close to the core of what is required for the basic application of ESM that will fit most research projects. Noteworthy is that while the platforms share similar features for the basic application of ESM, the user interfaces of different platform components (e.g., dashboard, app, wearable) differ substantially. Just as one can be a proponent of iOS or Android, one can prefer one ESM platform over another. Ultimately, this means that the choice for a platform is also one of personal preference secondary to research design, budget and legal restrictions. When multiple platforms fit within the budget and envisioned research design, it is recommended that the platforms be piloted first.

6.7 Conclusion

In this chapter, we provided insight into the different programming elements important for setting up an ESM study as well as associated legal considerations and software sustainability elements. At the end of the chapter, this information was aggregated into a table providing an overview of five excellent ESM platforms. With this chapter, we hope to have provided enough information for you to find a platform that will fit your aspirations.

CHAPTER 7

Briefing and Debriefing in an Experience Sampling Study

Aki Rintala, Silke Apers, Gudrun Eisele, Tessa Biesemans and Steffie Schoefs

This chapter aims to inform you about the importance of implementing a briefing and debriefing session with your participants within your ESM study. If you are designing a study using ESM, a briefing session will be one of the most important parts of the study. This is mainly because the participant is required to answer the ESM questionnaires without the researchers' presence in their daily life (Palmier-Claus et al., 2011); therefore, your participant needs to be informed quite extensively about the ESM procedure before the study starts. The lack of proper briefing might increase the risks of mistakes or violations of the study protocol. Therefore, participants must be properly informed about the procedures of the study. An effective briefing is also crucial to ensure compliance and data quality (Palmier-Claus et al., 2011; Rintala et al., 2019). A successful briefing session aims to motivate participants to follow the study protocol. Within this chapter, we will take you through all the necessary steps to establish an effective briefing session.

7.1 Briefing session

7.1.1 Preparation before the briefing session

During an ESM briefing session, the researcher needs to explain the purpose of the study to the participant. You as a researcher should also prepare all necessary equipment needed for the briefing session and make yourself familiar with the instructions you are planning to brief your participant (Palmier-Claus et al., 2011). We recommend making a checklist with information on what and how to brief your participant (an example of such a checklist can be found at the end of this chapter). You

could expect a proper briefing session to last 15 to 20 minutes. The exact duration will depend on the specificities and complexity of your ESM protocol and your study population as well as on how many questions your participant might ask during the briefing session. It is important to practice your briefing session before meeting your participants to evaluate the proper content (e.g., what is necessary to mention) and to ensure that you plan in enough time for the actual briefing session. If multiple researchers are involved in the study, make sure that they are all properly trained to do the ESM briefing session. We advise you to compile an ESM study manual as well as to practice the briefing session with each other.

During your entire study and in the study briefing, think about which term you will use to describe ESM to your participant. Depending on your study population and whether you are using other self-reported questionnaires alongside ESM, you may want to refer to ESM with terms such as 'diary' or 'electronic diary' to avoid confusion with other questionnaires. In this chapter, we use the term 'ESM' when referring to an ESM questionnaire or a diary.

The study briefing can be conducted either individually or in groups. We recommend briefing your participant individually as this gives you more time to ensure that your participant understands everything, increasing the chances of better compliance (Palmier-Claus et al., 2011). Some participants might also not feel comfortable asking questions in a group setting. However, if you are planning an ESM study with a large study sample, briefing in groups, either live or online, might be more efficient. Optimally, the briefing session should take place one day before the start of your ESM study (Delespaul, 1995; Palmier-Claus et al., 2011). If you are setting up a longitudinal study including several ESM periods, you also need to keep your participant motivated during follow-ups of your ESM study. To improve compliance with your study protocol, you can use different methods to ensure your participant's engagement by regularly contacting them using, for example, real-life meetings, video or phone calls, text messages or email reminders.

7.1.2 Starting the briefing session with your participant

Start by explaining what the ESM is about. If your participant asks why you are using this kind of assessment, you can answer that the ESM is specifically developed to assess momentary experiences or feelings in

their daily life. The purpose is to discover important new insights into their daily life experiences. For example, *'The fluctuations of feelings or mood during the day might help us to better understand your condition and to provide individual information to you to improve your quality of life'*. In sum, you can mention that the study might increase insights or knowledge about certain feelings or activities that will help enrich understandings of participants' condition or patterns of behaviour in the future.

Inform your participant about the number of study days and the number of assessments that will be delivered during each day. Your participant must know what to expect and understand how the questionnaires are shown or triggered. For example, '*We can do a lot of investigations within the clinic where we put people in a scanner or let people do tests or interviews. However, where things are really happening is in your normal daily life, in your normal daily routines. With the current study, we will use ESM to follow you in real life. We are interested to know how you feel, what you think, what activities you are involved i, and which people you connect to in your normal daily life. That is why we will ask you to fill out an ESM questionnaire on a regular basis'*.

Explain also your study period and how many assessments your participant will receive. For example, *'The ESM period will last six days, and every day you will get a maximum of 10 assessments'*.

We recommend explaining the content of the questionnaire in a general way. For example, *'The ESM questions are related to your sleep, feelings, activity, physical state, social interactions, stress and medication use'*. If you are conducting a study with a population with depressive symptoms, for instance, we advise you to explain the content in a general way instead of referring to their depressive mood.

You also need to explain to your participant when they can expect the first and the last assessment within one day. You don't need to explain the sampling scheme (e.g., random or fixed sampling scheme) in any detail to the participant. You can simply state the number of assessments they can expect within a day and the timeframe in which they can expect it. For example, *'ESM will trigger 10 assessments at random times between 08.30 and 22.00'*.

Participants should know what to do when they hear a notification and how to react to it in daily life. Go over different example scenarios with your participant. For example, '*We want to measure your daily life as it is, so it is important that you continue with your regular daily routines.*

When you hear a notification, open your screen and fill out the ESM questionnaire immediately. Please fill it out on as many occasions as you can. If you only fill it out when you are on your own in a quiet environment, that would not be very informative. We also need information when you are with others, when you are busy, when you are working or when you might feel stressed. It is fine if you miss a notification occasionally when you are not available – for example, when you are driving a car and you cannot stop safely or when you are engaging in an activity where you are not close to your smartphone, like swimming. It is normal to miss occasional notifications during the ESM study period because of the way ESM studies daily life. It would be impossible to answer each assessment, but we would like to ask you to fill in the questionnaire on as many occasions as you can'.

Emphasize that it is very important to keep their normal daily or nightly routines. For example, '*Keep your normal daily and nightly routines. If you want to go to sleep, take a nap, or if you do not want to be disturbed, you can keep your smartphone somewhere you cannot hear it (such as in the living room or kitchen) or switch the sound settings off'.*

If you are using a time window for answering the questionnaire in your study protocol, do not mention this exact timeframe to your participant (unless your study has a very short response window, for example, less than two minutes). For example, *'The ESM questionnaires will be available for a limited time and will disappear afterwards, so please complete them as soon as you receive the notification'.*

You can explain to your participant that the assessments will continue regardless of whether the previous assessment was filled or not. It is okay to miss an assessment; the purpose is to not change daily routines. For example, '*If you are not able to fill in the ESM questionnaire immediately, you can answer it a short period later on, but on those occasions, it is important that you try to remember where you were and what you did just before the notification went off. For example, when you hear a notification while you are in the shower, you could check if the assessment is still available after you get out of the shower* '. *It is important to answer all the questions based on the moment just before the ESM notification went off. So, when you hear a notification, try to take out your phone and answer the questions immediately'.*

Repeat several times the most important issues that your participant must be aware of. Remember, your participant will be on their own for the

entire ESM study period, so they need to fully understand and remember everything they need to do.

Explain to your participant that it is important to think of the moment just before the notification went off because otherwise, they might fill in the ESM questionnaire based on the feelings experienced while filling the questionnaire. This is not what we are looking for in the responses, and this, too, can be explained directly to the participant. It is also important that your participant always carries their phone with them and that they keep the volume turned on. You can also mention that they should check their phone from time to time if they are in a situation in which the phone needs to be switched on mute (e.g., in the theatre or at a concert). Another option is to keep the phone in vibrate mode to increase the chances of hearing or feeling the notification. To reduce reactivity, ask participants to avoid lingering on a question and to respond with the first thing that comes to mind. This helps minimize focus on what is being measured, preventing them from overthinking or becoming too deliberate about the content or purpose of the items (Kazdin, 2021). However, the questions have to be answered calmly rather than hurriedly. If your ESM application offers the possibility to change a given answer, you can mention that to your participant.

Talk to your participant about how to manage situations in which somebody asks what they are doing when they are filling an ESM questionnaire. This helps to decrease social pressure or disturbance in these situations. It is good for your participant to have an answer prepared for situations in daily life in which they might need to explain why they are using ESM. For example, '*You could tell people that you are part of a study, and they asked you to test out this new app. If you want to avoid questions about the aim of this study or the nature of the questions, you could tell people that you need to check if any technical issues occur, if the notifications are sent on time*' etc.

If you are using event sampling, it is important to explain to the participant what situations classify as an 'event'. For instance, if you want participants to fill in a questionnaire after every social interaction, it must be clear what situations count as social interactions. Make sure to explain any requirements that may qualify a situation as an 'event'; for example, if you only want participants to report social interactions of a particular length (e.g., it must have lasted at least five minutes), this should be stressed during the briefing.

7.1.3 Practice the demo ESM questionnaire with your participant

If the briefing is conducted in person, the researcher should hold the phone and sit next to the participant to avoid the participant from scrolling too fast through the questions before the researcher's instructions are finished. When you, the researcher, hold the phone, you can use all the time you need to explain every question, and you can make sure that the participant listens to the explanation. For example, *'We will now practice filling in the questions the moment after the app sends you a notification. All questions are about what you were thinking, feelin, or doing right before the notification.'* If the briefing is conducted online, the participant can be asked to open the questionnaire on their phone and read the questions aloud to make sure that the planned instructions can be given for every question. As an alternative, you can also make a PowerPoint presentation that contains all the ESM questions and answering options that you can share with your participant. The most important thing is that you come up with a practical way to go through the questionnaires when your ESM briefing needs to be conducted online.

When your participant is very interested in one particular aspect of the ESM, including something that you would normally explain later, explain that first before continuing. Read the questions to your participant (and provide additional information if needed). Check with your participant from time to time if they have understood all the questions. If your participant asks about the meaning of an item, first ask them, '*What do you think this item stand for?*'

Go through the ESM items one by one. It is important that your participant understands all the items correctly and understands how to answer questions. If you have decided to include an item where the participant has to choose between multiple answer options (e.g., a categorical item), make sure that all answer options are understood. Also make sure that your participant understands in which situations they are expected to select multiple answer options. One example of an item explanation is given below.

Question example: Feelings

Emphasize to your participant that by feelings you mean the feelings or emotions right before the notification went off. Explain the idea behind

the scale you are using (e.g., Likert scale, Visual Analog Scale, etc.). Make sure that your participant understands that they need to use the full scale and understand the anchor points. Repeat what your participant has entered. For example, for a 7-point Likert scale: *'You answered a 1 here, so this means that the statement 'I feel cheerful' does not apply to how you felt prior to filling in the questionnaire. Is that correct?'*

In Chapter 4, you can find more information about the ESM items and the answering options.

7.1.4 Additional information

7.1.4.1 Technical instructions

If participants receive a study phone or another device for their participation in the study, they need to understand how to use the device properly. For example, participants must know how often and how to charge the device, how to switch the device on and off, how to lock and unlock the screen, how to change sound settings and how to contact the researcher when experiencing any technical issues.

If participants are using their own smartphones for the data collection, some settings may need to be adjusted for the app to function correctly. For example, most apps need permission to send push notifications to the phone.

7.1.4.2 Talk about the upcoming period throughout which the study will occur

Does your participant expect that the next few days will represent their normal routine, or are there any special events planned? Will there be moments in which ESM may not be possible to fill? Discuss these moments (perhaps it is possible to fill in ESM) and negotiate if necessary (perhaps it is possible to find a way to fill in ESM even during difficult moments). Avoid starting an ESM study during holidays. Make sure to tell your participant to consistently keep their phone with them rather than, for example, leaving it at home when they go out.

7.1.4.3 Do's and don'ts!

- Make sure you explain your ESM questionnaire in a general way. For example, explain that your goal is to learn more about emotions, daily activities, social interactions and more without directly referring to the specific hypotheses of the study. In a study on stress recovery, for instance, you could explain that you're interested in how people feel and behave in different situations rather than focusing on stress or how participants respond to it.
- It is not advised to mention how much time participants have to answer the ESM questionnaire after a notification (unless your study has a very short response window – for example, less than two minutes).
- Explain the total number of assessments that the participant can be expected to receive randomly throughout the day, but don't mention the time-intervals (e.g., 90-minute time blocks). If your participant knows the time intervals, they might start changing their daily routines.
- Try to use real-life examples during the ESM briefing. Ask your participant to respond to the test questionnaire realistically, applied to that exact moment.
- Repeat several times that the participant needs to think about the moment right before the notification when they fill out a questionnaire.

7.1.4.4 Check how things are going

We strongly recommend calling your participant on the second ESM day to see how everything is going. In this call, it is recommended to ask how many assessments they have filled out that day and the day before just to get an idea of how well they are managing the ESM. If possible, you can also access the data on their compliance up to this point. If they have responded to fewer assessments than expected, you can encourage them to fill out more. Determine in advance how many assessments participants should have completed so that you can motivate them in a standardized way. Inquire about why they were unable to complete more assessments and explore ways to improve their participation. Log the information about the timing and the content of these telephone calls to your participant file. Also, give your participant your contact information and ensure your availability in case there are problems with ESM or with the study phone (if one has been given). During the contact call(s) with

your participant, check if the participant understood everything correctly. Participants should feel comfortable to contact you at any time if a problem occurs during the ESM period.

Repeat the most important issues to your participant during the telephone call:

- Keep up your normal daily routines, and do not cancel your daily appointments because of the ESM.
- Always carry the phone with you.
- Always answer the ESM assessment immediately after the notification (without creating an unsafe situation, of course).
- Mention that the ESM has a time period for answering without mentioning the exact time.

7.1.5 FAQs

7.1.5.1 What if some people find ESM stressful?

Ask the participant more specifically what content they perceive as stressful. Try to go through the questions once more and explain the reasoning behind the questions that they perceived as stressful. Highlight that it is very normal to occasionally miss notifications.

7.1.5.2 When should I contact the participant?

We recommend contacting your participant on the second day of the ESM period. If possible, monitor the data collection before reaching out. If there's poor compliance during the first few study days, you can address this during the call. Calling also gives you an opportunity to check for any technical issues.

If the compliance of participants is monitored in real-time, researchers also need to decide whether they want to contact participants if a drop in compliance is noticed later in the study. However, it is also not advisable to contact participants too often as doing so might further interfere with their daily routines.

7.1.5.3 Can I tell my participant how many assessments have been filled if they asked?

We do not recommend giving your participant an overview of the overall compliance per day during the study. It is also important that your participant does not change daily routines just to fill in the assessments. So, it is acceptable to occasionally miss one or two notifications if it is difficult for your participant to answer the ESM questionnaire.

7.1.5.4 What if a participant says they are unable to complete the ESM questionnaires during certain time periods or in certain situations?

Commonly, participants may question the relevance of filling in the assessments during situations where they feel reluctant to use the app. In these cases, try to approach your participant in a way that increases their understanding of the method and try to come up with creative solutions to ensure they can fill out as many questionnaires as possible. For example, if a participant mentions there is a three-hour window every evening when they cannot fill out the ESM due to hobbies, you could say that the notification might come while they are travelling to the location or just before starting the activity. Try to encourage the participant to check their phone just before and right after the activity and to complete the questionnaire if one is available.

If a participant indicates that their employer does not allow phone use at work, you could offer a letter as 'proof of participation in an ESM study'. Similarly, if you're conducting an ESM study in a school where students are not allowed to use their phones during class, you could arrange for the teacher to receive the notifications. The class could then fill out their responses on paper as a group (Camerman et al., 2024).

Essentially, try to work together with the participant to find an approach that ensures they continue providing data even in situations where it's more likely that notifications might be missed. Explain that it's important to gather information from all types of moments, including when they're at work or school or engaged in hobbies. The goal is to get a comprehensive overview of their daily contexts, which is why completing the ESM questionnaires in these situations is crucial.

7.1.6 ESM questionnaire for the researcher

To help you explain the content of an item to your participant, we recommend preparing a table with all the items in your ESM questionnaire and standardized explanations. Examples are given below in Table 7.1.

Table 7.1. Examples of ESM items and how to explain the item to your participant

ESM questionnaire	Scale option	Meaning/explanation
I feel content	Not at all 1 2 3 4 5 6 7 Very much	If you are feeling content, you are satisfied and happy
I am worrying	Not at all 1 2 3 4 5 6 7 Very much	If you feel you are constantly thinking about something worrying, and you do not know how to solve it
Think of the most important event that happened since the last assessment. This event was:	Very unpleasant -3 -2 -1 0 1 2 3 Very pleasant	This event can be anything that your participant feels was the most important event (even a small event like breakfast, conversation, morning routines, etc.).
This notification disturbed me	Not at all 1 2 3 4 5 6 7 Very much	If you experienced filling in the questionnaire as annoying or frustrating that time.

7.2 Debriefing session

Debriefing an ESM study can have two objectives: 1) gathering qualitative information on the participant's experience and 2) giving personal feedback on the results of the study.

7.2.1 Qualitative information

The main objective of this type of debriefing session is to gather information on how the study went from your participant's point of view. The debriefing session is recommended so you can start going through your participant's compliance over the ESM period and discuss missed notifications with your participant (Delespaul, 1995; Palmier-Claus et al., 2011). If you have set a minimum number of entries to be completed in order for a participant's participation to be considered compliant in your study, go

through your selected procedures with your participant to check whether minimum compliance was reached. Usually, if minimum numbers of entries are not met, researchers can ask the participant to complete additional days of assessments (Palmier-Claus et al., 2011). Debriefing is also important because it might allow your participant to report positive or negative issues they may have encountered, such as positive findings on EMS use, mistakes in their own reporting or short-term technical errors (Kimhy et al., 2012).

Several issues need to be clarified during the debriefing session:

1. What was the compliance of your participant with your study protocol?
2. What did your participant think of your study?
3. Were there any challenges with using the phone?
4. What did your participant think of the items?
5. Was there anything else that your participant discovered during the study?
6. Was the study period representative of the normal life of the participant?
7. Was there anything in particular that you as a researcher discovered during the study on this particular participant?

The debriefing session is also a good moment to introduce whatever feedback questionnaires or interviews related to study success you may have decided on. We strongly recommend using a validated feedback questionnaire that allows you to get some descriptives on the feasibility of your study.

Example of questionnaire on debriefing:

1. Was this a normal week: 1 (not at all) – 7 (very much)
2. Did any special events occur during the week? 1–7
3. What kind of events occurred during the week (open question)
4. Were you able to express your experiences via the app? 1–7
5. Did the ESM period influence your mood? 1–7
6. Did the ESM period influence your daily routine? 1–7
7. Did the ESM period influence your contact with other people? 1–7

7.2.1 Checklist for how to brief your participant in an ESM study

GENERAL	Check
Explain the purpose of the electronic diary and why it is useful to fill in the questionnaires	
Take your time to explain every item one by one	
Repeat the most important things (e.g., keeping the app with you at all times)	
Keep the instructions positive, and explain what you expect from the participant	

BRIEFING PARTICIPANT
The participant is asked to fill in the questionnaire immediately after receiving the notification on how he/she is feeling, doing, etc. right at the moment (i.e., just before the notification)
Keep normal daily or nightly routines
Explain the time frequency of the assessments
Practice the diary entry with the app
Contact calls
Procedure in case of a problem
Discuss situations that can occur during ESM (i.e., questions from other people)

CONCLUSION
What are your participants' thoughts on using the electronic diary?
Discuss upcoming period: normal routines vs. special events

7.2.2 Feedback on results

Researchers are increasingly exploring ways to provide participants with personalized feedback on their ESM data. Personalized feedback can make the study more personally relevant for the participant, increasing the perceived value of the study and, in turn, improving compliance (Hsieh et al., 2008). Research indicates that participants often expect personal feedback in ESM studies and that it can serve as an additional incentive beyond financial compensation (Kiekens et al., 2021). Additionally, presenting feedback examples and clarifying how missing data affects data quality may enhance participant compliance (Riese H., personal communication). However, randomized controlled trials are needed to further guide researchers on how, when and which type of feedback should be offered to the participants (Kiekens et al., 2021).

AFTER: THE ANALYSIS OF ESM DATA

CHAPTER 8

Structuring, Checking and Preparing the Data

Wolfgang Viechtbauer

ESM studies typically yield a substantial amount of data. Consider a study with 100 participants, each assessed 10 times per day over the course of six days using random time sampling. For variables that are measured at each assessment moment, we will therefore obtain 6,000 data points (although some of them will be missing due to participants being unable or unwilling to respond to the questionnaire at particular moments). Assuming that the questionnaire provided to the participants at each assessment moment contains 30 questions, we are therefore dealing with a dataset that contains up to 180,000 data points. A first challenge is choosing an appropriate structure for such a large amount of data.

8.1 Data Structure

When assessing a particular variable repeatedly within a group of participants, we could arrange the dataset so that each row corresponds to a single participant and multiple columns (i.e., variables) are used to represent the measurements at the various assessment moments (the response variables or 'items' for short). Further columns might provide subject-level information, such as their age and sex. In this case, the data are said to be in a 'wide format'. See Table 8.1 for an example of such a layout.

Table 8.1: Example of data structured in a 'wide format'

Subject	**Age**	**Sex**	**Item 1**				**Item 2**				**...**
			Time 1	**Time 2**	**...**	**Time 60**	**Time 1**	**Time 2**	**...**	**Time 60**	**...**
1	30	female	5	3	...	3	3	5	...	4	
2	32	male	2	1	...	3	2	2	...	5	
...											

However, this type of data structure quickly becomes unwieldy for studies using intensive longitudinal data collection techniques such as ESM. For example, for the design described above, we would need 60 columns for each variable that is assessed via the questionnaire and hence 1,800 columns for the set of 30 questions. Another 60 columns/variables would be required to record information about the exact time of the assessments (since the assessment times within a day will differ across participants when using random time sampling).

Instead, the preferred data structure for ESM data uses a 'long format'. Continuing with the example, the dataset will then contain 6,000 rows of data, where each row corresponds to a particular assessment moment for a given subject and the columns to the 'time-varying variables' measured at each assessment. These include the responses to the actual questionnaire items but might also cover additional measurements obtained via other sensors or data sources (e.g., accelerometer data, ambient noise/light levels, location information collected via GPS, temperature/weather data, physiological measurements). In addition, the dataset will include various 'design variables'. Most importantly, the dataset must contain a 'subject identifier' to indicate which rows (i.e., assessments) belong to the same participant. Other important design variables include a counter for the assessment day (1–6), a counter for the assessment number within each day (1–10) and the date and exact time of each assessment. Finally, the dataset will again also include some subject-level or 'time-invariant' variables (e.g., the age and sex of the participants) that are constant within each subject and that are typically collected once at baseline (such subject-level variables will often be stored initially in a separate dataset but ultimately will be merged together with the momentary assessments based on the subject identifier). See Table 8.2 for an example of this type of layout.

Table 8.2 Example of data structured in a 'long format'

Subject	Age	Sex	Assessment Day	Assessment Number	Item 1	Item 2	...
1	30	female	1	1	5	2	
1	30	female	1	2	3	1	
...	...	...	...	...	...	...	
1	30	female	6	10	3	3	

Subject	Age	Sex	Assessment Day	Assessment Number	Item 1	Item 2	...
2	32	male	1	1	3	2	
2	32	male	1	2	5	2	
...	...	...	...	...	...	...	
2	32	male	6	10	4	5	
...							

Note that some time-varying variables may not be measured at each assessment moment but rather once per day (e.g., when a subject's rating of their sleep quality the previous night is obtained only at the first assessment within each day) or at other sampling frequencies (e.g., when additional measurements are collected via passive sensors). Also, studies using an event-contingent design or a combination of event and time sampling do not yield a pre-planned number of rows of data (i.e., the number of rows in the dataset then depends on the number of events that occurred for the subjects), although the data structure is fundamentally the same. Finally, for pre-planned assessments, it is of course possible that a subject does not notice or respond to the signal prompting him or her to complete an assessment, in which case the questionnaire data will be missing for that assessment moment.

The long format can be easily extended to more complex data structures. For example, studies with multiple family members (see Chapter 14 for examples) and/or multiple ESM data collection phases – the latter is sometimes called a measurement-burst design (Sliwinski, 2008) – would simply require the addition of a family and/or phase identifier variable to the dataset. For example, if we would extend the design above to a pre- and post-treatment phase, each subject would contribute 120 rows of data where the two phases are distinguished by a single variable (e.g., coded 0 for the pre- and 1 for the post-treatment phase).

8.2 Example data and research questions

For illustration purposes, we will make use of a subset of the data from an ESM study including 328 participants who were asked to fill in a questionnaire assessing their mood and several contextual variables 10 times

per day over the course of six days. The participants fell into three groups based on their mental health status: 112 participants were healthy controls, 109 participants had a lifetime history of depression and current residual depressive symptoms while the remaining 107 participants had a diagnosis of a psychotic disorder (with most participants in this group suffering from schizophrenia). Participants were prompted to fill in the questionnaire at semi-random times within 90-minute blocks starting at 7:30 in the morning and ending at 22:30 in the evening (i.e., the first prompt of the day was delivered between 7:30 and 9:00, the second between 9:00 and 10:30, and so on). The exact times of the prompts were generated in such a way that adjacent prompts were at least 20 and at most 160 minutes apart. The average inter-prompt interval was around 90 minutes with a standard deviation of approximately 30 minutes. The prompt 'beeps' to fill in the questionnaire were delivered via wristwatches, and responses were entered into booklets that participants were asked to carry with them at all times.

The dataset is available on the book website (http://www.real-leuven.be). Aside from the subject identifier, it includes a selection of variables that can be grouped into three categories:

- **subject-level variables:** the age (in years), sex (female, male) and mental health status (control, depressed, psychotic) of the participants;
- **time-related variables:** the date and time when each beep occurred and when a subject filled in the questionnaire[1], the day number (1 to 6), the beep number within each day (1 to 10), the assessment number (1 to 60), the beep time in minutes after midnight on each day when a prompt was issued (e.g., a beep time of 477 corresponds to 7:57 in the morning since $7 \times 60 + 57 = 477$), the response time (again in minutes after midnight) when a participant filled in the questionnaire (e.g., 480 for 8:00), the response time in hours after midnight (e.g., 8.0 for a response time of 480) and the total number of hours that had passed since midnight of the first assessment day corresponding to each response time (e.g., for a response time of 1,104 on the fourth day, the value of this variable would be $3 \times 24 + 1104 / 60 = 90.40$);
- **response variables:** 3 items to assess 'positive affect' each rated on a 1 (not at all) to 7 (very much) scale (Right now, I feel cheerful / relaxed /

[1] Since the study used booklets (and not smartphones or some other device) for data collection, information about the time a questionnaire was filled had to be obtained from the subjects. This was asked at the end of the questionnaire, so to be precise, this variable indicates the time when a subject had finished filling in the questionnaire.

satisfied[2]), 6 items to assess 'negative affect' rated in the same manner (Right now, I feel irritated / anxious / down / guilty / insecure / lonely), a rating on a −3 to +3 bipolar scale of the (un)pleasantness of the most important event that had occurred since the previous beep (with −3 for a very unpleasant and +3 for a very pleasant event), whether the person indicated being alone at the time of the beep (1 = alone, 0 = not alone), a rating on a 1 (not at all) to 7 (very much) scale of how pleasant the company is that they are in (only applicable when the person indicated not being alone), a rating on a 1 (not at all) to 7 (very much) scale of whether participants would rather do something else than the activity they were engaged in at the time of the beep, whether they had consumed coffee or alcohol since the previous beep (1 = yes, 0 = no) and their location at the time of the beep (at 'home' or at some 'other' location).

When a participant did not fill in the questionnaire at a particular beep, the response time and all response variables are missing. Note that the dataset has been slightly adjusted from the original data for didactic purposes, but the findings described below should broadly reflect what was observed in this sample.

8.3 Software and code

In addition to the actual dataset, the interested reader will find on the book website R code corresponding to all of the following data checks, data preparation steps and the actual analyses to be described in the next chapter. We focus on R (https://www.r-project.org) since it is freely available, works across all major operating systems (i.e., Windows, MacOS, and Unix/Linux) and provides an extremely powerful platform for managing, visualizing and analysing ESM and other types of data. Several R packages and tools that facilitate the management and preprocessing of ESM data have also been developed (e.g., Revol, Carlier, et al., 2024; see also https://preprocess.esmtools.com/; Viechtbauer & Constantin, 2023). Resources focused on other statistical software packages will be provided at the end of the next chapter. However, it should be noted that the reader will

[2] The questionnaire was in Dutch and the most appropriate translation of the Dutch word *tevreden* that was used in the questionnaire is debatable. This could either be translated as satisfied or content. Both translations have been used in the literature.

not find any description or discussion of the R code in this chapter. The purpose of this chapter is to provide a conceptual understanding of the procedures and methods used in analysing ESM data and not to teach R (or some other software package). Those familiar with R should find the code on the website sufficiently documented so that it could be adapted to other datasets and analyses.

8.4 Data checks and preparation

Although the increased use of smartphones for data collection in ESM studies and the use of corresponding platforms for data synchronization and storage can reduce the risk of errors that might occur during manual data entry (which is necessary when the responses to the questionnaires are collected via paper booklets), various data checks should still be routinely conducted. The following describes some of the properties of the data one should examine as part of this process.

As a first step, design- and time-related variables (e.g., family/subject identifiers, phase identifiers as needed in measurement-burst designs and all variables related to the timing of the beeps) should be checked for missing data. Except for the response times (which of course will be missing for beeps not responded to), none of these variables should ever include a missing value. This is also the case within the illustrative dataset.

On the other hand, subject-level variables such as the participants' age and sex could very well include missing values (e.g., when a participant prefers not to disclose this information via the baseline questionnaire). However, in this case either all or none of the values should be missing for all rows corresponding to a particular subject. Moreover, if the values are not missing, then they should really be time-invariant (i.e., constant) within the subject.[3] Checking the variables age, sex and mental health status reveals no such inconsistencies in the illustrative dataset. These variables have been fully assessed in all participants.

As noted above, the values of variables assessed via the questionnaire administered at each beep will be missing when a study participant does

[3] Naturally, a subject's age could increase during the course of a study, but here we are referring to the age of the subject at the time this information was collected – for example, when the baseline questionnaire was administered. The same applies to other subject-level variables assessed through such a baseline questionnaire.

not respond to a particular beep. Hence, it makes little sense to check such variables for missing values. However, one should double-check that there are no out-of-range values for these variables, and it can never hurt to do the same for any design- and time-related variables as well. Simple frequency tables can be used for this purpose. This is also a good time to ensure that missing values in such variables are really treated as missing by the software and not indicated with a numeric code (e.g., −999) that inadvertently might end up being analysed as observed values.

Manual transfer of information from booklets to the database is tedious and error-prone work, and information from the same assessment moment might appear twice in the dataset. The same could even happen with the use of smartphones for data collection due to software glitches and data synchronization issues. Hence, one might want to check that there are no duplicated values in the time-related variables within subjects (e.g., the day and beep counters and the beep time variable). For example, in the illustrative dataset, the same day-beep combination (e.g., day = 2 and beep = 5) should only occur once within each subject, and the beep time values must be unique within each day for each subject. For designs where beeps are issued within time blocks, one might also want to confirm that the beep times are consistent with the range of the block times (e.g., in the illustrative dataset, all rows where beep = 1 must have a beep time that falls in the interval 450 to 540, corresponding to 7:30 and 9:00).[4]

The questionnaire administered to participants at each beep might also include 'branching items'. For example, in the study described above, participants were asked if they are alone or in the company of others at the time of the beep, and only in the latter case should they have rated the pleasantness of the company they are in. Therefore, the pleasantness rating should always be missing for moments where participants indicated being alone, which can be easily checked via a cross-tabulation of the branching item and the item(s) belonging to this branching structure. With paper booklets, one could also come across beeps where the branching item is missing (i.e., the person did not indicate whether they are alone or in company), but items for a particular branch are filled (i.e., a pleasantness rating was provided). One could then consider recoding the

[4] To be precise, for all first beeps the beep times should be ≥450 and <540, for all second beeps the beep times should be ≥540 and <630, and so on.

missing branching item (into 'not alone') or make the pleasantness rating missing – a choice to be documented, but which is ultimately up to the researcher. Fortunately, none of these types of inconsistencies are present in the illustrative dataset.

As part of the data checking, it may also be of interest to examine the response delay, that is, the amount of time that passed between the moment when a beep was issued and when a participant started to fill in the questionnaire (or finished filling it in). Figure 8.1(a) provides a bar plot of these response delay values (based on the difference between the response time and the beep time variables) for the illustrative dataset. The distribution is right skewed with a peak at 3 minutes and a range of 0 to 15 minutes (mean = 4.0, SD = 2.7).

When using smartphones (or other electronic devices) for data collection, typically a timeout is set such that the questionnaire becomes unavailable after a certain amount of time has passed since the beep. No such maximum can be automatically set when using paper booklets for data collection. Since the goal is to capture the participants' state and context at the time of the beep, concerns may arise when the delay between the beep and the response is quite long. In the study described above, responses that occurred more than 15 minutes after the beep were considered 'invalid' and marked as missing (which explains why there are no delays longer than 15 minutes in the data). This is of course an arbitrary choice, but so is any maximum allowed delay that is programmed into a smartphone app. In other words, either the choice is made before the study period or it can be enforced once the data have been collected.

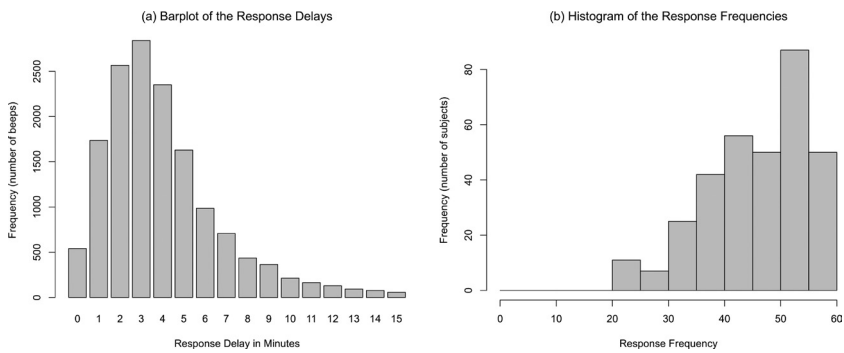


Figure 8.1 (a) Barplot of the response delays and (b) Histogram of the response frequencies

When using smartphones for data collection, it is also possible to obtain detailed information about the exact amount of time it took participants to fill out the entire questionnaire and even individual items. Although participants might become quite proficient in completing the questionnaire through practice (after all, they are asked to do so numerous times over the course of the study), very fast completion times might indicate careless responding and, consequently, unreliable data. Therefore, one could consider filtering out (i.e., marking as missing) responses from assessments where the completion time fell below some minimum threshold.

Finally, concerns can also be raised about participants that only respond to a very low number of the pre-planned beeps. Such participants may only choose to respond to beeps when it is convenient for them (e.g., when at home, when not stressed), which would make their responses unrepresentative for the states and contexts they experience in their daily lives. Figure 8.1(b) shows a histogram of the response frequencies for the 328 participants included in the illustrative dataset. In the study described above, 15 (out of 343 participants to begin with) that had responded to less than 1/3 (i.e., 20) of the 60 beeps were removed from the dataset. Despite the aforementioned reasoning, it should be noted that applying such a rule is in essence an arbitrary decision (and so is the specific cutoff point for the minimum required response frequency). Whenever such selection rules are applied, it is of course possible to conduct sensitivity analyses to examine whether the conclusions of the statistical analyses are unchanged when the rules are altered. Putting this issue aside, based on the response frequencies, we can compute the compliance rates (i.e., response frequencies / 60 × 100%) of the 328 participants included in the dataset. We find a mean compliance rate of 75.7% (SD = 15.2) with a range of 33.3% to 100% (note that these summary statistics were computed after removing the 15 subjects with very low compliance rates; mean compliance of all 343 participants was 73.1%).[5]

Other descriptive statistics about the participants can be computed as well. However, when the dataset is structured in the long format, we must be careful to first 'collapse' it to the subject level when reporting summary statistics about time-invariant variables (otherwise, the repeated values

[5] In event-sampling designs, such compliance rates cannot be computed since there is no predefined number of beeps to which participants should respond. In this case, summary statistics about the response frequencies can be provided instead.

of such a variable from the same subject would be treated as different subjects with identical values). In other words, if we want to obtain summary statistics about the age of the participants, we should first extract a single age value per subject. We then find that the mean age of the participants was 36.5 years (SD = 10.9) with a range of 18 to 65 years and that the majority (60.4%) of the participants were female (198 female, 130 male).

Another common data preparation step involves taking the sum or mean of several items (e.g., that are assumed to measure some common underlying construct) at each beep. For example, we might want to take the mean of the three items used to measure 'positive affect' and similarly the mean of the six items to measure 'negative affect' and add these as two new time-varying variables to the dataset. In doing so, one must consider how to handle beeps where the value of one or multiple items to be averaged is missing. This could, for example, happen when participants are allowed to skip items (which happens automatically when responses are collected via paper booklets but could also be an option when using a smartphone app for data collection) or when participants stopped filling in the questionnaire while in the middle of responding to the items to be averaged. The simplest option is to set the mean of the items to a missing value whenever at least one item is missing (this is also how we will compute these means for use in the analyses described below). Alternatively, one could take the mean of all non-missing items; they could also choose to do so only when at least a minimum number of items are available (e.g., we only take the mean of the 'positive affect' items when at least two of the three items have been responded to). More sophisticated techniques could also be used to impute the missing values but are beyond the scope of this chapter.

For reasons to be outlined in more detail later, we may also want to compute the mean of a time-varying variable within each subject and add these subject-level means as a time-invariant variable back to the dataset. Consider, for example, the 'event pleasantness' rating that participants were asked to provide about the most important event that had occurred since the previous beep (rated on a −3 to +3 scale). For each participant, we can compute the mean of this item (based on all non-missing values) and add this as a subject-level variable back to the dataset. In addition, by subtracting these subject-level means from the original variable, we can compute a 'within-person mean centred' version of the event pleasantness

variable. This is illustrated in Table 8.3 for part of the data from the first two subjects in the dataset.

Finally, there is yet another variable we may want to compute based on the original event pleasantness ratings, namely a 'lagged' version thereof. For this, we use the rating from the first beep as the value of this lagged variable at the second beep, the rating from the second beep then becomes the lagged value at the third beep, and so on. If the value is missing for a particular beep, then the lagged value will also be missing for the subsequent beep (as shown in the table for the second beep of the first subject).

Table 8.3 Part of the data for the first two subjects from the illustrative dataset.

Subject	Age	Sex	Day	Beep	Event Pleasantness	Mean Pleasantness	Centered Pleasantness	Lagged Pleasantness
C100	30	female	1	1	0	2.4	−2.4	
C100	30	female	1	2		2.4		0
C100	30	female	1	3	2	2.4	−0.4	
C100	30	female	1	4	−2	2.4	−4.4	2
C100	...	...	...	...	...	...	...	...
C100	30	female	6	10	3	2.4	0.6	3
C101	32	male	1	1	0	1.5625	−1.5625	
C101	32	male	1	2	0	1.5625	−1.5625	0
C101	32	male	1	3	2	1.5625	0.4375	0
C101	32	male	1	4	2	1.5625	0.4375	2
C101	...	...	...	...	...	...	...	...
C101	32	male	6	10	2	1.5625	0.4375	2

In essence, such a lagged variable can be created by making a copy of the original variable but shifting the data down by one row. However, when constructing such a variable, care must be taken not to use the value from the very last beep of the first subject as the lagged value for the very first beep of the second subject. Hence, the lagged value for the very first beep should always be missing for each subject. Moreover, as we will discuss in more detail later, we may not want to lag values across the night (e.g., to use the value from the 10th beep of the first day as the lagged value for the first beep on the second day). To avoid this, all values of the lagged variable for the first beep within each day (not just the first) should be set to a missing value.

8.5 Data visualization

An important first step when working with ESM data is data visualization, which can reveal patterns and subject-level differences that would be difficult to spot when looking at the raw data in tabular form. For example, Figure 8.2 shows the reported positive affect (on a 1–7 scale) over the course of the study for six randomly selected subjects from each of the three mental health status groups (c = control, d = depressed, p = psychotic). The x-axis gives the time of the assessment in hours after midnight starting from the first data collection day. The alternating grey and white shading within each subfigure corresponds to the six data collection days. Such a figure immediately reveals considerable differences between participants. For example, while participant 'c115' quite often rates their positive affect at the upper end of the continuum (i.e., at 7) and only shows occasional dips that never go lower than a 4, participant 'd215' shows the opposite pattern with many ratings at the lower end (i.e., at 1) interspersed with moments where positive affect is higher. Some participants show an increase in their positive affect over time (e.g., 'c119'), others a decrease (e.g., 'p373'). Finally, we see participants with very low variability in their ratings (e.g., 'p328') and others with substantial fluctuations over time (e.g., 'c154'). Therefore, such a figure can reveal the considerable diversity in the participants' experience of their affect levels over the course of the six study days.

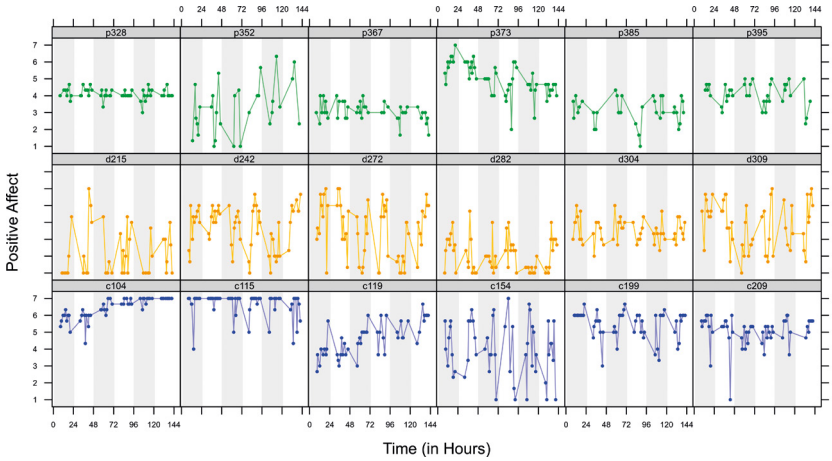


Figure 8.2 Positive affect over time for 6 randomly selected subjects from each of the three mental health status groups (c = control, d = depressed, p = psychotic)

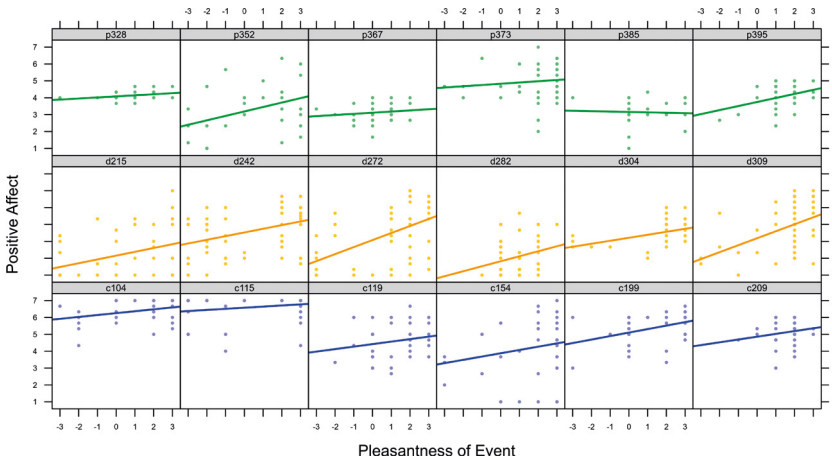


Figure 8.3 Positive affect as a function of the pleasantness of the most important event since the previous beep for 6 randomly selected subjects from each of the three mental health status groups (c = control, d = depressed, p = psychotic) with per-subject regression lines superimposed

Instead of time, one can also place some other time-varying variable on the x-axis. Figure 8.3 provides an illustration of this, with the rating of the pleasantness of the most important event since the previous beep placed

on the x-axis. Since the points are not ordered sequentially over time as in Figure 8.2, connecting them via lines is not sensible. However, one can add per-subject regression lines (based on a simple regression model using only the data of each individual subject) to such a figure, illustrating the (linear) association between the two variables within each participant. While the association between the two variables tends to be positive, we see some exceptions (e.g., 'p385') and differences in the slope of the regression line even if it is positive (c.f. 'p367', 'd272' and 'c119'). Again, the figure reveals considerable diversity across participants in how these two variables appear to be related to each other, which will also be relevant once we start considering the analysis of these data in more detail.

8.6 Conclusion

This chapter provided an overview of how ESM datasets are typically structured (i.e., in a long format), the type of variables they contain (i.e., subject-level, design and time-related and response variables), some important data checks, common data preparation steps and some examples of how to visualize ESM data. In the next chapter, we will examine statistical models for analysing such data.

CHAPTER 9

Statistical Methods for ESM Data

Wolfgang Viechtbauer

Before we elaborate on the statistical methods, it is worth considering what type of research questions can be addressed with ESM data. The goal here is not to provide an exhaustive list since the specific questions to be addressed will depend on the purposes of a study. However, many research questions fall into one of several categories (see also Chapter 2 and Bolger et al., 2003, for further details). First, on a more descriptive level, we may simply be interested in the mean level of a particular variable (e.g., positive affect) and its variability within and between study participants. A next step may be to compare differences in the mean level of a variable across groups (e.g., whether patients report, on average, lower levels of positive affect compared to healthy controls). The full strength of ESM, however, comes into play once we start to examine the within-person relationship between some outcome of interest (e.g., positive affect) and some time-varying predictor (e.g., stress) and how the strength of such a relationship may differ across groups (e.g., whether the relationship between positive affect and event pleasantness differs for patients and controls). Note that time itself can (unsurprisingly!) be considered a time-varying predictor, leading to questions about changes in some outcome over time and how the amount of change may differ across groups (e.g., whether those in a treatment group show larger increases in positive affect over time compared to those in a control group).

The possibility to examine research questions about within-person relationships is, in fact, one of the major benefits of using such an intensive data collection method. Suppose we are interested in the association between stress and affect (i.e., do we see decreased levels of positive affect when people report having experienced something stressful or unpleasant?). If we measure both these variables once in a large group of individuals, we can examine their cross-sectional association, and we might indeed find that people who report an elevated level of stress also tend to report lower positive affect. However, such a finding does not

allow us to differentiate the between- and the within-person relationships between these variables. In other words: Is it that individuals who are *on average* more stressed also tend to have lower positive affect? Or is it that when an individual is stressed *at particular moments*, he/she also tends to experience lower positive affect? While we are often interested especially in the latter, cross-sectional associations are an unknown mixture of these two phenomena and hence should not be used to draw any inferences about within-person associations. On the other hand, if we go through the trouble of repeatedly collecting information about an individual's level of affect across varying levels of stress, we can examine how these two variables are related to each other within the individual. If we collect such longitudinal data in an entire group of individuals, we can estimate and distinguish both the between- and within-person relationships. We will return to this topic in more detail further below.

9.1 Mixed-effects and multilevel models

Although an ESM study is essentially a repeated measures design, classical analysis procedures such as repeated measures or multivariate analysis of variance are not typically used in this context as they cannot easily handle the complexities involved (e.g., missing data, unequally spaced time points, time-varying covariates, autocorrelated observations). Instead, mixed-effects models (e.g., Harville, 1977; Henderson et al., 1959; Laird & Ware, 1982) are typically the method of choice for the analysis of ESM data (Bolger et al., 2003). Mixed-effects models extend the standard regression model by including additional 'random effects' which can be used to account for person-level differences in the model coefficients (i.e., intercepts and slopes). We will see in a moment how exactly this is done.

Mixed-effects models are also commonly used to analyse multilevel data (e.g., Goldstein, 2011; Hox et al., 2017; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999), which is also how we can think about the structure of ESM data, with the repeated assessments (level 1) nested within subjects (level 2), leading to a two-level model. One can extend this to a three-level model if we consider the repeated assessments (level 1) as nested within days (level 2) which in turn are nested within subjects (level 3). Using appropriate mixed-effects / multilevel models, we can then properly disentangle subject-level (and day-level) variability in an outcome

of interest (i.e., some subjects may tend to report higher/lower positive affect overall or on particular days) and within-subject variability (i.e., the degree of variability in the moment-to-moment assessments of positive affect). Similarly, when examining within-person relationships, we want to account for the fact that the strength of the relationship is likely to differ across subjects (and possibly also across days within subjects). Doing so requires allowing the slope of the regression line (relating the outcome to some predictor of interest) to vary across subjects (and possibly also across days within subjects).

9.2 Disentangling within- and between-person variability

As a first step toward this disentangling, we will examine to what extent the variability in some outcome of interest is due to within- and between-person differences. Let y_{ij} denote the observed outcome of the ith subject at the jth assessment moment (hence, in the illustrative dataset, $i = 1, \ldots, 328$ and $j = 1, \ldots, 60$). Then a two-level model to analyse the data is given by

$$y_{ij} = \mu + u_i + \epsilon_{ij}, \tag{1}$$

where μ is the average outcome (the model 'intercept'), u_i is a subject-level random effect that allows the outcome to be higher or lower on average for a particular participant, and ϵ_{ij} is a random effect that allows for the outcome at a particular moment to be higher or lower than the subject-specific average (the latter is often referred to as the 'error' term but really reflects within-person variability). Hence, we can think of $\mu + u_i$ as the average outcome of the ith subject and μ as the average of these subject-specific averages (so in essence, μ is the average of averages!).

This idea is illustrated in Figure 9.1, which shows the raw values of positive affect as reported by three participants (one from each mental health state group) from the illustrative dataset (the 5th, 183rd and 268th subject in the dataset). To better distinguish individual points, the data were slightly 'jittered' for this illustration; that is, a small amount of random noise was added to each point. The larger points with the black outline correspond to the (estimated) subject-specific averages, which deviate from the (estimated) overall average $\hat{\mu}$, which is indicated by the thick

gray horizontal line. The faint lines extending from the subject-specific averages to the individual observations represent the observed errors (also often referred to as the 'residuals').

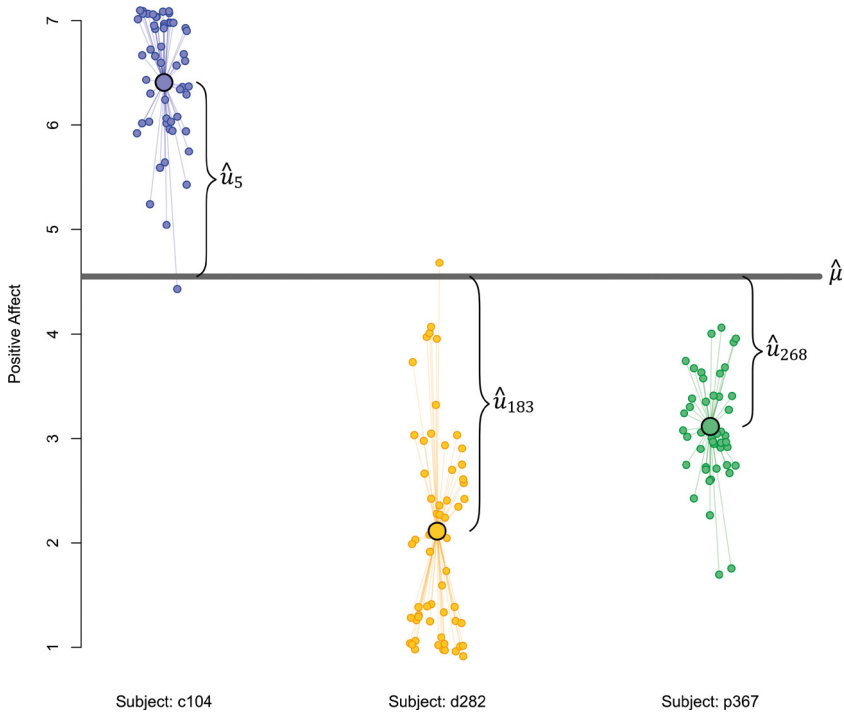


Figure 9.1: The values of positive affect for three subjects from the illustrative dataset, their estimated subject-specific averages, and the estimated overall average (c = control, d = depressed, p = psychotic)

In the context of the model above, we are (usually) not so interested in the subject-specific averages themselves but rather in their variability. In other words, how much do people differ from each other with respect to their average positive affect? For this, we assume that the deviations between μ and the subject-specific averages are normally distributed (i.e., $u_i \sim N(0, \tau^2)$), and hence τ^2 represents between-subject variability. Moreover, we are interested in how much the outcome varies within subjects. For this, we assume that the deviations of the actually observed values from the subject-specific averages are normally distributed (i.e.,

$\epsilon_{ij} \sim N\left(0, \sigma^2\right)$), and hence σ^2 denotes within-subject variability. Fitting this model to the illustrative dataset with positive affect as the outcome variable yields $\hat{\mu} = 4.55$, $\hat{\tau}^2 = 0.920$, and $\hat{\sigma}^2 = 1.015$.

Therefore, the estimated overall average level of positive affect on a 1–7 scale is $\hat{\mu} = 4.55$ although there is some uncertainty in this estimate as reflected by its standard error ($SE[\hat{\mu}] = 0.054$). An approximate 95% confidence interval (CI) for μ is given by $\hat{\mu} \pm 1.96 SE[\hat{\mu}]$, which yields 4.44 to 4.65. Hence, we can be fairly certain that this interval captures the true overall average level of positive affect in the population of individuals from which these 328 participants have come.[1]

The confidence interval above reflects the uncertainty in our estimate of μ, but it does not tell us anything about the distribution of the subject-specific averages themselves (it only says where we estimate the center of this distribution is located, under the assumptions of the model). To say something about the entire distribution of the subject-specific averages, we also need to consider their variance (and make an assumption about their distribution). Recall that $\hat{\tau}^2 = 0.920$ is the (estimated) variance of the subject-specific averages and hence, under the normality assumption, a rather different interval can therefore be computed, namely $\hat{\mu} \pm 1.96\hat{\tau}$, which estimates where the subject-specific average for approximately 95% of individuals is expected to fall. This fairly wide interval extends from 2.67 to 6.43 and therefore indicates considerable differences in the average level of positive affect across participants. Figure 9.2(a) shows a histogram of the estimated subject-specific averages for all 328 participants based on the model (i.e., the $\hat{\mu} + \hat{u}_i$ values), with a normal distribution with mean $\hat{\mu} = 4.55$ and variance $\hat{\tau}^2 = 0.920$ superimposed. Although the distribution of the estimated subject-specific averages is not exactly normal, it is fairly well approximated by a normal distribution.

[1] Before somebody sends angry emails about this interpretation of the confidence interval: Yes, this particular interval either covers the unknown but fixed value of μ or it does not, and probability statements about specific confidence intervals are, strictly speaking, not permissible in a frequentist framework. This doesn't change the fact that we can remain fairly certain that this particular interval captures μ.

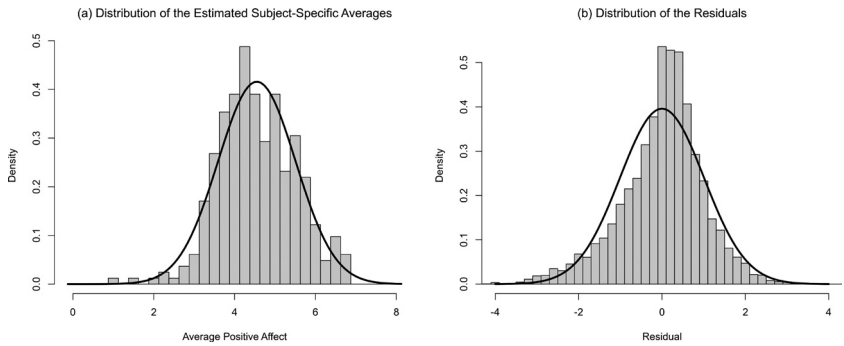


Figure 9.2: Histogram of (a) the estimated subject-specific averages and (b) the residuals, with normal distributions superimposed on the histograms

It should be noted that the $\hat{\mu} + \hat{u}_i$ values are not the same as the observed means of the positive affect values of each participant, which we can denote with $\overline{y}_i$. The former are so-called 'best linear unbiased predictions' (BLUPs) and are estimated from the mixed-effects model. Although the difference between these two sets of values is mostly negligible in these data, the BLUPs have an interesting property: They tend to be slightly pulled (or 'shrunken') towards $\hat{\mu}$, the more so when a participant has responded to a relatively low number of the assessments. For example, while the observed means of the three subjects shown in Figure 9.1 are 6.45, 2.07 and 3.08, their corresponding BLUPs are 6.41, 2.11 and 3.11, which are all pulled (just slightly) towards $\hat{\mu} = 4.55$. For participants that have responded to a lower number of beeps, this 'shrinkage effect' will tend to be more pronounced.

On an intuitive level, we can understand this phenomenon as follows. If little information is available about an individual's average level of positive affect (i.e., he or she has only responded to a low number of beeps), then we should pay relatively little attention to $\overline{y}_i$ (since this will be an inaccurate estimate) and instead give more weight to $\hat{\mu}$, the estimated overall average outcome of the entire group of participants. On the other hand, if a participant has responded to many assessments, then $\overline{y}_i$ is much more informative about this individual's average level of positive affect and the value of $\hat{\mu}$ is relatively unimportant. We can therefore think of the BLUPs as an optimal combination of $\hat{\mu}$ and $\overline{y}_i$ (depending on the amount

of information available about an individual) and hence $\hat{\mu} + \hat{u}_i$ will lie somewhere between these two extremes (i.e., between $\hat{\mu}$ and $\bar{y}_i$).

As noted earlier, the fluctuations of the actually observed outcomes around the subject-specific averages are the residuals.[2] Figure 9.2(b) shows their distribution with a normal distribution with mean 0 and variance $\hat{\sigma}^2 = 1.015$ superimposed. The distribution is slightly more peaked than would be expected under a normal distribution (the distribution is said to be 'leptokurtic') but otherwise is, again, fairly well approximated by a symmetric 'bell-shaped' distribution. Hence, approximately 95% of the residuals should fall within the interval $\pm 1.96\hat{\sigma}$, which in this case is given by −1.97 to 1.97 or roughly ±2. Hence, if we would consider an individual whose average positive affect is equal to 3 (so about 1.5 points lower than $\hat{\mu} \approx 4.5$), then at approximately 95% of the measurement occasions, this person's actually observed positive affect level would be expected to fall between 1 and 5.[3]

Based on the estimates of the between- and within-person variance components, we can compute the so-called intraclass correlation coefficient (ICC), which is given by

$$\text{ICC} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

and hence reflects how much of the total variance (which is simply the sum of the between- and within-person variances) is due to between-person differences. For our running example, both variance components are estimated to be of roughly equal size, and we therefore find ICC = 0.920 / (0.920 + 1.1015) ≈ 0.48. Hence, approximately half of the variability in positive affect is due to between-person differences and the other half due to momentary fluctuations.

The ICC derives its name from the fact that, under the two-level model outlined earlier, it describes the extent to which multiple observations from the same individual are correlated with each other. Large values of

[2] To be precise, these are the within-person residuals given by $e_i = y_i - (\hat{\mu} + \hat{u}_i)$. One could also compute residuals around the overall average (i.e., $e_i = y_i - \hat{\mu}$), but these conflate between- and within-person variability and are therefore usually not of interest.

[3] This calculation is only a rough approximation as it assumes normality of the residuals and ignores the fact that σ^2 might actually differ across individuals. In fact, as we noted earlier (cf. Figure 8.2), we actually see differences in how much positive affect fluctuates within individuals (see also the data of the three participants in Figure 9.1). Although models could be fitted that account for such differences, this is not common practice.

the ICC therefore indicate the need to use statistical models that account for the dependence in the outcome variable arising from the multilevel structure of the data. Although a test is sometimes conducted to examine if H_0: ICC = 0 can be rejected as a precondition to using multilevel modeling, we consider this practice unnecessary for ESM data, where, in our experience, the test is essentially always significant.[4]

9.3 Examining between-person differences

The two-level model introduced earlier is sometimes called the 'empty model' as it does not include any predictor variables. However, as we saw above, participants differ considerably in terms of their average level of positive affect. As a next step in our analysis, we might therefore be interested in examining which types of participants tend to report lower versus higher positive affect on average.[5] For this, we can extend the model by including one or multiple subject-level predictor variables. For example, suppose we want to examine if there are differences between male and female participants, younger versus older participants and those in the three different mental health status groups; a corresponding model would be

$$y_{ij} = \beta_0 + \beta_1 male_i + \beta_2\left(age_i - 35\right) + \beta_3 dep_i + \beta_4 psy_i + u_i + \epsilon_{ij}, \qquad (2)$$

where $mal\,e_i$ is a 'dummy variable' coded 0 for female and 1 for male participants, $ag\,e_i$ is the age of the ith participant, $de\,p_i$ is coded 1 for those in the depression group and 0 otherwise and $ps\,y_i$ is coded 1 for those in the psychosis group and 0 otherwise (hence, $de\,p_i = ps\,y_i = 0$ for those in the control group). Note that the constant 35 is subtracted from the participants' age values, which makes the model intercept (i.e., β_0) more

[4] Not surprisingly, a test of H_0: ICC = 0 for these data is highly significant ($p < .001$). It may seem that this is due to the rather large sample size of the study, but this is not so. Even if we were to test this null hypothesis based on only six randomly selected participants, we would still have more than 99% power to reject the null hypothesis.

[5] These might be exploratory analyses (in which case they should be designated as such) or one might have formulated a number of a priori hypotheses to be tested. In the remainder of the chapter, we will not draw further distinctions between these two cases, but in practice, one should do so.

interpretable.[6] Therefore, β_0 denotes the average level of positive affect for female participants that are 35 years of age and that are in the healthy control group, β_1 reflects the difference in average positive affect between male and female participants, β_2 denotes how the average level of positive affect differs for participants that are one year apart in age, β_3 denotes the difference between those in the depression group versus those in the control group and β_4 denotes the difference between those in the psychosis group versus those in the control group. Assumptions about u_i and ϵ_{ij} are the same as before.

The results after fitting this model to the illustrative dataset are given in Table 9.1. The table provides the estimates of the regression coefficients (i.e., $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_4$), the corresponding standard errors (SE), the z-values (i.e., z = estimate / SE) and the p-values for testing whether the estimated coefficients are significantly different from 0 (i.e., if a z-value is further away from 0 than ±1.96, then $p < .05$, and hence we reject the null hypothesis that the true coefficient is equal to 0).[7] Hence, the estimated average level of positive affect for 35 year old female participants from the healthy control group is 5.2. Male participants are, on average, estimated to have a positive affect level 0.04 points higher, but this difference is not significant ($p = .68$) and the coefficient is so small as to be practically meaningless (for an outcome measured on a 1–7 scale). Similarly, while the coefficient for age suggests a 0.005 difference in the average level of positive affect for participants one year apart in age, the coefficient is not significantly different from 0 ($p = .32$), and even participants 20 years apart would be estimated to differ by only 0.1 points. However, participants from the depression group have, on average, a lower level of positive affect by 1.2 points compared to those from the control group ($p < .001$), a sizable difference. Similarly, the results indicate a lower average positive affect level by 0.8 points for those in the psychosis group compared to the healthy controls ($p < .001$).

[6] A common practice is to subtract the mean age of the participants (i.e., 36.5), but this is not a requirement. In fact, β_0 would then not be interpretable unless the mean age is also known/reported. Instead, we suggest manually choosing a meaningful value for 'centering' the age variable – a value which does not change the fit of the model but makes the interpretation of the intercept explicitly clear.

[7] Some software treats the test statistics as having (approximate) t-distributions and computes the p-values accordingly. Given the size of the dataset, this distinction makes very little difference in the present case.

Table 9.1: Results for the model examining differences in average level of positive affect as a function of sex, age and mental health status (control, depressed and psychotic)

coefficient	estimate	SE	z-value	p-value
intercept	5.173	0.088	58.502	<.001
male	0.041	0.102	0.408	.683
age − 35	0.005	0.005	0.994	.321
depressed	−1.155	0.127	−9.115	<.001
psychotic	−0.809	0.119	−6.776	<.001

As before, the model also provides estimates of the variance of the u_i and ϵ_{ij} values, namely $\hat{\tau}^2 = 0.702$ and $\hat{\sigma}^2 = 1.015$. While the latter is essentially unchanged, the between-subject variance component has decreased noticeably (earlier we found $\hat{\tau}^2 = 0.920$ based on model 1). The reason for this is that $\hat{\tau}^2$ now estimates the variability in the subject-specific average positive affect levels after accounting for differences due to sex, age and mental health status. Especially the latter appears to account for some of the differences in the average affect levels between participants, leading to a proportional reduction of

$$R^2 = \frac{0.920 - 0.702}{0.920} = 0.24$$

in this variance component. As noted above, we can also think of this as an R^2-type measure, indicating how much of the between-subject variance is accounted for by the predictors included in the model.[8] Since these predictors are subject-level variables, they are not expected to account for momentary fluctuations in affect levels, and hence the within-person variance component is essentially unchanged.

9.4 Examining within-person associations

As discussed earlier, one of the great strengths of ESM is that it allows the examination of within-person associations, which we will consider next. For this, we will use the pleasantness of the most important event since

[8] For this and other R^2-type measures to be discussed in this chapter, it can happen that the value is negative, in which case we just set the value to 0.

the previous beep (recall that this was rated at each assessment moment on a −3 to +3 scale) as a predictor of positive affect. Now the model is given by

$$y_{ij} = \left(\beta_0 + u_{0i}\right) + \left(\beta_1 + u_{1i}\right)eventpl_{ij} + \epsilon_{ij}, \tag{3}$$

where *eventp* l_{ij} denotes the reported event pleasantness value of the *i*th subject at the *j*th assessment and, as before, y_{ij} denotes the corresponding reported level of positive affect at that moment. The goal of this analysis is to examine if we see differences in levels of affect when participants report that particularly pleasant or unpleasant events have occurred prior to the assessment. Of course, one could replace event pleasantness with any other time-varying predictor in this model, including some variable reflective of the passage of time, such as the day number, beep number within each day, the observation number or the amount of time in minutes or hours that has passed since the first assessment day.

Similar to a standard regression model, the model describes the (linear) relationship between the predictor and the response variable in terms of an intercept and slope. However, what distinguishes such a mixed-effects model from a regular regression model is that it allows for the intercept and slope to differ across participants. This is accomplished by the inclusion of two random effects, denoted by u_{0i} and u_{1i}, which allow the intercept and slope of the *i*th subject to differ from β_0 and β_1, which in turn denote the average intercept and slope (hence, $\beta_0 + u_{0i}$ and $\beta_1 + u_{1i}$ are the intercept and slope of the *i*th subject). Figure 9.3 illustrates this idea, showing the (slightly jittered) raw data for three participants (one from each mental health state group). The thick gray line corresponds to the line defined by the estimated values of β_0 and β_1. The estimated deviation between the intercept of the fifth subject from the average intercept is also indicated (i.e., $\hat{u}_{0,5}$) as is the estimated slope of the 183rd subject (i.e., $\hat{\beta}_1 + \hat{u}_{1,183}$), which is somewhat steeper than the average slope.

As in the models described earlier, we are typically not interested in the subject-specific intercepts and slopes themselves but rather in their variability. For the random effects, we assume $u_{0i} \sim N\left(0, \tau_0^2\right)$ and $u_{1i} \sim N\left(0, \tau_1^2\right)$, and hence τ_0^2 denotes the between-subject variance in the intercepts and τ_1^2 the between-subject variance in the slopes. Moreover, the random effects might be correlated with each other, and we use ρ to denote their correlation. A positive intercept-slope correlation would indicate

that those who have a higher-than-average intercept also tend to have a higher-than-average slope while a negative correlation would suggest that higher-than-average intercepts are associated with shallower slopes.

Fitting this model yields $\widehat{\beta}_0 = 4.298$ ($SE\left[\widehat{\beta}_0\right] = 0.054$), $\widehat{\beta}_1 = 0.194$ ($SE\left[\widehat{\beta}_1\right] = 0.008$), $\widehat{\tau}^2_0 = 0.908$, $\widehat{\tau}^2_1 = 0.0111$, $\widehat{\rho} = -0.44$ and $\widehat{\sigma}^2 = 0.893$. Therefore, for 'neutral' events (i.e., when event pleasantness is equal to 0), the average level of positive affect is estimated to be 4.3. Per one-unit increase in event pleasantness, the average level is estimated to change by almost 0.2 points ($p < .001$). For very unpleasant (i.e., −3) events, this implies an average positive affect level of around 3.7 points and a level of around 4.9 for very pleasant (i.e., +3) events (see also the grey line in Figure 9.3), a difference of roughly $6 \times 0.2 = 1.2$ points.

Note that $\widehat{\beta}_1$ is the estimate of the average slope. However, for particular participants, the relationship between event pleasantness and positive affect could be stronger or weaker. Analogous to what we did earlier, we can compute an interval, namely $\widehat{\beta}_1 \pm 1.96\,\widehat{\tau}_1$, that should capture the slope of approximately 95% of particular individuals. In this example, the bounds of this interval are −0.013 and 0.400, ranging from essentially no (or a slightly negative) association between the two variables to an association that is twice as strong as the average slope. Hence, for an individual where the association between event pleasantness and positive affect is this pronounced, the difference in positive affect between very pleasant and very unpleasant events would amount to approximately 2.4 points (i.e., 6×0.4).

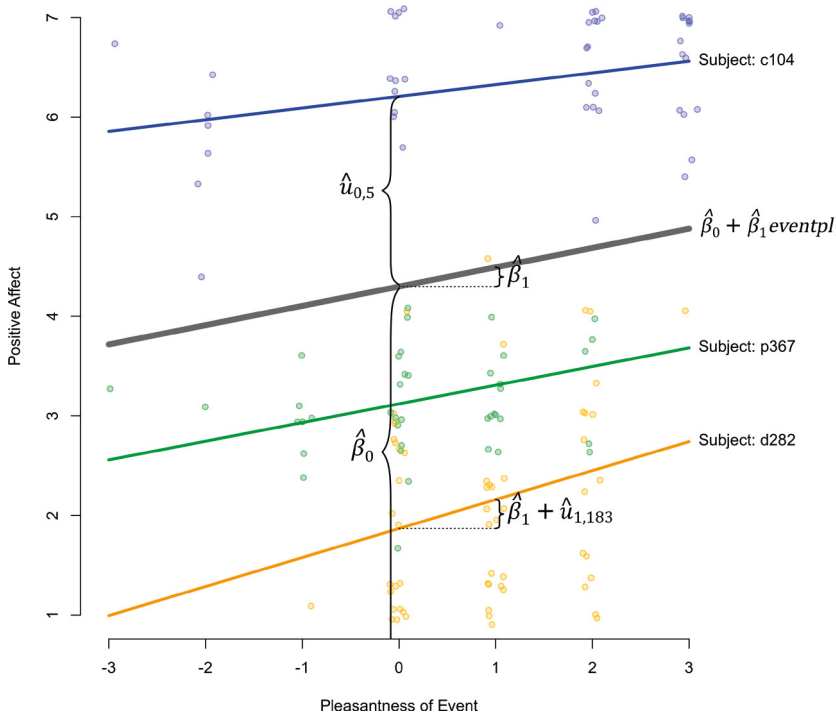


Figure 9.3: The values of positive affect as a function of event pleasantness for three subjects from the illustrative dataset, the estimated subject-specific regression lines and the overall regression line

Whether the assumption of normally distributed intercepts and slopes is (at least approximately) appropriate can be checked by examining histograms of the corresponding estimated values, as shown in Figure 9.4(a–b). These are, again, so-called BLUPs, which will exhibit the same shrinkage effect as described earlier; that is, the estimates of the subject-specific intercepts and slopes are pulled to some degree towards $\widehat{\beta}_0$ and $\widehat{\beta}_1$ when compared to the intercepts and slopes we would obtain when fitting simple regression models to the data from each individual subject (as shown in Figure 8.3). These distributions are not exactly normal, but the models described here are fairly robust to violations of the normality assumption anyway. Similarly, Figure 9.4(c) shows the distribution of the residuals, which are, again, a bit too peaked for a normal distribution but otherwise fairly symmetrically distributed around 0.

Finally, as noted earlier, intercepts and slopes are allowed to be correlated in this model. This is illustrated in Figure 9.4(d), showing a scatterplot of the estimated intercepts and slopes for the 328 participants. The estimated average intercept and slope are also indicated in the plot as dotted lines. The estimate $\hat{\rho} = -0.44$ suggests a negative relationship between these two sets of values, which we can also recognize in the plot. Hence, those with relatively high intercepts (i.e., above the average intercept) tend to have relatively shallow slopes (i.e., below the average slope). In other words, the level of positive affect of individuals with high average positive affect to begin with appears to be less sensitive to the (un)pleasantness of important events. This may reflect a certain insensitivity of individuals with high positive affect to the occurrence of unpleasantness/stressful events but might also be an artifact of a ceiling effect in the scale used to measure positive affect.

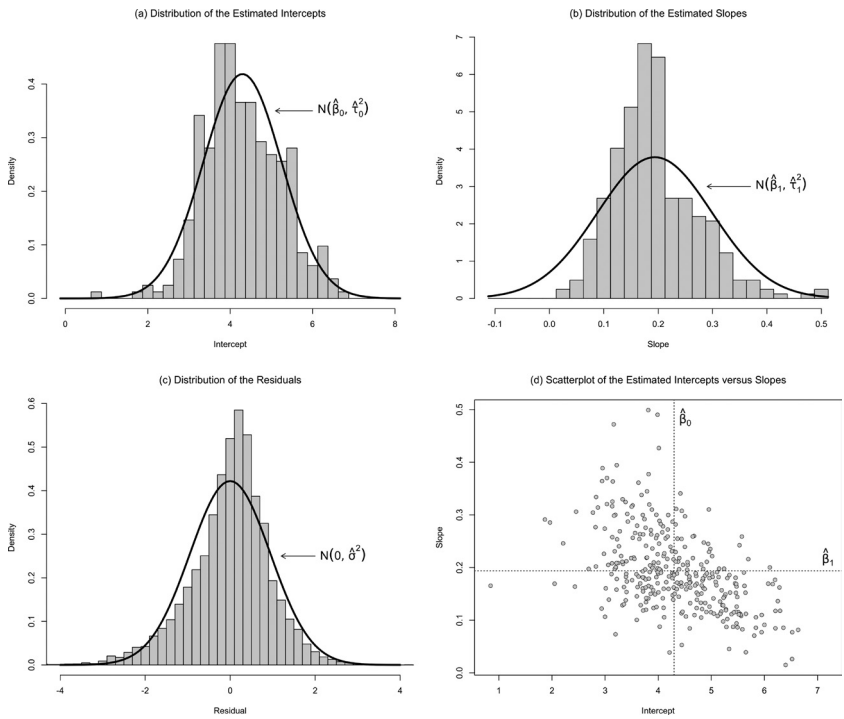


Figure 9.4: Histogram of (a) the estimated subject-specific intercepts, (b) the estimated subject-specific slopes, (c) the residuals and (d) a scatterplot of the estimated intercepts versus the slopes

We can think of model (3) as an extension of model (1) that attempts to account for fluctuations in the momentary levels of positive affect. Recalling that σ^2 reflects the within-person variability in the outcome variable (not already accounted for by any predictors included in the model), we can again calculate an R^2-type measure, now based on the proportional reduction in this variance component. For the example, we find

$$R^2 = \frac{1.015 - 0.893}{1.015} = 0.12$$

and hence event pleasantness accounts for approximately 12% of the variability in the momentary fluctuations in positive affect.

9.5 Examining between-person differences in within-person associations

The model above can be extended so that the average intercept and slope are allowed to differ as a function of one or multiple between-person variables. For example, let us examine if the strength of the association between positive affect and event pleasantness differs across the three mental health status groups. For this, we need to fit a model that includes dummy variables to distinguish the various groups plus their interaction with the time-varying predictor of interest. For this example, we would fit the model

$$\begin{aligned} y_{ij} = {} & (\beta_0 + \beta_1 dep_i + \beta_2 psy_i + u_{0i}) + \\ & (\beta_3 + u_{1i}) eventpl_{ij} + \beta_4 dep_i \times eventpl_{ij} + \beta_5 psy_i \times eventpl_{ij} + \epsilon_{ij}, \end{aligned} \tag{4}$$

where dep_i and psy_i are dummy variables coded as described earlier. Hence, β_0 denotes the average positive affect for neutral events for those in the control group, β_1 and β_2 indicate how the average level differs for those in the depression and psychosis groups compared to the control group, β_3 is the average slope of the association for those in the control group, and β_4 and β_5 indicate how the average slope differs for those in depression and psychosis groups again compared to the control group. Since dep_i and psy_i are person-level variables while $eventpl_{ij}$ is measured at the beep level, $dep_i \times eventpl_{ij}$ and $psy_i \times eventpl_{ij}$ are sometimes referred to as

'cross-level interactions' (e.g., Snijders & Bosker, 1999; Raudenbush & Bryk, 2002). The random effects u_{0i} and u_{1i} allow for between-person differences in the association between affect and event pleasantness not accounted for by group membership. Results for this model are given in Table 9.2. In addition, we find $\hat{\tau}^2_0 = 0.646$, $\hat{\tau}^2_1 = 0.0085$, $\hat{\rho} = -0.26$ and $\hat{\sigma}^2 = 0.893$.

Table 9.2: Results for the model examining differences between the mental health status groups with respect to the association between positive affect and event pleasantness

coefficient	estimate	SE	z-value	p-value
intercept	4.975	0.079	63.002	<.001
depressed	−1.227	0.112	−10.987	<.001
psychotic	−0.807	0.113	−7.132	<.001
event pleasantness	0.140	0.013	10.671	<.001
depressed × event pleasantness	0.124	0.018	6.920	<.001
psychotic × event pleasantness	0.025	0.019	1.296	.195

Hence, the results indicate significant differences between the average intercepts of the control and the depression and psychosis groups, with those in the latter two groups showing significantly lower average levels of positive affect for neutral events (both $p < .001$). This is most pronounced in the depression group, with an average intercept almost 1.3 points lower compared to that of the control group. On the other hand, the average slope of the depression group is significantly higher ($p < .001$), suggesting a 0.26 (= 0.140 + 0.124) increase in average positive affect per one-unit increase in event pleasantness, in contrast to the 0.14 increase in the control group. Interestingly, while the average slope of the psychosis group is a bit higher than that of the control group, the difference is not significant ($p = .20$). Hence, while the psychosis and especially the depression groups show lower levels of positive affect overall, the latter group shows a heightened sensitivity in their affect levels to the (un)pleasantness of important events in their lives.

The average slopes of the three groups and the estimated subject-specific regression lines of all 328 participants based on this model are illustrated in Figure 9.5. The figure emphasizes that while the statements above reflect our findings with respect to average levels and associations, there is considerable variability in intercepts and slopes within the groups and hence also overlap across groups, such that we could easily pick out

an individual from the depression group that has a higher estimated intercept and shallower slope than some individual from the control group. Hence, caution must always be exercised when stating the conclusions as associations observed at the group level may not apply to every individual within a particular group.

By allowing the average slope to differ across groups, we are essentially trying to explain why some individuals may exhibit a steeper or a shallower slope. If the person-level variable included in the model can explain such differences to a certain extent, we should see a decrease in the slope variance when comparing model (3) with model (4). Based on the respective estimates $\hat{\tau}^2_1 = 0.0111$ and $\hat{\tau}^2_1 = 0.0085$, we find

$$R^2 = \frac{0.0111 - 0.0085}{0.0111} = 0.233$$

and hence group membership accounts for approximately 23% of the variability in the degree to which event pleasantness and positive affect are related to each other.

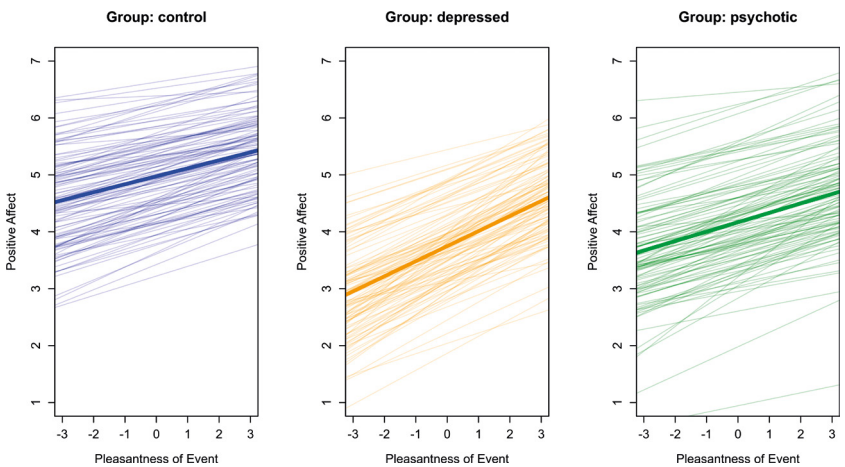


Figure 9.5: Estimated average and subject-specific regression lines for the association between event pleasantness and positive affect within the three mental health status groups

9.6 Disentangling within- and between-person associations

Models (3) and (4) above indicate that at moments where individuals report to have experienced a particularly pleasant event, their positive affect tends to be higher (and vice-versa). This might suggest that we have estimated the within-person relationship between the two variables, but it turns out that we are not quite there yet. In order to properly disentangle the within- and between-person relationships between event pleasantness and positive affect, we must take one further step (e.g., Curran & Bauer, 2011; Hoffman & Stawski, 2009; Wang & Maxwell, 2015). To do so, we need to fit the model

$$y_{ij} = (\beta_0 + u_{0i}) + \beta_1 \overline{eventpl}_i + \left(\beta_2 + u_{1i}\right)\left(eventpl_{ij} - \overline{eventpl}_i\right) + \epsilon_{ij}, \qquad (5)$$

where $\overline{eventpl}_i$ denotes the average of the event pleasantness ratings of the ith participant and hence $eventpl_{ij} - \overline{eventpl}_i$ corresponds to the within-person mean centered values of this variable (recall that we computed these values during the data preparation steps in the previous chapter). In this model, the intercept β_0 represents the average level of positive affect for participants whose average event pleasantness rating is equal to 0, and at moments when these participants' momentary event pleasantness rating is equal to their average level of event pleasantness, β_1 denotes the difference in the average level of positive affect for two individuals that differ in their average event pleasantness rating by one unit and β_2 is the average slope that describes how positive affect changes for a one-unit increase in the momentary event pleasantness rating. Fitting this model to our data yields $\widehat{\beta}_0 = 3.701$ ($SE\left[\widehat{\beta}_0\right] = 0.101$), $\widehat{\beta}_1 = 0.626$ ($SE\left[\widehat{\beta}_1\right] = 0.065$), $\widehat{\beta}_2 = 0.192$ ($SE\left[\widehat{\beta}_2\right] = 0.008$), $\widehat{\tau}^2_0 = 0.710$, $\widehat{\tau}^2_1 = 0.0111$, $\widehat{\rho} = -0.36$ and $\widehat{\sigma}^2 = 0.893$.

In this model β_1 corresponds to the between-person association between event pleasantness and positive affect. The size of this coefficient addresses the question: Do we see differences in average levels of positive affect between individuals who, on average, rate the events they experience as more or less pleasant? Figure 9.6 illustrates this idea by showing a scatterplot of the average event pleasantness ratings of the 328 participants and their corresponding estimated average positive affect values based on the model (the grey points). The grey dashed line corresponds to this between-person relationship, which in essence models the association between these two sets of averages.

On the other hand, β_2 corresponds to the within-person association between the two variables. The size of this coefficient addresses the question: Do we see differences in positive affect levels at particular moments when an individual rates the pleasantness of the events that occurred above or below their average event pleasantness rating? Figure 9.6 also illustrates this association as the solid grey line. To be precise, we should note that β_2 really corresponds to the average within-person association since the model again allows for the slope of this coefficient (and participants' intercepts) to vary across participants. The subject-specific regression lines for the three participants introduced earlier are also shown in Figure 9.6 to illustrate this point (their average pleasantness ratings and corresponding estimated average positive affect levels are also highlighted as the larger points with the black outline). Note that the raw data for the three participants are not shown (in contrast to Figure 9.3) as that would have made the figure hard to read.

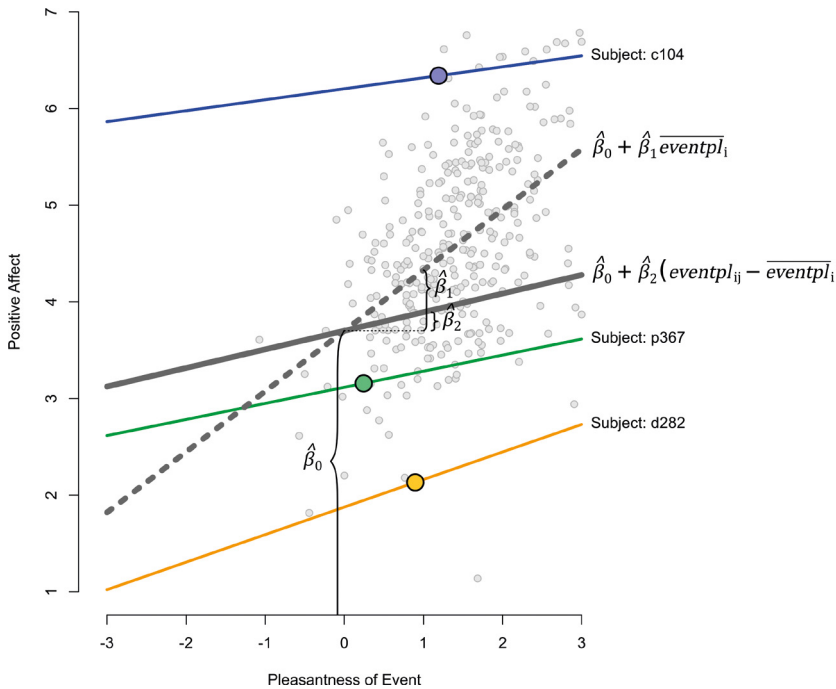


Figure 9.6: Scatterplot of the per-person average event pleasantness values versus the estimated average positive affect values, the between- and average within-person associations and the subject-specific regression lines for the within-person association for three subjects

Although both null hypotheses $H_0: \beta_1 = 0$ and $H_0: \beta_2 = 0$ are firmly rejected (both $p < .001$), the results indicate that the slope of the between-person association is more than three times larger than that of the within-person association. A proper test of the null hypothesis $H_0: \beta_1 = \beta_2$ also leads to rejection ($p < .001$), further emphasizing the usefulness of differentiating between these two types of relationships. In fact, for certain predictor-outcome combinations, not only the magnitude but also the direction (i.e., sign) of the two coefficients could differ, leading to very different conclusions about the relationship between the two variables at the person and at the beep (i.e., momentary) level.

In contrast, coefficient β_1 in model (3) can be shown to be a mixture of the between- and within-person associations (e.g., Snijders & Bosker, 1999) and hence does not provide a 'pure' estimate of the within-person association between the two variables. Therefore, model (5) is usually recommended when examining the association between a time-varying predictor and outcome in ESM data (e.g., Bolger & Laurenceau, 2013). We generally support this position, although in practice, one will typically find little difference in the respective estimates from the two models. For example, recall that we found $\widehat{\beta}_1 = 0.194$ based on model (3) and $\widehat{\beta}_2 = 0.192$ in model (5), and hence there is practically no difference between the two estimates. The reason for this is simple: Due to the large amount of data available within each individual, even coefficient β_1 in model (3) mostly reflects the within-person association. We can also see this reflected in the standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in model (5). The so-called 'relative efficiency' with which these two coefficients can be estimated is given by $SE\left[\widehat{\beta}_1\right]^2 / SE\left[\widehat{\beta}_2\right]^2$, which is approximately 64.5 in this case (hence, the precision of the estimate of the within-person association is more than 60 times larger than the precision of the estimate of the between-person association). Not coincidentally, the relative efficiency, that is, the ratio of the total number of observations in the dataset (which is roughly the relevant sample size for estimating the within-person association) over the number of participants (which in turn is more relevant for estimating the between-person association), is also close to 19680 / 328 = 60, although the exact link between these two ratios is more complex. In any case, we also support the use of model (5) as it not only provides us with estimates of the association at both levels but is only marginally more complex than model (3).

As an R^2-type value, we can again compare the estimate of the error variance from model (5) (i.e., $\widehat{\sigma}^2 = 0.893$) with that of model (1) (i.e., $\widehat{\sigma}^2 = 1.015$) just as we did earlier when comparing model (3) with (1). This yields $R^2 = 0.12$ and hence the same conclusion that event pleasantness accounts for approximately 12% of the momentary fluctuations in positive affect.

Analogous to model (4), we could now proceed to examine if both the within- and between-person associations differ between the three mental health status groups. For this, we include the dummy variables dep_i and psy_i in the model and allow these to interact with $\overline{eventpl}_i$ and $eventpl_{ij} - \overline{eventpl}_i$. For the same reasons explained above, the results with respect to the group differences in the within-person associations from this model will be almost identical to those from model (4), and hence we omit these results here. However, in addition, the results from this model now indicate whether the between-person relationship differs across the three groups. The results suggest no significant differences in the slopes at this level (p = .66 for the difference between the depression and the control group and p = .93 for the difference between the psychosis and the control group).

9.7 Examining lagged relationships

In models (3) through (5), the value of one time-varying variable (event pleasantness) was used as the predictor for another time-varying variable (positive affect). The results indicate that these two variables are associated with each other at both the between- and within-person levels and that the within-person association is on average stronger for participants in the depression and psychosis groups compared to those in the control group. So far, with respect to the within-person association, we have tried to avoid any kind of phrasing that would attribute causality, although we might be tempted to conclude based on the results so far that the occurrence of particularly pleasant/unpleasant events tends to lead to increases/decreases in positive affect (a more causally sounding framing of the findings). However, while participants were asked to rate the (un)pleasantness of the most important event that happened since the previous beep (i.e., they were asked to make an assessment of an event that happened before the assessment of their current level of positive affect), one might be concerned that a participant's level of positive affect at a given beep (which

could be high or low due to reasons unrelated to the (un)pleasantness of prior events) might impact how they rate the (un)pleasantness of an event that happened previously. In other words, doubts can be raised with respect to the directionality of the relationship between the two variables.

Simply reversing the predictor and outcome variables in the previous models does not help shed any light on this issue. We would then still be examining the concurrent association between the two variables (irrespective of how the question about event pleasantness was phrased). Instead, to more firmly establish the temporal sequence of events, we must place the hypothetical cause before our measurement of the alleged effect; that is, we use the event pleasantness rating from the *previous* beep as the predictor of the positive affect level at the *current* beep and so on. We therefore fit the model

$$y_{ij} = \left(\beta_0 + u_{0i}\right) + \left(\beta_1 + u_{1i}\right) eventpl_{ij-1} + \epsilon_{ij}, \tag{6}$$

where $eventp\,l_{ij-1}$ denotes the 'lagged' version of the event pleasantness variable that we created during the data preparation step. For this model, we find $\widehat{\beta}_0 = 4.464$ ($SE\left[\widehat{\beta}_0\right] = 0.055$), $\widehat{\beta}_1 = 0.080$ ($SE\left[\widehat{\beta}_1\right] = 0.008$), $\widehat{\tau}^2_0 = 0.933$, $\widehat{\tau}^2_1 = 0.0063$, $\widehat{\rho} = -0.37$ and $\widehat{\sigma}^2 = 0.981$. Hence, while we still find a significant association between (the lagged) event pleasantness rating and the level of positive affect ($p < .001$), the average slope now suggests a considerably weaker relationship between these two variables (recall that we earlier found $\widehat{\beta}_1 = 0.194$ based on model 3). Therefore, the difference in average positive affect for very pleasant versus very unpleasant events now amounts to only 0.5 points. Also, comparing the estimated error variance from this model with that of model (1) (i.e., $\widehat{\sigma}^2 = 1.015$) indicates a much smaller proportional reduction in the error variance (i.e., $R^2 = 0.03$ or 3%).

Several aspects of this analysis need highlighting. First, recall that the interval between assessments was on average approximately 90 minutes long. At a given beep, participants were asked to recall the most important event since the previous beep and rate its (un)pleasantness. If we assume that the recalled event was equally likely to occur at any point during the inter-prompt interval, then it would on average have occurred about 45 minutes before the beep.[9] Ignoring the directionality

[9] Participants may quite plausibly be inclined to recall more recent events (i.e., less than 45 minutes ago on average), especially if the event had a substantial impact on their mood, but we will ignore this intricacy.

issue discussed above, the results from model (3) therefore reflect the association between the (un)pleasantness of an event that happened on average 45 minutes ago and positive affect. On the other hand, the results from model (6) can then be thought of as the impact of events that happened on average 90 + 45 = 135 minutes ago on the participants' level of positive affect. Given this much longer lag, it is not surprising that the potential impact of very pleasant or unpleasant events on participants' mood might have dissipated to some extent in the meantime. It is actually quite remarkable to find that such effects may still linger on more than two hours later.

Related to the previous point, a second issue to consider is the fact that the interval between adjacent beeps was not actually constant in this study. Hence, the event pleasantness rating used to predict the subsequent level of positive affect might have been assessed 20 or up to 160 minutes ago (the possible range of values for the inter-prompt interval). Model (6) ignores this complication and essentially treats the assessments as evenly spaced. A possible recourse to address this issue would be to actually compute the inter-prompt interval between adjacent beeps and allow this to interact with the lagged event pleasantness predictor (Selig et al., 2012). Another approach is to make use of continuous-time models, which naturally take the unequally spaced measurements into consideration. The interested reader is referred to Ryan and colleagues (2018) and the references therein for further details on this approach.

To reduce concerns about treating the measurements as equally spaced, recall that the value of the first beep within each day was set to missing for the lagged event pleasantness variable. If we had not done so, we would also be using the event pleasantness rating from the very last beep of each day as the value of the predictor for the positive affect level on the following morning, roughly 13–14 hours later. Presumably, the lagged association spanning the entire night is of a rather different magnitude and nature than that within a day, and hence, by setting the lagged value of the first beep within each day to missing, we essentially remove all first beeps from the analysis. Therefore, $\widehat{\beta}_1$ from model (6) represents the average within-day lagged association between the two variables.

As a consequence of this step, the size of the dataset actually used to fit model (6) is reduced compared to model (3). The size of the usable data in such a lagged analysis is further reduced due to non-compliance by the participants to the assessment schedule. In particular, if a

participant does not respond to a particular beep, the event pleasantness rating and level of positive affect are, of course, missing, and hence this beep will also not be considered when fitting model (3). However, the subsequent beep, even if responded to by the participant, will also not be usable in a lagged analysis since the lagged event pleasantness rating will then be missing for this beep as well. Hence, while in fact 14,119 complete pairs of event pleasantness and positive affect values were used to fit model (3), only 10,923 rows of the dataset could be used to fit model (6). A total of 1,178 rows were lost as a consequence of setting the lagged event pleasantness to 'missing' for the first beep within each day, and the remaining 2018 rows were lost due to intermittent missing values. Hence, the number of observations used for the analysis was about 23% lower when fitting model (6) compared to model (3). Although steps could be taken to mitigate this loss of data to at least some extent (e.g., by setting the lagged event pleasantness value to the most recent non-missing lagged observation within a day), this would again require further adjustments to the model to account for the fact that some lagged values have come not from the prior beep but from an earlier time point (e.g., if beeps 5 and 7 have been responded to but beep 6 is missing, then the event pleasantness rating from beep 5 could be used to predict positive affect at beep 7, accounting for the increased lag between the beeps).

Note that model (6) could again be extended to examine group differences in the lagged association (by including dep_i and psy_i in the model and allowing these dummy variables to interact with $eventpl_{ij-1}$). We will skip reporting the detailed results here, but briefly note that we again find a significant group difference, with a significantly stronger lagged association for those in the depression group compared to the control group ($p < .001$).

One may wonder if one should disentangle the between- and within-person relationships also in the context of such lagged analyses. In principle, analogous steps could be taken as described earlier (i.e., using subject-level means and within-person mean centred values of the lagged event pleasantness variable as predictors), although for the same reasons outlined earlier, the size of the average within-person lagged coefficient in such a model is again virtually identical to that from model (6) for these data.

9.8 Controlling for autocorrelation

Lagged associations as described above play yet another role in the context of ESM analyses. One of the aspects not yet considered is the amount of 'autocorrelation' in the outcome variable. For the present analyses, this refers to the extent to which the level of positive affect at one assessment moment is predictive of the positive affect level at the subsequent assessment. Many variables of interest in ESM analyses will exhibit such autocorrelation, the more so the closer in time two measurements are taken. The amount of autocorrelation can be estimated by fitting model (6), this time using the lagged outcome variable itself as the predictor of interest. This model can then be further extended with other predictors, which may also be lagged variables. For example, in the model

$$y_{ij} = \left(\beta_0 + u_{0i}\right) + \left(\beta_1 + u_{1i}\right) y_{ij-1} + \left(\beta_2 + u_{2i}\right) eventpl_{ij-1} + \epsilon_{ij}, \tag{7}$$

coefficient β_1 denotes the average autocorrelation while β_2 corresponds to the average lagged relationship between event pleasantness and positive affect. In this model, we estimate the lagged relationship while 'controlling' for the autocorrelation in the outcome variable, which indicates the unique effect of event pleasantness on affect, over and beyond what is already accounted for by the autocorrelation in the outcome variable itself. Note that the model includes random effects corresponding to both these coefficients and hence allows the strength of the autocorrelation and of the lagged relationship to differ across participants. Model (7) is therefore a model with random intercepts and two random slopes, all of which have associated variances (i.e., τ_0^2, τ_1^2 and τ_2^2 are the variances of u_{0i}, u_{1i} and u_{2i}, respectively) and are allowed to be correlated (i.e., ρ_{01} is the correlation between u_{0i} and u_{1i}, ρ_{02} is the correlation between u_{0i} and u_{2i} and ρ_{12} is the correlation between u_{1i} and u_{2i}).

The degree of autocorrelation in the outcome variable may, in fact, be of interest in itself. For example, 'emotional inertia' is reflected by a high degree of autocorrelation in reported emotional states, which in turn may be predictive of lower well-being and psychological maladjustment (Kuppens, Allen, et al., 2010). Furthermore, the degree of autocorrelation may be useful to obtain evidence about an appropriate sampling frequency. Very high autocorrelation would indicate oversampling and hence redundancy in the information obtained from assessments close

in time. On the other hand, autocorrelation close to zero could be taken as evidence of under-sampling, suggesting that we are, in fact, obtaining 'snapshots' of daily life instead of a more continuous 'movie'.

As before, we can allow the size of the lagged relationship and also the degree of autocorrelation to differ across groups. The corresponding model is given by

$$\begin{aligned} y_{ij} &= (\beta_0 + \beta_1 dep_i + \beta_2 psy_i + u_{0i}) + \\ &= (\beta_3 + u_{1i}) y_{ij-1} + \beta_4 dep_i \times y_{ij-1} + \beta_5 psy_i \times y_{ij-1} + \\ &\quad (\beta_6 + u_{2i}) eventpl_{ij-1} + \beta_7 dep_i \times eventpl_{ij-1} + \beta_8 psy_i \times eventpl_{ij-1} + \epsilon_{ij}, \end{aligned} \tag{8}$$

where β_3 denotes the average autocorrelation of the control group, β_4 and β_5 allow the average autocorrelation for the depression and psychosis groups to differ from that of the control group, β_6 denotes the average lagged relationship of the control group and β_7 and β_8 allow for group differences with respect to the lagged relationship. We will not report the results from models (7) and (8) because there is yet one last complication that we need to address first.

9.9 Controlling for time trends

Strictly speaking, β_1 in model (7) can only be interpreted as the autocorrelation coefficient in the absence of time trends in the outcome variable. If such time trends are not accounted for, they can lead to bias (typically overestimation) in the estimate of β_1. Moreover, it may be important to account for time trends to reduce or avoid the potentially confounding effects of time itself on the (lagged) association between the time-varying predictor and outcome of interest (e.g., Curran & Bauer, 2011; Hoffman & Stawski, 2009; Wang & Maxwell, 2015). For example, if people tend to experience more unpleasant events early in the day and, for unrelated reasons, also tend to have lower levels of positive affect in the morning, then this might lead to an apparent association that might mistakenly be attributed to a (potentially) causal link between these variables. By including some measure of time in the model, we can try to mitigate this problem. For this, we will include the actual time of the assessments within each day (the response time variable in hours after midnight) as an additional predictor in the model. Moreover, since trends might differ

across groups, we will allow the time variable to interact with the mental health status dummy variables. Therefore, the model is given by

$$\begin{aligned} y_{ij} &= (\beta_0 + \beta_1 dep_i + \beta_2 psy_i + u_{0i}) + \\ &= (\beta_3 + u_{1i}) y_{ij-1} + \beta_4 dep_i \times y_{ij-1} + \beta_5 psy_i \times y_{ij-1} + \\ &\quad (\beta_6 + u_{2i}) eventp\, l_{ij-1} + \beta_7 dep_i \times eventp\, l_{ij-1} + \beta_8 psy_i \times eventp\, l_{ij-1} + \\ &\quad (\beta_9 + u_{3i}) tim\, e_{ij} + \beta_{10} dep_i \times tim\, e_{ij} + \beta_{11} psy_i \times tim\, e_{ij} + \epsilon_{ij} \end{aligned} \quad (9)$$

where β_9 now denotes the average linear trend in positive affect in the control group and β_{10} and β_{11} allow for different average trends in the depression and psychosis groups. Note that the model also includes a random effect for the time variable so that trends are allowed to differ across participants within the different mental health status groups. The results for this model are given in Table 9.3. Also, $\hat{\tau}^2_0 = 0.739$, $\hat{\tau}^2_1 = 0.0298$, $\hat{\tau}^2_2 = 0.0032$, $\hat{\tau}^2_3 = 0.0003$ and $\hat{\sigma}^2 = 0.866$ (we skip reporting the six correlations between the random effects).

Table 9.3: Results for the model examining differences between the mental health status groups with respect to the lagged association between positive affect and event pleasantness while controlling for autocorrelation and time trends in positive affect

	estimate	**SE**	**z-value**	**p-value**
intercept	3.476	0.144	24.139	<.001
depressed	−1.057	0.187	−5.641	<.001
psychotic	−0.556	0.204	−2.725	.007
positive affect$_{j-1}$	0.325	0.025	12.821	<.001
event pleasantness$_{j-1}$	−0.004	0.012	−0.335	.738
response time	0.001	0.004	0.299	.765
depressed × positive affect$_{j-1}$	0.060	0.034	1.788	.074
psychotic × positive affect$_{j-1}$	−0.003	0.037	−0.089	.929
depressed × event pleasantness$_{j-1}$	0.031	0.017	1.882	.060
psychotic × event pleasantness$_{j-1}$	0.019	0.018	1.025	.305
depressed × response time	0.005	0.006	0.748	.455
psychotic × response time	0.000	0.007	0.042	.967

Hence, we estimate a significant average autocorrelation of about 0.33 in positive affect for those in the control group ($p < .001$). With 0.39, the average autocorrelation for those in the depression group is slightly higher (by 0.06 points), but not quite significantly so ($p = .07$), while the average autocorrelation for those in the psychosis group is essentially identical to that of the control group. There is actually no evidence for an average linear time trend in the control group, and neither the depression nor the psychosis groups differ from the control group in this respect. Finally, while we now no longer see a significant lagged relationship between event pleasantness and positive affect in the control group (the coefficient is essentially zero and $p = .74$), the results still suggest a slightly higher average slope in the depression group, but again, the difference between this group and the control group just fails to be significant ($p = .06$). To some extent these results therefore call into question whether there really is, on average, an association between event pleasantness and positive affect, at least when researchers use the lagged event pleasantness variable as the predictor (when using the concurrent value of event pleasantness in model 9, we find very similar results as those reported in Table 9.2).

Model (9) only accounts for time trends within days in the outcome variable. To account for time trends over the course of the entire study, one could, for example, include the total number of hours that have passed since the first assessment day as an alternative or as an additional predictor in the model. Moreover, as noted above, we are modelling linear trends, which might not be entirely realistic given that more complex diurnal patterns have been found in positive affect (e.g., Clark et al., 1989), which might also differ in shape across groups with different mental health conditions (e.g., Peeters et al., 2006). Such non-linear patterns could also be accounted for in the context of the models discussed above, but that is beyond the scope of this chapter.

To conclude this section, we should mention that time trends and diurnal patterns may also be of inherent interest and not just a means to avoid confounding of other relationships. Furthermore, if ESM is used as part of a measurement burst design, a change across phases could then be captured by including the phase identifier as a predictor in the model (and allowing this to interact with a grouping variable if group differences in phase changes are of interest).

9.10 Conclusions

The present chapter serves as an initial introduction to some of the main issues and approaches to consider when analysing ESM data. For readers interested in further details on the use of mixed-effects models in the context of ESM research, we would suggest to consult the excellent text by Bolger and Laurenceau (2013). Singer and Willet (2003) also provide a very accessible introduction to mixed-effects models for longitudinal data analysis in general. Those interested in further technical details could consult Demidenko (2004) and Verbeke and Molenberghs (2000). The latter also covers the use of SAS for analysing longitudinal data. For those interested in the use of R for mixed-effects modelling, we recommend Pinheiro and Bates (2000) and Gałecki and Burzykowski (2013) while the two-volume book by Rabe-Hesketh and Skrondal (2012) provides very thorough coverage of the use of Stata in this context. Bolger and Laurenceau (2013) also provide Mplus and SPSS code for the examples covered in their book.

CHAPTER 10

Non-Normal, Higher-Level, and VAR(1) Models for the Analysis of ESM Data

Ginette Lafit

This chapter introduces statistical approaches to analysing ESM data that complement the data analysis methods presented in Chapter 9. In particular, we introduce extensions to the linear mixed-effects model introduced in the previous chapter. This model is widely used in ESM research because it allows partitioning the variability in the data into variance at the individual level and at the measurement level. However, the linear mixed model has certain limitations. For example, it assumes that within-person errors are normally distributed. Therefore, in this chapter, we present the generalized mixed-effects model, which extends the linear mixed-effects model to analyse non-normal data. In particular, we focus on three types of outcomes: (1) a dichotomous outcome representing the occurrence of an event; (2) a count outcome representing the number of times an event has occurred during an ESM period; and (3) a positive and continuous outcome. Subsequently, we illustrate how to use mixed-effects models to account for the multiday-structure of ESM data (i.e., repeated measurements nested within days nested in persons). Finally, we discuss the multilevel vector autoregressive model of order one (VAR(1)) to examine the dynamic relationships between a set of variables.

10.1 Non-normal data

The underlying assumption of the linear mixed-effects model is that within-person errors are normally distributed. In certain situations, this assumption does not hold due to the nature of the variable. For example, ESM items can be dichotomous or binary. The occurrence of an event is, for instance, represented by a dichotomous variable. Examples include items that measure when a person is alone or in company with others. The outcome can also be discrete, representing the number of times an event

has occurred during an ESM period (e.g., alcohol consumption, number of stressful events, number of social encounters). ESM data can also include continuous outcomes with non-normal errors (e.g., symptom severity, negative affect). In these circumstances, the linear mixed-effects model is not suited to model these data. The generalized mixed-effects model extends the linear mixed model by allowing different distributions for the outcome variable (Hedeker & Gibbons, 2006; McCulloch & Neuhaus, 2014). In this section, we illustrate how to implement this model when the target outcome is non-normally distributed.

10.1.1 Dichotomous outcome

Suppose that we collect ESM data for N participants that are observed at times $t = 1, \ldots, T$. The outcome of interest is a dichotomous variable $y_{it} = \{0,1\}$. A dichotomous outcome reflects the occurrence of an event that can only take two possible values. For example, in an ESM study, we may be interested in assessing moments in which participants are alone, the occurrence of a stressful event, or moments when participants have been smoking or consuming alcohol. To analyse a binary outcome, we can use the logistic mixed-effects regression model (see Demidenko, 2013; Hedeker & Gibbons, 2006; McCulloch & Neuhaus, 2014; Zhang et al., 2011). The logistic mixed regression models link the predictor of interest, denoted by x_{it}, to the probability that the outcome is one, assuming the following form

$$log\left(\frac{Pr(y_{it} = 1 \mid v_0, v_1)}{1 - Pr(y_{it} = 1 \mid v_0, v_1)}\right) = \beta_0 + \beta_1 x_{it} + v_{0i} + v_{1i} x_{it}$$

where v_{0i} is the random intercept, and v_{1i} the random slope. The random effects are assumed to be multivariate normally distributed with mean zero and (2x2) covariance matrix Σ_v. The diagonal elements of the covariance matrix of the random effects are given by the variances $\sigma^2_{v_0}$ and $\sigma^2_{v_1}$; meanwhile, the off-diagonal elements by the covariance between the random effects denoted are by $\sigma_{v_{01}}$.

The logistic mixed-effects model can be estimated using the lme4 package (Bates et al., 2015) in R (R Core Team, 2020). To illustrate how to estimate this model, we use the dataset *data_company.* The dataset includes 100 individuals that participated in an ESM study with 10 beeps per day over seven days. At each beep, participants were asked to indicate whether they were in company with others or alone (0 = alone; 1 = not

alone) as well as their positive affect, computed as the mean of the items 'I feel happy right now', 'I feel relaxed right now' and 'I feel satisfied right now'. The items were rated on a 7-point Likert scale ranging from 1 (not at all) to 7 (very much). All the data sets presented in this chapter are publicly available in the git repository https://github.com/ginettelafit/ESM.Synthetic.DataSets.

To upload the dataset in R or Rstudio, we use the command read.table. In addition to the variables *Company* and positive affect (*PA*), the dataset includes the variable *id* indicating the participants' identification number, *day* denoting the study day, *beep* the prompt or beep number within a day and *obs* the number of time points within an individual.

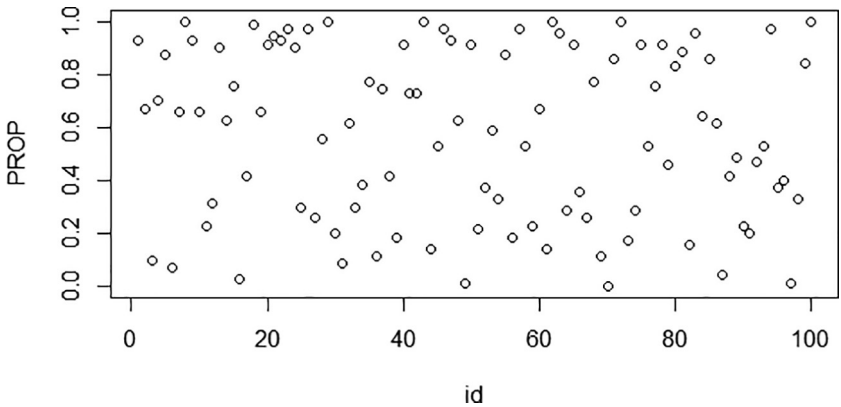
```
data_company = read.table(file="data_company.txt",header = TRUE, sep =
"")
head(data_company)
##   id day beep obs Company       PA
## 1  1   1    1   1       1 5.128097
## 2  1   1    2   2       1 4.481416
## 3  1   1    3   3       1 3.330492
## 4  1   1    4   4       1 3.920224
## 5  1   1    5   5       1 3.221645
## 6  1   1    6   6       0 3.176480
```

The descriptive statistics of the variable *Company* show that participants reported to be in company with others (*Company=1*) in approximately 60% of the total beeps.

```
summary(data_company$Company)
##    Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
##  0.0000  0.0000  1.0000  0.5744  1.0000  1.0000
```

We can also plot the distribution of the proportion of beeps at which participants reported to be in company with others. We first need to upload the library tidyverse (Wickham et al., 2019). Next, we plot the proportion of beeps where *Company* is one. The plot shows the distribution of the proportions of beeps where participants are not alone.

```
library(tidyverse)
data_company %>%
  group_by(id) %>%
  summarise(PROP = sum(Company)/n()) %>%
plot()
```



Before fitting the logistic mixed model, we person-mean centred the time-varying predictor *PA*. We use the library tidyverse, and for each participant's value of *PA* we subtract the participant's mean.

```
# Center predictor PA
data_company %>%
  group_by(id) %>%
  mutate(PA = PA - mean(PA))
```

The logistic mixed-effects model is fitted using the glmer function from the lme4 package. First, we estimate the intercept-only model. The first argument in glmer() is a formula that defines the structure of the fixed effects Company ~ 1. In this formula, Company is the dependent variable and 1 is the fixed intercept. The second argument corresponds to the random effect structure of the model (1|id), where 1|id corresponds to random intercept which is allowed to vary over participants (*id*). The next argument (data = data_company) indicates the data that will be used to estimate the model. To fit a logistic mixed-effects model, we need to

include the argument family = binomial(link=logit). The function summary() allows us to view the estimation results.

```
library(lme4)

# Intercept only model
fit1 = glmer(Company ~ 1 + (1|id),
data=data_company,family = binomial(link=logit))
summary(fit1)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Company ~ 1 + (1 | id)
##    Data: data_company
##
##    AIC   BIC   logLik deviance df.resid
##   6551.8   6565.5  -3273.9   6547.8      6998
##
## Scaled residuals:
##    Min   1Q  Median      3Q    Max
## -6.3624 -0.5529  0.1960  0.5474  5.9665
##
## Random effects:
##  Groups Name       Variance Std.Dev.
##  id     (Intercept) 4.545     2.132
## Number of obs: 7000, groups:  id, 100
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5757    0.2178   2.643  0.00821 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimation output displays the estimated variance of the random intercept (σ_{v_0}) equal to 4.545. To estimate the intra-class correlation coefficient (ICC) for generalized mixed-effects models, we use the package sjstats (Lüdecke, 2021). The ICC for generalized linear mixed-effects models with dichotomous outcomes is based on Wu and colleagues (2012).

```
library(sjstats)
performance::icc(fit1)
## # Intraclass Correlation Coefficient
##
##    Adjusted ICC: 0.580
##   Conditional ICC: 0.580
```

The ICC of 0.580 means that 58% of the variability in the outcome can be accounted for by the clustering structure of the data, in our case to between-person differences. Therefore, by using multilevel models, we can better model the variation in the outcome by allowing for individual differences in comparison to a model that does not take into account the multilevel structure of the data.

To investigate the relationship between the time-variant predictor *PA* and the dichotomous outcome, first we estimate a model where the slope is not allowed to vary across participants. The first argument in glmer() is a formula that defines the structure of the fixed effects Company ~ 1 + PA. In this formula, Company is the dependent variable, 1 is the fixed intercept and PA captures the effect of the time-variant predictor (i.e., *PA*). The second argument corresponds to the random effect structure of the model (1 | id), where 1 | id corresponds to random intercept, and it assumes the slope does not vary over participants. Similarly to the previous model, we set the argument family = binomial(link=logit).

```
fit2 = glmer(Company ~ 1 + PA + (1|id),
data=data_company,family = binomial(link=logit))
summary(fit2)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Company ~ 1 + PA + (1 | id)
##    Data: data_company
##
##    AIC   BIC   logLik deviance df.resid
##   6469.7   6490.2  -3231.8   6463.7      6997
##
## Scaled residuals:
##    Min   1Q  Median      3Q    Max
## -7.2639 -0.5425  0.1665  0.4985  7.0036
##
## Random effects:
##  Groups Name       Variance Std.Dev.
##  id     (Intercept) 4.661     2.159
## Number of obs: 7000, groups:  id, 100
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.83675    0.26873  -3.114  0.00185 **
## PA           0.40303    0.04406   9.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##    (Intr)
## PA -0.572
```

Subsequently, we estimate a model that includes a random slope for the time-varying predictor. Thus, in the random effects structure of the glmer() formula, we set (1 + PA|id), where 1 + PA|id corresponds to random intercept and the random slope.

```
fit3 = glmer(Company ~ 1 + PA + (1 + PA|id),
data=data_company,family = binomial(link=logit))
summary(fit3)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Company ~ 1 + PA + (1 + PA | id)
##    Data: data_company
##
##      AIC      BIC   logLik deviance df.resid
##   6441.9   6476.2  -3216.0   6431.9     6995
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -8.7765 -0.5370  0.1484  0.4954  6.0661
##
## Random effects:
##  Groups Name        Variance Std.Dev. Corr
##  id     (Intercept) 1.201    1.0958
##         PA          0.189    0.4347   0.40
## Number of obs: 7000, groups:  id, 100
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.07377    0.19859  -5.407 6.41e-08 ***
## PA           0.48475    0.06551   7.400 1.36e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##    (Intr)
## PA -0.452
```

To study if the inclusion of the random slope for *PA* improves the fit of the model, we use a likelihood ratio test. The likelihood ratio test is a statistical test that compares the fit of two models. A relatively more complex model is compared to a simpler one and is only valid if the simpler model is nested in the more complex one. To compute the likelihood ratio test we use the function anova.

```
anova(fit2, fit3, test="Chisq")
## Data: data_company
## Models:
## fit2: Company ~ 1 + PA + (1 | id)
## fit3: Company ~ 1 + PA + (1 + PA | id)
##      npar    AIC  BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fit2    3 6469.7 6490.2 -3231.8   6463.7
## fit3    5 6441.9 6476.2 -3216.0   6431.9 31.732  2  1.287e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

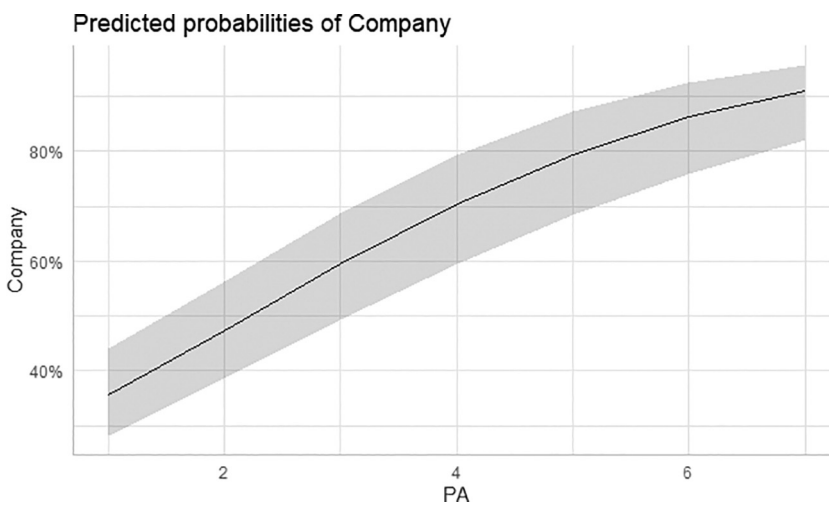
The results of the likelihood ratio test display: the AIC and BIC for models fit2 and fit3, respectively; log-likelihood when the parameters are restricted (logLik=- 3231.8); the log-likelihood of the unrestricted model (logLik=-3216.0); the value of the Likelihood Ratio Test statistic (Chisq=31.732); the degrees of freedom for the test (i.e., the difference in the number of parameters); and the p-value of the test (Pr(>Chisq)=1.287e-07). If Pr(>Chisq) is larger than the prespecified significant level (e.g., 0.05), we reject the null hypothesis. Thus, we can conclude that there is a significant improvement in the model fit when we include a random slope for *PA*.

The estimated fixed and random effects terms in the logistic mixed-effects model predict the change in log-odds. In this model, the fixed slope represents the change in the logit of the probability of being in company with others is associated with a unit change in the predictor *PA*. The glmer output also provides the z-statistic computed as the estimated coefficient value divided by its standard error and the p-value of a two-sided Wald test. From the results of the third model, we observe that *PA* is positively associated with the probability of being in company. Moreover, we can compute the odds ratio by exponentiating the estimated value of the fixed slope. To compute the odds ratio, we use the summ function from the jtools package (Long, 2020).

```
library(jtools)
summ(fit3.logit, exp = T)

MODEL INFO:
Observations: 7000
Dependent Variable: Company
Type: Mixed effects generalized linear regression
Error Distribution: binomial
Link function: logit

MODEL FIT:
AIC = 6441.92, BIC = 6476.19
Pseudo-R² (fixed effects) = 0.01
Pseudo-R² (total) = 0.61

FIXED EFFECTS:
----------------------------------------------------
                    exp(Est.)   S.E.   z val.      p
----------------- ----------- ------ -------- ------
(Intercept)              0.34   0.20    -5.41   0.00
PA                       1.62   0.07     7.40   0.00
----------------------------------------------------

RANDOM EFFECTS:
---------------------------------
 Group    Parameter    Std. Dev.
------- ------------- -----------
  id     (Intercept)     1.10
  id         PA          0.43
---------------------------------

Grouping variables:
-------------------------
 Group   # groups   ICC
------- ---------- ------
  id       100      0.27
-------------------------
```

We can interpret the output as follows: a one-unit increase in positive affect increases the odds of being in company with others by 62%. We note that the output also provides a value for the fixed intercept; however, this value lacks interpretation. Finally, we can also visualize the effect of the predictor *PA* on the probability of being in company with others.

```
library(magrittr)
library(ggeffects)
library(sjmisc)
ggpredict(fit3, "PA")
## Data were 'prettified'. Consider using `terms="PA [all]"` to get
smooth plots.
## # Predicted probabilities of Company
## # x = PA
##
## x | Predicted |       95% CI
## ----------------------------
## 1 |      0.36 | [0.28, 0.44]
## 2 |      0.47 | [0.39, 0.56]
## 3 |      0.59 | [0.49, 0.69]
## 4 |      0.70 | [0.59, 0.79]
## 5 |      0.79 | [0.68, 0.87]
## 6 |      0.86 | [0.76, 0.93]
## 7 |      0.91 | [0.82, 0.96]
##
## Adjusted for:
## * id = 0 (population-level)
# plot using the pipe
ggpredict(fit3, "PA") %>% plot()
```



The predicted probabilities show that when participants reported higher scores of positive affect, the probability of being in company with others increases.

10.1.2 Count outcome

ESM allows collecting data measuring the number of times an event occurs during a given period of time. For example, we might be interested in assessing the number of stressful events since the last beep or alcohol consumption in the last two hours. To analyse count data, we can use the Poisson mixed-effects regression model (see Gibbons et al., 2008; Hedeker & Gibbons, 2006). This approach allows modelling the conditional mean of the outcome variable y_{it} as a function of a predictor x_{it} assuming the following form:

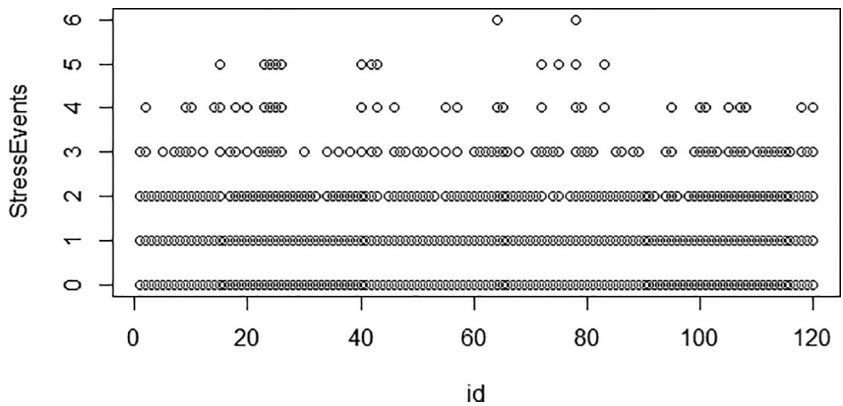
$$E\left[y_{it} \mid x_{it}, v_{0i}, v_{1i}\right] = \exp\left(\beta_0 + \beta_1 x_{it} + v_{0i} + v_{1i} x_{it}\right)$$

where $E(\cdot)$ denotes the mean, $\exp(\cdot)$ is the exponential function and v_{0i} and v_{1i} are the random effects of the random intercept and slope. The random effects are assumed to be multivariate normally distributed with mean zero and (2x2) covariance matrix Σ_v.

To estimate the Poisson mixed-effects model we use the glmer function from the lme4 package. To show how to estimate the model, we make use of the dataset *data_stress*, which includes data from 120 individuals that participated in an ESM study with ten beeps per day over seven days. At each beep, participants were asked to indicate the number of stressful events since the previous beep and the valence of their affect with the item 'How are you feeling right now' rated on a 7-point Likert scale ranging from 1 (not good) to 7 (very good).

The data include the outcome *StressEvents*, which represents the number of stressful events since the previous beep, and the predictor affective valence (*Valence*). The plot shows that across all beeps, the number of stressful events ranges from 0 to 6.

```
data_stress = read.table(file="data_stress.txt",header = TRUE, sep =
"")
head(data_stress)
##   id day beep obs StressEvents Valence
## 1  1   1    1   1            3       4
## 2  1   1    2   2            3       5
## 3  1   1    3   3            1       6
## 4  1   1    4   4            1       4
## 5  1   1    5   5            2       5
## 6  1   1    6   6            3       6
summary(data_stress$StressEvents)
##    Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
##  0.0000  0.0000  0.0000  0.5962  1.0000  6.0000
library(tidyverse)
data_stress %>%
  group_by(id) %>%
  summarise(StressEvents) %>%
  plot()
```



Before fitting the Poisson mixed model, we person-mean centred the time-varying predictor *Valence*.

```
# Center predictor Valence
data_stress %>%
  group_by(id) %>%
  mutate(Valence = Valence - mean(Valence))
```

The Poisson mixed-effects model is fitted using the glmer function. In contrast to the logistic model, we set in the argument family = poisson(link = “log”).

The output exhibits the estimated variance and correlations of the random effects and the estimated fixed effects of the Poisson regression along with the standard errors, z-statistic and p-values of a two-sided Wald test. In this model, we interpret the fixed slope as follows: if a participant increases the valence of their affect by one unit of change, the difference in the logs of expected counts would decrease by -0.20 units.

```
library(lme4)

fit.Poisson = glmer(StressEvents ~ Valence + (Valence|id),
data=data_stress,family = poisson(link = "log"))
summary(fit.Poisson)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: StressEvents ~ Valence + (Valence | id)
##    Data: data_stress
##
##    AIC   BIC   logLik deviance df.resid
##  16337.2  16372.4  -8163.6  16327.2      8395
##
## Scaled residuals:
##    Min   1Q  Median      3Q    Max
## -1.2871 -0.6774 -0.4907  0.5715  5.1566
##
## Random effects:
##  Groups Name       Variance Std.Dev. Corr
##  id     (Intercept) 0.003227 0.0568
##         Valence    0.011085 0.1053   1.00
## Number of obs: 8400, groups:  id, 120
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.30496    0.10689   2.853  0.00433 **
## Valence     -0.19911    0.02375  -8.383  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr)
## Valence -0.883
```

10.1.3 Non-normal positive continuous outcome

Throughout the previous sections, we illustrated how to implement generalized mixed-effects models to model dichotomous and count data. In certain situations, it is not straightforward to determine a priori the distribution of the outcome variable. For example, in many studies, scores are computed from a set of items, resulting in a positive skewed continuous variable. In this context, generalized linear models using the inverse Gaussian or Gamma distribution can be used to model positive continuous data, where the conditional variance of the outcome increases with its mean (see Dobson & Barnett, 2018; Lo & Andrews, 2015).

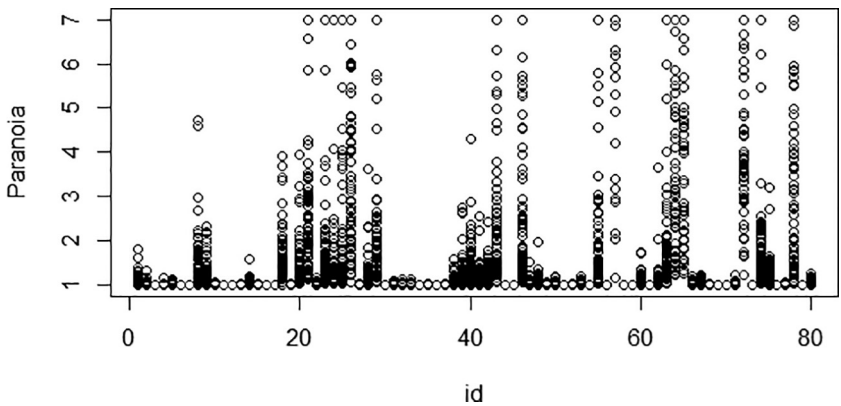
We briefly illustrate how to fit a generalized linear mixed model for modelling a positive skewed outcome using the Gamma distribution with a log link function:

$$E\left[y_{it} \mid x_{it}, v_{0i}, v_{1i}\right] = \exp\left(\beta_0 + \beta_1 x_{it} + v_{0i} + v_{1i} x_{it}\right)$$

where $E[\cdot]$ denotes the mean, $\exp(\cdot)$ is the exponential function and v_{0i} and v_{1i} are the random effects of the random intercept and slope. The random effects are assumed to be multivariate normally distributed. This model assumes multiplicative effects on the original outcome by the predictors.

We are interested in studying if negative affect predicts paranoia. To investigate this relation, we use the data set *Data_paranoia*, which includes data from 80 participants of an ESM study with 10 beeps per day over seven days. Paranoia was computed as the mean score of the items 'I feel that others dislike me', 'I feel that others might hurt me' and 'I feel suspicious'. Negative affect was defined as the mean score of the items 'I feel uncertain right now', 'I feel lonely right now', 'I feel guilty right now', 'I feel anxious right now' and 'I feel sad right now'. The items were rated on 7-point Likert scales, ranging from 1 (not at all) to 7 (very much).

```
data_paranoia = read.table(file="data_paranoia.txt",header = TRUE, sep
= "")
head(data_paranoia)
##   id day beep obs Paranoia   NegAff
## 1  1   1    1   1 1.066802 4.453393
## 2  1   1    2   2 1.121378 2.982072
## 3  1   1    3   3 1.008293 2.834387
## 4  1   1    4   4 1.028296 5.985223
## 5  1   1    5   5 1.067525 3.424695
## 6  1   1    6   6 1.088238 3.925255
summary(data_paranoia$Paranoia)
##    Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
##   1.000   1.001   1.016   1.461   1.181   7.000
library(tidyverse)
data_paranoia %>%
  group_by(id) %>%
  summarise(Paranoia) %>%
  plot()
```



From the plot, we observe that most values are clustered around the lower tail of the distribution, indicating that the variable is positively skewed. To estimate the model using the Gamma distribution, first, we person-mean centred the time-varying predictor:

```
# Center predictor negative affect
data_paranoia %>%
  group_by(id) %>%
  mutate(NegAff = NegAff - mean(NegAff))
```

Next, we use the glmer function to estimate a model setting the argument Gamma(link = "log"). The estimation output shows that the estimated fixed slope is 0.07, and it can be interpreted as follows: if a participant's negative affect increases by one unit of change, the logarithmic mean outcome increases by exp(0.08) = 1.08.

```
# Estimate Mixed Model Effects
library(lme4)

fit.paranoia = glmer(Paranoia ~ NegAff + (NegAff|id),
data=data_paranoia,family=Gamma(link = "log"))
summary(fit.paranoia)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: Gamma  ( log )
## Formula: Paranoia ~ NegAff + (NegAff | id)
##    Data: data_paranoia
##
##      AIC      BIC   logLik deviance df.resid
##   1771.4   1811.1   -879.7   1759.4     5594
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8623 -0.1292 -0.0392 -0.0054 11.9715
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.064407 0.25379
##           NegAff      0.006271 0.07919  -0.83
##  Residual             0.069158 0.26298
## Number of obs: 5600, groups:  id, 80
##
## Fixed effects:
##             Estimate Std. Error t value Pr(>|z|)
## (Intercept) -0.05456    0.04810  -1.134    0.257
## NegAff       0.07861    0.01921   4.093 4.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##        (Intr)
## NegAff -0.633
```

10.2 Three-level models

In this section, we describe how to use mixed-effects models to account for the multiday structure of ESM where repeated measurements (level 1) are clustered within days (level 2), which in turn are clustered within participants (level 3). We note that the modelling techniques introduced in this section can also be more generally applied to three-level models – for example, for beeps, nested in phases (e.g., pre- and post-treatment and follow-up) nested in participants or when collecting data from different countries or schools (see Broda, 2017; Walls et al., 2006).

Mixed-effects modelling is a flexible approach because it allows researchers to capture and quantify the variance at different clustering levels. In Chapter 9, we presented the two-level approach, which accounts for the fact that we have repeated measurements within persons. This model partitions the variance into variance at the person level and variance at the measurement level. We now propose an extended, three-level modelling approach (see de Haan-Rietdijk et al., 2016). As before, we consider the fact that the beeps are nested within persons, but in this case, we also account for the multi-day structure of the data by allowing for variation at the day level.

To show how to estimate the three-level model we make use of the data set *data_valence*, which includes data from 60 individuals that participated in an ESM study with 10 beeps per day over seven days. At each beep, participants were asked to indicate the valence of their affect with the item 'How are you feeling right now' rated on a 7-point Likert scale ranging from 1 (not good) to 7 (very good).

```
data_valence = read.table(file="data_valence.txt",header = TRUE, sep =
"")
colnames(data_valence) = c("id","day","beep","obs","Valence")
head(data_valence)
##   id day beep obs Valence
## 1  1   1    1   1       2
## 2  1   1    2   2       3
## 3  1   1    3   3       2
## 4  1   1    4   4       2
## 5  1   1    5   5       2
## 6  1   1    6   6       2
```

We consider the intercept-only model that partitions the variability in the outcome variable into variance at the measurement level, variance at the day level and variance at the participant level. For participant *i*, on the *t*-th beep of day *j*, the three-level model can be described as follows

$$\textit{Level 1: } \textit{Valence}_{ijt} = \gamma_{0ij} + \varepsilon_{ijt}$$

$$\textit{Level 2: } \gamma_{0ij} = \beta_{00i} + \nu_{0ij}$$

$$\textit{Level 3: } \beta_{00i} = \beta_{000} + \omega_{0i}$$

where γ_{0ij} represents the mean of individual *i* on day *j* and the error ε_{ijt} represents the deviation from the participant's affective valence on day *j* at beep *t*. The participant's mean level is denoted by β_{00i}, and ν_{0ij} is the deviation of the day mean level for this participant from the trait level. β_{000} denotes the grand mean of valence for the population, and ω_{0i} represents the deviation of each participant's affective valence from the population mean. In this model, the level 1 errors are normally distributed with mean zero and variance σ^2_ε. The random effects ν_{0ij} and ω_{0i} are normally distributed with mean zero and variance $\sigma^2_{\nu 0}$ and $\sigma^2_{\omega 0}$, respectively.

In a three-level model we can define two types of ICC (Hedges et al., 2012). The Level 2 ICC describes the proportion of the total variance of the outcome that is accounted for by the three-level structure of the data in which days are clustered within participants:

$$\rho_2 = \frac{\sigma^2_{\nu 0}}{\sigma^2_{\nu 0} + \sigma^2_{\omega 0} + \sigma^2_\varepsilon}$$

The Level 3 ICC describes the proportion of the total variance of the outcome that is accounted for by the participant clustering structure:

$$\rho_3 = \frac{\sigma^2_{\omega 0}}{\sigma^2_{\nu 0} + \sigma^2_{\omega 0} + \sigma^2_\varepsilon}$$

We first illustrate how to estimate a three-level model to account for the variability over days using the lme function from the nlme package (Pinheiro et al. 2017). The first argument in lme() is a formula that defines the structure of the fixed effects Valence ~ 1 where Valence is the dependent variable and 1 is the fixed intercept. The second argument corresponds to the random effect structure of the model random = ~ 1 | id/day where

1|id/day corresponds to random intercept, which are allowed to vary over days (day) nested within participants (id).

```
library(nlme)
library(lme4)
library(lmerTest)

fit.day.lme = lme(Valence ~ 1, random = ~ 1|id/day,
data=data_valence)
summary(fit.day.lme)
## Linear mixed-effects model fit by REML
##  Data: data_valence
##        AIC     BIC   logLik
##   9827.119 9852.49 -4909.56
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept)
## StdDev:   0.4217239
##
##  Formula: ~1 | day %in% id
##         (Intercept)  Residual
## StdDev:   0.1749989 0.7469523
##
## Fixed effects: Valence ~ 1
##                Value  Std.Error   DF  t-value p-value
## (Intercept) 2.009524 0.05630224 3780 35.69172       0
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.17960288 -0.62617970 -0.01641105  0.71622363  3.65030454
##
## Number of Observations: 4200
## Number of Groups:
##          id day %in% id
##          60         420
```

The estimation results show that the estimated level 1 standard deviation (σ_{ε}) is 0.75. The estimated standard deviation of the day level ($\sigma_{\upsilon 0}$) is 0.17, and the estimated standard deviation that accounts for the variability at the person level ($\sigma_{\omega 0}$) is 0.42. The Level 2 ICC of 0.04 means that 4% of the outcome variance can be accounted for by the three-level structure of the data. The Level 3 ICC of 0.232 means that 23.2% of the variability in the outcome can be accounted for by between-person differences. Therefore,

we observe that the Level 3 ICC is positive, and therefore, a three-level structure should be taken into consideration.

In a similar manner, we can also estimate the model using the lmer function from the lme4 package. To capture the variability at the day and person level, we set the random effects as (1|id/day). Comparing the two estimation procedures, we observe non-significant differences between them. We also note that the model presented in this chapter can be easily extended to include predictors and account for random slopes that vary at the day level (e.g., de Haan-Rietdijk et al., 2016).

```
fit.day.lmer = lmer(Valence ~ 1 + (1|id/day),
data=data_valence)
summary(fit.day.lmer)
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: Valence ~ 1 + (1 | id/day)
##    Data: data_valence
##
## REML criterion at convergence: 9819.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1796 -0.6262 -0.0164  0.7162  3.6503
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  day:id   (Intercept) 0.03063  0.1750
##  id       (Intercept) 0.17785  0.4217
##  Residual             0.55794  0.7470
## Number of obs: 4200, groups:  day:id, 420; id, 60
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)   2.0095     0.0563 58.9987   35.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

10.3 Multilevel vector autoregressive models

ESM data allow studying the dynamics of psychological functioning within individuals over time. The dynamic within-person framework (see Molenaar, 2004) has been used in many research areas. For example, in

psychopathology research, it has been applied to investigate the dynamic interaction between individual symptoms (Borsboom & Cramer, 2013) and to investigate emotional inertia and affect (Kuppens, Allen, et al., 2010).

The statistical approach that is widely used to model within-person dynamics is the vector autoregressive model of order one (VAR(1)) (see e.g., Bringmann et al., 2016; Pe et al., 2015). In this model, each variable is regressed on all variables (including itself) at the previous time point. Therefore, for each particular variable, the model allows estimating the effect of its past values on current values (i.e., autoregressive effects) as well as the effect of past values of the rest of the variables (i.e., cross-regressive effects). This model can be estimated at the individual level (i.e., person-specific VAR(1)) or over individuals using a multilevel framework that allows the VAR coefficients to differ across persons (see e.g., Bringmann et al., 2016; Bringmann et al., 2013; Bulteel et al., 2018).

In the VAR(1) model, the associated model parameters are typically interpreted as measures of specific psychological process features. For example, we might be interested in investigating the joint dynamics of positive affect (PA) and negative affect (NA). Figure 10.1 shows a path diagram of a bivariate VAR(1) model. The autoregressive effects (solid arrows) are considered measures of emotional inertia (Kuppens, Allen, et al., 2010), and the cross-regressive effects (dashed arrows) reflect the effect of past affect on current PA and NA.

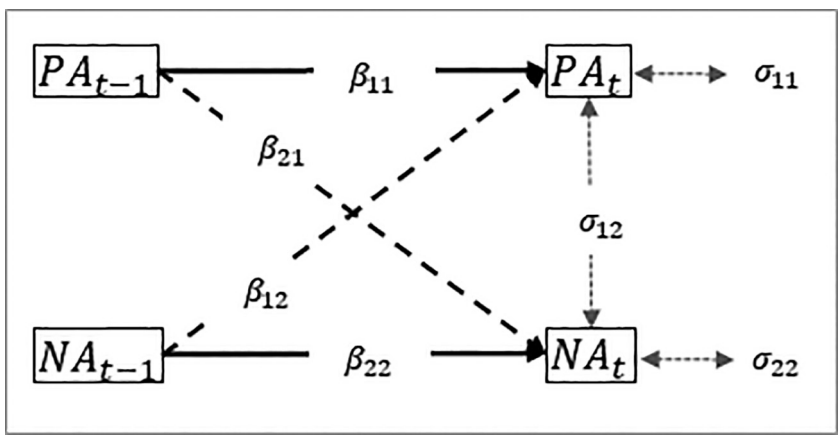


Figure 10.1 Bivariate VAR(1) model of PA and NA. The model assumes that current affect depends on previous affect through autoregressive and cross-regressive (solid and dashed arrows) effects. Their (co-)variances are represented by double-headed dotted arrows

Multilevel extensions of the VAR(1) model have been proposed to analyse intensive longitudinal data (Bringmann et al., 2013). These multilevel VAR(1) models allow capturing individual differences by including random autoregressive and cross-regressive effects that vary across persons. In such models, a linear mixed-effect model is estimated for each variable using all the lagged variables as predictors. The autoregressive and cross-regressive effects can be considered as random variables. As a result, the model estimates the within-individual autoregressive and cross-regressive effects while incorporating and using information about the distribution of these effects across individuals (e.g., Bringmann et al., 2013; Bulteel et al., 2018).

We illustrate how we can estimate a multilevel VAR(1) model using the data set *data_VAR*. The data include 200 individuals that participated in an ESM study with 10 beeps per day over six days. The data include the variables positive affect (*PA*) and negative affect (*NA.*). The variables are person-mean centred (centred on each participant's mean score). The data also include the lagged variables for PA and NA within person and within days and are denoted by *PA_lag* and *NA._lag*.

```
data_VAR = read.table(file="data_VAR.txt",header = TRUE, sep = "")
head(data_VAR)
##   Beep Day subjno Beepno      PA      NA.   PA_lag  NA._lag
## 1  1   1      1     1 3.608206 2.545103        NA       NA
## 2  2   1      1     2 4.617290 1.771382 3.608206 2.545103
## 3  3   1      1     3 2.696894 1.991890 4.617290 1.771382
## 4  4   1      1     4 2.492036 1.467974 2.696894 1.991890
## 5  5   1      1     5 3.319702 1.888963 2.492036 1.467974
## 6  6   1      1     6 3.450077 2.875780 3.319702 1.888963
```

We illustrate how to estimate the VAR(1) model using the lmer function from the lme4 package.

```
# Estimate Mixed Model Effects
library(lme4)
library(lmerTest)

fit.VAR.PA = lmer(PA ~ PA_lag + NA._lag + (PA_lag + NA._lag|subjno),
data=data_VAR)
summary(fit.VAR.PA)
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: PA ~ PA_lag + NA._lag + (PA_lag + NA._lag | subjno)
##    Data: data_VAR
##
## REML criterion at convergence: 29807
##
## Scaled residuals:
##    Min   1Q  Median      3Q    Max
## -3.6166 -0.6549 -0.0006  0.6482  3.8065
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  subjno   (Intercept) 0.91245  0.9552
##           PA_lag   0.02487  0.1577   -0.06
##           NA._lag  0.04527  0.2128   -0.04  0.08
##  Residual          0.82019  0.9056
## Number of obs: 10800, groups:  subjno, 200
##
## Fixed effects:
##              Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)   3.09299    0.08723 191.82863  35.460   <2e-16 ***
## PA_lag        0.38295    0.01419 212.50925  26.988   <2e-16 ***
## NA._lag      -0.03851    0.01874 197.91242  -2.055   0.0412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) PA_lag
## PA_lag  -0.347
## NA._lag -0.190  0.072
```

First, we fit a linear mixed-effect model to obtain the estimates of β_{11} and β_{12}. The first argument in lmer() is a formula that defines the structure of the fixed effects PA ~ 1 + PA_lag + NA._Lag. In this formula, PA is the dependent variable, 1 is the fixed intercept, PA_lag captures the effect of the past values of positive affect on its current values, and NA._lag captures the effect of the past values of negative affect on current values of positive affect. The second argument corresponds to the random effect structure of the model (1 + PA_lag + NA._Lag| subjno), where 1 + PA_lag + NA._Lag|subjno corresponds to random intercept, random autoregressive effect and the random cross-regressive effect which are allowed to vary over participants (subjno).

The estimation output includes the estimated value of the fixed autoregressive effect β_{11}, equal to 0.38, and the fixed cross-regressive effect β_{12}, equal to -0.04. Meanwhile, the estimated standard deviation of the within-person associated positive affect errors is 0.91.

In a similar fashion, we estimate a linear mixed-effect model where the outcome variable is negative affect. The results show that the estimated autoregressive effect β_{22} and cross-regressive β_{21} effects are 0.27 and -0.11, respectively. The estimated standard deviation of the within-person errors associated with negative affect is 0.70.

```
fit.VAR.NA. = lmer(NA. ~ PA_lag + NA._lag + (PA_lag + NA._lag|subjno),
data=data_VAR)
summary(fit.VAR.NA.)
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: NA. ~ PA_lag + NA._lag + (PA_lag + NA._lag | subjno)
##    Data: data_VAR
##
## REML criterion at convergence: 24450
##
## Scaled residuals:
##    Min   1Q  Median      3Q    Max
## -3.9638 -0.6657 -0.0062  0.6748  3.5334
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  subjno   (Intercept) 0.82459  0.9081
##           PA_lag   0.04135  0.2033   -0.04
##           NA._lag  0.04433  0.2105   -0.02  0.07
##  Residual          0.48612  0.6972
## Number of obs: 10800, groups:  subjno, 200
##
## Fixed effects:
##              Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)   1.90862    0.07784 179.74830  24.519  < 2e-16 ***
## PA_lag       -0.10796    0.01593 197.89925  -6.776 1.38e-10 ***
## NA._lag       0.27432    0.01728 214.86707  15.877  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##         (Intr) PA_lag
## PA_lag  -0.222
## NA._lag -0.136  0.069
```

Finally, we can compute the covariance between the within-person errors of positive and negative affect, denoted by σ_{12}.

```
# Compute the covariance between the within-person residuals
cov(residuals(fit.VAR.PA),residuals(fit.VAR.NA.))
## [1] 0.002785165
```

Figure 10.2 presents the dynamic network that results from estimating the multilevel VAR(1) model. We observe that both positive and negative affect are positively related to its past values and past values of negative affect

are negatively related to current values of positive affect, and a similar relationship occurs between past values of positive affect and current values of negative affect.

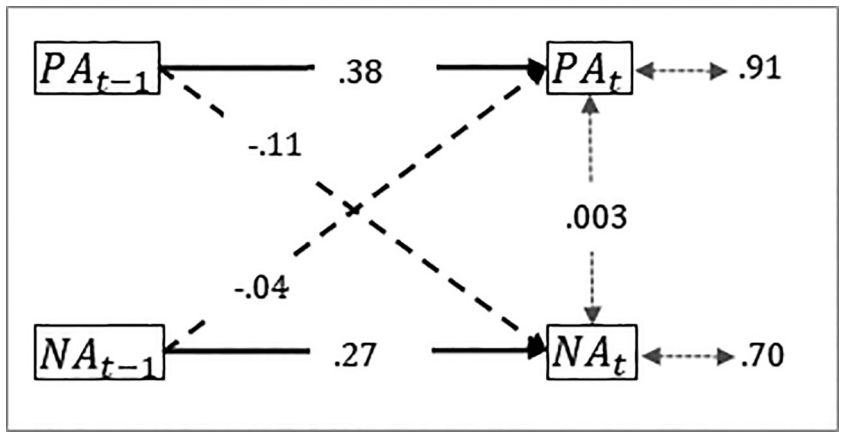


Figure 10.2 Estimated VAR(1) model of PA and NA

The procedure described above can be implemented when the multilevel VAR(1) models include more than two variables. We note that this is not the only approach to estimate VAR(1) models. We refer the reader to the following articles for more information: Hamaker and colleagues (2015) and Jongerling and colleagues (2015). Additional R packages to estimate network models and visualization are graphicalVAR (Epskamp, 2020) and mlVAR (Epskamp et al., 2019).

Overall, it may be said that the multilevel VAR(1) model has been extensively used to obtain the dynamic network structure of a set of variables. The network analysis provides information about the strength of the connections between the variables and the centrality of different variables (i.e., the relative importance a variable occupies in the network) (see Bringmann et al., 2013). However, the application of the VAR methodology in network analysis comes with caveats. First, the use of centrality indices has been highly criticized because the indices produce unstable estimates. Moreover, the interpretation of these indices is unclear in the context of psychological variables (see Bringmann et al., 2019).

Finally, it is worth mentioning that the application of VAR(1) models to ESM data is challenging because of the large number of parameters that need to be estimated. Most ESM studies usually include between 60 to 140 measurement occasions (see Chapter 3). Therefore, the question arises as to whether VAR(1) models are too complex to characterize the dynamics of psychological processes reliably. One way to investigate this question is to evaluate the predictive accuracy of a model (i.e., how well it generalizes to unseen data). Bulteel and colleagues (2018) used prototypical ESM data sets to show that person-specific VAR(1) models have the worst predictive accuracy in comparison to multilevel VAR(1) models. Furthermore, they showed that even the multilevel variants do not outperform the multilevel AR(1) model. Therefore, we suggest that researchers pay careful attention to the quality and quantity of the ESM data when selecting the statistical model to answer a specific research question.

10.4 Conclusions

In this chapter, we have presented the generalized mixed-effects models to analyse three types of non-normal responses. Specifically, we illustrate how to fit a logistic mixed-effects model to analyse dichotomous or binary outcomes. We also present the Poisson mixed-effects model for modelling count data and the inverse Gaussian mixed-effects model for analysing positive and continuous variables. Moreover, we have shown how to estimate a three-level model to account for the variability in ESM data with repeated measurements nested within days, which are nested within participants. Finally, we have presented the multilevel extension of the VAR(1) model. This model is widely used in psychological research to study how within-person processes evolve dynamically.

Besides these models, additional applications of the generalized mixed-effects model include the following distributions to model over-dispersed count data: negative binomial, zero-inflated Poisson and negative binomial models, and Hurdle Poisson and negative binomial models. Good general resources on over-dispersed count data include Brooks and colleagues (2017), Hall (2000), Molenberghs and Verbeke (2006), Zhang and Yi (2020) and Zuur and colleagues (2009). Finally, we note that ESM data often includes semicontinuous outcomes. In these

cases, a proportion of responses are equal to a single value. This value often represents whether an individual engaged in a behaviour. The remaining values follow a continuous and skewed distribution. In this case, two-part mixed-effect models have been proposed for modelling semicontinuous outcomes (see Blozis et al., 2020; Farewell et al., 2017; Tooze et al., 2002).

CHAPTER 11

Sample Size Selection in ESM Studies

Ginette Lafit

Sample size planning constitutes a crucial step in the design of an ESM study. The amount of collected data determines how much information is present to answer a research question and derive reliable conclusions. ESM typically allows the collection of ecologically valid intensive longitudinal data reflecting individuals' psychological processes over time. Data obtained from ESM studies have a multilevel structure in which repeated observations over consecutive days are nested within participants. The repeated measurements collected with ESM provide information to examine how individuals' psychological processes evolve in daily life while accounting for between-individual differences. The total sample size depends on the number of participants, the number of measured variables and the number of time points in which variables are measured for each participant.

When the ultimate goal of a study is to properly test a hypothesis, a criterion to select the sample size is statistical power. In ESM research, most published empirical studies do not report a statistical power analysis to justify the selection of the sample size (Trull & Ebner-Priemer, 2020). This does not come as a surprise as not much is known about statistical power in intensive longitudinal designs. Besides, software to aid researchers in conducting power analyses did not exist until recently.

This chapter provides an introduction to sample size planning for ESM studies. As the area of sample size planning in intensive longitudinal studies is constantly evolving, this chapter provides basic guidelines and methodological tools for conducting power analyses for several of the most common research questions in ESM. Therefore, we illustrate how to determine the necessary and sufficient sample sizes based on statistical power requirements for the most commonly used statistical model family in the study of individual differences in ESM studies: multilevel regression models. The set of methodological tools introduced in this chapter can be extended to more complex research questions with other complex considerations.

The chapter is organized as follows. First, we discuss the relation between sample size and power for multilevel models. Second, we present a brief overview of methodological approaches for conducting a priori power analysis for multilevel regression models. Third, we illustrate how to perform a power analysis to select the number of participants in an ESM study when the temporal design is predefined. Next, we illustrate how to conduct power analysis when the goal is to select the number of time points. Fifth, we discuss other considerations that have to be taken into account when deciding on the temporal design of the study. We conclude by discussing challenges concerning feasibility and sample size planning in ESM studies.

11.1 Power analysis in multilevel models

ESM studies are typically conducted to answer specific research questions. Commonly studied research questions are how individual characteristics (e.g., age, depression, neuroticism, psychiatric diagnoses, etc.) are related to characteristics of people's feelings, behaviour or thoughts over time in daily life or how these feelings, behaviours, and thoughts themselves are related to time-varying predictors (see also Chapter 2). The goal of all these research questions is to test if the effect of interest is present in the population under study. The selection of the sample size should allow the researcher to detect this effect reliably. Therefore, we can use statistical power as a criterion to select the sample size.

Statistical power is defined as the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true in the population under study (Cohen, 2013). The power to detect an effect is influenced by three factors. First, the size of the effect in the population under study. Second, the predetermined type I error (i.e., the probability of rejecting the null hypothesis when the findings have occurred by chance). Third, the standard error of the statistic used to test the hypothesis of interest. Holding the other two quantities fixed, the power increases with an increase in effect size, an increase in the nominal type I risk, and a decrease of the standard error. The standard error of a test statistic is inversely related to the sample size. Consequently, the relationship between the sample size and the standard error makes power a criterion to inform about sample size.

Power analysis in multilevel models can be used to determine the necessary number of participants and the necessary number of repeated measurements within participants. Performing a power analysis to test a statistical hypothesis in multilevel models is complex. The reason is that the hierarchical data structure determines two sources of variation, the within- and the between-person variability (Bolger, 2011) (see also Chapter 2). Furthermore, power calculations in intensive longitudinal designs, such as ESM studies, need to take the temporal dependencies of adjacent measurements into account (Lafit et al., 2021). In the next section, we review the main two frameworks to conduct power analysis in multilevel models: the analytical approach and the simulation-based approach.

11.2 Methodological approaches for power analyses in multilevel models

The literature on power analysis for multilevel models can be grouped into two approaches: an analytical and a simulation-based approach. The goal of the analytical approach is to obtain formulas where the sample size is a function of the effect of interest, the standard deviation and the test statistic. Snijders (2005), for example, derived formulas to compute power for basic research questions using multilevel models. As we mentioned before, the standard errors are a function of the sample size. Therefore, these formulas allow deriving the sample size to reach a predetermined value of power. The reader interested in learning more about analytical formulations to derive power in multilevel models is referred to the following studies: Hedeker and colleagues (1999); Moerbeek and colleagues (2000); Moerbeek and colleagues (2001); Moerbeek and Maas (2005); Raudenbush (1997); Raudenbush and Liu (2001); Snijders and Bosker (1993); Wang and colleagues (2015).

The main drawback of the analytical approach is that currently existing formulas are limited to simple research questions (see Snijders, 2005). To conduct power analysis for complex research questions, an alternative approach has been proposed – namely, simulation-based power analysis. The simulation-based framework uses the hypothesized population model and the effect of interest to generate a large number of synthetic data sets. Each of these data sets is then used to fit the model under study and to test

the statistical hypothesis of interest for significance. The power of a test is then calculated as the proportion of simulated data sets where the null hypothesis was rejected. Performing these calculations while varying the number of participants or the number of repeated measurements further allows determining the sample size necessary to reach a preferred power (e.g., 80%).

The simulation-based approach is especially useful when analytical formulations are not available or too difficult to derive. By now, a large set of methodological procedures have been developed for simulation-based power analysis in multilevel models (see Arend & Schäfer, 2019; Astivia et al., 2019; Bolger, 2011; Browne et al., 2009; Cools et al., 2008; P. Green & Macleod, 2016; Landau & Stahl, 2013; Lane & Hennes, 2018; Maas & Hox, 2005; Mathieu et al., 2012; Zhang, 2014; Zhang & Wang, 2009). Even though these methodological tools are useful, they do not consider a characteristic feature of ESM data: repeated occasions within individuals are likely to be correlated. To overcome this limitation, Lafit and colleagues (2021) proposed a user-friendly application, *PowerAnalysisIL*, for conducting simulation-based power analysis in multilevel models that account for temporal dependencies.

In this chapter, we are working with the user-friendly application *PowerAnalysisIL* which was developed in R (R Core Team, 2020) via the Shiny package (Chang et al., 2019). This app covers a set of research questions that are popular in the ESM literature and that can be assessed using multilevel regression models (see Chapter 2); it properly accounts for the temporal dependency that characterizes intensive longitudinal designs. Table 11.1 provides an overview of the 11 models included in the app. The app is available via a git repository hosted on GitHub at https://github.com/ginettelafit/PowerAnalysisIL. Users can download the app and run it locally on their computer in R or Rstudio (RStudio Team, 2015). A step-by-step tutorial on how to use the app is presented in Lafit and colleagues (2021).

Table 11.1 Overview of the population models of interest available in the PowerAnalysisIL application

Model	Description
Model 1	*Group differences in mean level*: estimates differences between two groups of individuals in the mean of the outcome variable.
Model 2	*Effect of a continuous time-invariant predictor on the mean level*: estimates whether an individual-specific time-invariant variable predicts individual differences in the mean level of the outcome variable.
Model 3	*Effect of a Level 1 continuous predictor (random slope):* estimates whether a time-varying variable predicts the outcome variable. This model includes a random slope to account for individual differences in the effect of the time-varying predictor on the outcome variable.
Model 4	*Effect of a Level 1 continuous predictor (fixed slope):* analogous to Model 3, but it assumes the effect of the time-varying predictor on the outcome variable is fixed across individuals.
Model 5	*Group differences in the effect of a time-varying continuous predictor (random slope)*: estimates differences between two groups of individuals with respect to the association between a time-varying predictor and the outcome of interest. This model includes a random slope to account for individual differences in the effect of the time-varying predictor on the outcome variable.
Model 6	*Group differences in the effect of a time-varying continuous predictor (fixed slope)*: analogous to Model 5, but it assumes the effect of the time-varying predictor on the outcome variable is fixed across individuals.
Model 7	*Cross-level interaction between two continuous predictors (random slope)*: this model estimates whether the effect of a time-varying predictor on the outcome variable is moderated by a continuous time-invariant predictor. This model includes a random slope to account for individual differences in the effect of the time-varying predictor on the outcome variable.
Model 8	*Cross-level interaction between two continuous predictors (fixed slope)*: analogous to Model 7, but it assumes the effect of the time-varying predictor on the outcome variable is fixed across individuals.
Model 9	*Multilevel autoregressive model*: estimates the effect of the lagged outcome variable (i.e., the observed outcome at the previous measurement occasion) on current values of the outcome variable.
Model 10	*Group differences in the mean autoregressive effect:* estimates the difference in the effect of the lagged outcome variable on current values of the outcome variable between two groups of individuals.
Model 11	*Cross-level interaction effect between a continuous time-invariant predictor and the lagged outcome variable*: estimates whether the time-invariant predictors moderate the autoregressive effect.

11.3 Illustrations

In this section, we illustrate how to perform a power analysis to decide on the number of participants needed to test group differences in the mean level of the outcome of interest. Subsequently, we show how to perform a power analysis to investigate the effect of the total number of repeated measurements on power.

11.3.1 Illustration I: power analysis to select the number of participants

We illustrate how to conduct a power analysis to decide on the number of participants when the goal is to estimate group differences in the mean level of the outcome of interest. We assume that a researcher is interested in conducting an ESM study to estimate meaningful differences in the mean level of negative affect between individuals diagnosed with major depressive disorder (MDD) and healthy control participants. We assume that the temporal design of the ESM studied is predetermined (i.e., there is a fixed number of at least approximately equidistant observations within individuals) and that the data includes 70 measurement occasions per individual.

To estimate the power, first we specify the multilevel regression model to test the hypothesis of interest. Second, we determine the value of the effect of interest (i.e., the difference in the mean level of negative affect between the two groups) to generate the datasets. Afterward, we conduct the power analysis via the *PowerAnalysis*IL shiny app to learn about the necessary number of participants.

Population Model. To estimate group differences in negative affect, we specify a two-level regression model. Let's denote the outcome variable $NegAff_{it}$ for the i-th individual at the t-th observation and $Diagnosis_i$, a dummy variable that is 1 for individuals diagnosed with MDD and 0 for control participants. Moreover, we denote N_0 to the number of healthy control participants and N_1 to the number of participants with MDD. The number of time points is denoted by T. Figure 11.1 shows a graphical representation of the multilevel model to estimate group differences in the mean level of negative affect between individuals with MDD and healthy control participants.

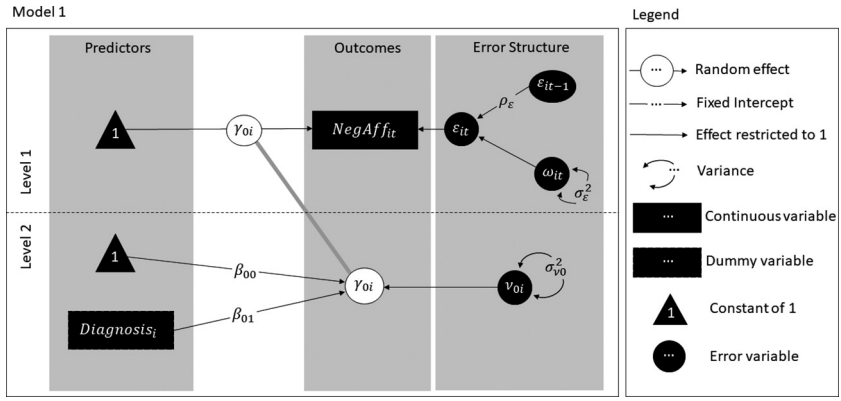


Figure 11.1 Graphical representation of the multilevel model to estimate group differences in the mean level of negative affect between individuals with MDD and healthy control participants

The multilevel regression model is written as follows:

Level 1: $NegAff_{it} = \gamma_{0i} + \varepsilon_{it}$

Level 2: $\gamma_{0i} = \beta_{00} + \beta_{01} Diagnosis_i + \upsilon_{0i}$

Inter-individual differences in negative affect are modelled by the random intercept γ_{0i}. The random intercept expresses the deviation of each participant's negative affect level from the subgroup-specific mean level. The random intercept is assumed to be normally distributed with standard deviation denoted by $\sigma_{\upsilon 0}$. The model can be interpreted as follows: for participants in the reference group (controls), the mean level of negative affect equals β_{00} for individuals diagnosed with MDD, and the mean level of negative affect is given by $\beta_{00} + \beta_{01}$.

To account for the temporal dependencies in ESM data, we allow for serially correlated errors. We assume that the level-1 errors ε_{it} follow a first-order autoregressive (AR(1)) process (Goldstein et al., 1994), where the correlation between two consecutive errors is denoted by ρ_ε, and σ_ε is the standard deviation of the Level 1 errors.

Specifying the parameter values of the population model. To specify the value of the parameters in the population model, we have three alternatives.

We can use values reported in the literature, data from a pilot study or data from previously conducted ESM studies (see Lane & Hennes, 2018; Maxwell et al., 2008).

In this illustration, we use information from a previously conducted ESM study.[1] The dataset includes 40 individuals that have been diagnosed with MDD and 60 control subjects. They all participated in an ESM study with 10 beeps per day over seven days. Therefore, the design includes 70 measurement occasions per participant. Participants were asked to repeatedly fill in a questionnaire containing 1 to 7 Likert-type scale items measuring negative affect. The dataset is publicly available in the git repository https://github.com/ginettelafit/PowerAnalysisIL.

To upload the dataset in R or Rstudio, we use the command read.table. In addition to the variables *NegAff* and *Diagnosis*, the dataset includes the variable *id*, representing participants' identification number, *day*, denoting the study day, *beep*, indicating the prompt number within a day, and *obs*, the number of time points within an individual.

```
data_pilot = read.table(file="data_pilot.txt",header = TRUE,
sep = "")
head(data_pilot)
##   id day beep obs NegAff Diagnosis
## 1  1   1    1   1      1         0
## 2  1   1    2   2      1         0
## 3  1   1    3   3      2         0
## 4  1   1    4   4      3         0
## 5  1   1    5   5      1         0
## 6  1   1    6   6      3         0
```

To estimate the parameters of the population model, which later will be used to perform the power analysis, we estimate a linear mixed-effect model. The multilevel model is fitted using the lme function from the nlme package (Pinheiro et al., 2017). We set REML as the estimation method,[2]

[1] To preserve the confidentiality of personal information, the ESM dataset used in the illustrations has not been made publicly available. Instead, we provided a synthetic dataset that mimics data from this ESM study.

[2] The lme function includes two optimization methods: maximum likelihood (ML) and restricted maximum likelihood (REML). ML assumes the fixed effects are known when estimating the variance components. Therefore, the estimates of the variance components are biased when the sample size is small. REML estimates unbiased variance components accounting for the degrees of freedom of the fixed effects estimates. As a result, when the number of participants is small, it is recommended to use REML.

and we specify an AR(1) structure for the Level-1 errors with the command correlation=corAR1(). The summary provides the estimated parameters using the synthetic dataset.

```
library(nlme)
fit.Model = lme(NegAff ~ Diagnosis, random = ~
1|id,na.action=na.omit, data=data_pilot,
method="REML",correlation=corAR1())
summary(fit.Model)
## Linear mixed-effects model fit by REML
##  Data: data_pilot
##        AIC     BIC   logLik
##   20251.12 20285.39 -10120.56
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept) Residual
## StdDev:   0.5247353 1.047873
##
## Correlation Structure: AR(1)
##  Formula: ~1 | id
##  Parameter estimate(s):
##    Phi
## 0.2705838
## Fixed effects: NegAff ~ Diagnosis
##                 Value  Std.Error   DF   t-value p-value
## (Intercept) 2.0601991 0.07099117 6900 29.020498  0.0000
## Diagnosis   0.3656786 0.11224690   98  3.257806  0.0015
##  Correlation:
##           (Intr)
## Diagnosis -0.632
##
## Standardized Within-Group Residuals:
##          Min         Q1          Med         Q3
      Max
## -3.625851595 -
0.667374206  0.008533682  0.621340636  3.465398449
##
## Number of Observations: 7000
## Number of Groups: 100
```

The model provides the estimates of the population parameters that will subsequently be used to perform the power analysis. The estimated fixed intercept β_{00} is 2.06; the differences in the mean level of negative affect

between the two groups β_{00} is 0.37. The standard deviation σ_{ε} and autocorrelation ρ_{ε} of the Level-1 errors are 1.05 and 0.27, respectively. And the standard deviation of the random intercept $\sigma_{\upsilon 0}$ is 0.52.

Power analysis. We use the *PowerAnalysisIL* app to conduct the power analysis to determine the necessary number of participants in each group. First, we select Model 1 and set the number of participants for the healthy controls (Group 0) and MDD group (Group 1) to 20, 30, 40, 60, 80 and 100. We set the number of measurement occasions to 70 (see Figure 11.2).

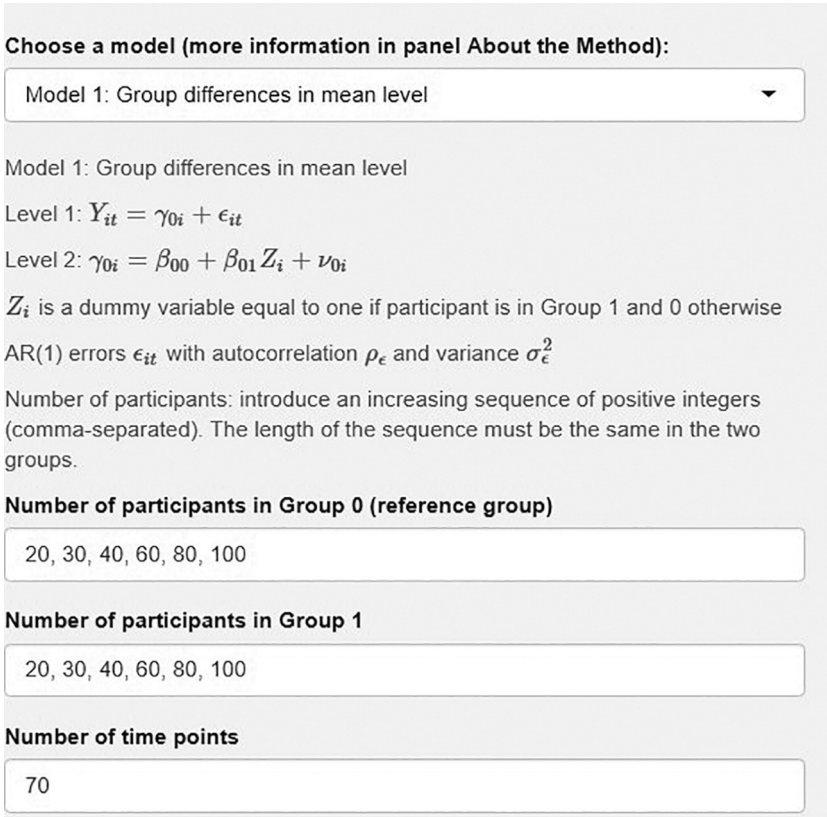


Figure 11.2 This screenshot of the PowerAnalysisIL app shows the window in which Model 1 has been selected and the sample size has been set

Subsequently, we set the values of the parameters of the population model (see Figure 11.3). We start with the fixed effects: the fixed intercept β_{00} is set

to 2.06, and the effect of the Level-2 dummy variable β_{01} is set to 0.37. We set the standard deviation σ_{ε} and autocorrelation of the Level 1 errors ρ_{ε} to 1.05 and 0.27, respectively. The standard deviation of the random intercept $\sigma_{\upsilon 0}$ is set to 0.52. We select the option *Estimated AR(1) correlated errors*. We set the Type I error to 0.05, and the number of Monte Carlo replicates to 1,000. To estimate the multilevel model, we choose the option *Maximizing the restricted log-likelihood*. Finally, we click on *Compute Power*.

The simulation-based procedure is computationally intensive. The computational time depends on the population model of interest, the number of participants, the number of measurement occasions, the number of Monte Carlo replicates, and the operating system. Therefore, to estimate how much time will be necessary for conducting the power analysis, the app provides the option 'Estimate Computational Time', which estimates the expected number of hours necessary to perform the analysis.

Figure 11.4 shows the power curve as a function of the number of participants. The power curve exhibits the estimated power to test mean differences in negative affect between individuals with MDD and healthy control participants. We observe that when the number of participants is 20 in both groups, the power for the effect of interest (i.e., β_{01}) is 55.4%. This result implies that in only 554 out of the 1,000 simulated datasets, the null hypothesis of no group differences in the mean level of negative affect was rejected. When the number of participants increases, the power increases as well. A power higher than 80% is achieved when the number of participants in both groups is greater than 40.

The app also provides information about the distribution of the estimates of the fixed effects across the Monte Carlo replicates. This summary includes: power; the average of the estimates of each fixed effect; the bias (i.e., the difference between the average of the estimates and the true value); the standard error; and the $(1 - \alpha)$ coverage proportion, computed as the proportion of Monte Carlo replicates for which the $(1 - \alpha)$ confidence interval includes the true value. Table 2 shows the summary statistics for the fixed effects.

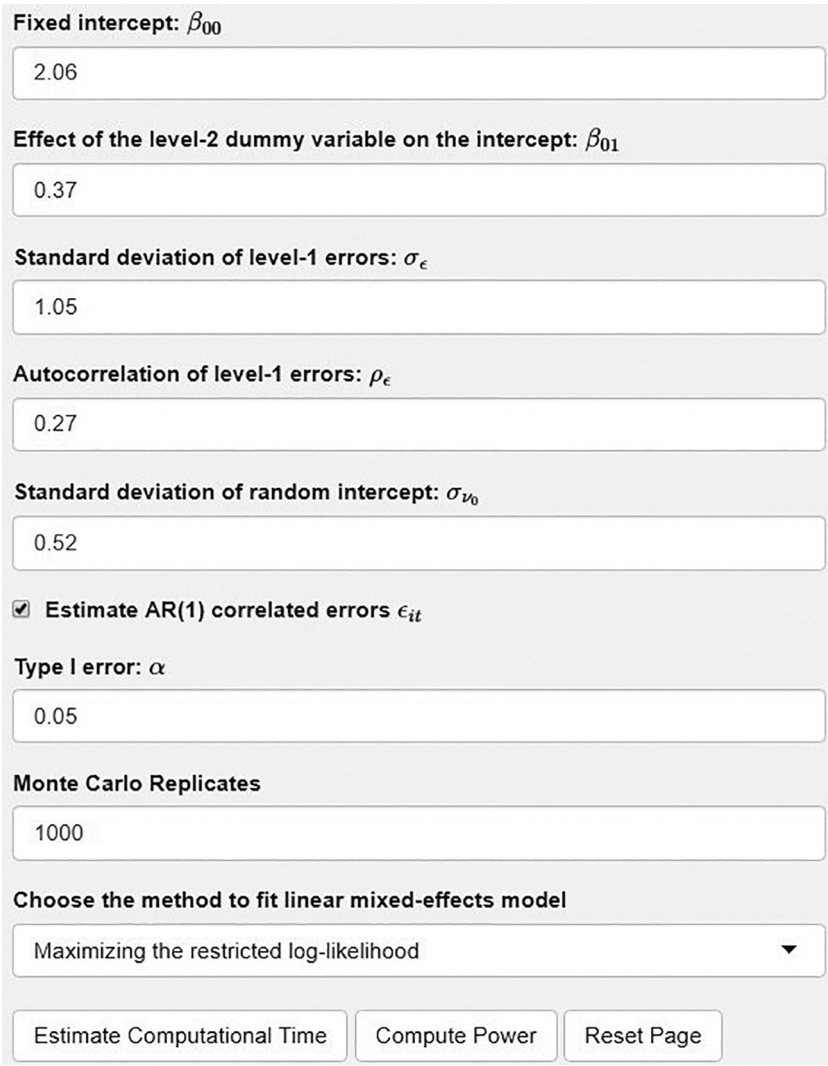


Figure 11.3 This screenshot of the PowerAnalysisIL app shows the window with the values to which the parameters of the model have been set

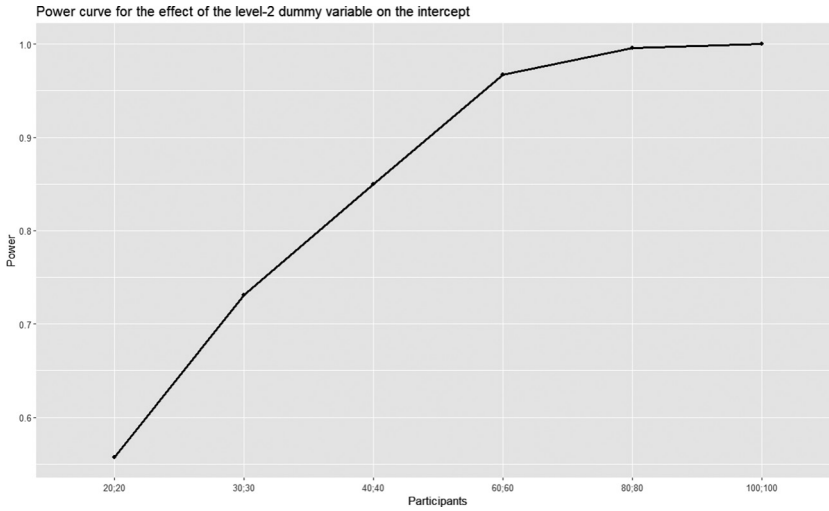


Figure 11.4 Power curve as a function of the number of participants for estimating mean differences in negative affect between individuals with MDD and healthy controls

Table 11.2 Summary of the fixed effects across 1,000 Monte Carlo replicates to estimate mean differences in negative affect between participants with MDD and healthy control participants

Fixed Effects	N_0	N_1	True Value	Mean	Std. error	Bias	(1-α) Coverage	Power
Fixed intercept (β_{00})	20	20	2.06	2.065	0.004	0.005	0.953	1.000
	30	30	2.06	2.057	0.003	-0.003	0.949	1.000
	40	40	2.06	2.058	0.003	-0.002	0.895	1.000
	60	60	2.06	2.063	0.002	0.003	0.934	1.000
	80	80	2.06	2.058	0.002	-0.002	0.955	1.000
	100	100	2.06	2.061	0.002	0.001	0.949	1.000
Effect of the level-2 dummy variable on the intercept (β_{01})	20	20	0.37	0.367	0.005	-0.003	0.968	0.554
	30	30	0.37	0.374	0.004	0.004	0.949	0.735
	40	40	0.37	0.373	0.004	0.003	0.902	0.873
	60	60	0.37	0.370	0.003	0.000	0.954	0.955
	80	80	0.37	0.371	0.003	0.001	0.949	0.991
	100	100	0.37	0.368	0.002	-0.002	0.961	0.997

11.3.2 Illustration II: power analysis to select the number of time points

Statistical power in multilevel models is a function of both the number of time points and the number of participants. In the previous illustration, we have focused on the number of participants while keeping the number and spacing of time points fixed. However, researchers might be interested not only in studying how the number of participants affects power but also how the number of time points might affect power. Therefore, we show how to conduct a power analysis to select the number of time points.

The current version of the *PowerAnalysisIL* app cannot display power curves when we vary the number of time points. However, it is possible to conduct a separate power analysis for each combination of the number of persons and the number of measurements occasions.

We illustrate how to use the app to investigate the effect of the number of time points on the power to estimate meaningful differences in the mean level of negative affect between individuals diagnosed with MDD and healthy control participants. We set the number of participants diagnosed with MDD to 20 and 40 and the number of healthy control participants to 20 and 40. We assume there is a fixed number of at least approximately equidistant observations within individuals. Thus, we set the total number of repeated measurements within individuals to 20, 40, 70, 100, 140 and 200.

To conduct the power analysis to select the number of time points, we perform six simulation analyses using the app. In each simulation, we fix the number of participants with MDD to 20 and 40, and the number of healthy control participants to 20 and 40; also for each simulation we specify the number of time points to 20, 40, 70, 100 and 140. Figure 11.5 shows the screenshots of the *PowerAnalysisIL* app. In each window, Model 1 has been selected and the sample size has been set for each combination of number of participants and number of time points. Subsequently, for each simulation study, we set the values of the parameters in the population model using the synthetic dataset (see Figure 11.3).

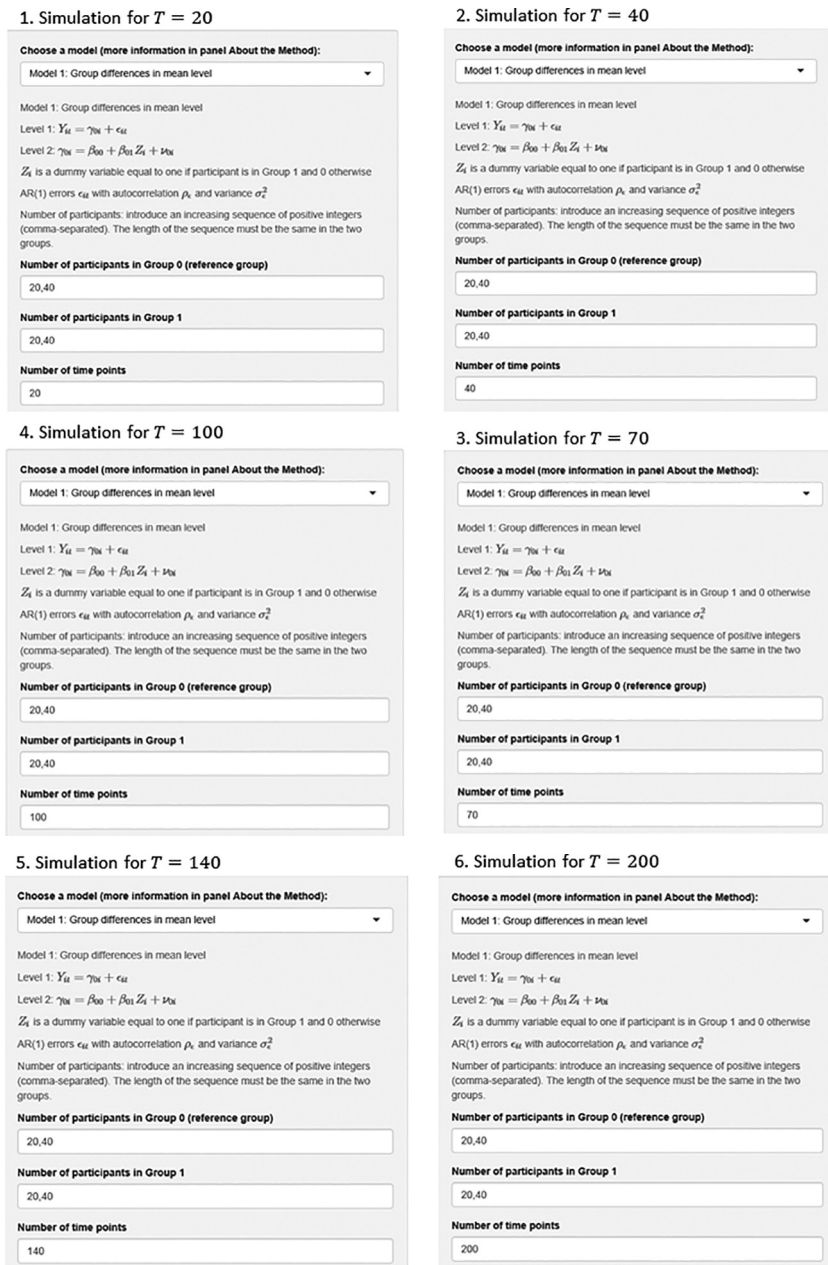


Figure 11.5 These screenshots of the PowerAnalysisIL app show the windows in which Model 1 has been selected and the sample size has been set for each simulation study

Table 11.3 shows the estimated power when the number of participants is 20 and 40 in each group and we vary the number of time points. We observe that increasing the number of time points has a lower impact on power than increasing the number of participants. The reason is that the hypothesis of interest concerns the effect of a Level 2 predictor (i.e., the dummy variable $Diagnosis_i$).). However, we note that whether it is preferable to increase the number of persons or the number of time points depends on the specific hypothesis of interest. For a broader discussion on the impact of the sample size at the different levels, we refer readers to Snijders (2005).

Table 11.3. Summary of the fixed effect 01 representing the Level 2 dummy variable on the intercept across 1,000 Monte Carlo replicates to estimate mean differences in negative affect between participants with MDD and healthy control participants

N_0, N_1	T	True Value	Mean	Std.error	Bias	(1-α) Coverage	Power
20,20	20	0.37	0.372	0.006	0.002	0.947	0.486
	40	0.37	0.364	0.006	-0.006	0.945	0.519
	70	0.37	0.362	0.005	-0.008	0.954	0.533
	100	0.37	0.362	0.005	-0.008	0.954	0.533
	140	0.37	0.368	0.005	-0.002	0.832	0.628
	200	0.37	0.375	0.005	0.005	0.958	0.569
40,40	20	0.37	0.366	0.004	-0.004	0.955	0.780
	40	0.37	0.372	0.004	0.002	0.955	0.832
	70	0.37	0.372	0.004	0.003	0.902	0.873
	100	0.37	0.373	0.004	0.003	0.949	0.864
	140	0.37	0.375	0.004	0.005	0.941	0.874
	200	0.37	0.369	0.004	-0.001	0.955	0.863

11.4 Additional consideration when selecting the temporal design

The temporal design of an ESM study is determined by the number of days in which ESM data is collected and the sampling frequency (i.e., the timing and distribution of questionnaire prompts) (see Chapter 3). Researchers interested in investigating how the total number of repeated measurements impacts power in multilevel models can use the *PowerAnalysisIL*

app, simulating datasets varying the number of time points while keeping the sample size constant. However, this approach has certain limitations that might compromise the reliability of the estimated power. First, the app simulates data assuming measurement occasions are equally spaced and that the data does not contain missing observations. However, ESM datasets will undoubtedly include missing assessments (e.g., Fuller-Tyszkiewicz et al., 2013; Santangelo et al., 2014; Stone, Broderick, et al., 2003). Therefore, when using the app, we recommend specifying the expected number of completed measurement occasions. After all, ESM studies include night breaks and unequal intervals between beeps (see Chapter 3). The effect of these factors can be evaluated using continuous-time models (de Haan-Rietdijk et al., 2017). Finally, we note that the simulation-based approach to conducting power analysis can be extended to three-level models in which measurement occasions are nested within days and individuals (de Haan-Rietdijk et al., 2016).

11.5 Feasibility and sample size planning in ESM studies

In the preceding pages, we introduced a methodological approach to performing power analysis for multilevel models. However, we did not consider additional constraints such as the feasibility of sampling participants from specific populations, the cost associated with enrolling extra participants, or the effect on participants' burden and compliance of increasing the number of measurement occasions. The goal of an optimal design in a longitudinal study is to select the sample size necessary to achieve high power while considering the cost related to including additional persons or increasing the total number of measurement occasions (see Brandmaier et al., 2015; Moerbeek, 2011). Even though the field of optimal design in ESM is still in its infancy, researchers facing limited resources to collect ESM data can first set different combinations for the number of persons and the total number of repeated measurements that are feasible to collect. Subsequently, power can be computed for each of them. For example, Revol, Lafit, Kirtley, et al. (2025) demonstrate how to formalize a cost function for ESM studies when conducting power analysis by determining four key parameters: fixed costs, cost of participants, cost of measurement occasions and specific costs attached to the participants' incentives.

11.6 Conclusions

In this chapter, we briefly reviewed current approaches for conducting power analysis in multilevel models. In particular, we illustrated how to perform a simulation-based power analysis for selecting the sample size using the PowerAnalysisIL app. The app implements a flexible approach based on Monte Carlo simulations to generate data that can be used to obtain power calculations when analytical formulations are not available. We note that this app includes an extensive set of models widely used in the ESM literature, and we refer to Lafit and colleagues (2021) for an overview of these models. For some of the models included in PowerAnalysisIL app, Lafit, Artner, et al. (2024) derived analytical power calculations which allow accelerating sample size calculations.

We note that power analysis requires that the investigator determine the values of the population model parameters. In certain situations, there is uncertainty related to the value of these parameters. The approach presented in the chapter can be extended to explore the effect of this uncertainty by performing a sensitivity analysis (see Lane & Hennes, 2018). In a sensitivity power analysis, power is computed for a range of plausible parameter values. In ESM research, it is common practice to rely on previously collected ESM data sets or pilot data to conduct the power analysis. To overcome the uncertainty associated with the use of previous studies, Lafit, Revol, et al. (2024) introduced an approach that implements safeguard power analysis to determine how much power-based sample size recommendations vary when considering differences in study design and preprocessing decisions.

We focused on power analysis for multilevel models applied to ESM data. However, open-source software and step-by-step tutorials have been developed for person-specific VAR(1) models (Revol, Lafit, & Ceulemans, 2024) and longitudinal actor-partner interdependence models for dyadic ESM designs (Lafit et al., 2022). It is also worth mentioning that, while in this chapter we focus on selecting the sample size to optimize statistical power, researchers can also focus on selecting the sample size to optimize predictive accuracy (i.e., the ability of a model to predict unseen data). In this regard, Revol, Lafit, & Ceulemans (2024) developed a novel sample size planning method to optimize predictive accuracy analysis for single-person VAR(1) models.

In sum, we highlight that the importance of conducting power analysis to justify the selection of the sample size in ESM studies not only relates to the reliability and replicability of psychological research (Munafò et al., 2017); as Button et al. (2013) state, there is an ethical dimension associated with low-powered studies, and therefore, we have the responsibility to avoid inefficient and wasteful research.

FUTURE: NEW DEVELOPMENTS IN ESM RESEARCH

CHAPTER 12

ESM as a Clinical Tool and Foundation for Intervention: From Research to Clinical Practice

Joanne R Beames, Lotte Uyttebroek, Evelien Schat, Glenn Kiekens and Inez Myin-Germeys

ESM is not only a research tool for studying psychological phenomena in daily life: it offers novel opportunities to advance clinical practice. This chapter focuses on four key topics related to the application of ESM in clinical practice. First, we explore the use of ESM as a clinical assessment tool, highlighting its potential to enhance person-centred care. Second, we examine how ESM can be integrated into Ecological Momentary Interventions (EMIs), which allow for timely and personalized therapeutic support. Third, we identify the critical role of co-design and lived experience in the development of ESM/EMI tools. Fourth, given that adoption of ESM/EMI tools in mental healthcare is limited, we discuss the role of implementation research in bridging the research-to-practice gap. Throughout the chapter, we give examples, identify design and methodological advancements and suggest avenues for future research.

12.1 ESM as a clinical tool

ESM can be a valuable clinical assessment tool in routine mental healthcare. By capturing real-time data on individuals' thoughts, feelings and behaviours in their natural environments, ESM can provide a nuanced understanding of the intensity and variability of symptoms (Myin-Germeys et al., 2018). ESM research has typically focused on identifying the role of candidate momentary mechanisms (as situated in context) in the development and maintenance of mental health problems (Schick et al., 2023). For example, applications of ESM in clinical populations have examined how symptoms relate to each other in the moment and

from one moment to the next, how symptoms vary across environments and how complex systems bounce back from disturbances (van Os et al., 2017). Other applications have examined ESM as a tool for tapering and dose-finding of psychotropic medications (van Os et al., 2017). This information is critical for adopting a person-centred approach; through ongoing digital monitoring, reporting and feedback, ESM can guide effective case conceptualization, symptom management and treatment planning (Myin-Germeys et al., 2018; Reininghaus & Myin-Germeys, 2023; von Klipstein et al., 2023). Idiographic examples illustrate the practical use and benefits of ESM as a clinical tool in relation to anxiety, sleep problems and depression, with detailed guides provided by Daniëls et al. (2023) and von Klipstein et al. (2023).

Accumulating evidence demonstrates that ESM in clinical practice can enhance person-centred care by 1) fostering patient engagement and empowerment, 2) supporting self-management and recovery, 3) guiding goal-oriented clinical assessment and care management, and 4) promoting shared decision-making. We discuss these four processes below (for more details, see Myin-Germeys et al., 2023; 2025) followed by a discussion of methodological advancements and clinical considerations.

12.1.1 Patient engagement and empowerment

ESM provides a unique way for patients to collect their own treatment and assessment data in routine clinical care. In doing so, ESM challenges traditional models of healthcare by positioning patients as experts or active partners in, rather than passive recipients of, their treatment (Myin-Germeys, 2020; van Os et al., 2017; Verhagen et al., 2022). Emerging evidence indicates that active involvement via self-monitoring increases patient engagement and sense of control or empowerment in the care process (Simons et al., 2015; Wichers et al., 2011). Qualitative studies investigating hypothetical use cases of ESM provide further indirect support, with clinicians (Bos et al., 2019; Piot et al., 2022; Weermeijer et al., 2024) and individuals with lived experience of psychosis (de Thurah et al., 2023) reporting that frequent assessments may enhance awareness and insight into symptoms and triggers. Relatedly, an integrated narrative review that includes lived experience perspectives from youth and clinicians found that ESM can enhance awareness and, in turn, reduce symptoms in the context of depression (Beames et al., 2021).

Although ESM has the potential to increase patient engagement and empowerment in clinical care, patients must be willing and able to engage in intensive self-reporting. Compliance with the ESM protocol is therefore a critical methodological consideration to ensure that ESM enhances quality of care. Recent reviews and meta-analyses show that patients can provide self-reports using ESM with average compliance levels of around 80% (Schick et al., 2023; Vachon et al., 2019). Despite decreases in compliance rates in some clinical populations or settings (Rintala et al., 2019; Zarbo et al., 2023), a pilot implementation study achieved a 50% compliance rate, indicating acceptable compliance in routine mental healthcare (Weermeijer et al., 2023). Active patient involvement in deciding which ESM items to assess might be an important consideration for compliance (Bos et al., 2019). Another consideration is determining the level of compliance required to achieve intended outcomes (Michie et al., (2017). Overall, these data support the potential of ESM to actively involve individuals in their own care.

12.1.2 Self-management and recovery

ESM data are ideally placed to identify the short-term patterns of mental health problems and the associations of these patterns with feelings, thoughts and behaviours. This information can inform self-management and recovery. For example, ESM can identify fluctuations in affect that indicate poor psychological well-being and the presence of mood disorders in children, adolescents and adults (Dejonckheere et al., 2019; Houben et al., 2015; Reitsema et al., 2022). Affect dynamics may also predict the naturalistic course of depression over a six-month period (Panaite et al., 2020), with changes reflecting early warning signs of subsequent depressed mood (Schreuder, et al., 2024; van de Leemput et al., 2014; Wichers, 2014; Wichers et al., 2020; see Helmich et al., 2024 for a critical perspective on early warning signals). In the context of psychosis, converging results indicate that negative emotions such as stress and anxiety are associated with the subsequent development of paranoia, particularly when such negative emotions are experienced during interpersonal situations (Bogudzińska et al., 2024; Lüdtke et al., 2023). Further, altered stress responses and poor sleep quality are associated with a range of mental health issues (Bogudzińska et al., 2024; Lachowicz, 2025; Littlewood et al., 2019; Lüdtke et al., 2023; van Winkel et al., 2015). Together, these examples

demonstrate the capability of ESM to facilitate early detection and prevention as well as to identify contextually relevant intervention targets.

A critical way that ESM facilitates self-management and recovery is through the provision of personalized feedback to patients about individual patterns of risk and behaviour. Such feedback functions to enhance understanding of experiences in context and signal the need for behavioural change or implementation of learned strategies. For example, ESM-based feedback can enhance resilience by strengthening the use of natural rewards (van Os et al., 2017), reduce depressive symptoms (Kramer et al., 2014) and induce lasting behavioural changes (Snippe et al., 2016). As this line of research progresses, personalized ESM-based feedback is emerging as a promising tool for embedding personalized insights into routine clinical care. For example, an ongoing trial is investigating personalized ESM monitoring and feedback via the Therap-i app to support psychological treatment for depression (Riese et al., 2021). Furthermore, PErsonalized Treatment by Real-time Assessment (PETRA), a web-based e-health application with ESM, has been integrated into the Dutch Electronic Health Record system since 2019 through several mental healthcare organizations (see https://umcgresearch.org/ontwikkeling-petra). A recent qualitative study of 15 ongoing ESM studies in clinical settings identified three necessary conditions for providing effective feedback: (1) training for healthcare professionals and researchers, (2) online interfaces and graphical visualizations to present data and (3) face-to-face interactions with patients to discuss in which contexts ESM is useful and potential implications (Bartels et al., 2023). These conditions align with other clinical implementation studies with mental health professionals (e.g., Weermeijer et al., 2023).

12.1.3 Goal-oriented clinical assessment and care management

The highly detailed and personalized data provided by ESM can shape both the assessment and management of symptoms in clinical care. ESM can help clinicians and their patients to identify focus areas and, in turn, lead to actionable and personalized therapy goals (Bartels et al., 2023; Myin-Germeys et al., 2024; van Os et al., 2017). Additional work is needed, however, to elucidate how therapy goals translate into specific ESM questionnaires and sampling schemes (Bos et al., 2019; Bos et al., 2022; Weermeijer et al., 2023). In addition to goal-setting, ESM can be used to assess treatment outcomes over time, providing insights over and above

traditional retrospective measures. For example, a recent review found that momentary measures of consummatory and anticipatory anhedonia were integral to capturing state-like changes over the course of psychological interventions (Beames et al., 2025). Another related benefit of ESM assessments in clinical care is that they can identify early processes of change and side effects (Myin-Germeys et al., 2018). The implication here is that ESM is a useful routine outcome monitoring tool that can guide treatment decisions (Bartels et al., 2023) and take us one step closer to precision medicine (Verhagen et al., 2022).

12.1.4 Shared decision-making

By integrating patients' subjective experiences with clinical expertise, it is theorized that ESM fosters collaborative treatment planning and decision-making. The idea is that ESM can provide the relevant qualitative day-to-day information on experiences and contextual factors that are needed for making joint treatment decisions (Myin-Germeys et al., 2024). One qualitative study with researchers implementing ESM in clinical care found that ESM-based feedback may provide relevant input for shared interpretation and care decisions (Bartels et al., 2023). However, scientific evidence about whether (and how) ESM improves shared decision-making in clinical mental healthcare is lacking.

12.1.5 Methodological advancements and considerations for clinical practice

A core part of using ESM as a clinical tool is presenting contextually informed, temporal feedback that is personalized to the individual. However, despite the potential of ESM-based feedback, research is still emerging about *how* to generate, use and implement it as a clinical tool. We dive deeper into recent advancements in the ESM field below, with a focus on identifying meaningful indicators of change in individual time-series data (personalization) and optimizing visualizations for easy interpretation.

12.1.5.1 Personalization

Qualitative research with clinicians and clients highlights the importance of personalizing ESM content to increase the potential usefulness of ESM

in practice (de Thurah et al., 2023; Piot et al., 2022; Weermeijer et al., 2024). Personalization could be achieved by allowing clinicians and patients to select or formulate relevant questions (e.g., von Klipstein, 2023; Weermeijer et al., 2024), using qualitative ESM items to generate rich descriptions of events and experiences, personalizing response options and self-learning items (von Klipstein, 2023) or allowing algorithms to recognize individual patterns and tailor questionnaires accordingly (Schneider et al., 2024).

12.1.5.2 Visualizations

ESM data are primarily visualized for researchers and are not always easy to interpret for clinicians and their patients. Qualitative studies indicate that clinicians want intuitive data visualizations such as line graphs depicting mood variability over time or pie charts displaying frequency of social contacts (Piot et al., 2022; Weermeijer et al., 2024). Clinicians also described the utility of depicting variation of symptoms and experiences according to context (Piot et al., 2022). ESM software and tools such as m-Path, FRED, and ESMvis can generate these types of visualizations for detailed feedback reports in clinical practice (Bringmann et al., 2021; Mestdagh et al., 2023; Rimpler et al., 2024). Person-specific symptom networks can also be visualized (von Klipstein et al., 2020), but they have limited reliability and are difficult for clinicians to interpret (Bastiaansen, Kunkels et al., 2020). Due to their complexity, some researchers advise against using networks in clinical practice (Wichers et al., 2021) or recommend exploratory use only (von Klipstein et al., 2020).

12.2 Ecological Momentary Interventions (EMIs)

ESM is increasingly being used in mental health research to inform and/or deliver Ecological Momentary Interventions (EMIs) in daily life via mobile devices (Myin-Germeys et al., 2018; Myin-Germeys et al., 2016). Although the technical definition of EMIs varies (Balaskas et al., 2021), there is a general consensus that they are 'treatments that are provided to people during their everyday lives (i.e., in real time) and in natural settings (i.e., real world)' using digital technology (Heron & Smyth, 2010). The underlying assumption of EMIs is that experiences and behaviours

are most amenable to change in the context within which they occur (Reininghaus, 2018). In this way, EMIs aim to integrate preventive and therapeutic methods into daily life, achieving sustainable change through ecological translation (Reininghaus, 2018; Schulte-Strathaus et al., 2023). For example, EMIs offer patients the opportunity to practice new skills and behaviours learned in therapy during their day-to-day life – arguably where it matters most. EMIs have been described as a 'therapist in your pocket' (Shiffman et al., 2008) and are part of the solution to reform clinical care through increased access and personalization (Reininghaus & Myin-Germeys, 2023).

In a typical EMI, intervention components are offered according to event-contingent, time-contingent or hybrid ESM design principles and sampling strategies (Myin-Germeys et al., 2018; Reininghaus, 2018). 'Event-contingent' means that interventions are delivered based on specific events, experiences or behaviours in daily life, all of which are defined by a-priori criteria or scores on an ESM survey. Examples include increased momentary event-related stress, increased negative affect or decreased physical activity. 'Time-contingent' means that interventions are delivered at predetermined or random times regardless of individuals' responses, activities or contextual circumstances. For example, a behavioural activation activity might be scheduled to occur every day at 3:00 p.m. and 6:00 p.m. Hybrid approaches combine event- and time-contingent approaches. EMIs therefore build on extensive ESM research and are designed to target candidate mechanisms and mental health outcomes in daily life; for example, common targets in the psychosis literature include stress sensitivity, threat anticipation, aberrant salience and self-esteem (Schulte-Strathaus et al., 2023). ESM-based EMIs involve some level of interaction from the individual and aim to provide timely support that is personalized based on their experiences and behaviour in daily life (Myin-Germeys et al., 2018; Reininghaus, 2018).

The content and delivery method of existing EMIs are highly varied (see Section 12.2.2 for examples). Research indicates that EMIs are mostly based on Cognitive Behavioural Therapy, Acceptance and Commitment Therapy, mindfulness, behavioural activation and relaxation (Reininghaus & Myin-Germeys, 2023; Versluis et al., 2016). Another review focused specifically on ESM-based EMIs found that they mostly consisted of coping strategies, motivational feedback (e.g., supportive messages) or informational feedback (e.g., personalized graphs) (Dao et al., 2021). These interventions show

great promise, particularly when adherence is promoted and social components such as blended care are incorporated (Dao et al., 2021; Versluis et al., 2016). Blended care involves combining digital tools such as EMIs, with traditional face-to-face therapy or professional support. Emerging evidence indicates that blended approaches may be as effective as standard care in treating mental health disorders, requiring fewer resources for delivery and potentially enhancing individuals' engagement in their recovery process (Erbe et al., 2017).

12.2.1 The emerging evidence base

EMIs and other related mobile health (mHealth) interventions are now being researched across a range of mental health domains, including depression, anxiety, trauma and stress-related disorders, psychosis and psychotic disorders, non-suicidal self-injury and substance misuse (Kiekens et al., 2023; Linardon et al., 2024; Miralles et al., 2020; Schulte-Strathaus et al., 2023). Emerging evidence shows that there is high acceptance and feasibility of EMIs in subclinical and clinical samples (e.g., Bell et al., 2017; Hanssen et al., 2020; Rauschenberg et al., 2021). There is also evidence that EMIs may be useful adjuncts to therapy, particularly when they are well integrated and tailored to individual patient needs (McDevitt-Murphy et al., 2018; Schueller et al., 2017). Although there is accumulating pilot data on effectiveness and efficacy (see Table 1), further research is needed to establish the evidence base (Gründahl et al., 2020). There is a need for high-quality randomized controlled trials (RCT) that include longer follow-up assessments, larger sample sizes and examination of whether candidate mechanisms lead to long-term improvements in mental health outcomes (Reininghaus et al., 2016; Schulte-Strathaus et al., 2023). Further, the effectiveness, cost-effectiveness and implementation of most EMIs in routine care settings remains unclear. Developing the evidence base should be a priority for researchers; it is critical to offer clinicians EMIs that they can safely use in their clinical practice.

12.2.2 Examples of EMIs

There are various examples of EMIs, some of which are provided in Table 12.1 below. One example is Acceptance and Commitment Therapy in

Daily Life (ACT-DL). ACT-DL uses a blended care approach that combines face-to-face ACT sessions with an EMI to enhance psychological flexibility beyond the therapy room (Reininghaus et al., 2019; Vaessen et al., 2019). Six core ACT skills, including acceptance, cognitive defusion, self-as-context, contact with the present moment, values and committed action (Hayes, 2016) are introduced in therapy sessions and then practiced in daily life with the ACT-DL EMI. The EMI provides users with on-demand visual, textual and audio materials along with prompts to complete short ESM questionnaires on current affect, context and activities. Each of the ESM questionnaires is followed by an ACT-exercise or metaphor. Two RCTs found preliminary efficacy of ACT-DL in improving clinical outcomes for individuals with emerging depressive and psychotic symptoms (Myin-Germeys et al., 2022; van Aubel et al., 2020). Both individual and group ACT-DL were feasible regarding treatment adherence, acceptability and usefulness; however, some participants found ESM burdensome (van Aubel et al., 2020; van Aubel et al., 2024). Qualitative user experience highlighted the need for flexibility in ESM design and personalized feedback to enhance the integration of the EMI within therapy sessions (Bouws et al., 2025). ACT-DL was subsequently optimized to allow personalization in the EMI (items, number of beeps, time window) and face-to-face intervention (order and number of sessions) and to integrate personalized feedback via a web-based dashboard for clinicians. Results from a pilot clinical implementation study are currently being analysed. Together, this body of work shows the importance of establishing efficacy and incorporating real-world use and lived experiences into the design and optimization of EMIs.

Table 12.1 Selective Overview of EMIs

EMI Description & Clinical Focus	Evidence
ZELF-i (Bastiaansen et al., 2018). Add-on ESM tool (5x/day, 28 days) for usual depression treatment with weekly digital feedback reports and one face-to-face session. Two ESM item and feedback modules: Do-Module (positive affect, activities); Think-Module (negative affect, thinking patterns).	Three-arm pragmatic RCT (N=161): 86% of completers recommended ZELF-i, but no changes in depression, social functioning or empowerment compared to control (Bastiaansen, Ornée, et al., 2020). Qualitative interviews (N=20): ZELF-i increased self-insight, awareness and management (Folkersma et al., 2021).
SELFIE (Daemen et al., 2021). Blended 6-week transdiagnostic intervention based on Cognitive Behavioural Therapy for improving self-esteem in youth exposed to childhood adversity. Includes three face-to-face sessions, three email contacts and an EMI with ESM (6x/day, 3 days during each intervention week) and tailored exercises.	Two-arm RCT (N=174): SELFIE improved self-esteem compared to control at post and 6-month follow-up, but no differences in symptoms or functioning (Reininghaus, Daemen et al., 2024).
EMIcompass (Schick et al., 2021). Blended 6-week compassion-focused transdiagnostic intervention for help-seeking youth. Includes four sessions with a psychologist and an EMI that aims to improve self-compassion, self-care and reduce stress reactivity. The EMI provides strategies each week according to different schemes e.g., based on optional ESM (6x/day, 3 days per intervention week) assessments of momentary stress or negative affect. Development included consultation with local experts.	Uncontrolled pilot (N=10) and exploratory RCT (N = 92): EMIcompass was feasible with low burden. Preliminary evidence for reduced stress sensitivity and improved momentary resilience and quality of life compared to control (Paetzold et al., 2023; Reininghaus et al., 2023).
Smartphone Assisted coping-focused intervention for Voices (SAVVy) (Bell et al., 2018a, 2018b). Blended 8-week coping-focused intervention for individuals with schizophrenia spectrum disorders with voice hearing experiences. Includes four therapy sessions and ESM (10x/day, 6 days) to perform functional analysis of voice hearing. Coping strategies identified are integrated into EMI and used as reminders (5x/day, 10 days). Development included lived experience feedback.	Pilot RCT (N = 34): supported feasibility, acceptability and improved clinical outcomes compared to control (Bell et al., 2020). Qualitative interviews (N = 12) showed that SAVVY was perceived as helpful for capturing and communicating experiences better to therapists, which in turn improved therapeutic relationships (Moore et al., 2020).

12.2.3 Methodological advancements

12.2.3.1 Just in Time Adaptive Interventions (JITAIs)

EMIs deliver real-time support as mobile interventions, extending the reach of traditional treatments (Myin-Germeys et al., 2018). Utilizing EMIs, individuals decide when to engage with interventions in their daily lives (i.e., pull mHealth interventions). However, this approach assumes that people know when to use EMIs, which may not always be the case. One sophisticated EMI that addresses this limitation and may have great promise for highly dynamic emotions, cognitions, and behaviours are JITAIs (Wang & Miller, 2020). JITAIs are an intervention design that adapts support based on an individual's changing status and context; they provide interventions when needed (i.e., push mHealth interventions) based on an individual's momentary state of vulnerability *and* when they are most receptive to engaging with the intervention (Nahum-Shani et al., 2018; Nahum-Shani et al., 2023). JITAIs consist of six key elements: distal outcomes, proximal outcomes, decision points, tailoring variables, intervention components and decision rules (Nahum-Shani et al., 2018; Nahum-Shani et al., 2023).

1. Distal outcome(s): the ultimate long-term outcomes that guide intervention development and are focused on long-term behaviour change.
2. Proximal outcome(s): the near-term outcomes that the intervention is designed to impact (e.g., short-term changes in affect and cognition) and which are immediate goals or mediators for the intervention's effectiveness.
3. Decision points: the moments when an intervention may or may not be delivered (e.g., every morning at 10:00 a.m. or between 8:00 p.m. and 9:30 p.m.).
4. Tailoring variables: the real-time data (e.g., feelings, thoughts, appraisals, behaviours, contextual variables) used to assess whether and which intervention is most appropriate at a particular decision point.
5. Intervention component and options: a set of intervention techniques (e.g., practicing skills) that can be provided (e.g., mindfully observing emotional sensations) at a decision point.
6. Decision rules: a set of rules that use tailoring variables to determine which intervention to offer and when. For example, if the risk level is 'Low,' no intervention is given; if 'Moderate,' intervention X is offered; if 'High,' intervention Y is recommended.

Once a meaningful distal outcome is defined (e.g., reduction of self-injurious behaviours), the anticipated primary mechanism of change toward that distal outcome must be formulated (e.g., reduction in emotional distress); this latter mechanism defines the proximal outcome(s) and requires a solid clinical and theoretical foundation. Similarly, the decision points (e.g., in the evening hours for self-injury; Kiekens et al., 2024), tailoring variables (e.g., increased negative emotions, reduced self-efficacy to resist self-injury; Hasking et al., 2017; Kuehn et al., 2022) and intervention components (e.g., skills training; Calvo et al., 2022) need to be determined based on prior research and theories of psychopathology. Finally, researchers must operationalize which intervention components to offer at different levels of risk and receptivity, which is governed by decision rules. Decision rules are evaluated (and optimized) in a micro-randomized control trial (Qian et al., 2022), an experimental design used to explore whether and under what circumstances interventions are effective for each individual (Bidargaddi et al., 2020). While few JITAIs for mental health have been developed and evaluated so far in micro-randomized control trials (Wang et al., 2023), this is a field in which we expect to see significant progress in the next few years as several studies are currently underway targeting behaviours such as cannabis use, smoking, gambling, self-injury and disordered eating behaviours (e.g., Coughlin et al., 2024; Dowling et al., 2024; Goldstein et al., 2021; Kiekens et al., 2023; Yang et al., 2023).

12.2.3.2 Detecting when to intervene

A large body of research has focused on retrospective analysis of ESM data to detect changes in responding that might indicate when an intervention is needed (e.g., change point detection methods; Cabrieto et al., 2018). However, for intervention purposes, it is more beneficial to detect changes in real time as ESM data are being collected. Statistical Process Control (SPC) procedures that have been developed in industry for monitoring production processes are particularly useful for this purpose (Montgomery, 2009; Shewhart, 1931). To illustrate the two distinct phases required for applying SPC in practice, we introduce the example of 'Alice'. In phase I, the in-control distribution of the monitored process is captured during a baseline phase of data collection. This involves estimating the mean and standard deviation of repeatedly assessed variables for an individual; measuring Alice's emotions is an example. These estimates are

then used to compute control limits (or cut-offs), which define the boundaries of Alice's normal emotions, making it a person-specific benchmark. In phase II, incoming emotions reported by Alice are monitored over time. If these scores remain within the control limits, Alice's emotions are considered 'normal' or in control. If a score falls outside these limits, it indicates a deviation from her normal emotional range, signalling that an intervention may be necessary. For a visual example of SPC, see Figure 1 in Schat et al. (2023). Both simulation and empirical studies have shown promising results for using SPC methods with ESM data to signal when a person is at risk of relapsing into depression (Schat et al., 2024; Schat et al., 2023; Schreuder et al., 2024; Smit et al., 2022; Smit & Snippe, 2023; Snippe et al., 2023).

In theory, SPC for intervention purposes can be implemented in various ways. Scores that fall outside of control limits, indicating that Alice deviates from her normal range, can be used to trigger ESM bursts, where multiple beeps are sent in a short period to investigate the deviation. EMIs can also be given at these moments, providing timely intervention when it is most needed. Additionally, SPC can aid in personalizing ESM; while patient input is valuable for deciding what to monitor, it can also reveal changes without the patient's (or clinician's) knowledge, providing data-driven insights. Further research is needed to clarify how SPC can inform the clinical implementation of EMIs, particularly in relation to the feasibility of collecting the large amount of data needed to estimate the control limits during phase I.

12.3 Co-design and stakeholder engagement

Co-design and stakeholder engagement are crucial for creating ESM-based assessment tools and interventions (referred to as 'experience sampling innovations' for the remainder of the chapter) that are effective and seamlessly integrated into clinical practice. Engaging clinicians and patients from the outset of design ensures that innovations are tailored to real-world needs and challenges, facilitating smoother implementation and greater acceptance (Deniz-Garcia et al., 2023). Clinicians' involvement can help identify practical barriers and opportunities for integration within healthcare settings while patient feedback can help to assess the acceptability, usability and relevance of the innovations (Heron & Smyth, 2010). For example, to

develop appropriate JITAIs, it is crucial that researchers identify meaningful distal outcomes in collaboration with mental health professionals and people with lived experience of mental health problems before selecting intervention components and designing intervention options. This collaborative approach addresses the high discontinuation rates of mHealth tools by incorporating stakeholder preferences and content requirements, thereby enhancing engagement and retention (Nicholas et al., 2017). The co-design process should involve mixed methodologies and precede large-scale trials to identify and resolve potential issues early, increasing the likelihood of successful, sustained implementation in clinical practice (Heron & Smyth, 2010). Co-design principles and stakeholder engagement have rarely been involved in the design of mHealth tools, however, let alone experience sampling innovations using mobile technologies (Veldmeijer et al., 2023). Stakeholders have mostly been involved in the final stages of digital innovation – for example, usability testing, where the ability to proactively influence change is severely limited (Veldmeijer et al., 2023).

12.4 Clinical implementation of ESM and EMIs

There is a growing body of evidence demonstrating the value of experience sampling innovations in enhancing person-centred mental healthcare. However, these innovations are hardly integrated into clinical practice, and when they are, setbacks and challenges are the norm (Myin-Germeys, 2020; Myin-Germeys et al., 2024). Implementation science is critical for understanding these challenges as well as for enhancing adoption and sustainability. In this section, we summarize clinical implementation science and research and outline next steps for evaluation (see Kip et al., 2025 for a comprehensive framework for the development, implementation and evaluation of eHealth technologies).

12.4.1 Overview of clinical implementation research

Most evidence to date identifies barriers and facilitators to implementing experience sampling innovations in clinical practice from the perspectives of clinicians and their patients. Studies typically explore the hypothetical use of ESM and/or EMIs in clinical practice although there is accumulating supporting evidence from pilot implementation studies. One factor

of critical importance to clinical implementation is the readiness of the healthcare professional to adopt ESM and/or EMIs in their practice. In a sense, clinicians are 'gatekeepers' of mHealth because they present options for care based on their personal training and expertise. A survey study of 375 practicing clinical psychologists found that they were hesitant to adopt ESM-based assessment and outcome-monitoring tools because they perceived them as less useful than existing instruments (Ellison, 2021). Other common barriers to implementation and scale-up reported by clinicians include concerns about patient burden, potential adverse effects, inadequate training and lack of guidelines and time constraints (Bos et al., 2019; Frumkin et al., 2021; Piot et al., 2022; Weermeijer et al., 2024; Weermeijer et al., 2023; Zimmermann et al., 2019). Perceived burden and adverse effects were also identified as potential barriers by patients (Bos et al., 2019; de Thurah et al., 2023; Weermeijer et al., 2023).

The tools of mHealth (in general) are more likely to be adopted by patients if they are personally relevant, seamlessly integrated into daily life and empowering for managing their illness through access to meaningful data (Greer et al., 2019; Torous et al., 2018). A realist evaluation of an EMI targeting self-esteem in youth exposed to childhood trauma (SELFIE) found that using personal smartphones was a facilitating context – it enhanced a sense of privacy, leading to increased disclosure and active participation (Postma et al., 2024). Clinicians also exhibit favourable attitudes towards using digital technology when they contribute to an improvement in the quality of care without a significant increase in workload (Kerst et al., 2020). Interoperability is a crucial consideration for integration of experience sampling innovations into existing healthcare systems and making them accessible for large-scale use. Most mHealth tools are custom-made for specific studies (Miralles et al., 2020) and therefore lack scalability. This highlights the need for specifications like Fast Healthcare Interoperability Resources for seamless data transfer and clinical implementation. Overall, it is essential to assess attitudes about experience sampling innovations, including readiness to adopt, before making decisions about dissemination and integration into routine care.

12.4.2 Next steps for evaluation

Assessing attitudes about experience sampling innovations is necessary but not sufficient to fully understand their clinical implementation.

Knowledge about broader system factors is essential because experience sampling innovations inherently necessitate change within the mental healthcare systems that they are being implemented in. We need to shift the status quo of evaluation to embrace methodologies from implementation science that can capture this complexity. One way forward is effectiveness-implementation hybrid trials (Curran et al., 2012; Landes et al., 2019).

12.4.2.1 Implementation hybrid trials

The scientific evaluation of the full life cycle of mHealth interventions (usability, feasibility, efficacy and effectiveness) can take several years throughout which it is difficult to keep up with rapid technological development. Implementation hybrid trial designs can accelerate the translation of research findings into practice, leading to more timely improvements in mental healthcare (Brown et al., 2017). These designs combine elements of traditional clinical trials with implementation science to evaluate both the effectiveness of an intervention and the strategies used to implement it (see Curran et al., 2012). Embedded mixed-methods process evaluations can be particularly informative, providing insight into why interventions work in some contexts and not others – they aim to demystify the 'black box' of complex intervention trials by considering contextual factors and adaptations made for intervention delivery into a particular system (Moore et al., 2015; Oakley et al., 2006). Such evaluations should be guided by implementation theory and frameworks (Papoutsi & Greenhalgh, 2024). The Consolidated Framework for Implementation Research (CFIR) provides a structured approach to understanding the multifaceted influences on implementation success such as intervention characteristics, inner and outer settings and individual factors (Damschroder et al., 2022). The Nonadoption, Abandonment, Scale-up, Spread, and Sustainability (NASSS) framework helps to identify challenges across multiple levels of a system that may hinder or facilitate the adoption and sustainability of innovations in healthcare (Greenhalgh et al., 2017). Furthermore, a recent review of mHealth interventions maps common factors influencing uptake in healthcare and provides a tool for policymakers to assess their healthcare system's readiness for integrating mHealth interventions (van Kessel et al., 2023). By using hybrid designs, researchers can identify and address implementation barriers early, ensuring that experience sampling

innovations are not only effective but also practical and sustainable in real-world settings (Curran et al., 2012; Landes et al., 2019).

Hybrid trials have occasionally been implemented in the context of experience sampling innovations. One example is the Implementing Mobile Mental Health Research in Europe (IMMERSE) effectiveness-implementation trial, which aims to investigate the reach, effectiveness, adoption, implementation and maintenance of an ESM-based monitoring tool in routine mental healthcare in Belgium, Germany, Scotland and Slovakia (Reininghaus, Schwannauer, et al., 2024). Clinical units in the experimental condition received the ESM tool for 12 months, beginning with a two-month focus period during which clinicians and patients were encouraged to use the tool for at least four weeks. Active implementation support was only provided during the first six months. Outcome data was collected at baseline, two months, six months and twelve months post-baseline. Participants from the experimental condition were also invited for a semi-structured interview during the first six months of ESM to identify key factors of successful implementation. The results from IMMERSE will build the foundation for translating digital ESM innovations into real-world practice, fostering true person-centred care.

12.4.2.2 Digital regulations

Data privacy, security, confidentiality and regulatory supervision of mHealth tools, including experience sampling innovations, remain critical concerns for clinicians and their patients (Alhammad et al., 2024; Giebel et al., 2023). These concerns are important because the potential of mHealth tools to transform healthcare ultimately hinges on trust in their safety and security (Sheppard, 2020). As part of the broader response to these concerns, European regulations like General Data Protection Regulation (GDPR) and the Medical Device Regulation (MDR 2017/745) have set stricter requirements for technology transfer. Developers, researchers, clinicians and their patients need to be aware of and comply with these regulations to ensure ethical and responsible use of mHealth tools in care. To facilitate implementation, van der Storm et al. (2023) have developed a checklist for using or developing an app as a medical device. However, compliance is complex, and the guidelines have been criticized as lacking practical effectiveness (Marelli et al., 2020; van der Storm et al., 2023; van Kolfschooten, 2022). For example, clinicians do not typically receive

standard training in mHealth, which is problematic because they can be held accountable for wrongful use (van der Storm et al., 2023). Further, there is no standardization in ethical and regulatory requirements across countries (or even within the European Union), meaning that widespread implementation can be time-consuming and difficult (Marelli et al., 2020). For further explanation and critique on digital regulations in Europe, see Myin-Germeys et al. (2024).

12.5 Conclusions

Experience sampling innovations show promise for enhancing mental healthcare through greater personalization and engagement. However, many open questions and challenges remain, including how to optimize the innovations and how to achieve clinical integration and sustainability as well as rapid methodological (and regulatory) advancements. In this chapter, we provided an overview of what experience sampling innovations are as well as key considerations for development and clinical implementation. We visualized these considerations in Figure 12.1; they are by no means exhaustive but aim to offer a starting point for those who are new to the field.

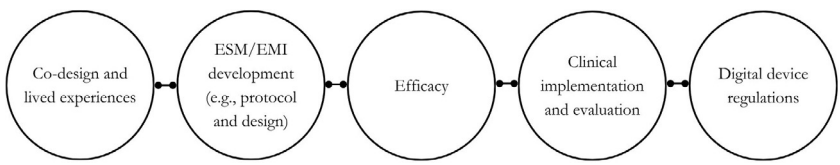


Figure 12.1 A graphical representation of five key considerations for the development and clinical implementation of ESM/EMIs. These components should not necessarily be considered linearly. For example, including perspectives from individuals with lived experience can be critical at all stages of the research-to-practice pipeline.

CHAPTER 13

Passive Sensing in ESM Research

Aleksandra M. Lachowicz, Laura Van Heck*, Koen Niemeijer and Thomas Vaessen*

*shared first author

Over the past decade, daily-life research has witnessed a steep increase in the use of passive sensing – data collection without the immediate active involvement of the user – to capture aspects of real-life behaviour, physiology and context. According to the Web of Science database, there was a more-than-tenfold increase between 2013 and 2023 in the number of scientific reports published annually on wearables (i.e., body-worn devices that measure bodily or environmental signals such as smartwatches). No doubt, this rise has to do with the technological advances in the fields of microelectronics and software development that make passive sensing using wearables possible but also with improvements in usability and aesthetic design that in turn improve usability and user experience. Likewise, the development of smartphones has taken flight, adding even more possibilities to the passive ecological investigation of daily life through use of smartphone sensors (also known as mobile sensing). These wearable and mobile sensing methods have become popular with experience sampling method (ESM) researchers who use them to complement and further develop their traditional protocols.

Commercial wearables have become mainstream products. In 2019, there were a projected total of 722 million wearables connected worldwide (Statista, 2019); 21% of American adults owned a smartwatch or fitness tracker, with projections for the next years indicating an increasing trend (Pew Research Center, 2020). Most commercial wearables can continually measure physiological variables for days without recharging. Whereas basic wearables can estimate heart rate (HR) through photoplethysmography and accelerometer-based movement (actigraphy), higher-end wearables complement these measures with, among others, electrocardiography (ECG) and electrodermal activity or skin temperature signals. The

widespread personal use of these wearables has lowered the threshold for scientific data collection – in terms not only of availability but also of user experience – paving the way for large-scale monitoring of physiological and motion variables in everyday life.

Even more so than wearables, the smartphone itself has become a critical tool for the passive collection of daily-life personal data due to its wide range of applications and its global availability. An estimated 4.69 billion people worldwide own a smartphone, with percentages of >94% for populations of developed countries (Statista, 2025). Not only the many built-in sensors but also the analysis of smartphone usage make it the richest data source of information on daily living. For instance, location and movement can be tracked relatively precisely using an accelerometer, gyroscope, GPS and Bluetooth – features that are available in almost all of today's smartphones. Social interactions can be monitored through Bluetooth, microphone, app usage, calls and text messages. Screen time, typing speed and patterns and Wi-Fi connectivity are only some of the other possibilities smartphones offer to capture behavioural and environmental markers of daily living. Given the pervasiveness of smartphones, mobile sensing data is continuously and freely generated, requiring only appropriate methods to capture and analyse it. These means of data collection offer unprecedented new opportunities for daily-life research.

Passive sensing is increasingly used in combination with ESM. Although the field is growing rapidly, tangible results have only started to come out recently. Results on the largest multimodal study combining passive and active ambulatory assessments to date – the RADAR-CNS project (Matcham et al., 2022) – are now being published. Combining passive sensing data (PSD) with ESM potentially holds several important benefits. First, adding PSD could add new information to ESM measures, enriching subjective reports with biological (e.g., HR), behavioural (e.g., step count) and contextual (e.g., geolocation) variables of interest. The addition of these variables allows for the investigation of interaction between multiple dynamic streams of real-time, real-life data or the assessment of single constructs at different levels (e.g., assessing both the perceived and cardiovascular stress response).

Second, although PSD may never be able to replace an individual's personal experience, passive sensing measures can be used as a proxy for variables that would previously have been assessed through ESM. In some cases, passive sensing even offers a richer or more direct measure

of the variable of interest while requiring minimal active input from the participant: Think a GPS location to replace questions and responses on a description of an individual's current location or actigraphy-based movement instead of inquiry and self-reporting about past-hour physical activity. As such, PSD can replace ESM items by assessing phenomena that can be directly measured using passive sensing. Following a different approach, PSD could be used to improve the precision of event-based ESM, increasing the sampling of moments of interest by triggering ESM questionnaires (or potentially ecological momentary interventions [EMIs]) when pre-set, passive-sensor-based thresholds are reached. Despite the wealth of potentially added value, additional sensors and protocols also add complexity and challenges to ESM research, potentially influencing participant compliance, data quality, ethical decisions, costs and time investment. In this chapter, we will briefly describe some of the most common use cases of combined ESM and passive sensing. Specifically, we will focus on passive sensing of physiological signals, such as HR, that tell us something about an individual's biological functioning, on actigraphy as an estimation of movement and on mobile sensing using smartphone sensors. For each, we will discuss the benefits and challenges of such an approach, illustrated with some examples, and end the chapter with a description of two hypothetical cases where PSD is combined with ESM.

13.1 Applications of passive sensing

13.1.1 Passive sensing of physiological signals

While wearables can monitor a variety of biological signals such as glucose levels (Mansour et al., 2024), cortisol (Ok et al., 2024), brain (Kaongoen et al., 2023) and muscle (Yamaguchi et al., 2023) activity or temperature (Butt et al., 2022), those typically measured in combination with ESM are parameters of the Autonomic Nervous System (ANS). Key examples include HR and heart rate variability (HRV) measured via ECG or photoplethysmography, electrodermal activity measured via electrodes assessing sweat gland activity and blood pressure estimated using photoplethysmography-based methods. The activity of the two branches of the ANS – sympathetic and parasympathetic – plays a major role in the physiological stress response (Chu et al., 2024) and is closely linked to a variety

of emotional and cognitive processes (Critchley et al., 2013). Consequently, one of the promises of integrating ESM with passive sensing of ANS activity is the possibility of examining the dynamic interplay between affect, cognition and physiology in ecologically valid settings. Arguably the most common application of this type of passive sensing is in stress research, where measuring ANS activity following an everyday stressful event can provide insight into the impact of a person-specific stressor rather than a stressor artificially induced in laboratory conditions. For instance, in our recent work, we examined the dynamic interplay between cognitive and biological processes in daily life by testing the mediating role of momentary perseverative cognition (i.e., ESM items on worry and rumination) in the relationship between stress and subsequent HRV; our results indicated the withdrawal of parasympathetic activity associated with a stress response (Lachowicz et al., 2025). Similarly, a recent study using ESM and wearable devices analysed the concurrent association between ANS parameters (i.e., HRV and electrodermal activity) averaged over time windows preceding ESM prompts and personality functioning reported through ESM (Sinnaeve et al., 2024). This approach allowed for the examination of the relationship between physiological parameters associated with physiological stress response and fluctuations in personality functioning on a within-person level in individuals with borderline personality disorder.

In addition to examining complex, dynamic relations, the effects of experimental manipulation on daily-life functioning can be better investigated using a combination of wearables and ESM. One study assessed how a psychosocial stress task influences daily-life affective and physiological states throughout one week following the stress induction, allowing for the assessment of the presence of adverse effects that could potentially extend beyond the lab (De Calheiros Velozo et al., 2024). Another study employed ESM and wearable ECG to assess the effects of a biobehavioural just-in-time adaptive intervention (JITAI – see also Chapter 12) initiated in response to a momentary increase in self-reported symptoms on daily-life levels of HRV (Lachowicz et al., 2026).

Another branch of research combining active and passive ambulatory assessment methods aims to determine whether passive sensing can be used to replace ESM in an attempt to decrease participant burden. For instance, Tomitani et al. (2022) measured blood pressure using wearables while simultaneously capturing momentary perceived stress levels

through ESM and found positive associations between blood pressure and perceived stress in daily life, indicating that cardiovascular signals may have some potential to serve as a proxy for ESM reports on perceived stress. Using more complex classification models, Tutunji et al. (2023), examined whether affect measures, ANS parameters (i.e., HRV and electrodermal activity) or their combination would successfully identify prolonged stress periods and found that although both modalities could successfully classify data into stressful and control periods, their combination outperformed the single features, implying that affective and physiological signals provide most accurate information when combined (Tutunji et al., 2023). Indeed, while the idea of substituting self-reported states with passive sensing to reflect the same construct is appealing, existing evidence does not fully support this. A recent study by Siepe and colleagues warns against the use of so-called 'stress scores' provided by commercial devices as the lack of transparency and the goal of good user experience may interfere with accuracy; this study found no overlap between the 'stress score' and self-report regardless of preprocessing choices (2025). A slight improvement can be found when taking cardiovascular signals instead of composite scores into account as illustrated by the results of a systematic review showing that the expected agreement between simultaneously assessed passive and active measures of stress is about 35% (Vaessen et al., 2021).

Despite weak evidence for the agreement between the two modalities on a momentary level, recent findings show that their combination might be successfully leveraged to facilitate the detection, prevention or even improved diagnosis of mental health conditions (Brietzke et al., 2019; Bufano et al., 2023; Insel, 2017). Such applications typically involve advanced analytical methods and integration of data from multiple sources beyond wearables-based physiological data and self-report to identify the so-called digital phenotypes indicative of risk for deterioration in physical or mental health. For example, a large-scale cross-sectional study using wearables identified a phenotype expressed in a blunted physiological response to daily stressors reported via ESM, linking this profile to poor mental and physical health (Smets et al., 2018). Likewise, another study found that wearable-based electrodermal activity measures showed the highest accuracy in classifying individuals into groups of low versus high perceived stress and low versus high mental health, outperforming the predictive abilities of mobile sensing and behavioural features, including

daily-diary data (Sano et al., 2018). In addition to data-driven group classifications, some studies adopt personalized approaches aiming to build individualized prediction models that can detect person-specific states of risk. For example, Shah et al. (2021) successfully predicted participants' depressed mood from a combination of 43 features including, among others, ESM and ambulatory HR recording (Shah et al., 2021).

Finally, passive physiological data can serve as a trigger for ESM questionnaires or, eventually, JITAIs (see Chapter 12). Triggering an ESM questionnaire in response to a change in the physiological signal can help examine the biological underpinnings of experiential states. For instance, Hoemann et al. (2021) developed a trigger for ESM prompts based on changes in ANS activity to capture physiological states underlying emotional granularity. Specifically, whenever a sustained change in the interval between subsequent heartbeats (i.e., inter-beat interval) was detected in the absence of movement, participants received an ESM prompt asking them to freely label their own emotions. Such research not only enhances the understanding of the experiential context of physiological states and vice versa but also lays the foundation for increased precision in the delivery of passively triggered JITAIs. With the aim of facilitating physiology-triggered JITAIs, Rominger & Schwerdtfeger (2023) applied such an algorithm based on nonmetabolic (additional) HRV reductions to detect negative psychosocial states in daily life. Although this attempt was not successful and such an approach is in its early stages, it paves the way for highly personalized interventions that can be triggered without the active engagement of a user. Furthermore, there are several published study protocols planning to use algorithms based on electrodermal activity (Bögemann et al., 2023) or HRV (Schwerdtfeger & Ofner, 2024) to directly test the efficacy of physiology-triggered JITAIs. However, despite some preliminary promising results – for instance, Bögemann et al., (2024) successfully implemented a combination of electrodermal activity and ESM to trigger a real-time emotion-regulation intervention – the field of passively triggered JITAIs is still in the early phases of development.

Despite rapid progress, the field of passive sensing of physiological data also faces challenges related to the complexity of physiological signals. Several conceptual and measurement issues need to be addressed to allow further advancements in the field. First, the mixed findings regarding physiological correlates of experiential states highlight the need for more fundamental research to clarify what is being measured and to draw

meaningful conclusions from physiological data. Furthermore, significant methodological heterogeneity across studies using passive sensing of physiological signals hinders reliable comparisons. These differences relate to both hardware and software including, among others, the type and version of the sensors, sampling frequency, and algorithms used for preprocessing and metrics calculation (Nelson et al., 2020; Roos & Slavich, 2023). Finally, accounting for the effect of relevant external factors which have a large impact on the quality and interpretation of physiological signals such as movement (Gashi et al., 2020; Littmann, 2021), posture (Grosprêtre et al., 2021) or speech (Saygin et al., 2024) remains a challenge outside of controlled laboratory settings.

13.1.2 Actigraphy

In addition to physiological phenomena, modern wearable sensors can inform on a person's activity and sleep patterns, offering opportunities to study these behaviours in ambulatory settings. The most common sensor used to measure movement is an accelerometer, a sensor that combines changes in speed, direction and (often in conjunction with a gyroscope) body position. Accelerometers are becoming especially popular in ambulant designs as they are integrated into most modern mobile phones and activity watches. A review by Zapata-Lamana and colleagues found that of all ESM studies between 2008 and 2018 that included self-reported physical activity, half also measured accelerometery, mostly via smartphone (2020). Despite the smartphone's popularity, accelerometery derived from body-worn sensors, for example in a wristband, provides more reliable information about bodily movements than smartphone sensors, which are often placed in a pocket or bag.

Acceleration, a simple variable that is often calculated from accelerometery data, can be used to determine when and to what extent a person is moving. More complicated algorithms aim to derive from the raw data which physical activity a person was engaged in as well as its intensity. Sometimes the data is converted into a proxy for movement (intensity), such as steps taken, energy expenditure and activity points. While open-source software is available to convert the data, a lot of commercial wearables only provide this proxy data without insight into their calculations; this complicates comparisons between and within studies when participants use different brands or even different models. In the

case of body-worn sensors, acceleration measured during the night can also inform about specific aspects of sleep such as duration and quality. These measures of activity and sleep, when collected during the course of an ESM study, offer valuable information in different ways.

Notably, actigraphy is a promising method to replace or substantiate subjective reports of movement and sleep in ESM studies by providing a more objective and continuous measurement. This is especially important in populations that are difficult to study accurately because they may struggle to complete electronic surveys, such as young children. In a study investigating physical activity in children between the ages of five and seven, accelerometer-measured activity was strongly associated with ESM completed by the parents (de Brito et al., 2020). The addition of passive sensing worked to confirm children's actual physical activity. Additionally, actigraphy can inform on more detailed aspects of (in) activity that participants are not consciously aware of such as the exact moment the participant fell asleep and woke up and how much the participant moved in their sleep (Staples et al., 2017), to name a few. However, while actigraphy could potentially capture information with less burden, without retrospective bias, and at higher frequency than self-reports, there are only a few papers that have investigated the overlap between so-called subjective and objective measurements (e.g., see Siepe et al., 2025). Additionally, while actigraphy can provide information about the length and intensity/objective qualities of movement and sleep, the subjective appraisal of (in)activity is often strongly related to outcomes and cannot be captured with wearables. Due to this and lack of evidence, future studies should not replace subjective reports with information solely from actigraphy but view both sources in tandem.

As ESM is traditionally used to study variable subjective phenomena such as affect, symptoms, and appraisals, another manner in which actigraphy can add value to an ESM design is by informing on the relation of these phenomena to movement and sleep and possibly predicting them. Several studies on the relation between accelerometer-measured movement and mood have found that engaging in moderate-to-vigorous physical activity is associated with subsequent increased positive affect and decreased negative affect (Dunton et al., 2014; Li et al., 2022; Pannicke et al., 2020). Additionally, actigraphy data can be used to study the relationship between activity levels and symptom severity. For example, Li and colleagues examined whether changes in the levels of physical activity,

operationalised through acceleration, are linked to depressive symptoms reported via ESM and found a negative association (2022). In a study by Hennig and Lincoln, actigraphy recorded during the night was used to objectively estimate the duration of sleep. They found a stronger relationship between the objectively estimated duration of sleep and symptoms of paranoia the morning after compared to self-reported hours of sleep (2018). Furthermore, continuous monitoring of activity provides valuable information on changes to individual patterns of rest and activity that can help predict acute negative experiences. By estimating the duration and quality of sleep from accelerometer data, Yap and colleagues found that shorter- and worse-than-usual nights of sleep are associated with higher stress the day after (2020). Another study looked at the within-person effect of physical activity – measured through step count – on pain in older adults and found that while those with higher step counts overall experienced less intense pain, taking more steps than usual increased the risk of experiencing pain (Davis et al., 2022). While actigraphy can be used to capture sleep-related measures as well as physical activity-related measures, it can also be used to infer intraday activity, i.e., fluctuation in activity levels, which is even more predictive of the presence and intensity of pain (Sarwar et al., 2022). As the field of psychological intervention research is turning towards individualized models, these recent studies indicate that actigraphy data can provide important information as, besides its low burden and richness in information, the derived individual patterns of activity are predictive of the events that JITAIs, for example, want to detect and tackle.

13.1.3 Mobile sensing (smartphone)

While the uptake of wearables is still rising, smartphones have already been seamlessly integrated into our daily lives, effectively becoming an extension of ourselves (Harari et al., 2024). This deep integration offers a unique opportunity to enhance ESM research by providing continuous, objective and real-time data on people's behaviour and context throughout daily life, all without requiring extra effort from participants (Insel, 2018). Collecting mobile sensing data places minimal burden on participants as they only need an app to collect such data – often the same app used for ESM research. Additionally, mobile sensing data is multimodal and continuous, allowing for new types of data to be collected and certain ESM

data (e.g., current location) to be gathered at a much higher frequency. In this section, we will differentiate between two types of data that can be collected using mobile sensing: behavioural data (what someone does) and contextual data (what happens around someone) (Niemeijer & Kuppens, 2024).

Firstly, there are several smartphone-derived sensors that allow for the collection of behavioural data. As already noted in the previous section on actigraphy, accelerometers and gyroscopes track physical activity and provide information into movement and sleep patterns. Beyond these measures, smartphones also track a wide range of device activities indicative of digital behaviour. For example, mobile sensing allows researchers to accurately track how long participants spend using their phones as well as which apps they use and how long. This can be useful to find connections between, for example, smartphone addiction and well-being (Meegahapola & Gatica-Perez, 2021) or social media usage and self-harm (Wu et al., 2024). Indeed, the scope of behavioural studies now extends beyond traditional realms to include virtual behaviour as well. Online social behaviour, encompassing activities like messaging, calling and engaging with social media, is often difficult to capture through questionnaires alone. Smartphone usage data provides a more reliable and accurate depiction of these behaviours, offering a clearer understanding of participants' habits and routines (Aalbers et al., 2023; Langener, Bringmann, et al., 2024). Furthermore, offline social behaviour may be tracked through features like Bluetooth (see Cummins et al., 2023), speech analysis of ambient noise recording (Schoedel et al., 2023; Zhang et al., 2021), and GPS, enabling researchers to capture a wide range of social interactions as well as both the frequency and nature of social engagements (see Mohr et al., 2017; Bitran et al., 2024).

In addition to behavioural data, smartphones also facilitate the collection of contextual information that provides information on the environment a participant is in. Data from Bluetooth and GPS, for example, can significantly enhance ESM by providing near-continuous measurements of various aspects of the environment that ESM alone might miss unless explicitly targeted. For instance, the specifics of an individual's daily commute are often omitted from ESM despite having a direct impact on their mood. A study combining GPS with ESM found that it may not be the destination but the mode of transportation that influences mood (Glasgow et al., 2019). People generally reported a more

positive mood when engaging in active transportation such as walking or biking compared to driving or taking a bus. Similarly, a commute through green spaces and near water bodies was found to be more beneficial for mood than one through dense urban areas. Other contextual sensors on smartphones measure light, noise and even air quality and weather (using Internet services), offering additional context about surroundings that might impact psychological and physiological states.

Despite the wealth of data that mobile sensing can provide, evidence regarding its effectiveness in directly predicting latent ESM variables remains mixed. While there are some correlations between sensor data and psychological variables – such as physical activity and well-being (Lathia et al., 2017) or GPS location and depression (Sandstrom et al., 2017) – these associations are often weak (De Angel et al., 2022). For example, while Bluetooth interactions may suggest social behaviour (Harari & Gosling, 2023), this is only possible when nearby devices have Bluetooth activated, and it may be challenging to accurately identify the interacting device (e.g., mistaking a printer for another smartphone can lead to erroneous conclusions). Another challenge is processing raw mobile sensing data to match the timescale of ESM data (Langener, Stulp, et al., 2024). These and other methodological challenges raise questions about the reliability of mobile sensing as a standalone tool for replacing ESM. While mobile sensing data offers valuable context, it may not always be sufficient to accurately infer experiences on its own. At the same time, one can acknowledge the large potential of this field: even though mobile sensing will not allow us to measure everything, it allows us to measure much more than ESM alone.

Rather than using mobile sensing as a direct indicator of behaviour or context, a promising application of mobile sensing is its use alongside ESM – particularly in triggering ESM questionnaires at moments of interest. By using mobile sensing data to identify significant changes in behaviour or environment, researchers can prompt participants to provide self-reports during key moments, potentially enhancing the relevance and accuracy of the collected data. This approach is especially useful in studies involving sensitive topics where participants may be unwilling or unable to accurately report their behaviour. For instance, in the case of substance abuse, rather than relying solely on self-reports, GPS data combined with physiological data can provide further evidence to substantiate participants' reports. GPS data can verify if a participant is near a location

where substance use is likely, such as a bar, and this information can be supplemented with physiological data, providing a more complete picture of the participant's behaviour. Moreover, integrating mobile sensing with ESM can improve participant compliance by triggering questionnaires at optimal times, such as after a phone call (Fischer et al., 2011) rather than in the middle of a sports activity or when driving, when answering a questionnaire would be difficult and in some cases dangerous. This integration could not only reduce the burden on participants by minimizing the frequency of prompts but also improve the quality of data by targeting moments that are likely to be more salient, leading to more informative insights into human behaviour (Murray et al., 2023).

13.2 How to combine ESM with Passive Sensing Data

While a design that combines ESM with PSD can offer unique insights, it also poses unique practical challenges before, during, and after data collection. The next part of this chapter gives an overview of what to consider at each stage, illustrated by two practical examples.

After choosing a signal that is validated to measure the variable of interest in ambulant circumstances, the main goal before data collection is the selection of a device that captures this signal. Specific considerations when selecting a wearable or smartphone (application) can be divided into three categories: measurement-related, practical and ethical. Because PSD is a unique type of data and may require an additional device compared to a regular ESM study, there are logistical challenges that arise during the data collection phase relating to participant recruitment, briefing and study management. As PSD is different in nature and often larger in size than ESM data, handling the data is not straightforward. There are considerations at each step to prepare the data for analysis: data storage, data cleaning and, finally, combining PSD and ESM data.

Because of the many associated challenges, it is important that the use of PSD is justified. The decision to measure a concept with PSD should be deliberate and fit with the theoretical framework. It is therefore recommended to carefully consider the research questions before data collection. Below are two concrete examples of researchers – Bob and Clara – who decided to conduct studies that combine ESM and PSD as it

made sense for their research goals. For additional clarity, Table 1 gives an overview of the different aspects of PSD for each example.

Table 1. Important concepts

		Bob	Clara
Construct	What do we want to investigate?	Arousal	Being in a green environment
Variable	How do we operationalize it?	HR	Geolocation
Signal	How do we measure it?	PPG	GPS
Metric	In which units?	BPM	Coordinates
Device	With which device?	Wearable	Smartphone (application)

HR: heart rate; PPG: photoplethysmography; GPS: global positioning system; BPM: beats per minute.

Example 1: Bob

Binge-eating episodes are often preceded by an increase in negative emotions. The ability to passively detect these negative affective states as they occur would enable the prediction and potentially prevention of the occurrence of maladaptive eating behaviours in real time. Bob aims to combine ESM and physiological data collected with wearables to passively detect these moments of risk in daily life. Specifically, his objective is to retrospectively examine whether an increase in self-reported negative affect (NA) preceding binge eating episodes can be linked to increases in HR during varying time windows before the episode. The choice of HR rather than more complex physiological signals is driven by the aim of identifying a biomarker that can be accurately measured with widely available commercial wearables. The study will last three weeks and involve 100 adolescents with bulimia nervosa.

Example 2: Clara

Spending time in a green environment – forests, parks, rural areas – has been related to happiness in retrospective studies. Clara aims to uncover whether this relation occurs at the between-person level, i.e., individuals who spend more time in green environments are happier, and/or at the within-person level i.e., individuals are happier when they are in green environments, or both. To test this, she sets up a study with a semi-random ESM design that prompts 50 participants to rate their positive emotions at

that moment 10 times per day for eight weeks. To obtain an objective and continuous measure of whether and how long participants are in a green environment before or during a prompt, Clara decides to collect data on the participants' location from their smartphone to combine with the ESM data after the study.

13.2.1 Before data collection

13.2.1.1 Measurement considerations

Two device-related factors frequently mentioned in the literature (Cho et al., 2021) that should be considered when selecting a device are the quality of the signal (hardware) and the quality of the algorithm (software). The quality of the signal applies to the reliability and validity of the device to measure the signal, of which Roos and Slavich mention three key considerations in selecting a device (2023). Firstly, the device should be validated to measure the signal of interest as evidenced by published studies comparing the device against the gold standard. Secondly, the device should be equitable, meaning it has been validated in different populations and under different circumstances. For example, movement in ambulatory conditions and skin tone can affect the accuracy of HR derived from photoplethysmography (respectively Gillinov et al., 2017; Colvonen, 2021). Thirdly, the sampling frequency of the signal should be high enough to reliably capture the metric. Two additional considerations for the reliability of the signal are whether the device is worn at the optimal location of the body to capture the signal (e.g., wrist-worn vs. chest-worn) and whether the device is calibrated correctly.

The quality of the algorithm is determined by the transparency of the metric and how the variable is calculated. Except in the case in which the raw data can be accessed, the most important consideration is whether the smartphone or wearables company grants insight into how their algorithm processes the raw data, including into often summarized data. This includes information about how the data was cleaned – how artefacts, low-quality data, and missing data were handled – and how summary scores are created. While processed summary scores may be preferred for their simplicity, they limit true understanding of the data and comparability with results from other devices as well as with those from the same device when the algorithm is updated.

13.2.1.2 Practical considerations

As all devices have practical limitations, whether they be technical or user-related (Cho et al., 2021), the researcher needs to select one that fits the requirements and possibilities of the study. Devices may run out of battery or storage during the study or be uncomfortable to wear for some populations while particular applications can be difficult to use or simply not be available on some smartphones. Limitations that can be identified before the selection of the device can be taken into account in the process of choosing it; however, it is highly recommended to run a pilot study with the selected device to identify additional limitations. If it is impossible to avoid a limitation, the researcher can mitigate data loss by anticipating necessary steps. Daily reminders to charge the device or upload data can help with battery and storage problems while clear instructions can prevent using a device incorrectly. Additionally, it is important to consider whether participants can interact with their own data as feedback from the device may influence their behaviour and inadvertently affect the study outcome (Roos & Slavich, 2023). Lastly, the pricing and availability of a device will often play an important if not deciding role during the selection process.

13.2.1.3 Ethical considerations

Collecting data using mobile sensing or wearables requires careful consideration of potential threats to data security and participants' privacy due to data breaches or selling data to third-party vendors (Sui et al., 2023). Given the varying data protection regulations across countries, determining server locations and data storage durations is crucial to ensure compliance with effective laws such as the General Data Protection Regulation (GDPR, 2016) within the European Union. Likewise, the privacy requirements of a dedicated application or mobile-sensing application, such as access to search history, camera roll or survey data collected in the study, must be critically examined to avoid the potential sale of sensitive data to third parties (Papageorgiou et al., 2018). It may even be illegal to collect certain types of data – for instance, recording phone calls in countries that require two-party consent. To mitigate these threats, priority should be given to wearables and applications that minimize unnecessary data sharing and only collect the data of interest. Moreover, it is advised

to remain mindful of and address potential changes in the company's privacy requirements in consultation with a local ethics committee when necessary.

13.2.1.4 Examples before data collection

Example 1: Bob

(Measurement) Bob chooses to use a validated commercial wrist-worn wearable that employs photoplethysmography technology to record HR. The data are saved in a dedicated mobile application. The wearable he chose samples pulse rate every 10 seconds. Bob notes that the output is provided in the form of beats per minute (bpm) in one-minute time windows rather than in a raw format. Nevertheless, the manufacturer of the wearable offers a transparent explanation of how the one-minute level HR is calculated – in this case using the mean within the one-minute time window. Additionally, since the wearable uses a green LED light, Bob makes sure to collect information on participants' skin tones to properly interpret the signal at the stage of the analysis. He also ensures that the wearable has a built-in accelerometer and gyroscope to accurately control movement influences.

(Practical) Parallel to ensuring that the chosen wearable offers an accurate and transparent measurement, Bob considers the practical aspects of its use. Firstly, to minimize the involvement of participants, he opts for a wearable that allows for automatically uploading data to the cloud when it connects with a mobile device with a dedicated application. Second, given the duration of the study, he acknowledges the necessity to regularly charge the wearable throughout the study period. To prevent participants from forgetting about charging, Bob sets up a daily reminder in the ESM app prompting them to charge it overnight. Finally, keeping participants' comfort in mind, Bob chooses a wearable with an adjustable wristband to ensure it fits wrists of varying sizes.

(Ethical) After selecting his preferred wearable, Bob takes time to carefully consider the ethical issues related to its use. He needs to be particularly mindful for two reasons. First, he has chosen a commercial wearable (i.e., a fitness tracker) with a dedicated application produced by a United States-based company, which is subject to different data protection regulations when compared to the EU's GDPR policy. Second, his study involves a doubly vulnerable population – minors with an eating disorder.

Therefore, Bob carefully reviews the privacy requirements of the app and informs the participants and their parents about the data-sharing policies – both in the informed consent form and during the initial briefing.

Example 2: Clara

(Measurement) Clara reviews different signals to measure location and decides that a GPS signal is the best way to recognize whether a participant is in a green environment. Modern smartphones have a built-in GPS, so instead of an additional device, she looks into an app that uses the participants' GPS data to determine whether the participant was in a green environment at a given time. For the sampling frequency, Clara opted to collect participants' locations once every minute in order not to drain their phones' batteries too quickly.

(Practical) As Clara's study does not require a secondary device to measure location, her focus is on selecting an ESM app that minimizes the power and storage necessary while still updating regularly. Some battery drainage is to be expected, so Clara makes a note to mention this during the briefing. The selected app is available only for Android software, so she will need to exclude participants without a smartphone that runs this system. To ensure that she selects an app that is user-friendly and runs sufficiently stably in the smartphone's background, Clara tests out several apps herself and runs a small pilot study to ensure the app works on a variety of phones.

(Ethical) The app Clara selects is from a European country and is GDPR-compliant. As GPS data is inherently personal and cannot be pseudonymised, Clara has selected an app that does not give her access to the raw GPS coordinates. Instead, the app automatically queries an online geographic database – for instance, OpenStreetMap or Google Maps – to find out whether it considers the coordinates of a participant's location to be in a green area such as a forest or park. The app then saves this binary processed data; it does not store the raw GPS coordinates.

13.2.2 During data collection

13.2.2.1 Recruitment

Using passive sensing might make recruitment and data collection more challenging and time-consuming compared to traditional ESM studies.

First, compliance with additional exclusion criteria related to passive sensing measures may reduce the number of eligible participants; for instance, some participants may be excluded due to the use of cardioactive medication in the case of HR monitoring. Likewise, if the app used to communicate with the wearable or collect the sensing data was developed for only some operating systems (e.g., Android only), participants whose smartphones do not meet these technical requirements have to be excluded or study phones need to be provided (Matcham et al., 2022). In addition, in the case of using external devices, the number of available wearables inherently influences the pace of recruitment. Although there is a possibility of recruiting only participants who already own a wearable capable of measuring a given signal, this naturally leads to a systematic bias due to the inclusion of only certain individuals (Canali et al., 2022). Finally, the need to collect and return the wearable at the beginning and end of the study requires additional time either from the participant, which can diminish their interest in the study, or from the researchers, which will increase the burden of the data collection.

13.2.2.2 Briefing

Proper briefing is essential in ambulatory studies because in such studies, researchers have limited control over the data collection process (for more information, see Chapter 7). Thus, unless the study's objective is to observe participants during the naturalistic use of a wearable or an application, comprehensive instructions regarding their use and the study protocol must be provided during a briefing session. The chosen approach should be reported in publications (Nelson et al., 2020). For example, with regard to wearables, it might be necessary to explain when and how to wear or remove them, how they should be taken care of and the rules for returning them upon the study's completion. In mobile sensing studies, the briefing could include a demonstration of how to set up the application and keep it running in the background, the application's functionality, and privacy concerns, including an explanation of whether the data is shared with the researcher in real time. Providing an information sheet following the briefing might help ensure the participants can review the key information when in doubt.

13.2.2.3 Study management

Conducting a study combining ESM and passive sensing requires detailed study protocols along with contingency plans to mitigate potential issues. Data management should be implemented from the early stages of data collection – for instance, to ensure correct transfer from external to permanent storage or to verify whether the data is collected properly. Incorporating time checkpoints, e.g., reporting the first time the wearable is put on by each participant, can help prevent mix-ups between participants and synchronize timestamps between wearable/application and ESM data. Additionally, if possible, scheduling a phone call with the participants after the first few days of data collection might help detect potential problems early on.

Importantly, using wearables introduces additional considerations compared to mobile (i.e., smartphone) sensing alone, such as planning the distribution of wearables within the study timeframe and ensuring adequate backup in case of malfunction, damage, or delayed returns by participants. Contingency plans should include decisions on whether to cease the study or replace the wearable should it malfunction. In the case of replacement, the researchers should decide, for instance, whether they would be available for technical support outside of working hours or how they can be contacted. This considerably increases the burden of the study, necessitating careful planning to minimize its impact on both the study and the researcher. Study management should also include maintaining a detailed log of numbered wearables and a record of any deviations from the protocol.

13.2.2.4 Examples during data collection

Example 1: Bob

(Recruitment) When planning the recruitment strategy, Bob decides to exclude individuals who regularly take medication affecting HR as well as those with cardiovascular or respiratory diseases as these factors would obstruct the interpretation of HR data. Additionally, he includes only individuals who do not report any skin condition or allergy that could prevent them from wearing the wearable continuously due to excessive skin irritation.

(Briefing) During the briefing, he instructs the participants to put on the wearable during waking hours and remove it only for showering,

swimming or overnight charging. To ensure consistency in the protocol both within and between participants, Bob asks each participant to always wear the wearable on a non-dominant hand and to refrain from changing the tightness of the wristband throughout the study. Next, Bob describes the return procedure for the wearable and asks participants to sign a document confirming they will keep it in their possession for the study's duration and will return it afterwards. Following the briefing session, he e-mails a concise information sheet to the participants to ensure they can review the necessary information at any time.

(Study management) In line with Bob's study management plan, on the third day of the study, each participant receives a short call from the research team to ensure participants are complying with the protocol. Moreover, at the same time, the researcher logs into the application using participants' anonymized study IDs to check if the data is being collected properly. As defined beforehand, if the participant reports a problem with the wearable during the study, the research team schedules a meeting with the participant to replace it. Bob always keeps aside several wearables as backups in case of such a technical failure. Such impromptu meetings take place outside of the lab and only in urgent cases. If the malfunction of the wearable occurs towards the end of the data collection period, the regular protocol is maintained. Bob reports each such episode and keeps a log of the wearables used by the participants.

Example 2: Clara

(Recruitment) Clara only includes participants with a smartphone that runs a recent version of Android so they can install the latest version of the app. Additionally, she decided to exclude participants with a smartphone from certain brands (e.g., Huawei or Xiaomi) as they are known for their extreme battery-saving mechanisms that force apps to shut down, which would prevent the app from collecting GPS data continuously.

(Briefing) During the briefing, Clara helps the participants install the app on their smartphones. She ensures that the participants give the app only the necessary permissions. She asks the participants not to update the software of their smartphones during the study as this may inadvertently lead to missing data.

(Study management) Clara makes an individual account on the app for each participant so she can access and download the data after data

collection is completed. As the location data is collected by the ESM app, there is no issue in matching the timestamps for both.

13.2.3 After data collection

13.2.3.1 Data storage

As passive sensing data is often collected continuously over the course of the study, the volume of raw data far exceeds that of a traditional ESM study. Handling big data poses specific challenges that most researchers may not be familiar with. Firstly, storing a large amount of data might not be possible on a local drive, rather requiring (several) external drive(s) or cloud space, which often increases costs. Secondly, manipulating and analysing big data requires more computational power. While it may be possible to use the data on a local computer, running a program will take longer; in more extreme cases it is advised to use a supercomputer, which often comes with a price. Thirdly, as these large datasets cannot be accessed easily, it is essential to keep detailed records of metadata, such as information about the device or application used, its specifications, codebooks with descriptions of the variables included in the dataset, etc. Researchers should anticipate these challenges beforehand and write an extensive data management plan that can be updated throughout the study.

13.2.3.2 Data cleaning

Passive sensing data requires careful cleaning and preprocessing. First, despite being labelled 'continuous', this type of data often contains a large number of missing values. This might be the result of non-wear periods in the case of wearables (Lin et al., 2020) or mobile sensing applications being shut down by the participant or the operating system to preserve the battery (Bähr et al., 2022; Niemeijer et al., 2023), among other reasons. The researchers should address the strategy for handling and reporting missing data early on in the data management plan (for an overview of strategies for handling missing data, see Darji et al. 2023). Second, the outliers and implausible values should be carefully eliminated from the recording. For certain physiological signals, applying moving averages can smooth the data and filter out noise (Cajas et al., 2020). Furthermore,

if possible, the passive sensing signal could be compared with another related measure to assess its plausibility. An example could be to compare the number of steps walked in a day with an ESM-based question about the level of physical activity to see if these two measures (approximately) correlate. Finally, it is advised to confirm the agreement between ESM and passive sensing timestamps to avoid any accidental drifts due to varying time zones or daylight savings. All data cleaning and preprocessing steps should be reported in detail to ensure reproducibility, preferably by following guidelines such as STROBE for reporting longitudinal data (von Elm et al., 2007).

13.2.3.3 Data combination

The final step before analysing the data is the combination of ESM data and passive sensing data into a single dataset. While both types of data are collected during the same study, matching them temporally is not straightforward. Given that passive sensing data is generally collected at a higher frequency than ESM surveys – often representing a continuous stream of data rather than a limited number of 'snapshots' per day – it may be necessary to down-sample one or up-sample the other. For example, the rating of 'since the last beep, I have felt stressed' can be imputed every minute since the previous measurement (up-sampling) or the total screen time can be summed over the interval between ESM-prompts. The choice depends on the research question and preferred analysis. More information and concrete recommendations about combining ESM and passive sensing data can be found in De Calheiros Velozo and colleagues (2022).

13.2.3.4 Examples after data collection

Example 1: Bob
(Data storage) The wearable and application used in Bob's study provide output (i.e., HR) at a one-minute level. Given the duration of the study and including observations missing due to non-wear periods, each participant provides up to 30,240 minute-level observations (21 days * 24 hours * 60 minutes). Bob decides that this data can be stored in a CSV file on the university cloud and does not require a separate data storage setup. He creates a file that includes metadata with information on the variables

obtained from the wearable, an explanation of the metrics calculation, the sampling frequency, and specifications of the wearable, including the version of the algorithm and the app used in the study.

(Data cleaning) Despite the extensive briefing, Bob expects a large amount of missing data resulting from charging periods, showering, participants forgetting to wear the wearable or incorrect placement of the wearable on their wrists. He reports the missing data and ensures it is coded in a non-numeric format to avoid introducing errors in the data. He also inspects the HR data for biologically unlikely values, such as an HR of 0 bpm, which might be recorded during non-wear periods. Besides using statistical methods, he performs visual checks to detect suspicious values. In addition, he examines the study log to identify participants who were included in the study during daylight savings changes to verify the timestamps were correctly recorded. Finally, considering that HR should increase during physical activity, he examines the correlation between HR and motion parameters to confirm the presence of the expected positive correlations and therefore the data's plausibility.

(Data combination) In the next step, Bob combines the two types of data: the continuous one-minute level physiological data and self-report ESM data sampled 10 times a day. Given his objective of investigating whether an increase in negative affect preceding binge eating episodes can be linked to increases in HR (HR), he needs to down-sample the collected continuous HR data to match each ESM questionnaire. Due to the exploratory nature of his research question, he chooses to average mean HR data over three different time windows of 5, 30 and 60 minutes to identify which correlates best with self-reported negative affect. He decides to use the time windows preceding the ESM questionnaire rather than those following it to avoid potential effects of reactivity to the ESM questionnaire.

Example 2: Clara

(Data storage) When data collection is completed for a participant, Clara logs into their account to download the data. In addition to a dataset with the ESM data, Clara receives a dataset with the binary 'green environment' data per minute, which amounts to almost 100.00 observations per participant (60 minutes * 24 hours * 7 days * 8 weeks = 80,640 data points). Clara saves the output for each participant separately on a hard drive. She creates a file with information for each participant that includes

information such as when they started and ended the study and what software version and device they were using.

(Data cleaning) Clara expects some missing data to software glitches, a spotty GPS signal and loss of signal when the phone was turned off or ran out of battery or when the app was forced to shut down. She makes a note of participants with a notably higher percentage of missing data to check whether this has to do with the version of the application or smartphone software.

(Data combination) Given that Clara is interested in whether a participant was in a green environment, she down-samples the binary minute-level location data to different levels: the percentage of the total time a participant spent in green environments during the course of the study for the intrapersonal research question and the percentage of time spent in green environments since the last prompt for the interpersonal research question. The percentage of time is taken beep to control for the semi-random sampling scheme that can cause the intervals between prompts to be different lengths as well as for the possibility of missing data, which can influence the total amount of data available. Lastly, Clara also calculates whether a participant was in a green environment at any time during the answering of the prompt to control for the immediate effects of being in a green area.

13.3 Conclusion

Although passive sensing holds great promise for the future of daily life research, combining ESM with passive sensing methods comes with both conceptual and practical challenges. In this chapter, we illustrated how studies have used passive sensing in combination with ESM for different purposes. Combining these two methods allows us to look into the interplay between different, highly dynamic processes in unprecedented temporal detail and with increasing reliability, all within the natural context of daily life. Increasingly, researchers are looking at ways to replace ESM with passive sensing, but although there are many cases where passive sensing may offer a good – or even better – alternative option to ESM, self-report will remain the most valid method to assess conscious experience. A better option may be to use passive sensing proxies of specific conscious experiences or predictors thereof to trigger ESM questionnaires assessing

the construct of interest through self-reports. Although such designs seem very promising, literature on their application is scarce. Similarly, passive sensing could be used to trigger EMIs or even their more flexible sibling, JITAIs, to deliver treatment to the precise moment in daily life where it is most helpful. Again, however, this idea remains mostly a future promise to date.

Future developments will undoubtedly assist progress in the field of combined ESM and passive sensing research. Exponential progress in hardware and software as well as new ideas and insights from research will facilitate the accuracy of existing signals, user experience and usability of passive sensing devices, analysis of the signals and the combination of data streams, and general methodology. Furthermore, this progress will allow us to assess constructs in daily life that have to date mainly been investigated under lab conditions; an example is the recent developments in continuous ambulatory assessment of cortisol (Hogenelst et al., 2019; Kusov et al., 2023). They will provide us with novel opportunities to capture highly subtle contextual information; an example is the use of complex speech analysis (Wadle et al., 2024). ESM itself is not necessarily bound to visual forms of assessment and written text but can also be audio- and speech-based, meaning that recording speech offers new opportunities to ESM methodology, increasing user experience and providing far richer datasets when combined with speech analysis.

Regardless of these developments and the endless possibilities they bring, researchers need to start from a solid theoretical framework and a clear notion of what they want to investigate and how they can achieve it. Developers of consumer wearables claim that their devices validly and reliably assess a multitude of measures, but they often do so without a solid evidence base. One common example of this is the claim of measuring HRV continuously using a photoplethysmography signal on smartwatches; this measure is generally unreliable due to the sensor's sensitivity to noise brought about by the movements involved in daily living. Therefore, a thorough understanding of the possibilities and limitations of passive sensing devices and their capacity to assess the variable of interest is important. In the case of wearables, the Stress in Action project has developed a database to inform researchers of the different wearables that are available and help them choose the right one for their study (Schoenmakers et al., 2025).

Finally, researchers should be aware of the additional challenges that combined ESM and passive sensing pose, some of which we described in this chapter. Allergic skin reactions to wearables, smartphone battery drainage and additional protocols on device handling are only a few examples of how these designs increase the burden for participants. The researchers have to deal with other issues such as possibly very complex ethical problems associated with constant monitoring and uploading of personal data (e.g., location, physiological data), handling and storage of large quantities of data and complex data preprocessing and analysis, and expensive and complicated study designs that require more time investment and resources. Thorough preparation and familiarization with the conceptual and practical challenges and possibilities that come with passive sensing research is therefore key.

CHAPTER 14

Dyadic Experience Sampling Research

Otto Versyp, Liesse Frérart and Martine Verhees

ESM aims to capture individuals' experiences, e.g., behaviours, emotions, thoughts, in their daily lives. Often, these experiences shape and are shaped by social interactions and relationships (e.g., Boiger & Mesquita, 2012; Van Lange & Balliet, 2015). Therefore, it can be relevant to assess such experiences in an explicitly interpersonal framework, i.e., to obtain ESM data from multiple mutually interacting people to map their emotions, behaviours and perceptions onto one another. Taking an interpersonal approach in ESM studies by including multiple interacting individuals allows a range of research questions to be explored; for example, how do emotions covary within romantic partners? (Sels, Cabrieto, et al., 2020); how do adolescents' and their parents' perceptions of parenting diverge? (Janssen et al., 2021); how is support reciprocated between co-workers (Zeijen et al., 2020)?

Besides the opportunities that such interpersonally oriented ESM studies provide, they also come with several particular considerations and challenges beyond those of individual ESM studies. In the current chapter, we discuss particularities of interpersonal ESM studies that should be considered before (e.g., design-related considerations), during (e.g., (de)briefing) and after (e.g., data structure and analysis) the study.

Note that most of what is discussed in previous chapters is also highly relevant for interpersonal ESM studies (e.g., questionnaire construction, study design considerations). We therefore strongly advise consulting previous chapters as well when setting up your interpersonal ESM study. Finally, although many of the topics discussed and recommendations provided in this chapter may apply to different types of ESM study designs that include multiple people within a relationship, this chapter's main focus is on studies that use a standard dyadic design – that is, studies in which each participant forms a dyad with only one other participant in the sample. Other designs (e.g., one-to-many design, social relations model design; Kenny et al., 2006) are possible though less frequently used to date

in ESM studies (but see Veenman et al., 2024; Verhees, Bodner, et al., 2024 for ESM research in triads).

14.1 Considerations before running the study

In the following sections, we discuss some necessary considerations when setting up your dyadic ESM study. Specifically, we will supplement Chapters 3 (*Designing an Experience Sampling study*), 4 (*Questionnaire Design and Evaluation*) and 5 (*Ethical Issues in Experience Sampling Method Research*) of this book with dyadic research by covering the foundational aspects of dyadic ESM study and questionnaire design as well as ethical issues such as privacy, burden and reactivity.

14.1.1 Sampling scheme

There are multiple sampling methods available for use in dyadic ESM studies; these methods can be used separately or in combination to suit the study's objectives. Firstly, if the goal is to relate a momentary variable of one member to that of another member of the dyad, you can opt for synchronized sampling. This method involves sampling both members at the same moments in time, ensuring that data points from both members are temporally aligned. The sampling scheme can follow a fixed, random or semi-random pattern, as discussed in Chapter 3. The specific sampling times can be strategically chosen to either increase or decrease the likelihood that members of a dyad are together. For example, to study emotional interdependence in adult romantic couples, Sels, Cabrieto, et al. (2020) sampled outside of Belgian office hours – between 5 p.m. and 10 p.m. during weekdays and between 10 a.m. and 10 p.m. on weekends – to increase the chance that the partners would be together. Note that if an identical sampling scheme (often assuming traditional office hours) is applied to *all* participants in a study, you are likely to encounter issues for those with non-traditional working hours such as night shifts or irregular hours. Similarly, if you recruit child-adult dyads from families where adults are co-parenting after separation or if you recruit couples who are not living together, you might often sample at moments where participants are not together or not interacting. If you want to optimize sampling to capture the moments where people are

together, you should consider how to resolve this (see the bottom of this section).

Secondly, if the focus is on capturing interactive dynamics or immediate responses to specific events, you can choose to let a specific response from one member trigger a beep for the other member. For instance, if one member reports feeling stressed, this could trigger a beep asking the other member about their mood or behaviour at that moment. This approach helps capture additional information at moments of interest (in this example, moments of stress), which can provide valuable insights into dyadic processes, and is implemented in some ESM data collection platforms (e.g., m-Path).

Thirdly, a beep can also be triggered when members of a dyad physically come close to each other. This can, for example, be achieved using proximity detection via Bluetooth. For instance, in a study by Janssen et al. (2024), a questionnaire asking about social interactions was triggered when the Bluetooth signal indicated that parents and adolescents had been in close proximity for more than 10 minutes and then moved apart.

In all of the above, it is important to balance the allowed response delay (i.e., how quickly the ESM beep expires after it is sent to the participants) to avoid using either too short or too long an allowance (if it is important for the study that dyad members report more or less simultaneously, at least). On the one hand, compliance might be affected when the allowed response delay is too short because participants might not have enough time to answer. On the other hand, longer response delays can reduce data reliability due to self-selection and recall biases (see Chapter 3; Eisele, Vachon, et al., 2021). Specifically for dyadic ESM studies, a long response delay can cause temporal misalignment between the answers of different members, meaning their responses may not correspond to the same event or moment. Therefore, to maintain the integrity and accuracy of the data, researchers should carefully consider which response window would be appropriate depending on their specific research question. If your research question requires that responses from both participants are temporally close (e.g., such as when examining momentary covariance in mood), minimizing response delays might be more important.

Moreover, dyadic ESM studies have a greater risk of missing data because whether a data point is useful for dyadic analyses/research questions depends on the responses of both members. In other words, the failure of one member to respond results in a missing data point for the

dyad. This risk necessitates designing the study with sufficient data points to mitigate the impact of missing data. Conducting power analyses can help determine an appropriate sample size. For instance, using tools like the PowerLAPIM application created by Lafit and colleagues (2022) can aid in estimating the required sample size to achieve adequate statistical power despite potential missing data (see also Chapter 11 about sample size selection).

Additionally, there are other ways to avoid structurally missing data. One option is to include specific inclusion criteria. These might include that participants must live together or that they have to frequently be together in person during weekends and evenings (i.e., at the moments you are sampling). This might, however, exclude specific groups of participants with non-traditional working hours (linked to SES; Winkler et al., 2018) or relationships (long-distance, etc.). You could therefore choose to include the participants anyway and be flexible with the sampling scheme to the extent possible or include them using the sampling scheme as is. For example, if someone works during the weekends but has days off during the week, you could opt to shift the sampling scheme for this dyad by a few days. Similarly, you could shift the sampling scheme by a few hours if one or both members of a dyad works at night. Note that you always must do this for both dyad members if they must receive beeps at the same time. If the (working) schedules for different members of a dyad differ too much, the latter is not an appropriate solution, and sticking to your initial sampling scheme might be best. You should not make any changes beyond shifting your scheme by a couple of hours/days (such as adding or removing beeps) to ensure that the scheme does not differ too much between dyads. Generally, it is important to consider beforehand that you might encounter these situations and decide how you will handle them so that you remain consistent across dyads (i.e., use inclusion or exclusion criteria and clearly communicate these to participants; offer flexibility to some degree; include these participants but accept that you will have lower compliance).

14.1.2 Questionnaire design

In addition to the principles of item construction discussed in Chapter 4, several specific types of items can be of relevance to use in dyadic designs. First, there are items focusing on the individual's own experiences, such

as 'How happy do you feel at this moment?' or variables concerning the dynamics of the dyad, such as 'How satisfied are you with your family interactions since the last beep?' In dyadic designs, both members often report on the same set of items, allowing you to examine possible covariation between dyad members' variables. Dyadic designs also provide the possibility to assess members' perceptions of each other. For instance, the item 'How happy do you feel at this moment?' can be rephrased to capture an individual's perception of their dyadic partner: 'How happy do you think your partner feels at this moment?' These perceptions can be compared to their self-reports to, for instance, examine interesting research questions related to the accuracy and possible biases in how dyad partners perceive one another (e.g., Verhees, Ceulemans, et al., 2024). Note, however, that these types of questions may increase reactivity (see 14.3.3).

Next, some items ask to report on significant positive or negative events either inside or outside the dyad. For example, if romantic partners are the study subjects, a question concerning events inside the dyad would be 'Did something nice happen between you and your romantic partner since the last beep?', and an example of an event outside the dyad would be 'Did you experience a conflict with your co-worker since the last beep?' Especially in case of events inside the dyad, it might be important to ask for more details about the specific event to ensure both members are reporting on the same event. Another approach to achieve this consistency is to use the response of one member to trigger a prompt for the other member, as previously discussed. This process may also involve sharing one member's response with the other to facilitate consistent reporting.

Lastly, since partner effects (the correlation between a variable of one member and the same of the other member) are often assumed to be larger when members are together (Frérart et al., 2024; Rauers et al., 2013; Sels et al., 2017), contact items are essential for many dyadic research questions. These items assess whether members are currently together, whether they have been in contact since the last beep (e.g., whether they have seen each other, texted and/or spoken over the phone) or whether individuals outside the dyad are present (e.g., family, friends, co-workers, strangers).

Chapter 4.2. discussed several matters that are important when constructing a questionnaire. Of these, particular emphasis must be placed

on the randomization of items and the implementation of branching questions within the context of dyadic designs. A risk inherent to dyadic ESM studies is the potential for participants to discuss their answers while completing the questionnaire since the two members often fill in the beeps at the same time. Such discussions might increase reactivity and can compromise the integrity of the data by introducing biases and reducing the independence of responses, as will be elaborated upon below (see 14.3.3). To mitigate this risk, one strategy is to randomize the order in which items are presented. By using this strategy, the likelihood that both members of the dyad encounter the same item at the same time is significantly reduced, thereby decreasing the chances of shared discussions influencing their answers.

Additionally, the use of branching questions is often necessary in dyadic designs because items are often dependent on the presence of the other member or on one member having had contact with the other member. For example, a question about whether a conflict occurred is irrelevant if the participants reported that they had no contact since the last beep. Branching logic ensures that participants only answer questions relevant to their specific context. When using branching logic (e.g., presenting different sets of questions to participants based on whether there was contact with a partner), it is important to ensure that each branch contains a similar number of questions. This approach minimizes the risk of participants deliberately avoiding the branch with a greater number of questions (see 4.2.2). Piloting the study is an essential step for identifying and addressing potential logical errors and ensuring the questionnaire functions as intended.

14.1.3 Ethical issues

As discussed previously (see Chapter 5), ESM offers plenty of possibilities but also introduces new ethical considerations and responsibilities. This is *particularly* true when sampling multiple individuals within a dyad. Note that all previously mentioned ethical principles for experience sampling studies remain relevant with dyad sampling but will not be repeated here.

14.1.3.1 Privacy and consent

An important ethical consideration concerning dyadic ESM is the privacy of each *individual* participant. While the sampling unit might be larger

in these studies (i.e., sampling a dyad as opposed to an individual), each individual participant still needs to be considered as such. You should therefore be careful when relating or sharing data of individuals among the dyad in any way, and also when obtaining informed consents or communicating results and debriefing.

As discussed in the Sampling scheme section of this chapter (see section 14.1), recently developed methods allow researchers to trigger beeps to one member in a manner that is conditional on the responses of another member of the dyad (e.g., elevated levels of stress reported by one person could trigger a beep with related questions to the other). However, the potential consequences of such questions and of any other form of data sharing with other participants must be carefully considered. Unless participants are explicitly informed, (indirectly) revealing the answers of one individual to the other member of the dyad can breach confidentiality and lead to unwanted consequences. To mitigate these risks, you should ensure that participants are fully informed about the (potential) sharing of their responses in any way. Overall, additional caution and attention to privacy are advised when conducting dyadic ESM studies and especially when (indirectly) sharing data between different members of a dyad.

Furthermore, informed consents also require additional attention in dyadic experience sampling studies. ESM studies are especially burdensome, requiring significant time investment and effort from participants. As a result, the monetary reward or other extrinsic incentive for participating in these studies can be relatively high, leading to an increased risk of coercion (Gabbidon et al., 2022), particularly in situations potentially involving power imbalances (e.g., in parent-child, employer-employee or some romantic relationships). Clear informed consent must thus be obtained from *each individual participant*. Wittenborn et al. (2013) suggest asking individual participants (particularly children) whether they feel that they 'have to' take part in the study to 'keep the peace' or 'make someone happy'. Beyond the ethical requirement for voluntary informed consent, a lack of participant motivation can result in low compliance and/or low-quality data. Ideally, you would therefore contact all participants on individual channels with information and informed consent, obtain everyone's individual consent (preferably in person) and verify that participation is voluntary and that everyone understands the study's demands and implications before allowing participation. Alternatively,

you could obtain the informed consents from each individual participant separately during a (dyadic) briefing session.

14.1.3.2 Burden

We are not aware of empirical studies testing whether burden is different in dyadic ESM studies compared to individual ESM studies. Burden may be greater in dyadic ESM. The shorter response delay (to ensure that responses from dyad members are temporally close; see 14.1) allowed by dyadic studies may increase burden for participants as they must respond faster to beeps. Furthermore, while individual ESM studies also disrupt the flow of daily life, this disruption can be more pronounced in dyadic ESM studies as sampling all parts of a dyad simultaneously can bring (common) activities to a complete halt. This can be particularly challenging in contexts where one may try to avoid extensive smartphone use (families, relationships, work or school environments) as remarked by a parent when asked what they disliked about their participation in a triadic (family) ESM study (Verhees, Bodner, et al., 2024): '*We try not to let our smartphone take up too much time in the family, and this study required the opposite*'. On the other hand, participating as a dyad can sometimes be considered an advantage in this regard as it eliminates the need for one member of the dyad to explain one's actions to the other member of the dyad when having to pause interactions. Regardless, the interruption of the flow of daily life also invites participants to discuss the questions and answers among each other, potentially introducing reactivity.

14.1.3.3 Reactivity

The mere act of sampling both members of a dyad is likely to have an impact on that dyad and how they respond to ESM questions. For example, you might be interested to know how supportive, disinterested or responsive people perceive their partners to be. When asking about these partner perceptions at a certain moment in time, participants are likely to become aware that their partner is currently receiving the same set of questions. As a result, participants might put *more* effort into being supportive, interested or responsive. Timmons and colleagues (2017) asked participants in a dyadic ESM study, '*How much did filling out the hourly phone surveys change the way you interacted with your romantic partner?*' Reported changes in

behaviour seemed minor in size (less than 10% of participants responded that their behaviour changed 'a lot' or 'extremely'), but most participants (69%) did indicate that they changed the way they interacted at least a little. While this could be beneficial for their relationship, it is likely to result in biased assessments, potentially obscuring the 'actual' relational dynamics. Similarly, earlier research on families suggests that ESM may help participants to reflect on routines and become more aware of certain interpersonal orientations or behaviours (Keijsers et al., 2022). This is something which also has been noted by participants in a previous triadic ESM study (Verhees, Bodner, et al., 2024) as a positive benefit of participation: *'[We had] more attention for each other and the family. It's nice to do something "together" and even talk about this topic'.* The participant indicated that the study and items led to increased familial attention to each other and the topic of the study, potentially altering relational dynamics during the study period. Additionally, the participant noted talking about the study topics, indicating another form of unintended reactivity. Such discussions could influence participants' behaviours regarding these items and potentially enhance their awareness of the other's states, thoughts and feelings. This is especially problematic when you are interested in how participants perceive each other and the accuracy of these perceptions (e.g., Sels, Ruan, et al., 2020; Verhees, Ceulemans, et al., 2024). If, for example, someone explicitly asked their partner how they are feeling while answering a beep, the data would be invalidated (unbeknownst to you) since you are not assessing the perception of the partner anymore. Moreover, drawing the attention of individuals to certain topics/constructs could possibly result in conflicts, specifically in (sensitive) interpersonal contexts such as between romantic partners, parents, employees and employers.

To minimize some forms of reactivity, you should explicitly and clearly brief participants on *not* discussing the experience sampling items or their answers for the duration of the study. While this might be hard, and participants might end up doing so to some degree anyway, it is important to stress this during the briefing session.

14.2 Considerations during the study

Once your dyadic study is set up, you will begin engaging with participants. This involves briefing and introducing them to dyadic experience

sampling research and debriefing them at the end of their participation. In the following section, we build upon Chapter 7 with some additional (dyadic) considerations.

14.2.1 Briefing and debriefing

Like individual ESM studies, it is preferred to brief each participant individually, ideally at a point in time close before the start of the ESM period. As mentioned before, when involving multiple individuals from the same dyad, obtaining informed consent separately from each participant is helpful to minimise risk of coercion. Additionally, in dyadic ESM studies, different participants might require different instructions and may possibly be presented with different items. For instance, in studies involving children and parents, employers and employees or depressed and non-depressed romantic partners, the way the study is explained, and which items are presented may vary among different members of the dyad.

While individual briefings are generally preferred and advised, dyad-level briefings are also possible. This approach has certain disadvantages, such as the complexity of obtaining informed consent in a dyadic setting or participants possibly feeling unable to share concerns or ask specific questions. However, dyad-level briefings also offer some specific benefits. In both individual *and* dyad-level briefings, it is important to assess whether participants will be interacting frequently throughout the upcoming study period, especially if this is relevant to your research questions. An advantage of dyad-level briefings is the convenience of immediately selecting a new starting date if any participants have upcoming separate holidays, extended school field trips, festivals or work-related trips. If such scheduling conflicts occur during individual briefings, it may be more difficult to immediately plan a new starting date that works for both dyad members. In any case, it is preferable that the new start date is not too long after the initial briefing session.

During the briefing session, you should also clarify specific aspects of your study to participants. As mentioned above, for most research questions, it is likely essential that the participants are instructed to *not* discuss the experience sampling items for the duration of the study and definitely not *while* answering the beeps. Furthermore, to avoid incongruent reports across members of a dyad, it is very important to clearly explain

how participants should interpret your seemingly standard ESM items. Research has shown that couples often do not agree when reporting on various relationship events via questionnaires (Christensen & Nies, 1980). For instance, individuals might have different interpretations of what constitutes a conflict, leading to instances where one person reports having had a conflict while the other does not. While this variability might not always be problematic if the construct is subjective to some degree, such differences can also arise when participants report on seemingly objective events such as whether they have had contact or are together with their partner. Take, for example, the item: *'Are you together with your partner at the moment?'* In this case, you should specify how participants should interpret this question – does 'yes' mean they are in the same room or space or simply in the same house? Discrepancies may stem from forgetfulness or (minimal) differences in the timing of beep responses (this depends on dyadic inter-response time; see 14.6 below), but it is best to minimize any differences arising from *misinterpretation*. Therefore, it is essential to provide clear instructions on how participants should interpret and answer your questions to avoid confusion or incongruencies for 'objective' events.

Moreover, when studying dyads, the focus oftentimes is on their relationship, parenting, work or family dynamics. This can result in reluctance to participate in studies due to concerns about privacy or fear of judgment or even the fear that participating in the study itself might fundamentally change something within that specific relationship. For example, in studies with romantic partners, we have experienced reluctance among couples due to their fear of changing their romantic relationship by 'putting it under the microscope'. In such cases, it might help to emphasize confidentiality and the participants' anonymity. Additionally, you could further inform participants about the sort of items/questions that you will be using as this may reassure them that the study is unlikely to fundamentally change anything about their relationship.

Lastly, debriefing sessions are also ideally done individually and at a moment of choice by the participant. Especially if you plan on providing individual feedback, reviewing time series of the collected data or sharing personal information, debriefing should not happen with other participants present. Alternatively, verbal debriefing can be done at the dyad level (i.e., sharing general study results with participants, asking participants about challenges or issues with their participation), with individual written feedback provided to each participant. In this case,

written feedback should be sent directly to the individual participants, ensuring no data from others is revealed. Note that it is not preferred to give any written feedback without additional verbal explanation. Furthermore, it is important to consider the appropriateness of providing written feedback at all; in some cases, verbal feedback or no feedback at all may be more suitable, particularly in situations where the data is sensitive and/or where power differences could pressure participants to disclose their feedback or debriefing information to others.

14.3 Considerations after the study

As discussed in Chapters 8 through 10, individual ESM studies require thinking about data structuring and data non-independence (as multiple assessments are nested within individuals) in analyses. Such considerations are even more numerous in dyadic ESM studies. Below we discuss dyadic ESM datasets in terms of types of variables and data structuring as well as matters relating to checking, preparing and analysing dyadic ESM data.

14.3.1 Dataset

Consider a study in which 50 mixed-gender romantic couples (i.e., 100 individuals) were assessed six times per day for one week (seven days). At each beep, they were asked, among others, how happy they felt at that moment (on a scale ranging from 1 = not at all happy to 7 = very happy) and whether they had had a conflict with their partner since the previous beep (0 = no conflict, 1 = conflict). A dataset containing the data from such a study includes different types of variables. Next to the subject identifier (which we also need in individual ESM studies), in a dyadic ESM dataset, a dyad identifier needs to be added, as is done in the first column of the example dataset in Table 14.1. The dyad ID is a non-time-varying dyad-level or between-dyad variable, i.e., it varies between dyads but not within dyads or across time points. Here we assume that beep numbers align for the two dyad partners; thus, beep number is a time-varying dyad-level variable, i.e., a variable that is the same for both dyad members but varies across time points. Other time-varying and non-time-varying dyad-level variables may be included; for example, relationship length at the start of the ESM study could be an example of a non-time-varying dyad-level

variable. Additionally, the dataset includes person-level variables that can differ between dyad members. For example, gender is a non-time-varying variable that, in our dataset example, will differ within dyads.[1] The example dataset in Table 14.1. also contains the variable 'Partner', which assigns the partners the number 1 or 2. In our example, partner 1 is always the female partner and partner 2 is the male partner. An example of a person-level time-varying variable is the extent to which an individual reports feeling happy at each ESM beep.[2]

Typically, three ways of structuring dyadic data from studies using a standard dyadic design are used (Kenny et al., 2006). Which data structure is preferred depends in part on the statistical analyses you plan to use (Iida et al., 2023; Kenny et al., 2006). First, in an individual dataset, each row contains a sampling moment for one individual.[3] In the case of our data example, this means that the dataset contains 4,200 rows of data. See Table 14.1. for an illustration of an individual dataset. Note that when using an individual dataset, a researcher cannot directly model the influence that partner variables have on an individual's outcome. Second, in a dyadic dataset, each row contains a sampling moment for a dyad. Thus, following the example, we would have 2,100 rows of data. The dyadic dataset contains more columns than the individual dataset, as all assessed person-level variables span two columns: one for each dyad member, e.g., $Gender_1$ for partner 1 (in our data example, the female partner) and $Gender_2$ for partner 2 (the male partner). See Table 14.2. for an example of a dyadic dataset. Third, we can have a pairwise dataset, which is a combination of the individual and dyadic datasets. It has a separate row for each individual and time point, and each row also contains the data for the individual's dyadic partner at that time point. This means all data are repeated twice, and we have as many rows as in the individual dataset and as many columns as in the dyadic dataset. See Table 14.3. for an example of a pairwise dataset.

[1] Variables that differ within the dyad but when averaged across partners all dyads have the same score – such as gender in mixed-gender couples – are also referred to as within-dyad variables (Kenny et al., 2006). A non-time-varying within-dyad variable such as gender can also be used to distinguish between the two dyad members and systematically test in analyses whether the roles of the members are equal or distinguishable (see also 14.7 below).

[2] Note that variables that differ both within- and between dyads are also referred to as mixed variables (Kenny et al., 2006).

[3] Note that for all dataset examples, we structure the ESM measurements using a 'long format' in which each row corresponds to a particular sampling moment for a particular individual or dyad, as recommended in Chapter 8.

Table 14.1 Example of an individual dataset

Dyad ID	Beep number	Subject ID	Partner	Gender	Happy	Conflict	...
1	1	1	1	female	6	0	
1	2	1	1	female	5	1	
....	....	....		....	....	....	
1	42	1	1	female	5	0	
1	1	2	2	male	5	0	
1	2	2	2	male	7	0	
....	....	....		....	....	....	
2	1	3	1	female	2	1	
2	2	3	1	female	7	0	
...							

Table 14.2 Example of a dyadic dataset

DID	Beep nr	SID_1	P_1	G_1	H_1	C_1	...	SID_2	P_2	G_2	H_2	C_2	...
1	1	1	1	female	6	0	...	2	2	male	5	0	
1	2	1	1	female	5	1	...	2	2	male	7	0	
....	....	....		....	...	....	...	....	....	....	....	....	
1	42	1	1	female	5	0	...	2	2	male	2	0	
2	1	3	1	female	2	1	...	4	2	male	2	1	

DID	Beep nr	SID_1	P_1	G_1	H_1	C_1	...	SID_2	P_2	G_2	H_2	C_2	...
2	2	3	1	female	7	0	...	4	2	male	5	0	
...													

Note: DID = Dyad ID, Beep nr = Beep number, SID = Subject ID, P = Partner, G = Gender, H = Happy, C = Conflict

Table 14.3. Example of a pairwise dataset

DID	Beep nr	SID_1	P_1	G_1	H_1	C_1	...	SID_2	P_2	G_2	H_2	C_2	...
1	1	1	1	female	6	0	...	2	2	male	5	0	
1	2	1	1	female	5	1	...	2	2	male	7	0	
....	....	....	....	....	....	....	...	....	....	....	....	....	
1	42	1	1	female	5	0	...	2	2	male	2	0	
1	1	2	2	male	5	0	...	1	1	female	6	0	
1	2	2	2	male	7	0	...	1	1	female	5	1	
....	....	....	....	....	....	....	...	....	....	....	....	....	
2	1	3	1	female	2	1	...	4	2	male	2	1	
2	2	3	1	female	7	0	...	4	2	male	5	0	
...													

Note: DID = Dyad ID, Beep nr = Beep number, SID = Subject ID, P = Partner, G = Gender, H = Happy, C = Conflict

14.3.2 Data checks and preparation

Readers interested in step-by-step instructions for checking and preparing ESM data in general are referred to Chapter 8 of this book and the book website (real-leuven.be) for related R code. Additionally, Revol, Carlier, et al. (2024) developed a step-by-step framework for preprocessing ESM data, an R package and a tutorial website containing R code (preprocess.esmtools.com), which also includes some steps specific to dyads. Similar to subject identifier variables in individual datasets, identification variables at the dyad level (e.g., DyadID in Table 14.1.) should not contain any missing data and should be consistent within each dyad and different between dyads. Often dyadic ESM studies aim to map experiences of dyad members onto one another, which means it is important that they report on the same time points. Therefore, in checking and preparing dyadic ESM datasets, dyadic compliance and dyadic inter-response time are of relevance. Dyadic compliance is the percentage of beeps that were answered by both dyad members at the same time point and gives an indication of how compliant the dyad as a whole is (and thus how much data is available at the dyad level). Dyadic inter-response time pertains to the time in between responses of dyad members to the same beep and should preferably be short (how long inter-response time can be depends on the maximum allowed response delay within the study, i.e., how quickly the ESM beep expires after it is sent to the participants; see 14.1).

Data preparation in dyadic datasets may concern adding, subtracting or averaging across the two dyad members' answers. For example, we may want to know whether both partners agreed on having a conflict or not. To this aim, we could sum both dyad members' responses to the 'conflict' variable. As conflict is coded '0' for 'no conflict' and '1' for 'conflict', the score '1' on the summed conflict variable would reflect that partners disagreed on having had a conflict while a score of '0' indicates they agreed on not having had a conflict and a score of '2' that they agreed on having had a conflict. When we look at the dataset structures examples above (Tables 14.1–Table 14.3), it becomes clear that we need a dyadic or pairwise dataset rather than an individual dataset to make such calculations.

Concerning data visualization, we can plot the time series data of both dyad members in one plot to allow visual exploration of the two dyad members' responses over time. In our example dataset, we could look at how both partners' happiness evolves over the course of the study.

Figure 14.1 shows the hypothetical happiness (on the y-axis) of the partners of two dyads (dyad 10 and dyad 12) over 15 beeps (on the x-axis). We can see that for both these dyads, partner 1 (the female partner) seems to have a higher happiness score overall than partner 2 (the male partner). Additionally, for dyad 12, the two partners' happiness scores are more strongly correlated than for dyad 10.

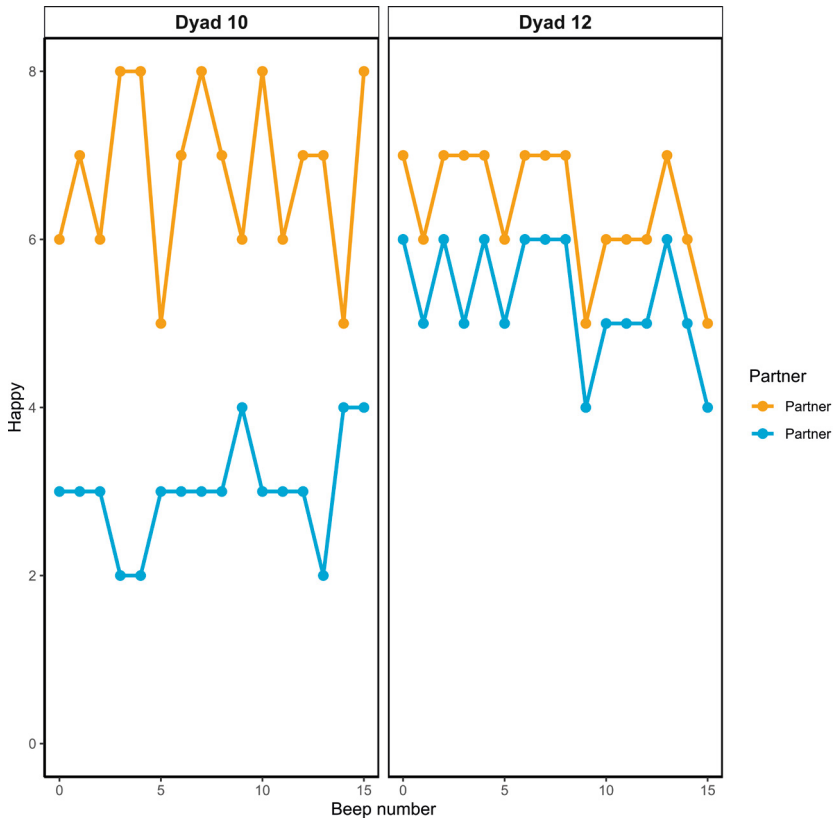


Figure 14.1 Happiness over time for two dyads

14.3.3 Analyses

ESM data obtained from individuals who form a dyad contain various sources of non-independence which must be accounted for in analyses. First, just like in individual ESM data, there is non-independence across

time (beeps are nested within individuals). Additionally, there is non-independence between dyad members, who can mutually influence each other, may experience similar events or can just be more similar in general as people tend to surround themselves with similar others (e.g., McPherson et al., 2001). This means that dyad members are often more similar to each other than randomly chosen people from the population. Thus, we can consider the structure of dyadic ESM data as follows: beeps are nested within individuals who are nested within dyads. Importantly, if dyad members are measured at the same time points, time and individuals are crossed, not nested: at each time point the level of time is the same for the relationship members. Therefore, in multilevel modelling of dyadic ESM data, data are typically considered to be two-level (e.g., Iida et al., 2023; Laurenceau & Bolger, 2012).

An important consideration in dyadic data analysis is whether dyadic partners can and should be distinguished from one another based on a certain variable. For example, mixed-gender couples are conceptually distinguishable based on gender while this variable cannot be used to distinguish partners from same-gender couples. It is relevant to not only consider distinguishability conceptually (i.e., is there a distinguishing variable?) but also investigate empirically whether dyad members should be treated as distinguishable (i.e., are there actual differences in parameters between dyad members based on the distinguishing variable?) (see Gistelinck et al., 2018; Kenny et al., 2006). Additionally, factors such as sample size may play a role in whether dyad members can be analysed as distinguishable since models in which dyad members are treated as distinguishable have more parameters and thus require a larger sample (Iida et al., 2023).

Different analytical approaches can be used to analyse dyadic time series data as obtained with dyadic ESM. An in-depth discussion of these approaches is beyond the scope of this chapter, but we describe some examples below and refer interested readers to the cited articles for more information. Which approach to use naturally depends on your research questions, and a relevant consideration is whether you are interested in outcomes at an individual level (where individuals within the dyad are interdependent and potentially influence one another) or at a dyad level (Iida et al., 2018). One approach that is often used and allows examining effects at an individual level is the longitudinal actor-partner interdependence model (L-APIM; e.g., Gistelinck & Loeys, 2019; Kenny et

al., 2006; Laurenceau & Bolger, 2012). In this model, the outcome of each dyadic partner is related to their own predictor (actor effect) as well as their partner's predictor (partner effect) across time. Thus, the mutual influence that may occur between individuals in a dyad is explicitly modelled and investigated. For example, you can investigate the extent to which both partners' sexual desire is predicted by their own mood (actor effect) as well as their partner's mood (partner effect; Frérart et al., 2024) or the extent to which both partners' perception of their partner's emotions is predicted by their own emotions (actor effect) and their partner's actual emotions (partner effect; Verhees, Ceulemans, et al., 2024). You could also only model actor effects and look at correlations between dyad members' intercepts, slopes or residuals using, for instance, dyadic growth curve modelling (Iida et al., 2023). For example, Rogers et al. (2018) examined partners' co-regulation of daily negative affect by looking at the covariance of residuals between dyad members after accounting for self-reported daily relationship transactions (actor effects) and potential linear effects of time. Specifically focusing on effects at a dyad level, within-couple covariation in emotions has also been examined by calculating correlations across partners' emotions over time points and then relating this to couple-level variables such as relationship duration (Sels, Cabrieto, et al., 2020; see also Carlier et al., 2024 for different ways in which dyadic similarity across multiple variables can be calculated). Another approach that allows to analyse effects at the dyad level is the dyadic score model (Iida et al., 2018). In this model, averages across and differences between partners in variables of interest serve as predictor and outcome variables. The longitudinal dyadic score model has recently been applied to dyadic experience sampling data by Stadler et al. (2023). They examined whether daily partner averaged companionship (mean of both partners) and partner differences in companionship (differences between both partners) could predict partner average experienced affect and partner differences in experienced affect.

Finally, in the context of dyadic analyses, you can evaluate whether an observed association genuinely reflects dyadic processes and does not arise due to confounding factors or stereotype accuracy by creating pseudo dyads (Corsini, 1956; Kenny et al., 2006). This approach includes creating dyads where one member is paired with an unrelated member from another dyad and then comparing these to the real sample; see Carlier et al. (in prep) for a more in-depth discussion and possible solutions.

The above short introduction to the modelling of dyadic data cannot cover all the possibilities and difficulties involved. For the reader interested in obtaining a more thorough introduction to the modelling of dyadic data, we can recommend Kenny et al. (2006) as well as Bolger and Laurenceau (specifically chapter 8; 2013).

14.4 Conclusion

Dyadic experience sampling studies allow researchers to further contextualize people's feelings, thoughts and behaviours in daily life by also directly examining their social environment. Recognizing that people function within larger social systems is important, and researchers should try to account for this when possible. While dyadic ESM studies offer many additional benefits, insights and potential in this regard for answering research questions, there are some additional aspects to consider. In this chapter, we discussed these different considerations that must be made before (sampling scheme, ESM items, ethical issues), during (briefing and debriefing participants) and after (structuring, checking and analysing the data) conducting a dyadic ESM study.

References

Aalbers, G., Hendrickson, A. T., Abeele, M. M. V., & Keijsers, L. (2023). Smartphone-tracked digital markers of momentary subjective stress in college students: Idiographic machine learning analysis. *JMIR mHealth and uHealth*, *11*(1), e37469. https://doi.org/10.2196/37469

Achterhof, R., Myin-Germeys, I., Bamps, E., Hagemann, N., Hermans, K. S. F. M., Hiekkaranta, A. P., Janssens, J. J., Lecei, A., Lafit, G., & Kirtley, O. J. (2025). COVID-19-related changes in adolescents' daily-life social interactions and psychopathology symptoms. *Journal of Nervous and Mental Disease*, *213*(4), 99–107. https://doi.org/10.1097/NMD.0000000000001826

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*(883). https://doi.org/10.3389/fpsyg.2014.00883

Adolf, J. K., Loossens, T., Tuerlinckx, F., & Ceulemans, E. (2021). Optimal sampling rates for reliable continuous-time first-order autoregressive and vector autoregressive modeling. *Psychological Methods*, *26*(6), 701–718. https://doi.org/10.1037/met0000398

Ajayi, O. O., Omotayo, A. A., Orogun, A. O., Omomule, T. G., & Orimoloye, S. M. (2018). Performance evaluation of native and hybrid Android applications. *Communication on Applied Electronics*, *7*(16), 1–9. https://doi.org/10.5120/cae2018652701

Alhammad, N., Alajlani, M., Abd-alrazaq, A., Epiphaniou, G., & Arvanitis, T. (2024). Patients' perspectives on the data confidentiality, privacy, and security of mhealth apps: Systematic review. *Journal of Medical Internet Research*, *26*, e50715. https://doi.org/10.2196/50715

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science: A call to action. *European Journal of Psychological Assessment*, *38*(1), 1–5. https://doi.org/10.1027/1015-5759/a000699

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.) (2014). Standards for educational and psychological testing. American Educational Research Association.

Arean, P. A., Hoa Ly, K., & Andersson, G. (2016). Mobile technology for mental health assessment. *Dialogues in Clinical Neuroscience*, *18*. https://doi.org/10.31887/dcns.2016.18.2/parean

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*(1), 1–19. https://doi.org/10.1037/met0000195

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2021). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*, *26*(2), 175–185. https://doi.org/10.1037/met0000294

Astivia, O. L. O., Gadermann, A., & Guhn, M. (2019). The relationship between statistical power and predictor distribution in multilevel logistic regression: A simulation-based approach. *BMC Medical Research Methodology*, *19*(1), 97–117. https://doi.org/10.1186/s12874-019-0742-8

Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2022). Missing data and other measurement quality issues in mobile geolocation sensor data. *Social Science Computer Review*, *40*(1), 212–235. https://doi.org/10.1177/0894439320944118

Bai, S., Babeva, K. N., Kim, M. I., & Asarnow, J. R. (2020). Future directions for optimizing clinical science & safety: Ecological momentary assessments in suicide/self-harm research. *Journal of Clinical Child & Adolescent Psychology*, *50*(1), 141–153. https://doi.org/10.1080/15374416.2020.1815208

Bak, M., Drukker, M., Hasmi, L., & van Os, J. (2016). Correction: An n=1 clinical network analysis of symptoms and treatment in psychosis. *PLoS One*, *11*(10), e0165762. https://doi.org/10.1371/journal.pone.0165762

Balaskas, A., Schueller, S. M., Cox, A. L., & Doherty, G. (2021). Ecological momentary interventions for mental health: A scoping review. *PloS One*, *16*(3), e0248152. https://doi.org/10.1371/journal.pone.0248152

Barker, R. G. (1975). *Ecological psychology: Concepts and methods for studying the environment of human behaviour*. Stanford University Press. (Original work published 1968.)

Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). The Guilford Press.

Bartels, S. L., van Knippenberg, R. J. M., Malinowsky, C., Verhey, F. R. J., & de Vugt, M. E. (2020). Smartphone-based experience sampling in people with mild cognitive impairment: Feasibility and usability study. *JMIR Aging*, 3(2), e19852. https://doi.org/10.2196/19852

Bartels, S. L., van Zelst, C., Melo Moura, B., Daniëls, N. E. M., Simons, C. J. P., Marcelis, M., Bos, F. M., & Servaas, M. N. (2023). Feedback based on experience sampling data: Examples of current approaches and considerations for future research. *Heliyon*, *9*(9), e20084. https://doi.org/10.1016/j.heliyon.2023.e20084

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., Ryan., O., … 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*, 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Bastiaansen, J. A., Meurs, M., Stelwagen, R., Wunderink, L., Schoevers, R. A., Wichers, M., & Oldehinkel, A. J. (2018). Self-monitoring and personalized feedback based on the experiencing sampling method as a tool to boost depression treatment: A protocol of a pragmatic randomized controlled trial (ZELF-i). *BMC Psychiatry*, *18*(1), 276. https://doi.org/10.1186/s12888-018-1847-z

Bastiaansen, J. A., Ornée, D. A., Meurs, M., & Oldehinkel, A. J. (2020). An evaluation of the efficacy of two add-on ecological momentary intervention modules for depression in a pragmatic randomized controlled trial (ZELF-i). *Psychological Medicine*, *52*(13), 1–10. https://doi.org/10.1017/s0033291720004845

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beames, J. R., Kikas, K., & Werner-Seidler, A. (2021). Prevention and early intervention of depression in young people: An integrated narrative review of affective awareness and Ecological Momentary Assessment. *BMC Psychology*, *9*(1), 113. https://doi.org/10.1186/s40359-021-00614-6

Beames, J. R., Uyttebroek, L., Edwards, C. J., Eisele, G. V., Kemme, N. D. F., Collier, O., van Roekel, E., Kwapil, T. R., Kirtley, O. J., & Myin-Germeys, I. (2025). Using the experience sampling methodology to measure anhedonia and its correlates in mental health research. *Clinical Psychology Review*, *119*, 102590. https://doi.org/10.1016/j.cpr.2025.102590

Belisario, J. S. M., Doherty, K., O'Donoghue, J., Ramchandani, P., Majeed, A., Doherty, G., Morrison, C., & Car, J. (2017). A bespoke mobile application for the longitudinal assessment of depression and mood during pregnancy: Protocol of a feasibility study. *BMJ Open*, *7*, e014469. https://doi.org/10.1136/bmjopen-2016-014469

Bell, I. H., Fielding-Smith, S. F., Hayward, M., Rossell, S. L., Lim, M. H., Farhall, J., & Thomas, N. (2018a). Smartphone-based ecological momentary assessment and intervention in a blended coping-fo-

cused therapy for distressing voices: Development and case illustration. *Internet Interventions*, *14*, 18–25. https://doi.org/10.1016/j.invent.2018.11.001

Bell, I. H., Fielding-Smith, S. F., Hayward, M., Rossell, S. L., Lim, M. H., Farhall, J., & Thomas, N. (2018b). Smartphone-based ecological momentary assessment and intervention in a coping-focused intervention for hearing voices (SAVVy): Study protocol for a pilot randomised controlled trial. *Trials*, *19*(1), 262. https://doi.org/10.1186/s13063-018-2607-6

Bell, I. H., Lim, M. H., Rossell, S. L., & Thomas, N. (2017). Ecological momentary assessment and intervention in the treatment of psychotic disorders: A systematic review. *Psychiatric Services*, *68*(11), 1172–1181. https://doi.org/10.1176/appi.ps.201600523

Bell, I. H., Rossell, S. L., Farhall, J., Hayward, M., Lim, M. H., Fielding-Smith, S. F., & Thomas, N. (2020). Pilot randomised controlled trial of a brief coping-focused intervention for hearing voices blended with smartphone-based ecological momentary assessment and intervention (SAVVy): Feasibility, acceptability and preliminary clinical outcomes. *Schizophrenia Research*, *216*, 479–487. https://doi.org/10.1016/j.schres.2019.10.026

Ben-Zeev, D., Young, M. A., & Madsen, J. W. (2009). Retrospective recall of affect in clinically depressed individuals and controls. *Cognition and Emotion*, *23*(5), 1021–1040. https://doi.org/10.1080/02699930802607937

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

Bentley, K. H., Maimone, J. S., Kilbury, E. N., Tate, M. S., Wisniewski, H., Levine, M. T., Roberg, R., Torous, J. B., Nock, M. K., & Kleiman, E. M. (2021). Practices for monitoring and responding to incoming data on self-injurious thoughts and behaviours in intensive longitudinal studies: A systematic review. *Clinical Psychology Review*, 90, 102098. https://doi.org/10.1016/j.cpr.2021.102098

Bentley, K. H., Millner, A. J., Bear, A., Follet, L., Fortgang, R. G., Zuromski, K. L., Kleiman, E. M., Coppersmith, D. D. L., Castro-Ramirez, F., Millgram, Y., Haim, A., Bird, S. A., & Nock, M. K. (2024). Intervening on high-risk responses during ecological momentary assessment of suicidal thoughts: Is there an effect on study data? *Psychological Assessment*, *36*(1), 66–80. https://doi.org/10.1037/pas0001288

Bidargaddi, N., Schrader, G., Klasnja, P., Licinio, J., & Murphy, S. (2020). Designing m-Health interventions for precision mental health support. *Translational Psychiatry*, *10*(1), 222. https://doi.org/10.1038/s41398-020-00895-2

Bitran, A. M., Sritharan, A., Trivedi, E., Helgren, F., Buchanan, S. N., Durham, K., Li, L. Y., Funkhouser, C. J., Allen, N. B., Shankman, S. A., Auerbach, R. P., & Pagliaccio, D. (2024). The effects of family support and smartphone-derived homestay on daily mood and depression among sexual and gender minority adolescents. *Journal of Psychopathology and Clinical Science*, *133*(5), 358–367. https://doi.org/10.1037/abn0000917

Blozis, S. A., McTernan, M., Harring, J. R., & Zheng, Q. (2020). Two-part mixed-effects location scale models. *Behavior Research Methods*, *52*(5), 1836–1847. https://doi.org/10.3758/s13428-020-01359-7

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioural research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bögemann, S. A., van Leeuwen, J. M. C., van Kraaij, A., Weermeijer, J., de Raedt, W., Kalisch, R., Myin-Germeys, I., & Hermans, E. J. (2024). Real-time analysis of psychological and physiological stress signals to trigger emotion-regulation interventions in daily life. *Psychoneuroendocrinology*, *160*, 106721. https://doi.org/10.1016/j.psyneuen.2023.106721

Bögemann, S. A., Riepenhausen, A., Puhlmann, L. M. C., Bar, S., Hermsen, E. J. C., Mituniewicz, J., Reppmann, Z. C., Uściłko, A., van Leeuwen, J. M. C., Wackerhagen, C., Yuen, K. S. L, Zerban, M., Weermeijer, J., Marciniak, M. A., Mor, N., van Kraaij, A., Köber, G., Pooseh, S., Koval, P., Arias-Vázquez, Binder, H., …Walter, H. (2023). Investigating two mobile just-in-time adaptive interventions to foster psychological resilience: Research protocol of the DynaM-INT study. *BMC Psychology*, *11*(1), 245. https://doi.org/10.1186/s40359-023-01249-5

Bogudzińska, B., Jaworski, A., Zajdel, A., Skrzypek, K., & Misiak, B. (2024). The experience sampling methodology in psychosis risk states: A systematic review. *Journal of Psychiatric Research*, *175*, 34–41. https://doi.org/10.1016/j.jpsychires.2024.04.050

Boiger, M., & Mesquita, B. (2012). The construction of emotion in interactions, relationships, and cultures. *Emotion Review*, *4*(3), 221–229. https://doi.org/10.1177/1754073912439765

Boker, S. M., Molenaar, P. C. M., & Nesselroade, J. R. (2009). Issues in intraindividual variability: Individual differences in equilibria and dynamics over multiple time scales. *Psychology and Aging*, *24*(4), 858–862. https://doi.org/10.1037/a0017912

Bolger, N. (2011). Power analysis for intensive longitudinal studies. In N. Bolger, G. Stadler, & J. P. Laurenceau (Eds.), *Handbook of research methods for studying daily life* (pp. 285–301). The Guilford Press.

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 (methodology in the social sciences)*. The Guilford Press.

Boon, B., Stroebe, W., Schut, H., & Ijntema, R. (2002). Ironic processes in the eating behaviour of restrained eaters. *British Journal of Health Psychology*, *7*(1), 1–10. https://doi.org/10.1348/135910702169303

Borah, T. J., Murray, A. L., Eisner, M., & Jugl, I. (2018). Developing and validating an experience sampling measure of aggression: The Aggression-ES Scale. *Journal of Interpersonal Violence*, 36 (11–12), 6166–6162. https://doi.org/10.1177/0886260518812068

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Bos, F. M., Snippe, E., Bruggeman, R., Wichers, M., & van der Krieke, L. (2019). Insights of patients and clinicians on the promise of the experience sampling method for psychiatric care. *Psychiatric Services*, *70*(11), 983–991. https://doi.org/10.1176/appi.ps.201900050

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). A web-based application for personalized ecological momentary assessment in psychiatric care: User-centered development of the PETRA application. *JMIR Mental Health*, *9*(8), e36430. https://doi.org/10.2196/36430

Bouisson, J., & Swendsen, J. (2003). Routinization and emotional well-being: An experience sampling investigation in an elderly French sample. *Journals Gerontology: series B Psychological Sciences and Social Sciences*, *58*(5), 280–282. https://doi.org/10.1093/geronb/58.5.P280

Bouws, J., Uyttebroek, L., Beames, J. R., de Koning, M., Schirmbeck, F., Henrard, A., Reininghaus, U., de Haan, L., & Myin-Germeys, I. (2025). Perspectives of patients with early psychosis on the use of an app in acceptance and commitment therapy: A qualitative study. *Early Intervention in Psychiatry*, *19*(8), e70073. https://doi.org/10.1111/eip.70073

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C., & Lindenberger, U. (2015). LIFESPAN: A tool for the computer-aided design of longitudinal studies. *Frontiers in Psychology*, *6*, *272.* https://doi.org/10.3389/fpsyg.2015.00272

Brans, K., Koval, P., Verduyn, P., Lim, Y. L., & Kuppens, P. (2013). The regulation of negative and positive affect in daily life. *Emotion*, *13*(5), 926–939. https://doi.org/10.1037/a0032400

Brietzke, E., Hawken, E. R., Idzikowski, M., Pong, J., Kennedy, S. H., & Soares, C. N. (2019). Integrating digital phenotyping in clinical characterization of individuals with mood disorders. *Neuroscience & Biobehavioural Reviews*, *104*, 223–230. https://doi.org/10.1016/j.neubiorev.2019.07.009

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T. W. & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892–903. https://doi.org/10.1037/abn0000446

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment*, *23*(4), 425–435. https://doi.org/10.1177/1073191116645909

Bringmann, L. F., van der Veen, D. C., Wichers, M., Riese, H., & Stulp, G. (2021). ESMvis: A tool for visualizing individual experience sampling method (ESM) data. *Quality of Life Research*, *30*(11), 3179–3188. https://doi.org/10.1007/s11136-020-02701-4

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS One*, *8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188

Broda, M. (2017). Using multilevel models to explore predictors of high school students' nonresponse in experience sampling method (ESM) studies. *Social Science Computer Review*, *35*(6), 733–750. https://doi.org/10.1177/0894439316667049

Broderick, J. E., & Vikingstad, G. (2008). Frequent assessment of negative symptoms does not induce depressed mood. *Journal of Clinical Psychology Medical Settings*, *15*(4), 296–300. https://doi.org/10.1007/s10880-008-9127-6

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/rj-2017-066

Brown, C. H., Curran, G., Palinkas, L. A., Aarons, G. A., Wells, K. B., Jones, L., Collins, L. M., Duan, N., Mittman, B. S., Wallace, A., Tabak, R. G., Ducharme, L., Chambers, D. A., Neta, G., Wiley, T., Landsverk, J., Cheung, K., & Cruden, G. (2017). An overview of research and evaluation designs for dissemination and implementation. *Annual Review of Public Health*, *38*, 1–22. https://doi.org/10.1146/annurev-publhealth-031816-044215

Browne, W. J., Mousa, G. L., & Parker, R. M. A. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package*. University of Bristol.

Bufano, P., Laurino, M., Said, S., Tognetti, A., & Menicucci, D. (2023). Digital phenotyping for monitoring mental disorders: systematic review. *Journal of Medical Internet Research*, *25*, e46778. https://doi.org/10.2196/46778

Bülow, A., Janssen, L. H. C., Dietvorst, E., Bij de Vaate, N. A. J. D., Hillegers, M. H. J., Valkenburg, P. M., & Keijsers, L. (2025) From burden to enjoyment: A user-centered approach to engage adolescents in intensive longitudinal research. *Journal of Adolescence*, *97*(4), 886–900. https://doi.org/10.1002/jad.12478

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

Burke, T. A., Fox, K., Kautz, M., Siegel, D. M., Kleiman, E., & Alloy, L. B. (2021). Real-time monitoring of the associations between self-critical and self-punishment cognitions and nonsuicidal self-injury. *Behaviour Research and Therapy*, *137*, 103775. https://doi.org/10.1016/j.brat.2020.103775

Butt, M. A., Kazanskiy, N. L., & Khonina, S. N. (2022). Revolution in flexible wearable electronics for temperature and pressure monitoring: A review. *Electronics*, *11*(5), 716. https://doi.org/10.3390/electronics11050716

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Buu, A., Yang, S., Li, R., Zimmerman, M. A., Cunningham, R. M., & Walton, M. A. (2020). Examining measurement reactivity in daily diary data on substance use: Results from a randomized experiment. *Addictive Behaviors*, *102*, 106198. https://doi.org/10.1016/j.addbeh.2019.106198

Cabrieto, J., Tuerlinckx, F., Kuppens, P., Wilhelm, F. H., Liedlgruber, M., & Ceulemans, E. (2018). Capturing correlation changes by applying kernel change point detection on the running correlations. *Information Sciences*, *447*, 117–139. https://doi.org/10.1016/j.ins.2018.03.010

Cajas, S. A., Landínez, M. A., & López, D. M. (2020). Modeling of motion artifacts on PPG signals for heart-monitoring using wearable devices. In *15th International Symposium on Medical Information Processing and Analysis*, *11330*, 320–334. SPIE. https://doi.org/10.1117/12.2540554

Calvo, N., García-González, S., Perez-Galbarro, C., Regales-Peco, C., Lugo-Marin, J., Ramos-Quiroga, J.-A., & Ferrer, M. (2022). Psychotherapeutic interventions specifically developed for NSSI in adolescence: A systematic review. *European Neuropsychopharmacology*, *58*, 86–98. https://doi.org/10.1016/j.euroneuro.2022.02.009

Camerman, E., Kuppens, P., Lavrijsen, J., & Verschueren, K. (2024). Real-time fluctuations in student emotions and relations with day of the week, time of the day, and teaching methods. *Frontiers in Education*, *9*, 1470565. https://doi.org/10.3389/feduc.2024.1470565

Canali, S., Schiaffonati, V., & Aliverti, A. (2022). Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness. *PLOS Digital Health*, *1*(10), e0000104. https://doi.org/10.1371/journal.pdig.0000104

Capon, H., Hall, W., Fry, C., & Carter, A. (2016). Realising the technological promise of smartphones in addiction research and treatment: An ethical review. *International Journal of Drug Policy*, *36*, 47–57. https://doi.org/10.1016/j.drugpo.2016.05.013

Carlier, C., Frérart, L., Ceulemans, E., & Kuppens, P. (forthcoming). *Developing a framework for testing dyadic associations using permutations*.

Carlier, C., Karch, J. D., Kuppens, P., & Ceulemans, E. (2024). A comparison of measures for assessing profile similarity in dyads. *Psychological Belgica*, *64*, 72–84. https://doi.org/10.5334/pb.1297

Castro-Alvarez, S., Bringmann, L. F., Back, J., & Liu, S. (2025). The many reliabilities of psychological dynamics: An overview of statistical approaches to estimate the internal consistency reliability of intensive longitudinal data. *Psychological Methods*. https://doi.org/10.1037/met0000778. Advance online publication.

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

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Aden-Buie, G., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2019). *Shiny: Web application framework for R*. https://CRAN.R-project.org/package=shiny

Chin, A., Markey, A., Bhargava, S., Kassam, K. S., & Loewenstein, G. (2017). Bored in the USA: Experience sampling and boredom in everyday life. *Emotion*, *17*(2), 359–368. https://doi.org/10.1037/emo0000232

Cho, S., Ensari, I., Weng, C., Kahn, M. G., & Natarajan, K. (2021). Factors affecting the quality of person-generated wearable device data and associated challenges: Rapid systematic review. *JIMR mHealth and uHealth*, *9*(3), e20738. https://doi.org/10.2196/20738

Christensen, A., & Nies, D. C. (1980). The spouse observation checklist: Empirical analysis and critique. *American Journal of Family Therapy*, *8*(2), 69–79. https://doi.org/10.1080/01926188008250357

Chu, B., Marwaha, K., Sanvictores, T., Awosika, A. O., & Ayers, D. (2023, May 7). *Physiology, stress reaction*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK541120/

Clark, L. A., Watson, D., & Leeka, J. (1989). Diurnal variation in the positive affects. *Motivation and Emotion*, *13*(3), 205–234. https://doi.org/10.1007/bf00995536

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *35*(3), 189. https://doi.org/10.1037/pas0001200

Cloos, L., Mestdagh, M., Vanpaemel, W., Ceulemans, E., & Kuppens, P. (2024). Measuring continuous affect in daily life with intensity profile drawings. *Assessment*, *32*(5), 689–704. https://doi.org/10.1177/10731911241266286

Coan, J. A. (2010). Emergent ghosts of the emotion machine. *Emotion Review*, *2*(3), 274–285. https://doi.org/10.1177/1754073910361978

Cohen, J. (2013). *Statistical power analysis for the behavioural sciences*. Academic Press.

Collip, D., van Winkel, R., Peerbooms, O., Lataster, T., Thewissen, V., Lardinois, M., Drukker, M, Rutten B. P. F., Van Os, J., Myin-Germeys, I. (2011). COMT Val158Met-stress interaction in psychosis: Role of background psychosis risk. *CNS Neuroscience & Therapeutics*, *17* https://doi.org/10.1111/j.1755-5949.2010.00213.x

Colombo, D., Suso-Ribera, C., Fernández-Álvarez, J., Cipresso, P., Garcia-Palacios, A., Riva, G., & Botella, C. (2020). Affect recall bias: Being resilient by distorting reality. *Cognitive Therapy and Research*, *44*, 906–918. https://doi.org/10.1007/s10608-020-10122-3

Colvonen, P. J. (2021). Response To: Investigating sources of inaccuracy in wearable optical heart rate sensors. *Nature Partner Journal Digital Medicine*, *4*(1), 38. https://doi.org/10.1038/s41746-021-00408-5

Conner, T. S., & Reid, K. A. (2012). Effects of intensive mobile happiness reporting in daily life. *Social Psychological and Personality Science*, *3*(3), 315–323. https://doi.org/10.1177/1948550611419677

Conner, T. S., & Mehl, M. R. (2015). Ambulatory assessment: Methods for studying everyday life. In R. A. Scott, M. C. Buchmann & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioural sciences*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/9781118900772.etrds0010

Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: A program for designing efficient multilevel studies. *Behaviour Research Methods*, *40*(1), 236–249. https://doi.org/10.3758/brm.40.1.236

Coppersmith, D. D. L., Fortgang, R. G., Kleiman, E. M., Millner, A. J., Yeager, A. L., Mair, P., & Nock, M. K. (2022). Effect of frequent assessment of suicidal thinking on its incidence and severity: High-resolution real-time monitoring study. *The British Journal of Psychiatry*, *220*(1), 41–43. https://doi.org/10.1192/bjp.2021.97

Cordier, R., Brown, N., Chen, Y. W., Wilkes-Gillan, S., & Falkmer, T. (2016). Piloting the use of experience sampling method to investigate the everyday social experiences of children with Asperger syndrome/high functioning autism. *Developmental Neurorehabilitation*, *19*(2), 103–110. https://doi.org/10.3109/17518423.2014.915244

Corsini, R. J. (1956). Understanding and similarity in marriage. *The Journal of Abnormal and Social Psychology*, *52*(3), 327–332. https://doi.org/10.1037/h0043556

Coughlin, L. N., Campbell, M., Wheeler, T., Rodriguez, C., Florimbio, A. R., Ghosh, S., Guo, Y., Hung, P.-Y., Newman, M. W., Pan, H., Zhang, K. W., Zimmermann, L., Bonar, E. E., Walton, M., Murphy, S. & Nahum-Shani, I. (2024). A mobile health intervention for emerging adults with regular cannabis use: A micro-randomized pilot trial design protocol. *Contemporary Clinical Trials*, *145*, 107667. https://doi.org/10.1016/j.cct.2024.107667

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

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

Critchley, H. D., Eccles, J., & Garfinkel, S. N. (2013). Interaction between cognition, emotion, and the autonomic nervous system. In G. Goldenberg (Ed.), *Handbook of clinical neurology* (Vol. 117, pp. 59–77). Elsevier. https://doi.org/10.1016/B978-0-444-53491-0.00006-7

Cruise, C. E., Broderick, J., Porter, L., Kaell, A., & Stone, A. A. (1996). Reactive effects of diary self-assessment in chronic pain patients. *Pain*, *67*(2–3), 253–258. https://doi.org/10.1016/0304-3959(96)03125-9

Csikszentmihalyi, M., & Larson, R. (1984). *Being adolescent: Conflict and growth in the teenage years*. Basic Books.

Cummins, N., Dineley, J., Conde, P., Matcham, F., Siddi, S., Lamers, F., Carr, E., Lavelle, G., Leightley, D., Withe K. M., Oetzmann, C., Campbell, E. L., Simblett, S., Bruce, S., Haro, J. M., Penninx, B. W J. H., Ranjan, Y., Rashid, Z., Stewart, C., Folarin, A. A., … Hotopf M. (2023). Multilingual markers of depression in remotely collected speech samples: A preliminary analysis. *Journal of Affective Disorders*, *341*, 128–136. https://doi.org/10.1016/j.jad.2023.08.097

Curran, G. M., Bauer, M., Mittman, B., Pyne, J. M., & Stetler, C. (2012). Effectiveness-implementation hybrid designs: Combining elements of clinical effectiveness and implementation research to enhance public health impact. *Medical Care*, *50*(3), 217–226. https://doi.org/10.1097/MLR.0b013e3182408812

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review Psychology*, *62*, 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

Daemen, M., Postma, M. R., Lindauer, R., Hoes-van der Meulen, I., Nieman, D., Delespaul, P., Breedvelt J. J. F., van der Gaag, M., Viechtbauer, W., Schruers, K., van den Berg, D., Bockting, C., van Amelsvoort, T., & Reinighaus, U. (2021). Efficacy of a transdiagnostic ecological momentary intervention for improving self-esteem (SELFIE) in youth exposed to childhood adversity: Study protocol for a multi-center randomized controlled trial. *Trials*, *22*(1), 641. https://doi.org/10.1186/s13063-021-05585-y

Damschroder, L. J., Reardon, C. M., Widerquist, M. A. O., & Lowery, J. (2022). The updated consolidated framework for implementation research based on user feedback. *Implementation Science*, *17*(1), 75. https://doi.org/10.1186/s13012-022-01245-0

Dao, K. P., De Cocker, K., Tong, H. L., Kocaballi, A. B., Chow, C., & Laranjo, L. (2021). Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviours: Systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. *JMIR Mhealth Uhealth*, *9*(11), e22890. https://doi.org/10.2196/22890

Darji, J., Biswas, N., Jones, L. D., & Ashili, S. (2023). Handling missing data in the time-series data from wearables. In *IntechOpen zoek via crossref*. https://doi.org/10.5772/intechopen.1002536

Davis, T. J., Hevel, D. J., Dunton, G. F., & Maher, J. P. (2022). Bidirectional associations between physical activity and pain among older adults: An ecological momentary assessment study. *Journal of Aging and Physical Activity*, *31*(2), 240–248. https://doi.org/10.1123/japa.2022-0014

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 Behavioural Assessment*, *42*(1), 13–24. https://doi.org/10.1007/s10862-019-09766-7

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D. C., Dobson, R., & Hotopf, M. (2022). Digital health tools for the passive monitoring of depression: A systematic review of methods. *npj Digital Medicine*, *5*(1), 1–14. https://doi.org/10.1038/s41746-021-00548-8

De Brito, J. N., Loth, K. A., Tate, A., & Berge, J. M. (2020). Associations between parent self-reported and accelerometer-measured physical activity and sedentary time in children: Ecological momentary assessment study. *JMIR mHealth and uHealth*, *8*(5), e15458. https://doi.org/10.2196/15458

de Bruin, L., Newen, A., & Gallagher, S. (2018). *The Oxford handbook of 4E cognition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198735410.001.0001

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. (2022). Designing daily-life research combining experience sampling method with parallel data. *Psychological Medicine*, *54*(1), 98–107. https://doi.org/10.1017/S0033291722002367

De Calheiros Velozo, J., Vaessen, T., Claes, S., & Myin-Germeys, I. (2024). Investigating adverse daily life effects following a psychosocial laboratory stress task, and the moderating role of psychopathology. *Stress*, *27*(1), 2380403. https://doi.org/10.1080/10253890.2024.2380403

de Haan, S. (2020). *Enactive psychiatry*. Cambridge University Press. https://doi.org/10.1017/9781108685214

de Haan-Rietdijk, S., Kuppens, P., & Hamaker, E. L. (2016). What's in a day? A guide to decomposing the variance in intensive longitudinal data. *Frontiers in Psychology*, *7*, 891. https://doi.org/10.3389/fpsyg.2016.00891

de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discrete-vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology*, *8*, 1849. https://doi.org/10.3389/fpsyg.2017.01849

de Thurah, L., Kiekens, G., Sips, R., Teixeira, A., Kasanova, Z., & Myin-Germeys, I. (2023). Using experience sampling methods to support clinical management of psychosis: The perspective of people with lived experience. *Psychiatry Research*, *324*, 115207. https://doi.org/10.1016/j.psychres.2023.115207

de Vries, L. P., Baselmans, B. M. L., & Bartels, M. (2020). Smartphone-based ecological momentary assessment of well-being: A systematic review and recommendations for future studies. *Journal of Happiness Studies*, *22*, 2361–2408. https://doi.org/10.1007/s10902-020-00324-7

De Vuyst, H.-J., Dejonckheere, E., Van der Gucht, K., & Kuppens, P. (2019). Does repeatedly reporting positive or negative emotions in daily life have an impact on the level of emotional experiences and depressive symptoms over time? *PLoS One*, *14*(6), e0219121. https://doi.org/10.1371/journal.pone.0219121

Dejonckheere, E., Bastian, B., Fried, E. I., Murphy, S. C., & Kuppens, P. (2017). Perceiving social pressure not to feel negative predicts depressive symptoms in daily life. *Depression and Anxiety*, *34*(9), 836–844. https://doi.org/10.1002/da.22653

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. https://doi.org/10.1037/pas0001178

Dejonckheere, E., Houben, M., Schat, E., Ceulemans, E., & Kuppens, P. (2021). The short-term psychological impact of the COVID-19 pandemic in psychiatric patients: Evidence for differential emotion and symptom trajectories in Belgium. *Psychologica Belgica*, *61*(1), 163–172. https://doi.org/10.5334/pb.1028

Dejonckheere, E., & Mestdagh, M. (2021). On the signal-to-noise ratio in real-life emotional time series. In C. Waugh & P. Kuppens (Eds.), *Affect dynamics* (pp. 131–152). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-82965-0_7

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Brock, B., & Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*(2), 323–341. https://doi.org/10.1037/pspp0000186

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

Dejonckheere, E., Verdonck, S., Andries, J., Röhrig, N., Piot, M., Kilani, G., & Mestdagh, M. (2024). Real-time incentivizing survey completion with game-based rewards in experience sampling research may increase data quantity, but reduces data quality. *Computers in Human Behaviour*, *160*, 108360. https://doi.org/10.1016/j.chb.2024.108360

Delespaul, P. A. E. G. (1995). *Assessing schizophrenia in daily life: The Experience Sampling Method*. Universitaire Pers Maastricht. https://doi.org/10.26481/dis.19950504pd

Demidenko, E. (2004). *Mixed models: Theory and applications*. Wiley. https://doi.org/10.1002/0471728438

Demidenko, E. (2013). *Mixed models: Theory and applications with R*. John Wiley & Sons. https://doi.org/10.1002/9781118651537

Deniz-Garcia, A., Fabelo, H., Rodriguez-Almeida, A. J., Zamora-Zamorano, G., Castro-Fernandez, M., Alberiche Ruano, M. D. P., Solvoll, T., Granja, C., Schopf, T. R., Callico G. M., Soguero-Ruiz, C., Wägner A. M., & WARIFA Consortium. (2023). Quality, usability, and effectiveness of mHealth apps and the role of artificial intelligence: Current scenario and challenges. *Journal of Medical Internet Research*, *25*, e44030. https://doi.org/10.2196/44030

Dewa, L. H., Lavelle, M., Pickles, K., Kalorkoti, C., Jaques, J., Pappa, S., & Aylin, P. (2019). Young adults' perceptions of using wearables, social media and other technologies to detect worsening mental health: A qualitative study. *PLoS One*, *14*(9), e0222655. https://doi.org/10.1371/journal.pone.0222655

Dickens, Y. L., Van Raalte, J., & Hurlburt, R. T. (2018). On investigating self-talk: A descriptive experience sampling study of inner experience during golf performance. *The Sport Psychologist*, *32*, 66–73. https://doi.org/10.1123/tsp.2016-0073

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. CRC Press. https://doi.org/10.1201/9781315182780

Dora, J., Kuczynski, A. M., Schultz, M. E., Acuff, S. F., Murphy, J. G., & King, K. M. (2024). An experimental investigation into the effect of negative affect on the behavioural economic demand for alcohol. *Psychology of Addictive Behaviours*, *38*(4), 409–423. https://doi.org/10.1037/adb0000984

Dowling, N. A., Rodda, S. N., & Merkouris, S. S. (2024). Applying the just-in-time adaptive intervention framework to the development of gambling interventions. *Journal of Gambling Studies*, *40*(2), 717–747. https://doi.org/10.1007/s10899-023-10250-x

Dubad, M., Winsper, C., Meyer, C., Livanou, M., & Marwaha, S. (2018). A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. *Psychological Medicine*, *48*(2), 208–228. https://doi.org/10.1017/S0033291717001659

Dunton, G. F., Huh, J., Leventhal, A., Riggs, N., Spruijt-Metz, D., Pentz, M. A., & Hedeker, D. (2014). Momentary assessment of affect, physical feeling states, and physical activity in Children. *Health Psychology*, *33*(3), 255–263. https://doi.org/10.1037/a0032640

Eisele, G., Hiekkaranta, A., Kunkels, Y. K., aan het Rot, M., van ballegooijen, W., Bartels, S. L., Bastiaansen, J. A., Beymer, P. N., Bylsma, L. M, Carpenter, R. W., Ellison, W. D., Fisher, A. K.; Forkmann, T., Frumkin M. R., Fulford, D., Naragon-Gainey, K., Greene, T., Heininga, V. E., Jones, A., Kalokernos, E. K., … Kirtley, O. J. (2024). ESM-Q: A consensus-based quality assessment tool for experience sampling method items. *Behaviour Research Methods*, *57*(4), 124. https://doi.org/10.31234/osf.io/sjynv

Eisele, G., Lafit, G., Vachon, H., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2021). Affective structure, measurement invariance, and reliability across different experience sampling protocols. *Journal of Research in Personality*, *92*, 1–11. https://doi.org/10.1016/j.jrp.2021.104094

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Eisele, G., Vachon, H., Myin-Germeys, I., & Viechtbauer, W. (2021). Reported affect changes as a function of response delay: Findings from a pooled dataset of nine experience sampling studies. *Frontiers in Psychology*, *12*, 580684. https://doi.org/10.3389/fpsyg.2021.580684

Ellison, W. D. (2021). An initial study of practicing psychologists' views of the utility of ecological momentary assessment for difficult psychotherapy cases. *Administration and Policy in Mental Health*, *48*(4), 597–607. https://doi.org/10.1007/s10488-020-01093-4

Epskamp, S. (2020). *graphicalVAR: Graphical VAR for Experience Sampling Data. R package version 0.2.4*. https://CRAN.R- project.org/package=graphicalVAR

Epskamp, S., Deserno, M. K., & Bringmann, L. F. (2019). *mlVAR: Multi-Level Vector Autoregression. R package version 0.4.4*. https://CRAN.R-project.org/package=mlVAR

Erbe, D., Eichert, H. C., Riper, H., & Ebert, D. D. (2017). Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: Systematic review. *Journal of Medical Internet Research*, *19*(9), e306. https://doi.org/10.2196/jmir.6588

Farewell, V. T., Long, D. L., Tom, B. D. M., Yiu, S., & Su, L. (2017). Two-part and related regression models for longitudinal data. *Annual Review of Statistics and Its Application*, *4*, 283–315. https://doi.org/10.1146/annurev-statistics-060116-054131

Fischer, J. E., Greenhalgh, C., & Benford, S. (2011). Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services—MobileHCI '11* (p. 181). https://doi.org/10.1145/2037373.2037402

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings National Academy of Sciences of USA*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303–315. https://doi.org/10.1086/209351

Fleeson, W., & Law, M. K. (2015). Trait enactments as density distributions: The role of actors, situations, and observers in explaining stability and variability. *Journal of Personality and Social Psychology*, *109*(6), 1090–1104. https://doi.org/10.1037/a0039517

Folkersma, W., Veerman, V., Ornee, D. A., Oldehinkel, A. J., Alma, M. A., & Bastiaansen, J. A. (2021). Patients' experience of an ecological momentary intervention involving self-monitoring and personalized feedback for depression. *Internet Interventions*, *26*, 100436. https://doi.org/10.1016/j.invent.2021.100436

Forkmann, T., Spangenberg, L., Rath, D., Hallensleben, N., Hegerl, U., Kersting, A., & Glaesmer, H. (2018). Assessing suicidality in real time: A psychometric evaluation of self-report items for the assess-

ment of suicidal ideation and its proximal risk factors using ecological momentary assessments. *Journal of Abnormal Psychology*, *127*(8), 758–769. https://doi.org/10.1037/abn0000381

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaoszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviours: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187–232. https://doi.org/10.1037/bul0000084

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition and Emotion*, *14*(4), 577–606. https://doi.org/10.1080/026999300402808

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, *65*(1), 45–55. https://doi.org/10.1037//0022-3514.65.1.45

Frérart, L., De Roovere, C., Sels, L., Ceulemans, E., Janssen, E., & Kuppens, P. (2024). In the mood: How sexual desire predicts and is predicted by romantic partners' mood. *The Journal of Sex Research*, *62*(5), 832–842. https://doi.org/10.1080/00224499.2024.2395482

Frijters, P., & Beatton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behaviour & Organization*, *82*(2–3), 525–542. https://doi.org/10.1016/j.jebo.2012.03.008

Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomyn, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment*, *29*(9), 1120–1128. https://doi.org/10.1037/pas0000411

Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., & Mills, J. (2013). Does the burden of the experience sampling method undermine data quality in state body image research? *Body Image*, *10*(4), 607–613. https://doi.org/10.1016/j.bodyim.2013.06.003

Frumkin, M. R., Piccirillo, M. L., Beck, E. D., Grossman, J. T., & Rodebaugh, T. L. (2021). Feasibility and utility of idiographic models in the clinic: A pilot study. *Psychotherapy Research*, *31*(4), 520–534. https://doi.org/10.1080/10503307.2020.1805133

Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, *34*(2), 244–254. https://doi.org/10.1177/0894439315575477

Gabbidon, K., Chenneville, T., & Rote, W. (2022). Ethical considerations for parent-adolescent dyadic research. *Ethics & Human Research*, *44*(3), 24–33. https://doi.org/10.1002/eahr.500127

Gałecki, A., & Burzykowski, T. (2003). *Linear mixed-effects models using R*. Springer. https://doi.org/10.1007/978-1-4614-3900-4_1

Gashi, S., Di Lascio, E., Stancu, B., Swain, V. D., Mishra, V., Gjoreski, M., & Santini, S. (2020). Detection of artifacts in ambulatory electrodermal activity data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(2), 1–31. https://doi.org/10.1145/3397316

General Data Protection Regulation (GDPR) (2016). REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119/1.

Gibbons, R. D., Segawa, E., Karabatsos, G., Amatya, A. K., Bhaumik, D. K., Brown, C. H., Kapur, K., Marcus, S. M., Hur, K., & Mann, J. J. (2008). Mixed-effects poisson regression analysis of adverse event reports: The relationship between antidepressants and suicide. *Statistics in Medicine*, *27*(11) 1814–1833. https://doi.org/10.1002/sim.3241

Gibson, J. J. (2015). *The ecological approach to visual perception*. Psychology Press. https://doi.org/10.4324/9781315740218

Giebel, G. D., Speckemeier, C., Abels, C., Plescher, F., Börchers, K., Wasem, J., Blase, N., & Neusser, S. (2023). Problems and barriers related to the use of digital health applications: Scoping review. *Journal of Medical Internet Research*, *25*, e43808. https://doi.org/10.2196/43808

Gillinov, S., Etiwy, M., Wang, R., Blackburn, G., Phelan, D., Gillinov, A. M., Houghtaling, P., Hoda Javadikasgari, H., & Desai, M. Y. (2017). Variable accuracy of wearable heart rate monitors during aerobic exercise. *Medicine & Science in Sports & Exercise*, *49*(8), 1697–1703. https://doi.org/10.1249/MSS.0000000000001284

Gistelinck, F., & Loeys, T. (2019). The actor–partner interdependence model for longitudinal dyadic data: An implementation in the SEM framework. *Structural Equation Modeling*, *26*(3), 329–347. https://doi.org/10.1080/10705511.2018.1527223

Gistelinck, F., Loeys, T., Decuyper, M., & Dewitte, M. (2018). Indistinguishability tests in the actor–partner interdependence model. *British Journal of Mathematical and Statistical Psychology*, *71*(3), 472–498. https://doi.org/10.1111/bmsp.12129

Glasgow, T. E., Le, H. T. K., Scott Geller, E., Fan, Y., & Hankey, S. (2019). How transport modes, the built and natural environments, and activities influence mood: A GPS smartphone app study. *Journal of Environmental Psychology*, *66*, 101345. https://doi.org/10.1016/j.jenvp.2019.101345

Glenn, C. R., Kleiman, E. M., Kearns, J. C., Santee, A. C., Esposito, E. C., Conwell, Y., & Alpert-Gillis, L. J. (2020). Feasibility and acceptability of ecological momentary assessment with high-risk suicidal adolescents following acute psychiatric care. *Journal of Clinical Child & Adolescent Psychology*, *51*(1), 32–48. https://doi.org/10.1080/15374416.2020.1741377

Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons. https://doi.org/10.1002/9780470973394

Goldstein, H., Healy, M. J., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, *13*(16), 1643–1655. https://doi.org/10.1002/sim.4780131605

Goldstein, S. P., Zhang, F., Klasnja, P., Hoover, A., Wing, R. R., & Thomas, J. G. (2021). Optimizing a just-in-time adaptive intervention to improve dietary adherence in behavioural obesity treatment: Protocol for a microrandomized trial. *JMIR Research Protocols*, *10*(12), e33568. https://doi.org/10.2196/33568

Gould, C. E., Ma, F., Loup, J. R., Juang, C., Sakai, E. Y., & Pepin, R. (2020). Technology-based mental health assessment and intervention. In N. Hantke, A. Etkin, & R. O'Hara (Eds.), *Handbook of mental health and aging* (pp. 401–415): Academic Press. https://doi.org/10.1016/b978-0-12-800136-3.00024-7

Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychological Methods*, *11*(1), 87–105. https://doi.org/10.1037/1082-989X.11.1.87

Green, P., & Macleod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210x.12504

Greenhalgh, T., Wherton, J., Papoutsi, C., Lynch, J., Hughes, G., A'Court, C., Hinder, S., Fahy, N., Procter, R. & Shaw, S. (2017). Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, *19*(11), e367. https://doi.org/10.2196/jmir.8775

Greenleaf, G. (2017). Global data privacy laws 2017: 120 national data privacy laws, including Indonesia and Turkey. *Privacy Laws & Business International Report*, *145*(10–13), 17–45. Available at SSRN: https://ssrn.com/abstract=2993035.

Greer, B., Robotham, D., Simblett, S., Curtis, H., Griffiths, H., & Wykes, T. (2019). Digital exclusion among mental health service users: Qualitative investigation. *Journal of Medical Internet Research*, *21*(1), e11696. https://doi.org/10.2196/11696

Groot, P. C. (2010). Patients can diagnose too: How continuous self-assessment aids diagnosis of, and recovery from, depression. *Journal of Mental Health*, *19*(4), 352–362. https://doi.org/10.3109/09638237.2010.494188

Grosprêtre, S., Marusic, U., Gimenez, P., Ennequin, G., Mourot, L., & Isacco, L. (2021). Stand up to excite the spine: Neuromuscular, autonomic, and cardiometabolic responses during motor imagery in standing vs. sitting posture. *Frontiers in Physiology*, *12*, 762452. https://doi.org/10.3389/fphys.2021.762452

Gründahl, M., Deckert, J., & Hein, G. (2020). Three questions to consider before applying ecological momentary interventions (EMI) in psychiatry. *Frontiers Psychiatry*, *11*, 333. https://doi.org/10.3389/fpsyt.2020.00333

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics*, *56*(4), 1030–1039. https://doi.org/10.1111/j.0006- 341x.2000.01030.x

Hall, M., Scherner, P. V., Kreidel, Y., & Rubel, J. A. (2021). A systematic review of momentary assessment designs for mood and anxiety symptoms. *Frontiers in Psychology*, *12*, 642044. https://doi.org/10.3389/fpsyg.2021.642044

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322. https://doi.org/10.1177/1754073915590619

Han, D., Zhang, C., Fan, X., Hindle, A., Wong, K., & Stroulia, E. (2012). Understanding android fragmentation with topic analysis of vendor-specific bugs. In *19th Working Conference on Reverse Engineering (WCRE), 2012* (pp. 83–92). IEEE. https://doi.org/10.1109/wcre.2012.18

Hanssen, E., Balvert, S., Oorschot, M., Borkelmans, K., van Os, J., Delespaul, P., & Fett, A.-K. (2020). An ecological momentary intervention incorporating personalised feedback to improve symptoms and social functioning in schizophrenia spectrum disorders. *Psychiatry Research*, *284*, 112695. https://doi.org/10.1016/j.psychres.2019.112695

Harari, G. M., & Gosling, S. D. (2023). Understanding behaviours in context using mobile sensing. *Nature Reviews Psychology*, *2*(12), 767–779. https://doi.org/10.1038/s44159-023-00235-3

Harari, G. M., Soh, S., & Kroencke, L. (2024). How to Conduct Mobile Sensing Research. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Priemer (Eds). *Mobile Sensing in Psychology: Methods and applications*. Guilford Press.

Harville, D. A (1977). Maximum likelihood approaches to variance component estimation and to related problems: Rejoinder. *Journal of the American Statistical Association*, *72*(358), 339–340. https://doi.org/10.2307/2286798

Hasking, P., Whitlock, J., Voon, D., & Rose, A. (2017). A cognitive-emotional model of NSSI: Using emotion regulation and cognitive processes to explain why people self-injure. *Cognition & Emotion*, *31*(8), 1543–1556. https://doi.org/10.1080/02699931.2016.1241219

Haslbeck, J. M. B., Jover Martínez, A., Roefs, A., Fried, E. I., Lemmens, L. H. J. M., Groot, E. , & Edelsbrunner, P. A. (2025). Comparing likert and visual analogue scales in ecological momentary assessment. *Behaviour Research Methods*, 57, 217. https://doi.org/10.31234/osf.io/yt8xw_v2

Haslbeck, J., 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. (2022). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behaviour Research Methods*, *54*(4), 1541–1558. https://doi.org/10.3758/s13428-021-01683-6

Hasselhorn, K., Ottenstein, C., Meiser, T., & Lischetzke, T. (2024). The effects of questionnaire length on the relative impact of response styles in ambulatory assessment. *Multivariate Behavioural Research*, *59*(5), 1043–1057. https://doi.org/10.1080/00273171.2024.2354233

Hayes, S. C. (2016). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioural and cognitive therapies—republished article. *Behaviour Therapy*, *47*(6), 869–885. https://doi.org/10.1016/j.beth.2016.11.006

Hecht, M., & Zitzmann, S. (2020). Sample size recommendations for continuous-time models: Compensating shorter time series with larger numbers of persons and vice versa. *Structural Equation Modeling. 28*(2), 229–236. https://doi.org/10.1080/10705511.2020.1779069

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley. https://doi.org/10.1002/0470036486

Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-Related Contrasts Between Two Groups. *Journal of Educational and Behavioural Statistics 24*(1), 70–93. https://doi.org/10.3102/10769986024001070

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of intraclass correlations in three- and four-level models. *Educational and Psychological Measurement*, *72*(6), 893–909. https://doi.org/10.1177/0013164412445193

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., & Kuppens P. (2019). The dynamical signature of anhedonia in major depressive disorder: Positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry*, *19*(1), 59. https://doi.org/10.1186/s12888-018-1983-5

Hektner, J., Schmidt, J. A., & Csikszentmihalyi, M. (2007). Experience sampling method: Measuring the quality of everyday life. Sage Publications, Inc. https://doi.org/10.4135/9781412984201

Henderson, C. R., Kempthorne, O., Searle, S. R., & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, *15*(2), 192–218. https://doi.org/10.2307/2527669

Hennig, T., & Lincoln, T. M. (2018). Sleeping paranoia away? An actigraphy and experience-sampling study with adolescents. *Child Psychiatry & Human Development*, *49*(1), 63–72. https://doi.org/10.1007/s10578-017-0729-9

Hermans, K., Achterhof, R., Myin-Germeys, I., Kasanova, Z., Kirtley, O., & Schneider, M. (2019). Improving ecological validity in research on social cognition. In K.E. Lewandowski & A.A. Moustafa (Eds.), *Social Cognition in Psychosis* (pp. 249–268). Academic Press. https://doi.org/10.1016/b978-0-12-815315-4.00010-0

Hermans, K. S. F. M., Myin-Germeys, I., Gayer-Anderson, C., Kempton, M. J., Valmaggia, L., McGuire, P., Murray, R. M., Garety, P., Wykes, T., Morgan, C., Kasanova, Z., & Reininghaus, U. (2021). Elucidating negative symptoms in the daily life of individuals in the early stages of psychosis. *Psychological Medicine*, *51*(15), 2599–2609. https://doi.org/10.1017/S0033291720001154

Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: Incorporating mobile technology into psychosocial and health behaviour treatments. *British Journal of Health Psychology*, *15*(1), 1–39. https://doi.org/10.1348/135910709X466063

Heron, K. E., & Smyth, J. M. (2013). Is intensive measurement of body image reactive? A two-study evaluation using Ecological Momentary Assessment suggests not. *Body Image*, *10*(1), 35–44. https://doi.org/10.1016/j.bodyim.2012.08.006

Hillbrand, M., & Waite, B. M. (1994). The everyday experience of an institutionalized sex offender: An idiographic application of the experience sampling method. *Archives of Sexual Behavior*, *23*(4), 453–463. https://doi.org/10.1007/BF01541409

Himmelstein, P. H., Woods, W. C., & Wright, A. G. C. (2019). A comparison of signal- and event-contingent ambulatory assessment of interpersonal behaviour and affect in social situations. *Psychological Assessment*, *31*(7), 952–960. https://doi.org/10.1037/pas0000718

Hoelscher, E. C., Victor, S. E., Kiekens, G., & Ammerman, B. (2025). Ethical considerations for the use of ecological momentary assessment in non-suicidal self-injury research. *Ethics & Behaviour*, *35*(8), 593–610. https://doi.org/10.1080/10508422.2025.2456714

Hoemann, K., Khan, Z., Feldman, M. J., Nielson, C., Devlin, M., Dy, J., Barrett, L. F., Wormwood, J. B., & Quigley, K. S. (2020). Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific Reports*, *10*(1), 12459. https://doi.org/10.1038/s41598-020-69180-y

Hoemann, K., Khan, Z., Kamona, N., Dy, J., Barrett, L. F., & Quigley, K. S. (2021). Investigating the relationship between emotional granularity and cardiorespiratory physiological activity in daily life. *Psychophysiology*, *58*(6), e13818. https://doi.org/10.1111/psyp.13818

Hoemann, K., Lee, Y., Kuppens, P., Gendron, M., & Boyd, R. L. (2023). Emotional granularity is associated with daily experiential diversity. *Affective Science*, 4(2), 291–306. https://doi.org/10.1007/s42761-023-00185-2

Hoffman, L. B., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*(2–3), 97–120. https://doi.org/10.1080/15427600902911189

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Hogenelst, K., Soeter, M., & Kallen, V. (2019). Ambulatory measurement of cortisol: Where do we stand, and which way to follow? *Sensing and Bio-Sensing Research*, *22*, 100249. https://doi.org/10.1016/j.sbsr.2018.100249

Hollenstein, T. (2021). Affect dynamics and time scales: Pictures of movies. In C. E. Waugh & P. Kuppens (Eds.), *Affect dynamics* (pp. 117–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-82965-0_6

Hong, J. I. (2023). Designing for privacy in mobile sensing systems. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications* (pp. 25–42). The Guilford Press.

Hopwood, C. J., Bleidorn, W., & Wright, A. G. C. (2022). Connecting theory to methods in longitudinal research. *Perspectives on Psychological Science*, *17*(3), 884–894. https://doi.org/10.1177/17456916211008407

Hormuth, S. E. (1986). The sampling of experiences in situ. *Journal of Personality*, *54*, 262–293. https://doi.org/10.1111/j.1467-6494.1986.tb00395.x

Houben, M., Claes, L., Vansteelandt, K., Berens, A., Sleuwaegen, E., & Kuppens, P. (2017). The emotion regulation function of nonsuicidal self-injury: A momentary assessment study in inpatients with borderline personality disorder features. *Journal of Abnormal Psychology*, *126*(1), 89–95. https://doi.org/10.1037/abn0000229

Houben, M., & Kuppens, P. (2019). Emotion dynamics and the association with depressive features and borderline personality disorder traits: Unique, specific, and prospective relationships. *Clinical Psychological Science*, *8*(2), 226–239. https://doi.org/10.1177/2167702619871962

Houben, M., Mestdagh, M., Dejonckheere, E., Obbels, J., Sienaert, P., van Roy, J., & Kuppens, P. (2021). The statistical specificity of emotion dynamics in borderline personality disorder. *Journal of Personality Disorders*. *35*(6), 819–840. https://doi.org/10.1521/pedi_2021_35_509

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Hsieh, Gary & Li, I. & Forlizzi, J. & Hudson, S. (2008). Using visualizations to increase compliance in experience sampling. In *UbiComp 2008—Proceedings of the 10th International Conference on Ubiquitous Computing* (pp. 164–167). https://doi.org/10.1145/1409635.1409657

Hurlburt, R. T. (1993). *Sampling inner experience in disturbed affect*: Springer US. https://doi.org/10.1007/978-1-4899-1222-0

Husky, M., Olie, E., Guillaume, S., Genty, C., Swendsen, J., & Courtet, P. (2014). Feasibility and validity of ecological momentary assessment in the investigation of suicide risk. *Psychiatry Research*, *220*(1–2), 564–570. https://doi.org/10.1016/j.psychres.2014.08.019

Huynh, M. Q., Ghimire, P., & Truong, D. (2017). Hybrid app approach: Could it mark the end of native app domination? *Issues in Informing Science and Information Technology*, *14*, 49–65. https://doi.org/10.28945/3723

Iida, M., Savord, A., & Ledermann, T. (2023). Dyadic longitudinal models: A critical review. *Personal Relationships*, *30*(2), 356–378. https://doi.org/10.1111/pere.12468

Iida, M., Seidman, G., & Shrout, P. E. (2018). Models of interdependent individuals versus dyadic processes in relationship research. *Journal of Social and Personal Relationships*, *35*(1), 59–88. https://doi.org/10.1177/0265407517725407

Impett, E. A., Strachman, A., Finkel, E. J., & Gable, S. L. (2008). Maintaining sexual desire in intimate relationships: The importance of approach goals. *Journal of Personality and Social Psychology*, *94*(5), 808–823. https://doi.org/10.1037/0022-3514.94.5.808

Ingram, R. E., & Siegle, G. J. (2009). Methodological issues in the study of depression. In I. H. Gotlib & C. L. Hammen (Eds.), *Handbook of depression* (pp. 69–92). The Guilford Press.

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behaviour. *Jama*, *318*(13), 1215–1216. https://doi.org/10.1001/jama.2017.11295

Insel, T. R. (2018). Digital phenotyping: A global tool for psychiatry. *World Psychiatry*, *17*(3), 276–277. https://doi.org/10.1002/wps.20550

International Telecommunication Union (ITU) (2025, February 19). Facts and figures 2023: Mobile phone ownership. https://www.itu.int/itu-d/reports/statistics/2023/10/10/ff23-mobile-phone-ownership/

Jacobs, N., Nicolson, N. A., Derom, C., Delespaul, P., van Os, J., & Myin-Germeys, I. (2005). Electronic monitoring of salivary cortisol sampling compliance in daily life. *Life Sciences*, *76*(21), 2431–2443. https://doi.org/10.1016/j.lfs.2004.10.045

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

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, *13*(4), 354–375. https://doi.org/10.1037/a0014173

Janssen, L. H. C., Verkuil, B., van Houtum, L. A. E. M., Wever, M. C. M., & Elzinga, B. M. (2021). Perceptions of parenting in daily life: Adolescent-parent differences and associations with adolescent Affect. *Journal of Youth and Adolescence*, *50*(12), 2427–2443. https://doi.org/10.1007/s10964-021-01489-x

Janssen, L. H. C., Verkuil, B., Nedderhoff, A., van Houtum, L. A. E. M., Wever, M. C. M., & Elzinga, B. M. (2024). Tracking real-time proximity in daily life: A new tool to examine social interactions. *Behaviour Research Methods*, *56*, 7482–7497. https://doi.org/10.3758/s13428-024-02432-1

Janssens, J. J., Kiekens, G., Jaeken, M., & Kirtley, O. J. (2024). A systematic review of interpersonal processes and their measurement within experience sampling studies of self-injurious thoughts and behaviours. *Clinical Psychology Review*, *113*, 102467. https://doi.org/10.1016/j.cpr.2024.102467

Janssens, J. J., Myin-Germeys, I., Lafit, G., Achterhof, R., Hagemann, N., Hermans, K. S. F. M., Hiekkaranta, A. P., Lecei, A., & Kirtley, O. J. (2023). Lifetime and current self-harm thoughts and behaviours and their relationship to parent and peer attachment. *Crisis-the Journal of Crisis Intervention and Suicide Prevention*, *44*(5), 424–432. https://doi.org/10.1027/0227-5910/a000878

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

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*, *27*(6), 958–998. https://doi.org/10.1037/met0000312

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

Jongerling, J., Laurenceau, J. P., & Hamaker, E. L. (2015). A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*(3), 334–349. https://doi.org/10.1080/00273171.2014.1003772

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, *306*(5702), 1776–1780. https://doi.org/10.1126/science.1103572

Kalokerinos, E. K., Erbas, Y., Ceulemans, E., & Kuppens, P. (2019). Differentiate to regulate: Low negative emotion differentiation is associated with ineffective use but not selection of emotion-regulation strategies. *Psychological Science*, *30*(6), 863–879. https://doi.org/10.1177/0956797619838763

Kalokerinos, E. K., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., Gleeson, J., Thompson, R. J., Van Rijjckeghem, D. M. L., Kuppens, P., & Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proceedings National Academy of Sciences USA*, *117*(17), 9270–9276. https://doi.org/10.1073/pnas.1919934117

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science*, *18*(7), 614–621. https://doi.org/10.1111/j.1467- 9280.2007.01948.x

Kaongoen, N., Choi, J., Choi, J. W., Kwon, H., Hwang, C., Hwang, G., Kim, B. H., & Jo, S. (2023). The future of wearable EEG: A review of ear-EEG technology and its applications. *Journal of Neural Engineering*, *20*, 051002. https://doi.org/10.1088/1741-2552/acfcda

Karthick, S., & Binu, S. (2017). Android security issues and solutions. In *International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 686–689). https://doi.org/ 10.1109/ICIMIA.2017.7975551

Kasanova, Z., Hajduk, M., Thewissen, V., & Myin-Germeys, I. (2020). Temporal associations between sleep quality and paranoia across the paranoia continuum: An experience sampling study. *Journal of Abnormal Psychology*, *129*(1), 122–130. https://doi.org/10.1037/abn0000453

Kasanova, Z., Oorschot, M., & Myin-Germeys, I. (2018). Social anhedonia and asociality in psychosis revisited: An experience sampling study. *Psychiatry Research*, *270*, 375–381. https://doi.org/10.1016/j.psychres.2018.09.057

Kazdin, A. E. (2021). *Research design in clinical psychology* (5th ed.). Cambridge University Press. https://doi.org/10.1017/9781108993647

Keijsers, L., Boele, S., & Bülow, A. (2022). Measuring parent–adolescent interactions in natural habitats. The potential, status, and challenges of ecological momentary assessment. *Current Opinion in Psychology*, *44*, 264–269. https://doi.org/10.1016/j.copsyc.2021.10.002

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. The Guilford Press.

Kerst, A., Zielasek, J., & Gaebel, W. (2020). Smartphone applications for depression: A systematic literature review and a survey of health care professionals' attitudes towards their use in clinical practice. *European Archives of Psychiatry and Clinical Neuroscience*, *270*(2), 139–152. https://doi.org/10.1007/s00406-018-0974-3

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Kiekens, G., Claes, L., Kleiman, E. M., Luyckx, K., Coppersmith, D. D. L., Fortgang, R. G., Myin-Germeys, I., & Nock, M. K. (2024). The short-term course of non-suicidal self-injury among individuals seeking psychiatric treatment. *The Journal of the American Medical Association Network Open*, *7*(10), e2440510. https://doi.org/10.1001/jamanetworkopen.2024.40510

Kiekens, G., Claes, L., Schoefs, S., Kemme, N. D. F., Luyckx, K., Kleiman, E. M., Nock, M. K., & Myin-Germeys, I. (2023). The detection of acute risk of self-injury project: Protocol for an ecological momentary assessment study among individuals seeking treatment. *Journal of Medical Internet Research Research Protocol*, *12*, e46244. https://doi.org/10.2196/46244

Kiekens, G., Hasking, P., Nock, M. K., Boyes, M., Kirtley, O., Bruffaerts, R., Myin-Germeys, I., & Claes, L. (2020). Fluctuations in affective states and self-eddicacy to resist non-suicidal self-injury as real-time predictors of non-suicidal self injurious thoughts and behaviors. *Frontiers Psychiatry*, *11*, 214. https://doi.org/10.3389/fpsyt.2020.00214

Kiekens, G., Robinson, K., Tatnell, R., & Kirtley, O. J. (2021). Opening the black box of daily life in non-suicidal self-injury research: With great opportunity comes great responsibility. *Journal of Medical Internet Research Mental Health*, *8*(11), e30915. https://doi.org/10.31234/osf.io/yp86x

Kimhy, D., Myin-Germeys, I., Palmier-Claus, J., & Swendsen, J. (2012). Mobile assessment guide for research in schizophrenia and severe mental disorders. *Schizophrenia Bulletin*, *38*(3), 386–395. https://doi.org/10.1093/schbul/sbr186

King, K. M., Feil, M. C., Gomez Juarez, N., Moss, D., Halvorson, M. A., Dora, J., Upton, N. F., Bryson, M. A., Seldin, K., Shoda, Y., Lee, C. M., & Smith, G. T. (2025). Negative urgency as a state-level process. *Journal of Personality*, *93*, 529–552. https://doi.org/10.1111/jopy.12961

Kip, H., Beerlage-de Jong, N., van Gemert-Pijnen, L. J. E. W. C., & Kelders, S. M. (2025). The CeHRes roadmap 2.0: Update of a holistic framework for development, implementation, and evaluation of ehealth technologies. *Journal of Medical Internet Research*, 27, e59601. https://doi.org/10.2196/59601

Kirtley, O. J. (2022). Advancing credibility in longitudinal research by implementing open science practices: Opportunities, practical examples, and challenges. *Infant and Child Development*, 31(1), Article e2302. https://doi.org/ARTNe230210.1002/icd.2302

Kirtley, O., Achterhof, R., Hagemann, N., Hermans, K. S. F. M., Hiekkaranta, A. P., Lecei, A., Boets, B., Henquet, C., Kasanova, Z., Schneider, M., van Winel, R., Reininghaus, U., Viechtbauer, W., & Myin-Germeys, I. (2021). Initial cohort characteristics and protocol for SIGMA: An accelerated longitudinal study of environmental factors, inter- and intrapersonal processes, and mental health in adolescence. *preprint*. https://doi.org/10.31234/osf.io/jp2fk

Kirtley, O. J., Claes, S., Schoefs, S., Vermaelen, N., Kemme, N. D. F., Portzky, G., & Myin-Germeys, I. (2025, February 19). *Investigating social interactive pathways between exposure to suicide attempts and*

suicidal thoughts and behaviours in young adults: The Social Connections amOng yoUth in disTress (SCOUT-Clinical) Project. https://doi.org/10.17605/OSF.IO/AQH4K

Kirtley, O. J., Eisele, G., Kunkels, Y. K., Hiekkaranta, A. P., Van Heck, L., Pihlajamäki, Kunc, B., Schoefs, S., Vermaelen, N., & Myin-Germeys, I. (2025). *The Experience Sampling Method (ESM) Item Repository* 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

Kirtley, O. J., Lafit, G., Wampers, M., & Myin-Germeys, I. (2020, 07/12/20 - 08/12/20). Establishing a low-threshold data checkout system using REDCap to facilitate preregistration and Registered Reports for pre-existing data CSPD 2020: Sharing Psychological Research Data: Best Practices and New Developments, https://www.conference-service.com/CSPD2020/xpage.html?xpage=226&lang=en

Kirtley, O. J., Sohier, B., Šimsa, B., Achterhof, R., Myin-Germeys, I., & Lafit, G. (2025). Reactivity to experience sampling among adolescents with and without a lifetime or current history of self-harm thoughts or behaviours. *Psychological Assessment*, *37*(11), 585–598. https://doi.org/10.1037/pas0001375

Kivelä, L. M. M., Fiss, F., van der Does, W., & Antypa, N. (2024). Examination of acceptability, feasibility, and iatrogenic effects of Ecological Momentary Assessment (EMA) of suicidal ideation. *Assessment*, *31*(6), 1292–1308. https://doi.org/10.1177/10731911231216053

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., & Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of Abnormal Psychology*, *126*(6), 726–738. https://doi.org/10.1037/abn0000273

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Picard, R. W., Huffman, J. C., & Nock, M. K. (2018). Digital phenotyping of suicidal thoughts. *Depression and Anxiety*, *35*(7), 601–608. https://doi.org/10.1002/da.22730

Klippel, A., Viechtbauer, W., Reininghaus, U., Wigman, J., van Borkulo, C., MERGE, Myin-Germeys, I., & Wichers, M. (2018). The cascade of stress: A network approach to explore differential dynamics in populations varying in risk for psychosis. *Schizophrenia Bulletin*, *44*(2), 328–337. https://doi.org/10.1093/schbul/sbx037

Koval, P., Kalokerinos, E. K., Greenaway, K. H., Medland, H., Kuppens, P., Nezlek, J. B., Hinton, J. D. X., & Gross, J. J. (2023). Emotion regulation in everyday life: Mapping global self-reports to daily processes. *Emotion*, *23*, 357–374. https://doi.org/10.1037/emo0001097

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

Koval, P., & Kuppens, P. (2024). Changing feelings: Individual differences in emotional inertia. In A. C. Samson, D. Sander, & U. Kramer (Eds.), *Change in emotion and mental health* (pp. 3–21). Academic Press. https://doi.org/10.1016/b978-0-323-95604-8.00007-1

Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, *13*(6), 1132–1141. https://doi.org/10.1037/a0033579

Koval, P., Sutterlin, S., & Kuppens, P. (2015). Emotional inertia is associated with lower well-being when controlling for differences in emotional context. *Frontiers in Psychology*, *6*, 1997. https://doi.org/10.3389/fpsyg.2015.01997

Kramer, I., Simons, C. J., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., Schruers, K., van Bemmel, A. L., Myin-Germeys, I., Delespaul, P., van Os, J., & Wichers, M. (2014). A therapeutic

application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry*, *13*(1), 68–77. https://doi.org/10.1002/wps.20090

Kratz, A. L., Ehde, D. M., Bombardier, C. H., Kalpakjian, C. Z., & Hanks, R. A. (2017). Pain acceptance decouples the momentary associations between pain, pain interference, and physical activity in the daily lives of people with chronic pain and spinal cord injury. *The Journal of Pain*, *18*(3), 319–331. https://doi.org/10.1016/j.jpain.2016.11.006

Kuehn, K. S., Dora, J., Harned, M. S., Foster, K. T., Song, F., Smith, M. R., & King, K. M. (2022). A meta-analysis on the affect regulation function of real-time self-injurious thoughts and behaviours. *Nature Human Behaviour*, *6*(7), 964–974. https://doi.org/10.1038/s41562-022-01340-8

Kuppens, P. (2019). Improving theory, measurement, and reality to advance the future of emotion research. *Cognition and Emotion*, *33*(1), 20–23. https://doi.org/10.1080/02699931.2018.1536037

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

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

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Current Opinion Psychology*, *17*, 22–26. https://doi.org/10.1016/j.copsyc.2017.06.004

Kusov, P. A., Kotelevtsev, Y. V., & Drachev, V. P. (2023). Cortisol monitoring devices toward implementation for clinically relevant biosensing in vivo. *Molecules*, *28*(5), 2353. https://doi.org/10.3390/molecules28052353

Lachowicz, A., Houben, M., Ottaviani, C., Van Diest, I., Wampers, M., Cornelis, J., Myin-Germeys, I., & Vaessen, T. (2026). Efficacy of slow-paced breathing as a just-in-time adaptive intervention for anxiety: A randomized controlled study. Applied Psychopathology and Biofeedback, https://doi.org/10.1007/s10484-025-09760-8

Lachowicz, A. M., Lafit, G., Ottaviani, C., Van Diest, I., Cornelis, J., Myin-Germeys, I., Vaessen, T. & Houben, M. (forthcoming). *The role of perseverative cognition in recovery from daily stress in subclinical anxiety: A within-person moderated-mediation analysis*.

Lachowicz, A. M., Vaessen, T., Lafit, G., Achterhof, R., Akcaoglu, Z., Bamps, E., Hagemann, N., Hiekkaranta, A. P., Hermans, K. S. F. M., Janssens, J. J., Lecei, A., Kirtley, O., Myin-Germeys, I., & Houben, M. (2025). The link between delayed affective recovery from daily stressors and anxiety symptoms in youth. *International Journal of Stress Management*, *32*(2), 151–162. https://doi.org/10.1037/str0000347

Lafit, G., Adolf, J., 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 to perform 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

Lafit, G., Artner, R., & Ceulemans, E. (2024). Enabling analytical power calculations for multilevel models with autocorrelated errors through deriving and approximating the precision matrix. *Behavior Research Methods*, *56*(7), 8105–8131. https://doi.org/10.3758/s13428-024-02435-y

Lafit, G., Revol, J., Cloos, L., Kuppens, P., & Ceulemans, E. (2025). The effect of different construct operationalizations, study duration, and preprocessing choices on power-based sample size recommendations in intensive longitudinal research. *Assessment*, *32*(2). 206–223 https://doi.org/10.1177/10731911241286868

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

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974. https://doi.org/10.2307/2529876

Lamont, A. (2008). Young children's musical worlds: Musical engagement in 3.5-year-olds. *Journal of Early Childhood Research*, *6*, 247–261. https://doi.org/10.1177/1476718X08094449

Landau, S., & Stahl, D. (2013). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research*, *22*(3), 324–345. https://doi.org/10.1177/0962280212439578

Landes, S. J., McBain, S. A., & Curran, G. M. (2019). An introduction to effectiveness-implementation hybrid designs. *Psychiatry Research*, *280*, 112513. https://doi.org/10.1016/j.psychres.2019.112513

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationship*, *35*(1), 7–31. https://doi.org/10.1177/0265407517710342

Langener, A. M., Bringmann, L. F., Kas, M. J., & Stulp, G. (2024). Predicting mood based on the social context measured through the experience sampling method, digital phenotyping, and social networks. *Administration and Policy in Mental Health and Mental Health Services Research*, *51*(4), 455–475. https://doi.org/10.1007/s10488-023-01328-0

Langener, A. M., Siepe, B. S., Elsherif, M., Niemeijer, K., Andresen, P. K., Akre, S., Bringmann, L. F., Cohen, Z. D., Choukas, N. R., Drexl, K., Fassi, L., Green, J., Hoffmann, T., Jagesar, R. R., Kas, M. J. H., Kurten S., Schoedel, R., Stulp. G., Turner, G., & Jacobson, N. C. (2024). A template and tutorial for preregistering studies using passive smartphone measures. *Behaviour Research Methods*, *56*(8), 8289–8307. https://doi.org/10.3758/s13428-024-02474-5

Langener, A. M., Stulp, G., Jacobson, N. C., Costanzo, A., Jagesar, R. R., Kas, M. J., & Bringmann, L. F. (2024). It's all about timing: Exploring different temporal resolutions for analysing digital-phenotyping data. *Advances in Methods and Practices in Psychological Science*, *7*(1), 1–22. https://doi.org/10.1177/25152459231202677

Lathia, N., Sandstrom, G. M., Mascolo, C., & Rentfrow, P. J. (2017). Happier people live more active lives: Using smartphones to link happiness and physical activity. *PLOS ONE*, *12*(1), e0160589. https://doi.org/10.1371/journal.pone.0160589

Laurenceau, J.-P., & Bolger, N. (2012). Analysing diary and intensive longitudinal data from dyads. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 407–422). The Guilford Press.

Law, M. K., Furr, R. M., Arnold, E. M., Mneimne, M., Jaquett, C., & Fleeson, W. (2015). Does assessing suicidality frequently and repeatedly cause harm? A randomized control study. *Psychological Assessment*, *27*(4), 1171–1181. https://doi.org/10.1037/pas0000118

Leahey, T. M., Crowther, J. H., & Mickelson, K. D. (2007). The frequency, nature, and effects of naturally occurring appearance-focused social comparisons. *Behaviour Therapy*, *38*(2), 132–143. https://doi.org/10.1016/j.beth.2006.06.004

Lei Wang, Lo, B. P., & Yang, Guang-Zhong (2007). Multichannel reflective PPG earpiece sensor with passive motion cancellation. *IEEE Transactions on Biomedical Circuits and Systems*, *1*(4), 235–241. https://doi.org/10.1109/tbcas.2007.910900

Levine, L. J., & Safer, M. A. (2002). Sources of bias in memory for emotions. *Current Directions in Psychological Science*, *11*(5), 169–173. https://doi.org/10.1111/1467-8721.00193

Li, Y.-M., Konstabel, K., Möttus, R., & Lemola, S. (2022). Temporal associations between objectively measured physical activity and depressive symptoms: An experience sampling study. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.920580

Lin, S., Wu, X., Martinez, G., & Chawla, N. V. (2020). Filling missing values on wearable-sensory time series data. In *Proceedings of the 2020 SIAM International Conference on Data Mining* (pp. 46–54). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611976236.6

Linardon, J., Torous, J., Firth, J., Cuijpers, P., Messer, M., & Fuller-Tyszkiewicz, M. (2024). Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. A meta-analysis of 176 randomized controlled trials. *World Psychiatry*, *23*, 139–149. https://doi.org/10.1002/wps.21183

Littlewood, D. L., Kyle, S. D., Carter, L. A., Peters, S., Pratt, D., & Gooding, P. (2019). Short sleep duration and poor sleep quality predict next-day suicidal ideation: An ecological momentary assessment study. *Psychological Medicine*, *49*(3), 403–411. https://doi.org/10.1017/s0033291718001009

Littmann, L. (2021). Electrocardiographic artifact. *Journal of Electrocardiology*, *64*, 23–29. https://doi.org/10.1016/j.jelectrocard.2020.11.006

Liu, S.-H., Wang, J.-J., Su, C.-H., & Tan, T.-H. (2018). Development of a patch-type electrocardiographic monitor for real time heartbeat detection and heart rate variability analysis. *Journal of Medical and Biological Engineering*, *38*(3), 411–423. https://doi.org/10.1007/s40846-018-0369-y

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The history and philosophy of ecological psychology. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02228

Löchner, J., Santangelo, P. S., Ansell, E., Bolger, N., Ebner-Priemer, U., Fried, E., Gawrilow, C., Hamaker, E., Hepp. J., Kaurin, A., Kirtley, O. J., Kubiak, T., Kuppens, P., Laurenceau, J.-P., Myin-Germeys, I., Neubauer, A., Schneider, S., Schuller, B., Shiffman, S., & Smyth, J. M. (2025, February 3). *Ambulatory assessment in mental health research: Expert consensus on current practices and future directions*. https://doi.org/10.31234/osf.io/mhb5g_v1

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98. https://doi.org/10.1037/0033-2909.116.1.75

Long, J. A. (2020). *jtools: Analysis and presentation of social scientific data. R package version 2.1.0*. https://cran.r- project.org/package=jtools

Lüdecke, D. (2021). *sjstats: Statistical functions for regression models (Version 0.18.1)*. https://CRAN.R-project.org/package=sjstats

Lüdtke, T., Hedelt, K. S., & Westermann, S. (2023). Predictors of paranoia in the daily lives of people with non-affective psychosis and non-clinical controls: A systematic review of intensive longitudinal studies. *Journal of Behaviour Therapy and Experimental Psychiatry*, *81*, 101885. https://doi.org/10.1016/j.jbtep.2023.101885

Luff, P., & Heath, C. (2012). Some 'technical challenges' of video analysis: Social actions, objects, material realities and the problems of perspective. *Qualitative Research*, *12*(3), 255–279. https://doi.org/10.1177/1468794112436655

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioural and Social Sciences*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Magno, M., Salvatore, G. A., Jokic, P., & Benini, L. (2019). Self-sustainable smart ring for long-term monitoring of blood oxygenation. *IEEE Access*, *7*, 115400–115408. https//doi.org/10.1109/ACCESS.2019.2928055

Maher, J. P., Rebar, A. L., & Dunton, G. F. (2018). Ecological momentary assessment is a feasible and valid methodological tool to measure older adults' physical activity and sedentary behaviour. *Frontiers in Psychology*, *9*, 1485. https://doi.org/10.3389/fpsyg.2018.01485

Mansour, M., Darweesh, M. S., & Soltan, A. (2024). Wearable devices for glucose monitoring: A review of state-of-the-art technologies and emerging trends. *Alexandria Engineering Journal*, *89*, 224–243. https://doi.org/10.1016/j.aej.2024.01.021

Marelli, L., Lievevrouw, E., & Van Hoyweghen, I. (2020). Fit for purpose? The GDPR and the governance of European digital health. *Policy Studies*, *41*(5), 447–467. https://doi.org/10.1080/01442872.2020.1724929

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., de Girolamo, G., Difrancesco, S., Haro, J. M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardinie, F., Mohr, D. C., Narayan, V. A., Oetzmann, C., Penninx, B. W. J. H., Bruce, S., Nica, R., … Hotopf, M., (2022). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC Psychiatry*, *22*(1), 136. https://doi.org/10.1186/s12888-022-03753-1

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

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*(5), 951–966. https://doi.org/10.1037/a0028380

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McCabe, K. O., Mack, L., & Fleeson, W. (2012). A guide for data cleaning in experience sampling studies. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 321–338): The Guilford Press.

McCulloch, C. E., & Neuhaus, J. M. (2014). *Generalized linear mixed models*. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat07540

McDevitt-Murphy, M. E., Luciano, M. T., & Zakarian, R. J. (2018). Use of ecological momentary assessment and intervention in treatment with adults. *Focus (American Psychiatry Publishing)*, *16*(4), 370–375. https://doi.org/10.1176/appi.focus.20180017

McLean, D. C., Nakamura, J., & Csikszentmihalyi, M. (2017). Explaining system missing: Missing data and experience sampling method. *Social Psychological and Personality Science*, *8*, 434. https://doi.org/10.1177/1948550617708015

McNeish, D. (2022). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behaviour Research Methods*, *55*, 4269–4290. https://doi.org/10.3758/s13428-022-02016-x

McNeish, D., & Hamaker, E. L. (2019). A primer on two-level dynamic structural equation models for intensive longitudinal data in M*plus*. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meegahapola, L., & Gatica-Perez, D. (2021). Smartphone sensing for the well-being of young adults: A review. *IEEE Access*, *9*, 3374–3399. https://doi.org/10.1109/ACCESS.2020.3045935

Meers, K., Dejonckheere, E., Kalokerinos, E. K., Rummens, K., & Kuppens, P. (2020). mobileQ: A free user-friendly application for collecting experience sampling data. *Behavior Research Methods*, *52*(4), 1510–1515. https://doi.org/10.3758/s13428-019-01330-1

Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A method for the naturalistic observation of daily social behaviour. *Current Directions in Psychology Science*, *26*(2), 184–190. https://doi.org/10.1177/0963721416680611

Mehl, M. R., Eid, M., Wrzus, C., Harari, G. M., & Ebner-Priemer, U. (2021). *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

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kilani, G., Tuerlinckx, F., Kuppens, P., & Dejonckheere E. (2023). m-Path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioural research and clinical practice. *Frontiers in Digital Health*, *5*, 1182175. https://doi.org/10.3389/fdgth.2023.1182175

Miralles, I., Granell, C., Díaz-Sanahuja, L., Van Woensel, W., Bretón-López, J., Mira, A., Castilla, D., & Casteleyn, S. (2020). Smartphone apps for the treatment of mental disorders: Systematic review. *Journal of Medical Internet Research Mhealth Uhealth*, *8*(4), e14897. https://doi.org/10.2196/14897

Moerbeek, M. (2011). The effects of the number of cohorts, degree of overlap among cohorts, and frequency of observation on power in accelerated longitudinal designs. *Methodology: European Journal of Research Methods for the Behavioural and Social Sciences*, *7*(1), 11–24. https://doi.org/10.1027/1614-2241/a000019

Moerbeek, M., & Maas, C. J. M. (2005). Optimal experimental designs for multilevel logistic models with two binary predictors. *Communications in Statistics – Theory and Methods*, *34*(5), 1151–1167. https://doi.org/10.1081/STA-200056839

Moerbeek, M., & van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioural Statistics*, *25*(3), 271–284. https://doi.org/10.3102/10769986025003271

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *50*(1), 17–30. https://doi.org/10.1111/1467-9884.00257

Mohr, D. C., Shilton, K., & Hotopf, M. (2020). Digital phenotyping, behavioural sensing, or personal sensing: Names and transparency in the digital age. *Nature Partner Journal Digital Medicine*, *3*, 45. https://doi.org/10.1038/s41746-020-0251-5

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

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

Molenberghs, G., & Verbeke, G. (2006). *Models for discrete longitudinal data*: Springer Science & Business Media. https://doi.org/10.1007/0-387-28980-1

Monk, T. H., Houck, P. R., & Shear, M. K. (2006). The daily life of complicated grief patients: What gets missed, what gets added? *Death Studies*, *30*(1), 77–85. https://doi.org/10.1080/07481180500348860

Montgomery, D. C. (2009). *Statistical quality control* (8th ed., Vol. 7). Wiley.

Moore, E., Williams, A., Bell, I., & Thomas, N. (2020). Client experiences of blending a coping-focused therapy for auditory verbal hallucinations with smartphone-based ecological momentary assessment and intervention. *Internet Interventions*, *19*, 100299. https://doi.org/10.1016/j.invent.2019.100299

Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., Moore, L., O'Cathain, A., Tinati, T., Wight, D., & Baird, J. (2015). Process evaluation of complex interventions: Medical research council guidance. *British Medical Journal*, *350*, h1258. https://doi.org/10.1136/bmj.h1258

Morren, M., van Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: A systematic review. *European Journal of Pain*, *13*(4), 354–365. https://doi.org/10.1016/j.ejpain.2008.05.010

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Munsch, S., Meyer, A. H., Milenkovic, N., Schlup, B., Margraf, J., & Wilhelm, F. H. (2009). Ecological momentary assessment to evaluate cognitive-behavioural treatment for binge eating disorder. *International Journal of Eating Disorders*, *42*(7), 648–657. https://doi.org/10.1002/eat.20657

Murray, A. L., Brown, R., Zhu, X., Speyer, L. G., Yang, Y., Xiao, Z., Ribeaud, D., & Eisner, M. (2023). Prompt-level predictors of compliance in an ecological momentary assessment study of young adults' mental health. *Journal of Affective Disorders*, *322*, 125–131. https://doi.org/10.1016/j.jad.2022.11.014

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). Taylor & Francis. https://doi.org/10.4324/9780203848852.ch2

Myin-Germeys, I. (2020). Digital technology in psychiatry: Towards the implementation of a true person-centered care in psychiatry? *European Archives of Psychiatry and Clinical Neuroscience*, *270*(4), 401–402. https://doi.org/10.1007/s00406-020-01130-1

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry*, *17*(2), 123–132. https://doi.org/10.1002/wps.20513

Myin-Germeys, I., Klippel, A., Steinhart, H., & Reininghaus, U. (2016) Ecological momentary interventions in psychiatry. *Current Opinion in Psychiatry*, *29*(4), 258–263. https://doi.org/10.1097/YCO.0000000000000255

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: Opening the black box of daily life. *Psychological Medicine*, *39*(9), 1533–1547. https://doi.org/1 10.1017/S0033291708004947

Myin-Germeys, I., Schick, A., Ganslandt, T., Hajdúk, M., Heretik, A., Van Hoyweghen, I., Kiekens, G., Koppe, G., Marelli, L., & Nagyova I., Weemeijer, J., Wensing, M., Wolters, M., Beames, J., de Allegri, M., di Folco, S., Durstewitz, D., Katreniaková, Lievevrouw, E., Nguyen, H., … Reininghaus, U. (2024). The experience sampling methodology as a digital clinical tool for more person-centered mental health care: An implementation research agenda. *Psychological Medicine*, *54*(11), 2785–2793. https://doi.org/10.1017/s0033291724001454

Myin-Germeys, I., van Aubel, E., Vaessen, T., Steinhart, H., Klippel, A., Lafit, G., Viechtbauer, W., Batink, T., Van Winkel, R., van der Gaag, M., van Amelsvoort, T., Marcelis, M., Schirmbeck, F., de Haan L., & Reininghaus, U. (2022). Efficacy of acceptance and commitment therapy in daily life in early psychosis: Results from the multi-center INTERACT randomized controlled trial. *Psychotherapy Psychosomatics*, *91*(6), 411–423. https://doi.org/10.1159/000522274

Myin-Germeys, I., van Os, J., Schwartz, J. E., Stone, A. A., & Delespaul, P. A. (2001). Emotional reactivity to daily life stress in psychosis. *Archives of General Psychiatry*, *58*(12), 1137–1144. https://doi.org/10.1001/archpsyc.58.12.1137

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-Time Adaptive Interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behaviour support. *Annals of Behavioral Medicine*, *52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

Nahum-Shani, I., Wetter, D. W., & Murphy, S. A. (2023). Chapter 7—Adapting just-in-time interventions to vulnerability and receptivity: Conceptual and methodological considerations. In N. Jacobson, T. Kowatsch, & L. Marsch (Eds.), *Digital therapeutics for mental health and addiction* (pp. 77–87). Academic Press. https://doi.org/10.1016/B978-0-323-90045-4.00012-5

Nelson, B. W., Low, C. A., Jacobson, N., Areán, P., Torous, J., & Allen, N. B. (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioural research. *Nature Partner Journal Digital Medicine*, *3*(1), 90. https://doi.org/10.1038/s41746-020-0297-4

Neubauer, A. B., & Schmiedek, F. (2020). Studying within-person variation and within-person couplings in intensive longitudinal data: Lessons learned and to be learned. *Gerontology*, *66*(4), 332–339. https://doi.org/10.1159/000507993

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

Nicholas, J., Fogarty, A. S., Boydell, K., & Christensen, H. (2017). The reviews are in: A qualitative content analysis of consumer perspectives on apps for bipolar disorder. *Journal of Medicine Internet Research*, *19*(4), e105. https://doi.org/10.2196/jmir.7273

Niemeijer, K., & Kuppens, P. (2024). Emotion detection with mobile sensing. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications* (pp. 561–580). The Guilford Press.

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. *Journal of Medical Internet Research Formative Research*, *7*(1), e43296. https://doi.org/10.2196/43296

Nock, M. K., Kleiman, E. M., Abraham, M., Bentley, K. H., Brent, D. A., Buonopane, R. J., Castro-Ramirez, F., Cha, C. B., Demsey W., Draper, J., Glenn, C. R., Harkavy-Friedman, J., Hollander, M. R., Huffman, J. C., Lee H. S, Millner, A. J. Mou, D., Onnela J.-P., Picard, R. W., Quay, H. M., … Pearson, J. L. (2021). Consensus statement on ethical & safety practices for conducting digital monitoring studies with people at risk of suicide and related behaviours. *Psychiatric Research & Clinical Practice*, *3*, 2. https://doi.org/10.1176/appi

Nusbaum, E. C., Silvia, P. J., Beaty, R. E., Burgin, C. J., Hodges, D. A., & Kwapil, T. R. (2014). Listening between the notes: Aesthetic chills in everyday music listening. *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 104–109. https://doi.org/10.1037/a0034867

Nutt, D., Wilson, S., & Paterson, L. (2008). Sleep disorders as core symptoms of depression. *Dialogues in Clinical Neuroscience*, *10*(3), 329–336. https://doi.org/10.31887/dcns.2008.10.3/dnutt

Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal*, *332*(7538), 413–416. https://doi.org/10.1136/bmj.332.7538.413

OJ L 117. *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. 2017 May 05*. https://eur-lex.europa.eu/legal- content/EN/TXT/?uri=CELEX%3A32017R0745

Ok, J., Park, S., Jung, Y. H., & Kim, T. I. (2024). Wearable and implantable cortisol-sensing electronics for stress monitoring. *Advanced Materials*, *36*(1), 2211595. https://doi.org/10.1002/adma.202211595

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. *Journal of Medicine Internet Research*, *21*(2), e11398. https://doi.org/10.2196/11398

Oosterwegel, A., Field, N., Hart, D., & Anderson, K. (2001). The relation of self-esteem variability to emotion variability, mood, personality traits, and depressive tendencies. *Journal of Personality*, *69*(5), 689–708. https://doi.org/10.1111/1467- 6494.695160

Ottenstein, C., & Werner, L. (2022). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

Paetzold, I., Schick, A., Rauschenberg, C., Hirjak, D., Banaschewski, T., Meyer-Lindenberg, A., Boehnke, J. R., Boecking, B., & Reininghaus, U. (2023). Exploring putative therapeutic mechanisms of change in a hybrid compassion-focused, ecological momentary intervention: Findings from the emicompass trial. *Behaviour Research and Therapy*, *168*, 104367. https://doi.org/10.1016/j.brat.2023.104367

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

Panaite, V., Rottenberg, J., & Bylsma, L. M. (2020). Daily affective dynamics predict depression symptom trajectories among adults with major and minor depression. *Affective Science*, *1*(3), 186–198. https://doi.org/10.1007/s42761-020-00014-w

Pannicke, B., Reichenberger, J., Schultchen, D., Pollatos, O., & Blechert, J. (2020). Affect improvements and measurement concordance between a subjective and an accelerometric estimate of physical activity. *European Journal of Health Psychology*, *27*(2), 66–75. https://doi.org/10.1027/2512-8442/a000050

Papageorgiou, A., Strigkos, M., Politou, E., Alepis, E., Solanas, A., & Patsakis, C. (2018). Security and privacy analysis of mobile health applications: The alarming state of practice. *Ieee Access*, *6*, 9390–9403. https://doi.org/10.1109/ACCESS.2018.2799522

Papoutsi, C., & Greenhalgh, T. (2024). Innovation, improvement, and implementation: Conceptual frameworks for thinking through complex change. In H. Kip, N. Beerlage-de Jong, L. van Gemert-Pijnen, R. Sanderman, & S. M. Kelders (Eds.), *eHealth research theory and development* (pp. 220–236). Routledge. https://doi.org/10.4324/9781003302049-15

Park, M., Thom, J., Mennicken, S., Cramer, H., & Macy, M. (2019). Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nature Human Behaviour*, *3*(3), 230–236. https://doi.org/10.1038/s41562-018- 0508-z

Passler, S., Müller, N., & Senner, V. (2019). In-ear pulse rate measurement: A valid alternative to heart rate derived from electrocardiography?. *Sensors*, *19*(17), 3641. https://doi.org/10.3390/s19173641

Pe, M. L., Koval, P., & Kuppens, P. (2013). Executive well-being: Updating of positive stimuli in working memory is associated with subjective well-being. *Cognition*, *126*(2), 335–340. https://doi.org/10.1016/j.cognition.2012.10.002

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion*, *6*(3), 383–391. https://doi.org/10.1037/1528-3542.6.3.383

Pew Research Center. (2019). *Mobile fact sheet*. https://www.pewresearch.org/internet/fact-sheet/mobile/

Pew Research Center. (2025, February 19). *Who is smartphone dependent*. https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Pew Research Center. (2020). *About one-in-five Americans use a smart watch or fitness tracker*. https://www.pewresearch.org/

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., & Maintainer, R. (2017). *Package 'nlme'. Linear and nonlinear mixed effects models, R package version 3.1-149*. https://CRAN.R-project.org/package=nlme

Pinheiro, J. C., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer.

Piot, M., Mestdagh, M., Riese, H., Weermeijer, J., Brouwer, J. M. A., Kuppens, P., Dejonckheere, E., & Bos, F. M. (2022). Practitioner and researcher perspectives on the utility of ecological momentary assessment in mental health care: A survey study. *Internet Interventions*, *30*, 100575. https://doi.org/10.1016/j.invent.2022.100575

Postma, M. R., Vrancken, S., Daemen, M., Hoes-van der Meulen, I., Volbragt, N., Delespaul, P., de Haan, L., van der Pluijm, M., Breedvelt, J. J. F., van der Gaag, M., Lindauer, R., van den Berg, D., Bockting, C., VAN Amelsvoort, T., Schwannauer, M., Doi, L., & Reininghaus, U. (2024). Working mechanisms of the use and acceptability of ecological momentary interventions: A realist evaluation of a guided self-help ecological momentary intervention targeting self-esteem. *BMC Public Health*, *24*(1), 1633. https://doi.org/10.1186/s12889-024-19143-z

Poulton, A., Pan, J., Bruns, L. R., Jr., Sinnott, R. O., & Hester, R. (2019). A smartphone app to assess alcohol consumption behaviour: Development, compliance, and reactivity. *Journal of Medical Internet Research Mhealth Uhealth*, *7*(3), e11157. https://doi.org/10.2196/11157

Provenzano, J., Bastiaansen, J. A., Verduyn, P., Oldehinkel, A. J., Fossati, P., & Kuppens, P. (2018). Different aspects of the neural response to socio-emotional events are related to instability and inertia of emotional experience in daily life: An fMRI-ESM study. *Frontiers in Human Neuroscience*, *12*, 501. https://doi.org/10.3389/fnhum.2018.00501

Quinlivan, L., Cooper, J., Meehan, D., Longson, D., Potokar, J., Hulme, T., Marsden, J., Brand, F., Lange, K., Riseborough, E., Page, L., Metcalfe, C., Davies, L., O'Connor, R., Hawton, K., Gunnell, D., & Kapur, N. (2017). Predictive accuracy of risk scales following self-harm: Multicentre, prospective cohort study. *The British Journal of Psychiatry: The Journal of Mental Science*, *210*(6), 429–436. https://doi.org/10.1192/bjp.bp.116.189993

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. Stata Press.

Rah, M. J., Walline, J. J., Lynn Mitchell, G., & Zadnik, K. (2006). Comparison of the experience sampling method and questionnaires to assess visual activities in pre-teen and adolescent children. *Ophthalmic and Physiological Optics*, *26*(5), 483–489. https://doi.org/10.1111/j.1475-1313.2006.00372.x

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Research in Human Development*, *14*(3), 253–270. https://doi.org/10.1080/15427609.2017.1340052

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods 2*(2), 173–185. https://doi.org/10.1037/1082-989X.2.2.173

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

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401. https://doi.org/10.1037/1082-989x.6.4.387

Rauers, A., Blanke, E., & Riediger, M. (2013). Everyday empathic accuracy in younger and older couples: Do you need to see your partner to know his or her feelings? *Psychological Science*, *24*(11), 2210–2217. https://doi.org/10.1177/0956797613490747

Rauschenberg, C., Boecking, B., Paetzold, I., Schruers, K., Schick, A., van Amelsvoort, T., & Reininghaus, U. (2021). A compassion-focused ecological momentary intervention for enhancing resilience in

help-seeking youth: Uncontrolled pilot study. *Journal of Medical Internet Research Mental Health*, *8*(8), e25650. https://doi.org/10.2196/25650

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Taylor & Francis. https://doi.org/10.4324/9780203841624

Reichenberger, J., Kuppens, P., Liedlgruber, M., Wilhelm, F. H., Tiefengrabner, M., Ginzinger, S., & Blechert, J. (2018). No haste, more taste: An EMA study of the effects of stress, negative and positive emotions on eating behaviour. *Biological Psychology*, *131*, 54–62. https://doi.org/10.1016/j.biopsycho.2016.09.002

Reininghaus, U. (2018). [Ecological momentary interventions in psychiatry: The momentum for change in daily social context]. *Psychiatrische Praxis*, *45*(2), 59–61. https://doi.org/10.1055/s-0044-101986 (Ambulatorische Interventionen in der Psychiatrie: das Momentum für Veränderung im alltäglichen sozialen Kontext.)

Reininghaus, U., Daemen, M., Postma, M. R., Schick, A., Hoes-van der Meulen, I., Volbragt, N., Nieman, D., Delespaul, P., de Haan, L., van der Pluijm, M., Breedvelt, J. J. F., van der Gaag, M., Lindauer, R., Boehnke, J. R., Viechtbauer, W., van den Berg, D., Bockting, C., & van Amelsvoort, T. (2024). Transdiagnostic ecological momentary intervention for improving self-esteem in youth exposed to childhood adversity: The SELFIE randomized clinical trial. *The Journal of the American Medical Association Psychiatry*, *81*(3), 227–239. https://doi.org/10.1001/jamapsychiatry.2023.4590

Reininghaus, U., Depp, C. A., & Myin-Germeys, I. (2016). Ecological interventionist causal models in psychosis: Targeting psychological mechanisms in daily life. *Schizophrenia Bulletin*, *42*(2), 264–269. https://doi.org/10.1093/schbul/sbv193

Reininghaus, U., Klippel, A., Steinhart, H., Vaessen, T., van Nierop, M., Viechtbauer, W., Baztink, T., Kasanova, Z., van Aubel, E., van Winkel R., Marcelis, M., van Amelsvoort, T., van der Gaag, M., de Haan L., & Myin Germeys, I. (2019). Efficacy of acceptance and commitment therapy in daily life (ACT-DL) in early psychosis: Study protocol for a multi-centre randomized controlled trial. *Trials*, *20*(1), 769. https://doi.org/10.1186/s13063-019-3912-4

Reininghaus, U., & Myin-Germeys, I. (2023). Mental health reform, ecological translation and the future of public mental healthcare. In M. Wensing & C. Ullrich (Eds.), *Foundations of health services research: Principles, methods, and topics* (pp. 223–233). Springer Cham. https://doi.org/10.1007/978-3-031-29998-8_18

Reininghaus, U., Paetzold, I., Rauschenberg, C., Hirjak, D., Banaschewski, T., Meyer-Lindenberg, A., Boehnke, J. R., Boecking, B., & Schick, A. (2023). Effects of a novel, transdiagnostic ecological momentary intervention for prevention, and early intervention of severe mental disorder in youth (EMIcompass): Findings from an exploratory randomized controlled trial. *Schizophrenia Bulletin*, *49*(3), 592–604. https://doi.org/10.1093/schbul/sbac212

Reininghaus, U., Schwannauer, M., Barne, I., Beames, J. R., Bonnier, R. A., Brenner, M., Breznoščáková, D., Dančik, De Allegri, M., Di Folco, S., Durstewitz, D., Gugel, J., Hajdúk, M., Heretik, A., Izáková, L., Katreniakova, Z., Kiekens, G., Koppe, G., Kurilla, A., Marelli, L., … Schick, A. (2024). Strategies, processes, outcomes, and costs of implementing experience sampling-based monitoring in routine mental health care in four European countries: Study protocol for the immerse effectiveness-implementation study. *BMC Psychiatry*, *24*(1), 465. https://doi.org/10.1186/s12888-024-05839-4

Reitsema, A. M., Jeronimus, B. F., van Dijk, M., & de Jonge, P. (2022). Emotion dynamics in children and adolescents: A meta-analytic and descriptive review. *Emotion*, *22*(2), 374. https://doi.org/10.1037/emo0000970

Revol, J., Carlier, C., Lafit, G., Verhees, M., Sels, L., & Ceulemans, E. (2024). Preprocessing experience-sampling-method data: A step-by-step framework, tutorial website, R package, and reporting templates. *Advances in Methods and Practices in Psychological Science*, *7*(4). https://doi.org/10.1177/25152459241256609

Revol, J., Lafit, G., & Ceulemans, E. (2024). A new sample-size planning approach for person-specific VAR(1) studies: Predictive accuracy analysis. *Behavior Research Methods*, *56*(7), 7152–7167. https://doi.org/10.3758/s13428-024-02413-4

Revol, J., Lafit, G., Kirtley, O. J., & Ceulemans, E. (2025). Cost-Effective ESM studies: Integrating budget constraints into sample size decisions. *Psychological Assessment*, *37*(10):479–492. https://doi.org/10.1037/pas0001409

Riese, H., von Klipstein, L., Schoevers, R. A., van der Veen, D. C., & Servaas, M. N. (2021). Personalized ESM monitoring and feedback to support psychological treatment for depression: A pragmatic randomized controlled trial (*Therap-i*). *BMC Psychiatry.* 21, 143. https://doi.org/10.1186/s12888-021-03123-3

Rimpler, A., Siepe, B. S., Rieble, C. L., Proppert, R. K. K., & Fried, E. I. (2024). Introducing FRED: Software for generating feedback reports for ecological momentary assessment data. *Administration and Policy in Mental Health*, *51*(4), 490–500. https://doi.org/10.1007/s10488-023-01324-4

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Robbins, M. L. (2017). Practical suggestions for legal and ethical concerns with social environment sampling methods. *Social Psychological and Personality Science*, *8*(5), 573–580. https://doi.org/10.1177/1948550617699253

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1–25. https://doi.org/10.1037/0033- 2909.132.1.1

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353–366. https://doi.org/10.1017/S0033291719003404

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*(6), 934–960. https://doi.org/10.1037/0033-2909.128.6.934

Rogers, A. A., Ha, T., Updegraff, K. A., & Iida, M. (2018). Adolescents' daily romantic experiences and negative mood: A dyadic, intensive longitudinal study. *Journal of Youth and Adolescence*, *47*(7), 1517–1530. https://doi.org/10.1007/s10964-017-0797-y

Rominger, C., & Schwerdtfeger, A. R. (2023). The real-time application of an additional HRV reduction algorithm to detect negative psychosocial states: Are we ready yet? *Zeitschrift für Psychologie*, *231*(4), 291–301. https://doi.org/10.1027/2151-2604/a000537

Roos, L. G., & Slavich, G. M. (2023). Wearable technologies for health research: Opportunities, limitations, and practical and conceptual considerations. *Brain*, *Behaviour*, *and Immunity*, *113*, 444–452. https://doi.org/10.1016/j.bbi.2023.08.008

Roth, A. M., Felsher, M., Reed, M., Goldshear, J. L., Truong, Q., Garfein, R. S., & Simmons, J. (2017). Potential benefits of using ecological momentary assessment to study high-risk polydrug use. *Mhealth*, *3*, 46. https://doi.org/10.21037/mhealth.2017.10.01

RStudio Team. (2015). *RStudio: Integrated development environment for R*. RStudio, Inc.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*, 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Ryan, O., Kuiper, R. M., & Hamaker, E. L. (2018). A continuous-time approach to intensive longitudinal data: What, why, and how? In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioural and related sciences* (pp. 27–54). Springer. https://doi.org/10.1007/978-3-319-77219-6_2

Ryff, C., Almeida, D. M., Ayanian, J. Z., Carr, D. S., Cleary, P. D., Coe, C., Davbidson, R. J., Krueger, R. F., Lachman, M. E., Marks, N. F., Mroczek, D. K., Seeman, T. E., Seltzer, M. M., Singer, B. H., Sloan, R. P., Tun, P. A., Weinstein, M., & Williams, D. R. (2017). *Midlife in the United States (MIDUS 2), 2004–2006*. https://doi.org/10.3886/ICPSR04652.v8

Safer, M. A., & Keuler, D. J. (2002). Individual differences in misremembering pre-psychotherapy distress: Personality and memory distortion. *Emotion*, *2*(2), 162–178. https://doi.org/10.1037/1528-3542.2.2.162

Sandstrom, G. M., Lathia, N., Mascolo, C., & Rentfrow, P. J. (2017). Putting mood in context: Using smartphones to examine how people feel in different locations. *Journal of Research in Personality*, 69, 96–101. https://doi.org/10.1016/j.jrp.2016.06.004

Sano, A., Taylor, S., McHill, A. W., Phillips, A. J. K., Barger, L. K., Klerman, E., & Picard, R. (2018). Identifying objective physiological markers and modifiable behaviours for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study. *Journal of Medical Internet Research*, *20*(6), e210. https://doi.org/10.2196/jmir.9410

Santangelo, P., Bohus, M., & Ebner-Priemer, U. W. (2014). Ecological momentary assessment in borderline personality disorder: A review of recent findings and methodological challenges. *Journal of Personality Disorders*, *28*(4), 555–576. https://doi.org/10.1521/pedi_2012_26_067

Santangelo, P. S., Reinhard, I., Koudela-Hamila, S., Bohus, M., Holtmann, J., Eid, M., & Ebner-Priemer, U. W. (2017). The temporal interplay of self-esteem instability and affective instability in borderline personality disorder patients' everyday lives. *Journal of Abnormal Psychology*, *126*(8), 1057–1065. https://doi.org/10.1037/abn0000288

Sarsenbayeva, Z., Fleming, C., Tag, B., Withana, A., van Berkel, N., & McEwan, A. (2023). A review on mood assessment using smartphones. In J. Abdelnour Nocera, M. Kristín Lárusdóttir, H. Petrie, A. Piccinno, & M. Winckler (Eds.), *Human-Computer Interaction—INTERACT* (pp. 385–413). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-42283-6_22

Sarwar, A., Agu, E. O., Polcari, J., Ciroli, J., Nephew, B., & King, J. (2022). PainRhythms: Machine learning prediction of chronic pain from circadian dysregulation using actigraph data – a preliminary study. *Smart Health*, *26*, 100344. https://doi.org/10.1016/j.smhl.2022.100344

Saygin, M., Schoenmakers, M., Gevonden, M., & de Geus, E. (2024). Controlling speech confounding in psychophysiology research: Speech detection via respiratory inductance plethysmography, thoracic impedance, accelerometers and gyroscopes. *Authorea Preprints*. https://doi.org/10.22541/au.172114399.90044532/v1

Schat, E., Tuerlinckx, F., De Ketelaere, B., & Ceulemans, E. (2024). Real-time detection of mean and variance changes in experience sampling data: A comparison of existing and novel statistical process control approaches. *Behaviour Research Methods*, *56*(3), 1459–1475. https://doi.org/10.3758/s13428-023-02103-7

Schat, E., Tuerlinckx, F., Smit, A. C., De Ketelaere, B., & Ceulemans, E. (2023). Detecting mean changes in experience sampling data in real time: A comparison of univariate and multivariate statistical process control methods. *Psychological Methods*, *28*(6), 1335–1357. https://doi.org/10.1037/met0000447

Schembre, S. M., Liao, Y., O'Connor, S. G., Hingle, M. D., Shen, S.-E., Hamoy, K. G., Huh, J., Dunton, G. F., Weiss, R., Thomson, C. A. & Boushey, C. J. (2018). Mobile ecological momentary diet assessment methods for behavioural research: Systematic review. *Journal of Medical Internet Research Mhealth Uhealth*, *6*(11), e11170. https://doi.org/10.2196/11170

Schick, A., Paetzold, I., Rauschenberg, C., Hirjak, D., Banaschewski, T., Meyer-Lindenberg, A., Boehnke, J. R., Boecking, B., & Reininghaus, U. (2021). Effects of a novel, transdiagnostic, hybrid ecological momentary intervention for improving resilience in youth (EMIcompass): Protocol for an explor-

atory randomized controlled trial. *Journal of Medical Internet Research Research Protocols*, *10*(12), e27462. https://doi.org/10.2196/27462

Schick, A., Rauschenberg, C., Ader, L., Daemen, M., Wieland, L. M., Paetzold, I., Postma, M. R., Schulte-Strathaus, J. C. C., & Reininghaus U. (2023). Novel digital methods for gathering intensive time series data in mental health research: Scoping review of a rapidly evolving field. *Psychological Medicine*, *53*(1), 55–65. https://doi.org/10.1017/S0033291722003336

Schiepek, G., Aichhorn, W., Gruber, M., Strunk, G., Bachler, E., & Aas, B. (2016). Real-time monitoring of psychotherapeutic processes: Concept and compliance. *Frontiers in Psychology*, *7*, 604. https://doi.org/10.3389/fpsyg.2016.00604

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, *2*. https://doi.org/10.3389/fnagi.2010.00027

Schneider, S., Junghaenel, D. U., Smyth, J. M., Fred Wen, C. K., & Stone, A. A. (2024). Just-in-time adaptive ecological momentary assessment (JITA-EMA). *Behaviour Research Methods*, *56*(2), 765–783. https://doi.org/10.3758/s13428-023-02083-8

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology*, *125*(6), 1442–1471. https://doi.org/10.1037/pspp0000469

Schoenmakers, M., Saygin, M., Sikora, M., Vaessen, T., Noordzij, M., & de Geus, E. (2025) Stress in action wearables database: A database of noninvasive wearable monitors with systematic technical, reliability, validity, and usability information. *Behaviour Research*, 57, 171. https://doi.org/10.3758/s13428-025-02685-4

Schoevers, R. A., van Borkulo, C. D., Lamers, F., Servaas, M. N., Bastiaansen, J. A., Beekman, A. T. F., van Hemert, A. M., Smit, J. H., Penninx, B. W. J. H., & Riese, H. (2020). Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. *Psychological Medicine*, *51*(11), 1906–1915. https://doi.org/10.1017/S0033291720000689

Schorrlepp, L., Stadel, M., Bringmann, L. F., Hesselink, M., & Maciejewski, D. (2025). Utilizing qualitative methods to detect validity issues in clinical experience sampling methodology (ESM). *Psychological Assessment*, *37*(11), 599–613. https://doi.org/10.1037/pas0001380

Schreuder, M. J., Schat, E., Smit, A. C., Snippe, E., & Ceulemans, E. (2024). Monitoring emotional intensity and variability to forecast depression recurrence in real time in remitted adults. *Journal of Consulting and Clinical Psychology*, *92*(8), 505–516. https://doi.org/10.1037/ccp0000871

Schroeders, U., Schmidt, C., & Gnambs, T. (2021). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 82(1). https://doi.org/10.1177/00131644211004708

Schueller, S. M., Aguilera, A., & Mohr, D. C. (2017). Ecological momentary interventions for depression and anxiety. *Depression and Anxiety*, *34*(6), 540–545. https://doi.org/10.1002/da.22649

Schulte-Strathaus, J. C. C., Rauschenberg, C., Baumeister, H., & Reininghaus, U. (2023). Ecological momentary interventions in public mental health provision. In C. Montag & H. Baumeister (Eds.), *Digital phenotyping and mobile sensing: New developments in psychoinformatics* (pp. 427–439). Springer International Publishing. https://doi.org/10.1007/978-3-030-98546-2_25

Schuurman, N. K. (2023, February 26). A "within/between problem" primer: About (not) separating within-person variance and between-person variance in psychology. PsyArXiv *preprint*. https://doi.org/10.31234/osf.io/7zgkx

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, S., Schultz, S., Reider, A., & Saunders, E. F. H. (2016). Daily mood monitoring of symptoms using smartphones in bipolar disorder: A pilot study assessing the feasibility of ecological momentary assessment. *Journal of Affective Disorders*, *191*, 88–93. https://doi.org/10.1016/j.jad.2015.11.013

Schwerdtfeger, A., & Ofner, S. (2024). *The effects of a just-in-time adaptive intervention (JITAI) on physiological activity and well-being in everyday life*. https://doi.org/10.17605/OSF.IO/7FGCK

Scollon, C. N., Prieto, C.-K., & Diener, E. (2009). Experience sampling: Promises and pitfalls, strengths and weaknesses. In E. Diener (Ed.), *Assessing well-being: The collected works of Ed Diener* (pp. 157–180). Springer Netherlands. https://doi.org/10.1007/978-90-481-2354-4_8

Scott, K. M., & Kline, M. (2019). Enabling confirmatory secondary data analysis by logging data checkout. *Advances in Methods and Practices in Psychological Science*, *2*(1), 45–54. https://doi.org/10.1177/2515245918815849

Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling time-dependent association in longitudinal data: A lag as moderator approach. *Multivariate Behavioural Research*, *47*(5), 697–716. https://doi.org/10.1080/00273171.2012.715557

Sels, L., Cabrieto, J., Butler, E., Reis, H., Ceulemans, E., & Kuppens, P. (2020). The occurrence and correlates of emotional interdependence in romantic relationships. *Journal of Personality and Social Psychology*, *119*(1), 136–158. https://doi.org/10.1037/pspi0000212

Sels, L., Ceulemans, E., & Kuppens, P. (2017). Partner-expected affect: How you feel now is predicted by how your partner thought you felt before. *Emotion*, *17*(7), 1066–1077. https://doi.org/10.1037/emo0000304

Sels, L., Ruan, Y., Kuppens, P., Ceulemans, E., & Reis, H. (2020). Actual and perceived emotional similarity in couples' daily lives. *Social Psychological and Personality Science*, *11*(2), 266–275. https://doi.org/10.1177/1948550619845927

Serre, F., Fatseas, M., Debrabant, R., Alexandre, J.-M., Auriacombe, M., & Swendsen, J. (2012). Ecological momentary assessment in alcohol, tobacco, cannabis and opiate dependence: A comparison of feasibility and validity. *Drug and Alcohol Dependence*, *126*(1–2), 118–123. https://doi.org/10.1016/j.drugalcdep.2012.04.025

Shah, R. V., Grennan, G., Zafar-Khan, M., Alim, F., Dey, S., Ramanathan, D., & Mishra, J. (2021). Personalized machine learning of depressed mood using wearables. *Translational Psychiatry*, *11*(1), 338. https://doi.org/10.1038/s41398-021-01445-0

Sheppard, M. K. (2020). mhealth apps: Disruptive innovation, regulation, and trust-a need for balance. *Medical Law Review*, *28*(3), 549–572. https://doi.org/10.1093/medlaw/fwaa019

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Macmillan. https://books.google.be/books?id=JtVnAAAAMAAJ

Shiffman, S., Hufford, M., Hickcox, M., Paty, J. A., Gnys, M., & Kassel, J. D. (1997). Remember that? A comparison of real-time versus retrospective recall of smoking lapses. *Journal of Consulting and Clinical Psychology*, *65*(2), 292–300. https://doi.org/10.1037/0022-006x.65.2.292.a

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., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. A. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). The Guilford Press.

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

Siepe, B. S., Tutunji, R., Rieble, C. L., Proppert, R. K. K., & Fried, E. I. (2025). Associations between ecological momentary assessment and passive sensor data in a large student sample. *Journal of Psychopathology and Clinical Science*, *16*, *134*(8), 912–925 https://doi.org/10.1037/abn0001013

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, *31*(4), 471–481. https://doi.org/10.1177/0894439313479902

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

Simons, C. J. P., Hartmann, J. A., Kramer, I., Menne-Lothmann, C., Höhn, P., Van Bemmel, A. L., Myin-Germeys, I., Delespaul, P., van Os, J., & Wichers, M. (2015). Effects of momentary self-monitoring on empowerment in a randomized controlled trial in patients with depression. *European Psychiatry*, *30*(8), 900–906. https://doi.org/10.1016/j.eurpsy.2015.09.004

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

Sinnaeve, R., Van Diest, I., Claes, S., Myin-Germeys, I., Van den Bosch, L M. C., Kamphuis, J. H., Vansteelandt, K., Van Hoof, C., Cornelis, J., & Houben M. (2024). Stress-related fluctuations in personality functioning in daly life: Pilot data from an ambulatory monitoring study in outpatients diagnosed with borderline personality disorder. *Clinical Psychology & Psychotherapy*, *31*(1), e2951. https://doi.org/10.1002/cpp.2951

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass*, *2*(1), 245–261. https://doi.org/10.1111/j.1751-9004.2007.00043.x

Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., Cornelis, J., Janssens, O., Van Hoecke, S., Claes, S., Van Diest, I., & Van Hoof, C. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *Nature Partner Journal Digital Medicine*, *1*, 67. https://doi.org/10.1038/s41746-018-0074-9

Smit, A. C., Schat, E., & Ceulemans, E. (2022). The exponentially weighted moving average procedure for detecting changes in intensive longitudinal data in psychological research in real-time: A tutorial showcasing potential applications. *Assessment*, *30*(5), 1354–1368. https://doi.org/10.1177/10731911221086985

Smit, A. C., & Snippe, E. (2023). Real-time monitoring of increases in restlessness to assess idiographic risk of recurrence of depressive symptoms. *Psychological Medicine*, *53*(11), 5060–5069. https://doi.org/10.1017/S0033291722002069

Smit, A. C., Snippe, E., & Wichers, M. (2019). Increasing restlessness signals impending increase in depressive symptoms more than 2 months before it happens in individual patients. *Psychotherapy and Psychosomatics*, *88*(4), 249–251. https://doi.org/10.1159/000500594

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioural science*: John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa492

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*(3), 237–259. https://doi.org/10.3102/10769986018003237

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Snippe, E., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Kramer, I., Booij, S. H., Viechtbauer, W., Delespaul, P., Myin-Germeys, I., & Wichers, M. (2016). Change in daily life behaviours and depression: Within-person and between-person associations. *Health Psychology*, *35*(5), 433–441. https://doi.org/10.1037/hea0000312

Snippe, E., Smit, A. C., Kuppens, P., Burger, H., & Ceulemans, E. (2023). Recurrence of depression can be foreseen by monitoring mental states with statistical process control. *Journal of Psychopathology and Clinical Sciences*, *132*(2), 145–155. https://doi.org/10.1037/abn0000812

Snir, A., Rafaeli, E., Gadassi, R., Berenson, K., & Downey, G. (2015). Explicit and inferred motives for nonsuicidal self-injurious acts and urges in borderline and avoidant personality disorders. *Personal Disorders*, *6*(3), 267–277. https://doi.org/10.1037/per0000104

Soderberg, C. K., Sallans, A., Clyburne-Sherin, A., Spitzer, M., Sullivan, I., Smith, J. F., & Mellor, D. T. (2019). *IRB and consent form examples*. https://osf.io/g4jfv

Soyster, P. D., Bosley, H. G., Reeves, J. W., Altman, A. D., & Fisher, A. J. (2019). Evidence for the feasibility of person-specific ecological momentary assessment across diverse populations and study designs. *Journal for Person-Oriented Research*, *5*(2), 53–64. https://doi.org/10.17505/jpor.2019.06

Stadler, G., Scholz, U., Bolger, N., Shrout, P. E., Knoll, N., & Lüscher, J. (2023). How is companionship related to romantic partners' affect, relationship satisfaction, and health behaviour? Using a longitudinal dyadic score model to understand daily and couple-level effects of a dyadic predictor. *Applied Psychology: Health and Well-Being*, *15*(4), 1530–1554. https://doi.org/10.1111/aphw.12450

Staples, P., Torous, J., Barnett, I., Carlson, K., Sandoval, L., Keshavan, M., & Onnela, J.-P. (2017). A comparison of passive and active estimates of sleep in a cohort with schizophrenia. *Nature Partner Journal Schizophrenia*, *3*(1), 37. https://doi.org/10.1038/s41537-017-0038-0

Statista. (2025). *Number of smartphone users worldwide from 2014 to 2029*. https://www.statista.com/

Stawski, R. S., MacDonald, S. W. S., & Sliwinski, M. J. (2015). Measurement burst design. In S. K. Whitbourne (Ed.), *The encyclopedia of adulthood and aging* (pp. 1–5). Wiley. https://doi.org/10.1002/9781118521373.wbeaa313

Steeg, S., Quinlivan, L., Nowland, R., Carroll, R., Casey, D., Clements, C., Cooper, J., Davies, L., Knipe, D., Ness, J., O'Connor, R. C., Hawton, K., Gunnell, D., & Kapur, N. (2018). Accuracy of risk scales for predicting repeat self-harm and suicide: A multicentre, population-level cohort study using routine clinical data. *BMC Psychiatry*, *18*(1), 113. https://doi.org/10.1186/s12888-018-1693-z

Steele, C. (2019). *What is the digital divide?* http://www.digitaldividecouncil.com/what-is-the-digital-divide/

Stein, K. F., & Corte, C. M. (2003). Ecologic momentary assessment of eating-disordered behaviours. *International Journal of Eating Disorders*, *34*(3), 349–360. https://doi.org/10.1002/eat.10194

Stone, A. A., Broderick, J. E., Schwartz, J. E., Shiffman, S., Litcher-Kelly, L., & Calvanese, P. (2003). Intensive momentary reporting of pain with an electronic diary: Reactivity, compliance, and patient satisfaction. *Pain*, *104*(1–2), 343–351. https://doi.org/10.1016/s0304-3959(03)00040-x

Stone, A. A., Broderick, J. E., Shiffman, S. S., & Schwartz, J. E. (2004). Understanding recall of weekly pain from a momentary assessment perspective: Absolute agreement, between- and within-person consistency, and judged change in weekly pain. *Pain*, *107*(1–2), 61–69. https://doi.org/10.1016/j.pain.2003.09.020

Stone, A. A., Schneider, S., & Harter, J. K. (2012). Day-of-week mood patterns in the United States: On the existence of 'Blue Monday', 'Thank God it's Friday' and weekend effects. *The Journal of Positive Psychology*, *7*, 306–314. https://doi.org/10.1080/17439760.2012.691980

Stone, A. A., & Shiffman, S. (1994). Ecological Momentary Assessment (EMA) in behavioural medicine. *Annals of Behavioural Medicine*, *16*(3), 199–202. https://doi.org/10.1093/abm/16.3.199

Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2003). Patient compliance with paper and electronic diaries. *Control Clinical Trials*, *24*(2), 182–199. https://doi.org/10.1016/s0197-2456(02)00320-3

Sui, A., Sui, W., Liu, S., & Rhodes, R. (2023). Ethical considerations for the use of consumer wearables in health research. *Digital Health*, *9*. https://doi.org/10.1177/20552076231153740

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136. https://doi.org/10.1177/0146167298242002

Sun, J., Rhemtulla, M., & Vazire, S. (2021). Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin*, *47*(11), 1535–1549. https://doi.org/10.1177/0146167220964639

Thewissen, V., Bentall, R. P., Lecomte, T., van Os, J., & Myin-Germeys, I. (2008). Fluctuations in self-esteem and paranoia in the context of daily life. *Journal of Abnormal Psychology*, *117*(1), 143–153. https://doi.org/10.1037/0021-843X.117.1.143

Thewissen, V., Bentall, R. P., Oorschot, M., A Campo, J., van Lierop, T., van Os, J., & Myin-Germeys, I. (2011). Emotions, self-esteem, and paranoid episodes: An experience sampling study. *British Journal of Clinical Psychology*, *50*(2), 178–195. https://doi.org/10.1348/014466510X508677

Timmons, A. C., Baucom, B. R., Han, S. C., Perrone, L., Chaspari, T., Narayanan, S. S., & Margolin, G. (2017). New frontiers in ambulatory assessment: Big data methods for capturing couples' emotions, vocalizations, and physiology in daily life. *Social Psychological and Personality Science*, *8*(5), 552–563. https://doi.org/10.1177/1948550617709115

Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., Vittinghoff, E., Lee, E. S., Fan, S. M., Gladstone, R. A., Mikell, C., Sohoni, N., Hsieh, J., & Marcus G. M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. *The Journal of the American Medical Association Cardiology*, *3*(5), 409–416. https://doi.org/10.1001/jamacardio.2018.0136

Tomitani, N., Kanegae, H., & Kario, K. (2022). Self-monitoring of psychological stress-induced blood pressure in daily life using a wearable watch-type oscillometric device in working individuals with hypertension. *Hypertension Research*, *45*(10), 1531–1537. https://doi.org/10.1038/s41440-022-00946-9

Tooze, J. A., Grunwald, G. K., & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medicine Research*, *11*(4), 341–355. https://doi.org/10.1191/0962280202sm291ra

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *Journal of Medical Internet Research Mental Health*, *3*(2), e16. https://doi.org/10.2196/mental.5165

Torous, J., Nicholas, J., Larsen, M. E., Firth, J., & Christensen, H. (2018). Clinical review of user engagement with mental health smartphone apps: Evidence, theory and improvements. *Evidence Based Mental Health*, *21*(3), 116–119. https://doi.org/10.1136/eb-2018-102891

Torous, J., Wisniewski, H., Liu, G., & Keshavan, M. (2018). Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients: Comparative survey study. *Journal of Medical Internet Research Mental Health*, *5*(4), e11715. https://doi.org/10.2196/11715

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review Clinical Psychology*, *9*, 151–176. https://doi.org/10.1146/annurev-clinpsy-050212-185510

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*(4), 457–462. https://doi.org/10.1037/a0017653

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

Tutunji, R., Kogias, N., Kapteijns, B., Krentz, M., Krause, F., Vassena, E., & Hermans, E. J. (2023). Detecting prolonged stress in real life using wearable biosensors and ecological momentary assessments:

Naturalistic experimental study. *Journal of Medical Internet Research*, *25*, e39995. https://doi.org/10.2196/39995

Ulitzsch, E., Viechtbauer, W., Lüdtke, O., Myin-Germeys, I., Nagy, G., Nestler, S., & Eisele, G. V. (2025). Investigating the effect of experience sampling study design on careless and insufficient effort responding identified with a screen-time-based mixture model. *Psychological Assessment*, *37*(8), 347–359, https://doi.org/10.1037/pas0001379

US Food and Drug Administration. (2019). *Policy for device software functions and mobile medical applications guidance for industry and Food and Drug administration staff; 2019*. https://www.fda.gov/media/80958/download

Vachon, H., Bourbousson, M., Deschamps, T., Doron, J., Bulteau, S., Sauvaget, A., & Thomas-Ollivier, V. (2016). Repeated self-evaluations may involve familiarization: An exploratory study related to Ecological Momentary Assessment designs in patients with major depressive disorder. *Psychiatry Research*, *245*, 99–104. https://doi.org/10.1016/j.psychres.2016.08.034

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

Vaessen, T., Kasanova, Z., Hernaus, D., Lataster, J., Collip, D., van Nierop, M., & Myin-Germeys, I. (2018). Overall cortisol, diurnal slope, and stress reactivity in psychosis: An experience sampling approach. *Psychoneuroendocrinology*, *96*, 61–68. https://doi.org/10.1016/j.psyneuen.2018.06.007

Vaessen, T., Rintala, A., Otsabryk, N., Viechtbauer, W., Wampers, M., Claes, S., & Myin-Germeys, I. (2021). The association between self-reported stress and cardiovascular measures in daily life: A systematic review. *PloS One*, *16*(11), e0259557. https://doi.org/10.1371/journal.pone.0259557

Vaessen, T., Steinhart, H., Batink, T., Klippel, A., Van Nierop, M., Reininghaus, U., & Myin-Germeys, I. (2019). Act in daily life in early psychosis: An ecological momentary intervention. *Psychosis*, *11*(2), 93–104. https://doi.org/10.1080/17522439.2019.1578401

van Aubel, E., Bakker, J. M., Batink, T., Michielse, S., Goossens, L., Lange, I., Schruers, K, Lieverse, R., Marcelis, M., van Amelsvoort, T., van Os, J., Wichers, M., Vaessen, T., Reininghaus, U., & Myin-Germeys, I. (2020). Blended care in the treatment of subthreshold symptoms of depression and psychosis in emerging adults: A randomised controlled trial of acceptance and commitment therapy in daily-life (ACT-DL). *Behaviour Research and Therapy*, *128*, 103592. https://doi.org/10.1016/j.brat.2020.103592

van Aubel, E., Vaessen, T., Uyttebroek, L., Steinhart, H., Beijer-Klippel, A., Batink, T., van Winkel, R., de Haan, L., van der Gaag, M., van Amelsvoort, T., Marcelis, M., Schirmbeck, F., Reininghaus, U., & Myin-Germeys, I. (2024). Engagement and acceptability of acceptance and commitment therapy in daily life in early psychosis: Secondary findings from a multicenter randomized controlled trial. *Journal of Medical Internet Research Formative Research*, 8, e57109. https://doi.org/10.2196/57109

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D. D., Smit, J. H., & Riper, H. (2016). Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): Protocol of a randomised controlled trial. *BMC Psychiatry*, *16*(1), 359. https://doi.org/10.1186/s12888-016-1065-5

van Berkel, N., Ferreira, D., & Kostakos, V. (2018). The experience sampling method on mobile devices. *ACM Computing Surveys*, *50*(6), 1–40. https://doi.org/10.1145/3123988

van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., & Kostakos, V. (2019). Context-informed scheduling and analysis: Improving accuracy of mobile self-reports. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 51 (pp. 1– 12). https://doi.org/10.1145/3290605.3300281

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L. J., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110

Van den Bergh, O., & Walentynowicz, M. (2016). Accuracy and bias in retrospective symptom reporting. *Current Opinion Psychiatry*, *29*(5), 302–308. https://doi.org/10.1097/yco.0000000000000267

Van der Gucht, K., Dejonckheere, E., Erbas, Y., Takano, K., Vandemoortele, M., Maex, E., Raes, F., & Kuppens, P. (2019). An experience sampling study examining the potential impact of a mindfulness-based intervention on emotion differentiation. *Emotion*, *19*(1), 123–131. https://doi.org/10.1037/emo0000406

van der Steen, Y., Gimpel-Drees, J., Lataster, T., Viechtbauer, W., Simons, C. J. P., Lardinois, M., Michel, T. M., Janssen, B., Bechdolf, A., Wagner, M., & Myin-Germeys, I. (2017). Clinical high risk for psychosis: The association between momentary stress, affective and psychotic symptoms. *Acta Psychiatrica Scandinavica*, *136*(1), 63–73. https://doi.org/10.1111/acps.12714

van der Storm, S. L., Jansen, M., Meijer, H. A. W., Barsom, E. Z., & Schijven, M. P. (2023). Apps in healthcare and medical research; European legislation and practical tips every healthcare provider should know. *International Journal of Medical Informatics*, *177*, 105141. https://doi.org/10.1016/j.ijmedinf.2023.105141

van Kessel, R., Roman-Urrestarazu, A., Anderson, M., Kyriopoulos, I., Field, S., Monti, G., Reed, S. D, Pavlova, M., Wharton, G., & Mossialos, E. (2023). Mapping factors that affect the uptake of digital therapeutics within health systems: Scoping review. *Journal of Medical Internet Research*, *25*, e48000. https://doi.org/10.2196/48000

van Knippenberg, R. J. M., de Vugt, M. E., Ponds, R. W., Myin-Germeys, I., & Verhey, F. R. J. (2018). An experience sampling method intervention for dementia caregivers: Results of a randomized controlled trial. *American Journal of Geriatric Psychiatry*, *26*(12), 1231–1243. https://doi.org/10.1016/j.jagp.2018.06.004

van Kolfschooten, H. (2022). The mHealth power paradox: Improving data protection in health apps through self-regulation in the European Union. In I. G. Cohen, T. Minssen, W. N. Price II, C. Robertson, & C. Shachar (Eds.), *The future of medical device regulation: Innovation and protection* (pp. 63–76). Cambridge University Press. https://doi.org/10.1017/9781108975452.006

Van Lange, P. A. M., & Balliet, D. (2015). Interdependence theory. In M. Mikulincer, P. R. Shaver, J. A. Simpson, & J. F. Dovidio (Eds.), *APA handbook of personality and social psychology* (Vol. 3. Interpersonal relations) (pp. 65–92). American Psychological Association. https://doi.org/10.1037/14344-003

van Os, J., Verhagen, S., Marsman, A., Peeters, F., Bak, M., Marcelis, M., Drukker, M., Reininghaus, U., Jacobs, N., Lataster, T., Simons, C., Lousber R., Gülöksüz, S., Leue, C., Groot, P. C., Viechtbauer, W., & Delespaul, P. (2017). The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. *Depression and Anxiety*, *34*(6), 481–493. https://doi.org/10.1002/da.22647

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

van Winkel, M., Nicolson, N. A., Wichers, M., Viechtbauer, W., Myin-Germeys, I., & Peeters, F. (2015). Daily life stress reactivity in remitted versus non-remitted depressed individuals. *European Psychiatry*, *30*(4), 441–447. https://doi.org/10.1016/j.eurpsy.2015.02.011

Veenman, M., Janssen, L. H. C., van Houtum, L. A. E. M., Wever, M. C. M., Verkuil, B., Epskamp, S., Fried, E. I., & Elzinga, B. M. (2024). A network study of family affect systems in daily life. *Multivariate Behavioural Research*, *59*(2), 371–405. https://doi.org/10.1080/00273171.2023.2283632

Veldmeijer, L., Terlouw, G., Van Os, J., Van Dijk, O., Van ,t Veer, J., & Boonstra, N. (2023). The involvement of service users and people with lived experience in mental health care innovation through design: Systematic review. *Journal of Medical Internet Research Mental Health*, *10*, e46590. https://doi.org/10.2196/46590

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer. https://doi.org/10.1007/978-1-4419-0300-6

Verhagen, S., van Os, J., & Delespaul, P. (2022). Ecological momentary assessment and other digital technologies for capturing daily life in mental health. In D. J. Stein, N. A. Fineberg, & S. R. Chamberlain (Eds.), *Mental health in a digital world* (pp. 81–108). Academic Press. https://doi.org/10.1016/b978-0-12-822201-0.00017-4

Verhagen, S. J. W., Hasmi, L., Drukker, M., van Os, J., & Delespaul, P. A. E. G. (2016). Use of the experience sampling method in the context of clinical trials. *Evidence Based Mental Health*, *19*(3), 86–89. https://doi.org/10.1136/ebmental-2016-102418

Verhees, M. W. F. T., Bodner, N., Bosmans, G., & Ceulemans, E. (2024). Examining stress in adolescents' daily lives: Feasibility of triadic paradigms. *Journal of Family Psychology*. *38*(6), 989–994. https://doi.org/10.1037/fam0001251

Verhees, M. W. F. T., Ceulemans, E., Sels, L., & Kuppens, P. (2024). Attachment and perceptual accuracy of hard and flat partner emotions in everyday life. *European Journal of Social Psychology*, *54*(4), 946–958. https://doi.org/10.1002/ejsp.3064

Versluis, A., Verkuil, B., Spinhoven, P., van der Ploeg, M. M., & Brosschot, J. F. (2016). Changing mental health and positive psychological well-being using ecological momentary interventions: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *18*(6), e152. https://doi.org/10.2196/jmir.5642

Victor, S. E., Scott, L. N., Stepp, S. D., & Goldstein, T. R. (2019). I want you to want me: Interpersonal stress and affective experiences as within-person predictors of nonsuicidal self-injury and suicide urges in daily life. *Suicide Life Threat Behaviour*, *49*(4), 1157–1177. https://doi.org/10.1111/sltb.12513

Viechtbauer, W. (2017). Reliability of ESM assessments of mood and mood sensitivity. In C. Vögele (Ed.), *Digital health in ambulatory assessment—Abstract book of the 5th Biennial Conference of the Society for Ambulatory Assessment* (pp. 25). University of Luxembourg.

Viechtbauer, W. & Constantin, M. (2023). *esmpack: A package to facilitate preparation and management of ESM/EMA data (version 01.20) (Computer software)*. https://wviechtb.github.io/esmpack/

Vogelsmeier, L. V. D. E., Cloos, L., Kuppens, P., & Ceulemans, E. (2023). Evaluating dynamics in affect structure with latent Markov factor analysis. *Emotion*, *24*(3), 782–794. https://doi.org/10.1037/emo0001307

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research*, *33*, 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

von Baeyer, C. (1994). Reactive effects of measurement of pain. *The Clinical Journal of Pain*, *10*, 18–21. https://doi.org/10.1097/00002508-199403000-00004

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *The Lancet*, *370*(9596), 1453–1457. https://doi.org/10.1016/S0140-6736(07)61602-X

von Klipstein, L., Riese, H., van der Veen, D. C., Servaas, M. N., & Schoevers, R. A. (2020). Using person-specific networks in psychotherapy: Challenges, limitations, and how we could use them anyway. *BMC Medicine*, *18*(1), 345. https://doi.org/10.1186/s12916-020-01818-0

von Klipstein, L., Servaas, M. N., Schoevers, R. A., van der Veen, D. C., & Riese, H. (2023). Integrating personalized experience sampling in psychotherapy: A case illustration of the Therap-i module. *Heliyon*, *9*(3), e14507. https://doi.org/10.1016/j.heliyon.2023.e14507

von Klipstein, L., Stadel, M., Bos, F. M., Bringmann, L F., Riese, H., & Servaas, M. (2025). Opening the contextual black box: A case for idiographic experience sampling of context for clinical applications. *Quality of Life Research*, *34*(3), 595–604. https://doi.org/10.31234/osf.io/nv7c6

Wadle, L.-M., Ebner-Priemer, U. W., Foo, J. C., Yamamoto, Y., Streit, F., Witt, S. H., Frank, J., Zillich, L., Limberger, M. F., Ablimit, A., Schultz, T., Gilles, M., Rietschel, M., & Sirignano L. (2024). Speech features as predictors of momentary depression severity in patients with depressive disorder undergoing sleep deprivation therapy: Ambulatory assessment pilot study. *Journal of Medical Internet Research Mental Health*, *11*, e49222. https://doi.org/10.2196/49222

Walls, T. A., Jung, H., & Schwartz, J. E. (2006). Multilevel models for intensive longitudinal data. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 3–33). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195173444.003.0001

Walsh, C. G., Xia, W., Li, M., Denny, J. C., Harris, P. A., & Malin, B. A. (2018). Enabling open-science initiatives in clinical psychology and psychiatry without sacrificing patients' privacy: Current practices and future challenges. *Advances in Methods and Practices in Psychological Science*, *1*(1), 104–114. https://doi.org/10.1177/2515245917749652

Walsh, E., & Brinker, J. K. (2016). Short and sweet? Length and informative content of open-ended Responses using SMS as a research mode. *Journal of Computer-Mediated Communication*, *21*(1), 87–100. https://doi.org/10.1111/jcc4.12146

Wang, C., Hall, C. B. & Kim, M. (2015). A comparison of power analysis methods for evaluating effects of a predictor on slopes in longitudinal designs with missing data. *Statistical Methods in Medical Research*, *24*(6), 1009–1029. https://doi.org/ 10.1177/0962280212437452

Wang, J., Wu, Z., Choi, S. W., Sen, S., Yan, X., & Miner, J. A., Sander, A. M., Lyden, A. K., Troost, J. P., & Carlozzi, N. E. (2023). The dosing of mobile-based just-in-time adaptive self-management prompts for caregivers: Preliminary findings from a pilot microrandomized study. *Journal of Medical Internet Research Form Res*, *7*, e43099. https://doi.org/10.2196/43099

Wang, L., & Miller, L. C. (2020). Just-in-the-moment adaptive interventions (JITAI): A meta-analytical review. *Health Communication*, *35*(12), 1531–1544. https://doi.org/10.1080/10410236.2019.1652388

Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychological Methods*, *20*(1), 63–83. https://doi.org/10.1037/met0000030

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Weermeijer, J., Kiekens, G., Wampers, M., Kuppens, P., & Myin-Germeys, I. (2024). Practitioner perspectives on the use of the experience sampling software in counseling and clinical psychology. *Behaviour & Information Technology*, *43*(3), 540–550. https://doi.org/10.1080/0144929X.2023.2178235

Weermeijer, J. D. M., Wampers, M., de Thurah, L., Bonnier, R., Piot, M., Kuppens, P., Myin-Germeys, I., & Kiekens, G. (2023). Usability of the experience sampling method in specialized mental health care: Pilot evaluation study. *Journal of Medical Internet Research Formative Research*, *7*, e48821. https://doi.org/10.2196/48821

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. https://doi.org/10.1037/em00000226

Welling, J., Fischer, R.-L., & Schinkel-Bielefeld, N. (2021). Is it possible to identify careless responses with post-hoc analysis in EMA studies? *UMAP '21: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (p. 150–156). https://doi.org/10.1145/3450614.3462237

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, *44*(7), 1349–1360. https://doi.org/10.1017/s0033291713001979

Wichers, M., Groot, P. C., & Psychosystems, ESM Group, EWS Group. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116. https://doi.org/10.1159/000441458

Wichers, M., Riese, H., Hodges, T. M., Snippe, E., & Bos, F. M. (2021). A narrative review of network studies in depression: What different methodological approaches tell us about depression. *Frontiers in Psychiatry*, *12*, 719490. https://doi.org/10.3389/fpsyt.2021.719490

Wichers, M., Simons, C. J. P., Kramer, I. M. A., Hartmann, J. A., Lothmann, C., Myin-Germeys, I., van Bemmel, A. L., Peeters, F., Delespaul, P., & van Os, J. (2011). Momentary assessment technology as a tool to help patients with depression help themselves. *Acta Psychiatrica Scandinavica*, *124*(4), 262–272. https://doi.org/10.1111/j.1600-0447.2011.01749.x

Wichers, M., Smit, A. C., & Snippe, E. (2020). Early warning signals based on momentary affect dynamics can expose nearby transitions in depression: A confirmatory single-subject time-series study. *Journal of Person-Oriented Research*, *6*(1), 1–15. https://doi.org/10.17505/jpor.2020.22042

Wickham, H, Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686, https://doi.org/10.21105/joss.01686

Wigman, J. T. W., van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., MERGE, Viechtbauer, W., Myin-Germeys, I., & Wichers, M. (2015). Exploring the underlying structure of mental disorders: Cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. *Psychological Medicine*, *45*(11), 2375–2387. https://doi.org/10.1017/S0033291715000331

Winkler, M. R., Mason, S., Laska, M. N., Christoph, M. J., & Neumark-Sztainer, D. (2018). Does non-standard work mean non-standard health? Exploring links between non-standard work schedules, health behaviour, and well-being. *SSM—Population Health*, *4*, 135–143. https://doi.org/10.1016/j.ssmph.2017.12.003

Wittenborn, A. K., Dolbin-MacNab, M. L., & Keiley, M. K. (2013). Dyadic research in marriage and family therapy: Methodological considerations. *Journal of Marital and Family Therapy*, *39*(1), 5–16. https://doi.org/10.1111/j.1752-0606.2012.00306.x

Wolf, G. I., & De Groot, M. (2020). A conceptual framework for personal science. *Frontiers in Computer Science*, *2*, 21. https://doi.org/10.3389/fcomp.2020.00021

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

Wrzus, C., & Schoedel, R. (2023). Transparency and reproducibility in mobile sensing research. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications* (p. 53). The Guilford Press.

Wu, H., Zhou, X., Chen, D., Zheng, Y., & You, J. (2024). Longitudinal association between social media exposure and nonsuicidal self-injury among adolescents: Investigating the directionality by within-person effects. *Current Psychology*, *43*, 9744–9754. https://doi.org/10.1007/s12144-023-05128-5

Wu, S., Crespi, C. M., & Wong, W. K. (2012). Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. *Contemporary Clinical Trials*, *33*(5), 869–880. https://doi.org/10.1016/j.cct.2012.05.004

Xiao, Y., Wang, P., & Liu, H. (2023). Assessing intra- and inter-individual reliabilities in intensive longitudinal studies: A two-level random dynamic model-based approach. *Psychological Methods*, 10.1037/met0000608. Advance online publication. https://doi.org/10.1037/met0000608

Yamaguchi, T., Mikami, S., Maeda, M., Saito, T., Nakajima, T., Yachida, W., & Gotouda, A. (2023). Portable and wearable electromyographic devices for the assessment of sleep bruxism and awake bruxism: A literature review. *CRANIO® The Journal of Craniomandibular & Sleep Practice*, *41*(1), 69–77. https://doi.org/10.1080/08869634.2020.1815392

Yang, M.-J., Sutton, S. K., Hernandez, L. M., Jones, S. R., Wetter, D. W., Kumar, S., & Vinci, C. (2023). A just-in-time adaptive intervention (JITAI) for smoking cessation: Feasibility and acceptability findings. *Addictive Behaviours*, *136*, 107467. https://doi.org/10.1016/j.addbeh.2022.107467

Yap, Y., Slavish, D. C., Taylor, D. J., Bei, B., & Wiley, J. F. (2020). Bi-directional relations between stress and self-reported and actigraphy-assessed sleep: A daily intensive longitudinal study. *Sleep*, *43*(3), zsz250. https://doi.org/10.1093/sleep/zsz250

Zapata-Lamana, R., Lalanza, J. F., Losilla, J.-M., Parrado, E., & Capdevila, L. (2020). mHealth technology for ecological momentary assessment in physical activity research: A systematic review. *PeerJ*, *8*, e8848. https://doi.org/10.7717/peerj.8848

Zarbo, C., Zamparini, M., Nielssen, O., Casiraghi, L., Rocchetti, M., Starace, F., & de Girolamo, G. (2023). Comparing adherence to the experience sampling method among patients with schizophrenia spectrum disorder and unaffected individuals: Observational study from the multicentric diapason project. *Journal of Medical Internet Research*, *25*, e42093. https://doi.org/10.2196/42093

Zeijen, M. E. L., Petrou, P., & Bakker, A. B. (2020). The daily exchange of social support between coworkers: Implications for momentary work engagement. *Journal of Occupational Health Psychology*, *25*(6), 439–449. https://doi.org/10.1037/ocp0000262

Zelenski, J. M., & Larsen, R. J. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality*, *34*(2), 178–197. https://doi.org/10.1006/jrpe.1999.2275

Zettle, R. D. (2016). *The Wiley Handbook of Contextual Behavioural Science*. Wiley Blackwell. https://doi.org/10.1002/9781118489857

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L., & Tu, X. M. (2011). On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, *30*(20), 2562–2572. https://doi.org/10.1002/sim.4265

Zhang, X., & Yi, N. (2020). NBZIMM: Negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics*, *21*(1), 488. https://doi.org/10.1186/s12859-020-03803-z

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Laiou, P., Matcham, F., Oetzmann, C., Lamers, F., Siddi, S., Simblett, S., Rintala, A., Mohr, D. C., Myin-Germeys, I., Wykes, T., Haro, J. M., Penninx, B. W. J. H., … RADAR-CNS, Consortium. (2021). Predicting depressive symptom severity through individuals' nearby Bluetooth device count data collected by mobile phones: Preliminary longitudinal study. *Journal of Medical Internet Research mHealth and uHealth*, *9*(7), e29840. https://doi.org/10.2196/29840

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: methods and software. *Behaviour Research Methods*, *46*(4), 1184-1198. https://doi.org/10.3758/s13428-013-0424-0

Zhang, Z., & Wang, L. (2009). Statistical power analysis for growth curve models using SAS. *Behaviour Research Methods*, *41*(4), 1083–1094. https://doi.org/10.3758/brm.41.4.1083

Zimmermann, J., Woods, W. C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., Spitzer, C., & Wright, A. G. C. (2019). Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. *Psychological Assessment*, *31*(4), 516–531. https://doi.org/10.1037/pas0000625

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

About Real

The Centre for **R**esearch on **E**xperience Sampling and **A**mbulatory Methods **L**euven (REAL) brings together researchers from KU Leuven who are experts in the use and application of experience sampling or ecological momentary assessment methods. ESM or EMA methods have become the state of the art in the study of emotional, behavioural and clinical phenomena in the context of everyday life, offering a unique window into what people do, want, feel, experience and encounter in their normal daily life. KU Leuven hosts a critical mass of researchers who are widely regarded as leading experts in methods for the study of daily life in the Research Group of Quantitative Psychology and Individual Differences (https://ppw.kuleuven.be/okp/home/), the Center for Contextual Psychiatry (http://www.ccp-leuven.be), and the Methodology of Educational Sciences Research Group (https://ppw.kuleuven.be/mesrg). Together they join forces in REAL. The aim is to be a leading centre for research into ecological and ambulatory methods for the study of daily life, advancing both fundamental science and clinical practice and offering training and consultancy for other researchers interested in these methods.

About the Authors

Inez Myin-Germeys is a psychologist and Professor of Contextual Psychiatry at KU Leuven – University of Leuven in Belgium, where she heads the Center for Contextual Psychiatry. Her research focuses on the interaction between the person and the environment in the development and maintenance of psychopathology in general and psychosis in particular. Next to this fundamental research on daily life processes associated with psychopathology, she investigates the implementation of ESM as a clinical tool in routine mental healthcare. She likes shaping her garden, playing the conga, being on the road with her bike – both for short and long trips – and collecting travel guides.

Peter Kuppens is Professor of Psychology at KU Leuven – University of Leuven in Belgium. His research focuses on studying the nature, regulation and dynamics of emotional experience both within and between individuals and how this relates to psychological well-being and mood disorder. His work is inspired by componential (e.g. appraisal) and dimensional theoretical perspectives on emotions and makes use of mathematical modelling of intensive longitudinal data collected in daily life and in the lab. He likes cats, but in a cruel twist of fate and to the dismay of his children, is allergic to them, and he has an unexplained interest in the vanitas theme in medieval art.

Leonie Cloos is a postdoctoral researcher at the Research Group of Quantitative Psychology and Individual Differences at KU Leuven. Her work focuses on the operationalization and measurement of non-observable experiences, particularly affect and emotions, using innovative methods. As part of the scientific integrity team, she promotes rigorous research practices and adaptation to technological advancements in psychological measurement. When she is not busy asking people how they feel, you will find her on the yoga mat, on a climbing wall or out in nature with her dog.

Gudrun Eisele is a clinical psychologist and postdoctoral researcher at the Center for Contextual Psychiatry at KU Leuven in Belgium. She is interested in methodological developments in psychological assessment and open science. Currently, she investigates the effects of design choices on data collected in experience sampling studies. She is also involved in the ESM Item Repository project, an open science initiative that aims to facilitate the sharing and evaluation of ESM items. When not thinking about ESM study design, she enjoys getting lost in nature and learning about other cultures.

Egon Dejonckheere is one of the co-founders of m-Path and assistant professor at Tilburg University in The Netherlands. Substantively, his research focuses on the dynamic properties of emotions and their role in mental well-being. He is particularly interested in the relation between positive and negative feelings in daily life and how these affective states are altered in people who suffer from clinical disorders (e.g. depression or borderline personality disorder). Methodologically, he investigates ways to assess the data quality in experience sampling research, and he hopes to find ways to improve good measurement practices in this field. In a previous life, his dance skills earned him a victory in the Belgian hip-hop championship, which he regularly likes to show off at parties. Although he has a cat, he thinks Shiba Inus are the cutest animals alive.

Yasemin Erbas obtained her PhD in Psychology at KU Leuven, Belgium and is now an Assistant Professor at the Department of Developmental Psychology at Tilburg University, The Netherlands. She studies the complexity of emotions, both in the lab and in daily life. She loves travelling and reading. To the dismay of her neighbours, she also loves playing the violin.

Marlies Houben obtained her PhD in Psychology at KU Leuven and is an Assistant Professor at the Department of Medical and Clinical Psychology of Tilburg University. Her research focuses on affective dynamics in daily life, potential mechanisms underlying affective changes over time, and concurrent and prospective relationships with mental health (changes) in the general population and in persons with psychiatric diagnoses. Next to being a researcher, she is also the proud mother of two kids, one cat and two chickens, is a die-hard vegetarian, and has a rock ‘n’ roll side as she

is into metal music, music festivals, tattoos and piercings and true crime podcasts.

Olivia J. Kirtley is an Assistant Research Professor and Co-Director of the Center for Contextual Psychiatry, KU Leuven – University of Leuven. Olivia's research focuses on understanding the factors that contribute to suicidal and non-suicidal self-harm, in particular social interaction, exposure to suicide and future thinking. As well as co-leading the CCP's research line on suicidal and non-suicidal self-injurious thoughts and behaviours, Olivia also co-leads the research line on ESM methodology, statistics and open science. As part of this, she also leads numerous open science initiatives to increase transparency, reproducibility and replicability in ESM and suicide research, including the registration template for ESM research and the ESM Item Repository. Olivia enjoys cooking, running, swimming and buying books she never has time to read.

Jeroen Weermeijer is a postdoctoral research psychologist at the Center for Contextual Psychiatry at KU Leuven. His research focuses on the clinical implementation of the ESM. He has a particular interest in user-experience perspectives, statistics and software development.

Glenn Kiekens is a senior FWO fellow at KU Leuven and an Assistant Professor of Clinical Psychology at Tilburg University. His research focuses on advancing scientific knowledge about why (non-suicidal) self-injurious thoughts and behaviours (e.g. cutting, hitting oneself) emerge and how to predict and prevent their occurrence. Glenn loves to travel and explore new places and cultures and is the proud owner of an adorable Great Dane.

Martien Wampers obtained a PhD in Psychology and works as a research psychologist in Universitair Psychiatrisch Centrum KU Leuven and as a database manager in the Research Group Psychiatry at KU Leuven. There, she is involved in most studies using the ESM to ensure data quality and consistency. She likes her cat, pointless activities, non-functional biking, and making things (from cookies to sweaters and everything in between).

Aki Rintala works as a senior lecturer in physiotherapy at LAB University of Applied Sciences, Lahti, Finland and collaborates with the research

team at the Center for Contextual Psychiatry, KU Leuven, Belgium. His research interest is in monitoring daily life in people with neurological conditions to understand the link between daily life experiences and treatment strategies using ESM. He likes to cheer for Finland in all kinds of competitions where Finland could make it to the finals, from sports to Eurovision.

Silke Apers works as a lab manager and financial coordinator at the Center for Contextual Psychiatry at KU Leuven in Belgium. She coordinates CCP projects that examine real-time and real-world person-environment interactions in the field of mental health by using the ESM. She provides support and oversees the methodological and practical management of the various ongoing studies. Silke is a lover of naps, melted ice-cream and Netflix. She's also a proud mom of two (mostly) wonderful daughters.

Tessa Biesemans was a Research Coordinator at the Center for Contextual Psychiatry at KU Leuven. Previously, she worked as a research assistant involved in multiple ESM studies, including being responsible for organizing briefings and debriefings. In her role, she coordinated ESM research projects exploring real-time, real-world interactions between individuals and their environment, focusing on stress, psychosis and other mental health difficulties. Outside of work, she finds joy in exploring nature and challenging herself with games and puzzles.

Steffie Schoefs works as Research Coordinator at the Center for Contextual Psychiatry at KU Leuven and is a clinical psychologist by training. She coordinates research projects that investigate real-time, real-world person-environment interactions in mental health using the Experience ESM and supports the methodological and practical aspects of various ongoing studies. In her free time, she enjoys festivals and concerts and hiking with her two dogs.

Wolfgang Viechtbauer is associate professor of methodology and statistics in the Department of Psychiatry and Neuropsychology and the Mental Health and Neuroscience Research Institute at Maastricht University in the Netherlands and the Center for Contextual Psychiatry at the University of Leuven in Belgium. His research is primarily focused on the statistical methods for meta-analysis, but his interests more generally encompass

the design and analysis of longitudinal and multilevel studies using appropriate mixed-effects models. In addition, he supports his colleagues in their research on the mechanisms through which social, genetic and environmental factors interact and contribute to the development, persistence and treatment of psychiatric disorders. We can neither confirm nor deny that Wolfgang is actually a cat.

Ginette Lafit obtained a PhD in Business Economics and Quantitative Methods at Universidad Carlos III de Madrid, Spain. Currently, Ginette is an Assistant Professor at the Methodology Research Group of Educational Sciences at KU Leuven. Ginette works in the field of statistics applied to psychology and mental health. Her research is focused on addressing methodological complexities and developing statistical methods to perform sample size planning for intensive longitudinal designs as well as delineating methodological pitfalls in the implementation of open science practices in intensive longitudinal research. To make the statistical methods easily available, Ginette develops open-source software. She loves theatre and dance, and she would like to have a second life to create a performance based on the history of feminism in Latin America for her daughter.

Joanne Beames is a Marie Skłodowska-Curie Postdoctoral Research Fellow and Clinical Psychologist who recently moved from Australia to join the Center for Contextual Psychiatry at KU Leuven. She has expertise in the prevention and early intervention of depression in young people, the development and evaluation of e-health technologies and methods to implement solutions at scale. The aim of her fellowship is to explore experiences of anhedonia, a lack of pleasure and/or interest, in daily life during individuals' adolescence and early adulthood. Her future research vision as a scientist–practitioner is to develop interventions targeting anhedonia that are more effective (and accessible) than those currently available. Joanne likes hiking and board games and thinks that you should go to Australia (despite that irrational fear of 'scary' animals).

Lotte Uyttebroek is a PhD student at the Center for Contextual Psychiatry at KU Leuven and clinical psychologist by training. Lotte is passionate about exploring the clinical potential of the ESM and integrating it into mental healthcare. In her free time, she enjoys concerts, triathlon

training and painting by numbers. She also loves fashion and analogue photography.

Evelien Schat is a postdoctoral researcher at the Research Group of Quantitative Psychology and Individual Differences at KU Leuven. Her research focuses on developing and evaluating statistical process control methods for the real-time detection of early warning signals of depression using ESM data. Next to being a researcher, Evelien is a yoga teacher who enjoys hiking, climbing and spending time in the mountains.

Aleksandra Lachowicz was a doctoral candidate at the Center for Contextual Psychiatry at KU Leuven when this handbook was written. Her research focused on the process of psychophysiological recovery from daily stressors and its connection to anxiety symptoms. Aleksandra is excited about what the future holds for the field of wearable technology and hopes to contribute to its development. Outside of her academic work, Aleksandra enjoys reading nineteenth-century novels and is a huge podcast enthusiast – she once estimated that over the past few years, she has spent nearly 2,000 hours listening to them.

Laura Van Heck is a doctoral student at the Center for Contextual Psychiatry at KU Leuven. Her research focuses on identifying interindividual differences in the experience of stress and how this relates to psychopathology. Additionally, she has an interest in improving methods to measure stress using wearable sensors in daily life. In her free time, Laura enjoys backpacking, swimming, painting and curling up on the couch with her cat and a good book.

Koen Niemeijer is a postdoctoral researcher at Belgium's KU Leuven – University of Leuven. His research focuses on using mobile sensing to capture and predict moment-to-moment affective time dynamics and mood disorders. He is particularly interested in predicting core affect and depression by leveraging both mixed-effect and machine-learning models to create personalised profiles that can be used to detect critical changes in emotion.

Thomas Vaessen is an assistant professor at the Centre for eHealth and Well-being Research at the University of Twente, The Netherlands, and

a senior research fellow at the Center for Contextual Psychiatry and the Mind Body Research Center at KU Leuven University, Belgium. His research focuses on the role of stress and stress recovery in the development of psychopathology and early interventions focused on coping behaviour in the context of stress. In his leisure time, despite a striking lack of talent, he likes to play football, guitar and social deduction games. Future research should uncover why he persists in engaging in these activities.

Otto Versyp is a doctoral student at the research group of Quantitative Psychology and Individual Differences at KU Leuven. His research focuses on (dysregulated) interpersonal emotional processes, but he's broadly interested in emotions, psychopathology and dyadic experience sampling. He loves (wild) camping, mountains, movies and reading fiction. Although he was once sceptical of cats, he now worships the ground his cat walks on.

Liesse Frérart is a PhD student at KU Leuven. Her research focuses on the relationship between emotions and sexuality in romantic couples, with a particular interest in how sexual activity and behaviour interact with dysregulated emotions. She pays special attention to using ESM to investigate the dynamic interplay between partners and how these processes unfold in daily life. Liesse enjoys browsing (way-too-expensive) real estate listings and having deep, one-sided conversations with her cat about the mystery of her missing canine tooth.

Martine Verhees is a post-doctoral researcher at KU Leuven whose research centres around interpersonal relationships in both parent–child and romantic contexts. Her work focuses on attachment and interpersonal emotion dynamics and their relations with well-being. She is also interested in advancing methods for interpersonal research. Martine enjoys running, reading and cooking, but since becoming a parent, her main hobby has become building impressive block towers – though they tend to have rather short lifespans thanks to some enthusiastic 'help' from her kid or cat.