

---

The Open Handbook of

# **EXPERIENCE SAMPLING METHODOLOGY**

---

A step-by-step guide to designing,  
conducting, and analyzing ESM studies

**Second Edition**

Edited by

**INEZ MYIN-GERMEYS & PETER KUPPENS**

The center for Research on Experience sampling and  
Ambulatory methods Leuven (REAL) - Belgium

Copyright © 2022 Germeys, Kuppens  
All rights reserved.  
ISBN: 9798417682889

*The advent of smartphones, apps and wearables has ultimately integrated real-life and real-time methodology, such as the Experience Sampling Methodology (ESM), into the toolbox of psychological science. Whereas books on how to analyze intensive longitudinal data are numerous, wisdom on designing ESM studies was for a long time only transferred by word-of-mouths from senior experts to their PhD students. The Open Handbook of Experience Sampling Methodology, edited by Inez and Peter, does close this gap entirely. As a step-by-step guide, the book paves the way on all elementary questions, like assessment frequency, sampling scheme, questionnaire density, ethical questions, and debriefing, to name a few. This handbook will be an abundant resource in educating upcoming generations.*

Prof. Dr. Ulrich Ebner-Priemer, Karlsruhe Institute of Technology, Central Institute of Mental Health, Mannheim

*This book delivers perfectly on its title. The Open Handbook is an immensely practical step-by-step guide to running studies using experience sampling methods (ESM). Combining foundational knowledge of ESM with the latest evidence-based developments, this handbook sympathetically walks readers through a decisional journey of what to do “before”, “during” and “after” conducting their ESM study. Readers will gain valuable insights such as why briefer-is-better in designing repeated mobile surveys, the difference between ideal versus realistic studies, tricky ethical considerations, a list of top five technology platforms that won’t let you down, the importance of piloting, and how to wrangle your gigabytes of data. Deftly edited by leading ESM experts Inez Myin-Germeys and Peter Kuppens, The Open Handbook is a must-read for graduate and postgraduate students, early career researchers, and seasoned researchers wanting to innovate their methods and quell their ESM FOMO. I highly recommend this resource.*

Tamlin S. Conner, Ph.D., Department of Psychology, University of Otago  
Co-editor of the Handbook of Research Methods for Studying Daily Life (Mehl & Conner, 2012)

# CHAPTERS

Chapter 1: Experience Sampling Methods, an introduction

*Inez Myin-Germeyns & Peter Kuppens*

Chapter 2: Research questions that can be answered with ESM research

*Peter Kuppens & Inez Myin-Germeyns*

Chapter 3: Designing an Experience Sampling study

*Egon Dejonckheere & Yasemin Erbas*

Chapter 4: Questionnaire design and evaluation

*Gudrun Eisele, Zuzana Kasanova & Marlies Houben*

Chapter 5: Ethical issues in Experience Sampling Method research

*Olivia J. Kirtley*

Chapter 6: Experience Sampling platforms

*Jeroen Weermeijer, Glenn Kiekens & Martien Wampers*

Chapter 7: Briefing and debriefing in an experience sampling study

*Aki Rintala, Silke Apers, Gudrun Eisele & Davinia Verboeven*

Chapter 8: Structuring, checking, and preparing ESM data

*Wolfgang Viechtbauer*

Chapter 9: Statistical methods for ESM data

*Wolfgang Viechtbauer*

Chapter 10: Non-normal, higher-level, and VAR(1) models for the analysis of ESM data

*Ginette Lafit*

Chapter 11: Sample size selection in ESM studies

*Ginette Lafit*

Chapter 12: Interventions

*Ana Teixeira*

Chapter 13: Passive sensing

*Joana De Calheiros Vellozo, Koen Niemeijer & Thomas Vaessen*

# TABLE OF CONTENTS

1	Experience sampling methods, an introduction	7
1.1	What are experience sampling methods	10
1.2	The scientific roots of ESM	14
1.2.1	Ecological psychology	14
1.2.2	Quantified self	15
1.2.3	Embedded and embodied cognition and contextual science	16
1.2.4	Within and between-person differences and personalized approaches	17
1.3	Conclusion	18
	<b>Before: research questions and designing an ESM study</b>	<b>20</b>
2	Research questions that can be answered with ESM research	21
2.1	Research questions in observational ESM research	23
2.1.1	Research questions related to the behavior of one time-varying variable	24
2.1.2	Questions related to the behavior of multiple time-varying variables	28
2.1.3	Research questions related to person-level characteristics	30
2.2	Research questions involving non-natural variation	31
3	Designing an experience sampling study	33
3.1	A selection of five ESM studies as an example	36
3.1.1	Study 1: The ‘typical’ ESM study, example 1	36
3.1.2	Study 2: The ‘typical’ ESM study, example 2	37
3.1.3	Study 3: The ‘measurement burst’ ESM study	38
3.1.4	Study 4: The ‘single case’ ESM study	39
3.1.5	Study 5: The ‘short & intensive’ ESM study	40
3.2	The most focal design parameters of an ESM protocol	43
3.2.1	Study duration	43
3.2.2	Assessment frequency	44
3.2.3	Sampling scheme	45
3.2.4	Questionnaire density	49
3.2.5	Study device	50
3.3	Fine-tuning the ESM parameters of your study: A guiding framework	53
3.3.1	Force 1: Answering your ESM research question in an ideal world	54
3.3.2	Force 2: Answering your ESM research question in a world with practical constraints	55
3.3.3	Defining validity in an ESM context	57

3.3.3.1	Quantitative indicators of validity in ESM	58
3.3.3.2	Qualitative indicators of validity in ESM	59
3.3.4	Study factors that determine the optimal value of your ESM parameters	61
3.3.4.1	Sample characteristics	62
3.3.4.2	Construct features	64
3.3.4.3	Statistical analyses	65
3.3.5	Interdependencies between ESM parameters	67
3.3.6	Conclusion: pilot your study	69
4	Questionnaire design and evaluation	71
4.1	Constructing individual ESM items	73
4.1.1	Capture dynamic phenomena	73
4.1.2	Different timeframes	74
4.1.3	Wording	77
4.1.4	Response scale options	80
4.2	Constructing a questionnaire	81
4.2.1	Order of questions	81
4.2.2	Length of questionnaires	82
4.2.3	Control questions	83
4.3	Assessing measurement quality	84
4.3.1	Pilot testing	84
4.3.2	Intra-class correlation	85
4.3.3	Reliability	85
4.3.4	Different forms of validity	87
4.4	The ESM Item Repository	89
4.5	Beyond self-report	89
5	Ethical issues in Experience Sampling Method research	91
5.1	Inclusivity in research	93
5.2	Privacy and consent	95
5.3	Real-time data, real-time responsibility?	96
5.4	Participant burden	98
5.5	Reactivity	99
5.6	Conclusion	100
<b>During: Conducting an ESM study</b>		<b>102</b>
6	Experience sampling platforms	103
6.1	The online dashboard	105
6.1.1	ESM questionnaires	106
6.1.2	Sampling schedules	106
6.1.3	Enrollment of participants	107
6.1.4	Data analytics	108
6.1.5	Data download	108
6.2	ESM apps	108
6.2.1	Native or hybrid	109

6.2.2	Push notifications: a warning	109
6.2.3	Helpful app features	110
6.3	Wearables	110
6.4	Legal considerations	111
6.4.1	Data privacy and electronic devices	111
6.4.2	Clinical use of ESM software, a medical device?	112
6.5	Sustainability of ESM software and hardware	113
6.6	Recommended ESM platforms	114
6.6.1	Overview of ESM platform features	114
6.6.2	Practical advice	116
6.7	Conclusion	117
7	Briefing and debriefing in an experience sampling study	119
7.1	Briefing session	121
7.1.1	Preparation before the briefing session	121
7.1.2	Starting the briefing session with a participant	122
7.1.3	Practice the demo ESM questionnaire with your participant	125
7.1.4	Additional information	127
7.1.5	FAQs	129
7.1.6	ESM questionnaire for the researcher	130
7.2	Debriefing session	131
7.2.1	Checklist for how to brief your participant in an ESM study	133
	<b>After: The analysis of ESM data</b>	135
8	Structuring, checking, and preparing ESM data	137
8.1	Data structure	139
8.2	Example	141
8.3	Software and code	143
8.4	Data checks and preparation	143
8.5	Data visualization	150
9	Statistical methods for ESM data	153
9.1	Mixed-effects and multilevel models	156
9.2	Disentangling within- and between-person variability	157
9.3	Examining between-person differences	162
9.4	Examining within-person associations	164
9.5	Examining between-person differences in within-person associations	169
9.6	Disentangling within- and between-person associations	172
9.7	Examining lagged relationships	176
9.8	Controlling for autocorrelation	179
9.9	Controlling for time trends	181
9.10	Conclusions	183

10	Non-normal, higher-level, and VAR(1) models for the analysis of ESM data	185
10.1	Non-normal data	187
10.1.1	Dichotomous outcome	188
10.1.2	Count outcome	197
10.1.3	Non-normal positive continuous outcome	200
10.2	Three-level models	203
10.3	Multilevel vector autoregressive models	208
10.4	Conclusions	214
11	Sample size selection in ESM studies	217
11.1	Power analysis in multilevel models	220
11.2	Methodological approaches for power analyses in multilevel models	221
11.3	Illustrations	224
11.3.1	Illustration I: power analysis to select the number of participants	224
11.3.2	Illustration II: power analysis to select the number of time points	232
11.4	Additional considerations when selecting the temporal design	235
11.5	Feasibility and sample size planning in ESM studies	236
11.6	Conclusions	236
	<b>Future: New developments in ESM research</b>	238
12	Ecological Momentary Interventions: from research to clinical practice	239
12.1	Ecological Momentary Interventions	241
12.2	Research and assessment of EMIs	243
12.3	Benefits of EMIs for clinical practice	246
12.4	Needs and barriers for clinical implementation	248
12.5	Conclusion	249
13	Passive sensing	251
13.1	Using passive sensing to enhance ESM research	254
13.1.1	Smartphone sensing	255
13.1.2	Wearable sensing	256
13.2	Using passive sensing to substantiate ESM research	257
13.3	Triggering ESM questionnaires based on passive sensors	259
13.4	Participant burden	260
13.5	Conclusions	261
	Index	263
	References	269
	About REAL	302
	About the authors	303





# CHAPTER 1

## EXPERIENCE SAMPLING METHODS, AN INTRODUCTION

Inez Myin-Germeys and Peter Kuppens

*Taking a single snapshot is usually not the best approach to understand 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 behavior 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 behavioral 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 especially over the last 10 years. Whereas ESM originally was based on paper-and-pencil approaches, where participants were signaled by a programmable watch or an alternative signaling system 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 over the last 2 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 you take a closer look at the enormous amount 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 questionnaire, incentives given, sampling schemes or feedback provided. Whereas this heterogeneity may underscore the unique nature of ESM, with researchers using this technique for 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 on how to properly conduct an ESM study. A study by Janssen 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 definitely is problematic for research, for example reducing the possibility for replication, it is even more problematic when we start to develop clinical applications based on these methods.

The current book aims to help to overcome this problem by providing a thorough and careful description of all the decisions one needs to take when designing an ESM study and the consequences of making specific choices, as such providing an overview of the current state-of-the-art. The ESM design strongly depends on the phenomenon of interest, the expected time frame of its occurrence, and the research question. Therefore, ESM research by definition will come in different forms. Yet, it is important to make informed decisions taking the consequences of specific design choices into account as well as to use similar research designs for similar questions in order to allow and 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 are providing data in the real world. In contrast to an experimental approach, where one zooms in on one specific aspect of experience or behavior 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 uses naturally occurring changes to investigate their effects on the variable of interest. A third defining characteristic is that ESM 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 experiences as closely in time as possible to their actual occurrence. Therefore, ESM typically includes more assessment points per day, although some authors would include end-of-day diary studies in ESM.

*Box 1.1. Main characteristics of ESM*

<b>Real-world and real-time</b>
<b>Prospective measurement</b>
<b>Self-report</b>
<b>Structured diary</b>
<b>Appraisal of context</b>

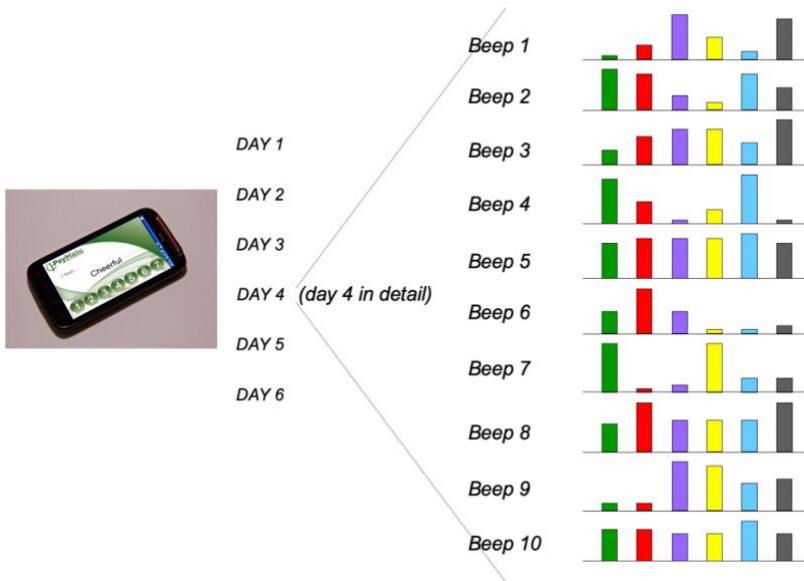
With ESM, participants are assessed prospectively in their normal daily life with questions about their actual mental state, thus reducing retrospective biases. Retrospective recall has been defined as an active reconstruction process which is subjected to cognitive biases (Stone et al., 2004), notably when recalling 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 the disproportionate importance of the individual's current state and most intense experience over the assessed period, i.e., peak-and-end effect in the 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. Yet, depending on the sampling scheme (number of assessments per day) and on the topic of interest, ESM questionnaires could still include retrospective questions (see chapter 3), in order to capture as much information as possible. However, they 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 (including timeframes of weeks to months or even a lifetime). Chapter 4 will focus on choices in questionnaire development and its' potential consequences.

The focus of ESM studies is on the subjective experience of the individual. The individual is considered the privileged observer and provides information about his or her mental state, mood, symptoms or context, using self-reports. This is in contrast with an observational approach, where one would

use for example video analysis to observe behavior in context (Luff & Heath, 2012). Whereas 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 at moments when they want to, ESM is a structured diary method, usually with a limited number of open questions. It consists of a structured questionnaire assessing specific experiences, typically inquiring about experiences in that very moment (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 being 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 will depend on the research question, but will also aid in reducing reactivity to the method. Usually, ESM researchers want to capture daily life processes without altering them (unless of course ESM is used in clinical practice where altering processes usually is the aim see chapter 12). This is not easy, when you ask 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 some ethical questions. What are the repercussions of doing this, can everybody do this and what questions are appropriate to be asked in that context? Chapter 5 will discuss the ethical aspects of ESM research.

Finally, ESM allows to include subjective appraisals of the context. Objective information about the context is useful, but does not necessarily reflect all relevant aspects. A snake being present will 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 day during a number of days.*

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



## 1.2 The scientific roots of ESM

### 1.2.1 *Ecological psychology*

In the 1970's, 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 behavior or experiences are radically situated, meaning that they can only be understood in relation to the context (Lobo et al., 2018). As a consequence, in order to fully understand experiences and behavior, they 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 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 that affected the behavior of the occupants. He described these 'behavioral 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).

The Experience Sampling Method 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 state of individuals and how these come about or interact with contextual factors. Mihaly Csikszentmihaly and Reed Larson developed the method with beeping people in the real-world as they were interested in what teenagers think and how they feel as they live their normal lives, spend time with their friends, interact with their parents or spend time at school (Csikszentmihalyi & Larson, 1984). Around the same time, Hurlburt developed his thought sampling, a structured way of capturing the stream of thinking 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 thinking and feelings such as loneliness or having low self-esteem, are not necessarily open to observation, the ESM researchers turned towards the privileged observer – the individual that is having

these experiences. Self-report became the standard of ESM research, including questions related both to the inner experiences and the context in which they occur.

So, ESM is rooted in ecological psychology with a focus on experiences as they occur in the real-world context. Adjacent to that, ESM is also claimed to be a research 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 therefore high in representativeness as it measures experiences in the real world. Generalizability, on the other hand, refers to how well an experimental task is actually predictive of its associated behavior or functioning in real-life. A generalizable task is not necessarily ‘real-world’-like. For ESM, the question pertains more to the generalizability of momentary behavior or experiences to overall functioning and behavior. 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 provides 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 quantified self is that you can gain self-knowledge through self-tracking of whatever variable you deem important, using technology (Wolf & De Groot, 2020). With apps, sensors or wearable devices, a diversity of things 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, kilometers travelled, goals achieved, but also mental state including mood or mental health symptoms. The goal of the quantified self-movement initially was to investigate what kind of self-tracking tools were available, what they could measure, and how that could be analyzed to be meaningful to the individual. Since then, the Quantified Self Movement has gained enormous momentum, with quantified self-labs and the Quantified Self Institute being established. Their current mission has now been expanded to improve quality of life by generating and sharing knowledge on Quantified Self (Wolf & De Groot, 2020).

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

Still, the advantage of being able to monitor continuously would be an important addition to ESM, as currently, people are only assessed at a limited number of times (e.g., 10 times per day). The use of wearable technology, using different sensors to capture aspects of the behavior, bodily experiences as well as the context provides a possible way forward to continuously capture relevant characteristics of the real-time and real-world interactions. Sensor information, derived from smartphones or wearables, can not only offer extensive information on bodily and behavioral features, tracking activity, heart rate or breathing, it 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 behavioral sensing, which would be the common denominator of these approaches (Mohr et al., 2020), could thus complement ESM, to create a meaningful understanding of the person-environment interactions. It would allow to capture additional contextual information, but 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 in itself 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, the real-world context is also consistent with a 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/brain is actively engaging with an everchanging environment in particular ways (de Haan, 2020; Myin-Germeys et al., n.d.). In order to understand these experiences, one needs to capture that dynamic interaction in 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, the thoughts and beliefs as well as the 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 to investigate mental states within a 4E-framework (Myin-Germeys et al., n.d.; Myin-Germeys et al., 2018). With ESM providing multiple assessments per person capturing different contexts, it allows to unravel temporal associations and identify 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 provides 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 compared to 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 is focusing on between-person differences, it has become increasingly clear that within-person variation is important as well (and that understanding the mechanisms involved in between-person differences are not necessarily the same as those involved in within-person differences; e.g., (Fisher et al., 2018). Chapter 2 will further discuss the different research questions that can be answered with ESM, relating to both within and between-person variability. As ESM per 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 is different 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 to outline the very specific patterns of behavior of one individual. ESM therefore is also useful for single-case studies, for example 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, for example investigating 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, but it also provides enormous opportunities for clinical application. ESM could be used to provide personalized feedback, thus enabling patients to get a better insight in their current mental states and behavior patterns that impact their symptoms and functioning. These personalized data could be shared with clinicians, helping in identifying personalized targets for treatment, as well as facilitating true shared decision-making (Myin-Germeys, 2020). Furthermore, it could be helpful in identifying moments when treatment is most needed, opening the way for ecological momentary interventions, bringing the therapy out of the office into the real life of people (Myin-Germeys et al., 2016) (see also chapter 12). ESM could thus contribute to a real needs- and patient-led personalized psychiatry.

### 1.3 Conclusion

ESM is an intriguing and interesting research method that has been around for over 40 years. Despite its exponential growth in many research fields including clinical and differential psychology and mental health, there is still a lack of methodological rigor and expertise, leading to a high degree of heterogeneity and lack of replicability. This book aims to help to overcome that 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 as well as analyzing ESM data. It therefore aims at both the beginning and more advanced ESM researcher, providing more practical advice (e.g., on briefing chapter 7, as well as on relevant digital platforms, chapter 6), as well as in-depth theoretical

discussions on the several research decisions. The ultimate goal of the book is not to direct everyone to one, unique ESM approach, but rather to provide a much clearer insight in the choices at stake 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 with experiment or data. While most of this volume is concerned with how to collect data in the most reliable, valid, and ethical way possible (see chapters 3, 4, and 5 using the experience sampling method, we should not forget that the data are there to answer our questions about the world in the first place. 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 of a generic nature, meaning that it can be applied to any domain of study, we will give concrete examples of research in order to illustrate how in practice the variety of research questions are being answered.

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

## **2.1 Research questions in observational ESM research**

Most of the existing, classic ESM research can be considered to be of an observational nature. In observational ESM research, one is interested in observing the natural behavior and ebb and flow of phenomena in the context of daily life. These phenomena are typically about how people feel, think, or behave, but can be as varied as ranging from lower order drives such as sex (e.g., Impett et al., 2008) and eating (e.g., Reichenberger et al., 2018), over higher faculties of the mind like morality (e.g., Hofmann et al., 2014), aesthetic experience (e.g., Nusbaum et al., 2014) or just letting the mind wander (e.g., Kane et al., 2007). The phenomena can be common and mundane (e.g., 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 behavior of one time-varying variable under study, (2) questions related to relations between several time-varying variables under study (within-person), and (3) questions about the role of person characteristics in time-varying variables (between-person).

In what follows, we will elaborate on and give concrete examples of these research questions. Of particular note is that this set of research questions follows the same buildup as chapter 9 about statistical modeling of ESM data, in the sense that they map onto (1) models with one time-varying variable, (2) models with more than one time-varying variable, and (3) models with person-level variables. As such, the reviewed research questions can be readily mapped onto the statistical models laid out in that chapter.

### ***2.1.1 Research questions related to the behavior of one time-varying variable***

We start with the simplest example, where a researcher has or is planning the collection of data related to a certain phenomenon of how people feel, think, or behave in the context of daily life. Several basic (but important) questions can be asked about the behavior of this phenomenon:

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

Of course, with ESM data, one is often interested in the changes, dynamics, temporal behavior of phenomena over time (why otherwise do ESM research in the first place right?). Yet, that does not imply that people's average feelings, thoughts, or behavior are not important to consider or study. To the contrary! One important reason is that several core concepts in the behavioral sciences are meant to reflect typical, habitual, "stable" forms of feeling, thinking and behaving and average levels of a variable in ESM data allow to get to these concepts in the context of daily life. Personality is a prime example of such a concept, and indeed prominent theories of personality describe people's behavior as a distribution of behavior of which the average or median is the most typical instantiation (e.g., Fleeson & Law, 2015). Relatedly, average levels are often

assessed in standard, trait-like measures of people's feelings, thoughts, and behaviors. An important type of research question therefore consists of examining the validity of these measures by relating them to average levels of people's feelings, thoughts and behavior, obtained in daily life (for an example of such research in the domain of emotion regulation, see e.g., Koval et al., submitted). 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 due to the fact that 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 mostly 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 modeling, 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 weird question on first sight, but is actually a crucial question in ESM research. The idea of assessing people's feelings, thoughts and behavior (or any other phenomenon) multiple times during daily lives, starts from the assumption that people's feelings, thoughts, and behaviors vary across time. As you know by now this is called within-person variability, and should be sizeable if ESM research wants to be meaningful (there is not much sense measuring the color of your eyes 6 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 mostly varies within persons, or rather between persons. An often-used fraction here is intra-class correlation coefficient, which is calculated between-person variance (within + between-person variance) and thus reflects the between-person variance over the total variance. If this is relatively large (e.g., higher than .5), this means that individual differences in the phenomenon under study are larger compared than within-person differences. Yet, if this is relatively low (e.g., lower than .5), this means that people differ more from themselves than people differ from each other.

Even in domains that study purportedly stable characteristics of the mind, this proportion can be surprising. For instance, personality is considered to be the stable set of characteristics that determines who you are as a person, and to lie at the base of individual differences in how people behave, think and feel. Yet, work by Fleeson and colleagues (e.g., Fleeson & Law, 2015) has shown that when repeatedly assessing personality-related behavior in daily life, the variance observed within persons is as large or even larger compared to the variance observed between persons (and is even comparable to the variance observed within persons in 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 to ask is what this variability looks like. Identifying and trying to understand the patterns and regularities that characterize the fluctuations of your phenomenon will without doubt contribute greatly to understand its nature and underlying mechanisms (for an example of this exercise in the domain of emotion or affect, see e.g., Kuppens & Verduyn, 2017).

The type 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, seasons, or over time more generally. For instance, Park and colleagues (M. Park et al., 2019) showed how music preferences change reliably over the course of a day (with less intense music being preferred in the mornings versus evenings). As another example, 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 time itself. This function can be linear, like for instance the personality traits of agreeableness and conscientiousness that show a more or less steady increase in the population throughout adulthood (Roberts et al., 2006) or non-linear, like for instance the alleged U-shaped relation between happiness and age (Frijters & Beaton, 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 often (but not always) 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 in their data. This is often one of the reasons why researchers sometimes shy away from using the first few assessments in the analyses of intensive longitudinal data.

Second, next to being predicted by day, week, or time itself, one may ask whether a particular phenomenon is predicted by itself over time. This refers to 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 it 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, like for instance emotion or affect (Kuppens, Allen, et al., 2010; Suls et al., 1998). The extent to which something is self-predictive, says a lot about the nature of it: If something is highly auto-correlated, it means that it is generated by a system that is not very much impacted by outside influences but is running its own course. If on the other hand, something is not very highly auto correlated, it means that it is very much under 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 being 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 behavior. Yet, a person who feels good the first part of the week and bad the second part of the week, shows the same variability yet 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 (that does not take the ordering into account) with indices of instability (that looks at the size of consecutive changes and therefore takes 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 e.g., Jahng et al., 2008). For instance, Thewissen and colleagues (2008) examined the importance of this distinction with respect to self-esteem and paranoia. As another example, Koval and colleagues (2015), examined this question for emotion in relation to depression.

### 2.1.2 *Questions related to the behavior of multiple time-varying variables*

Of course, often you are not interested in observing just one single variable, but in observing multiple variables, and how they relate to or even predict each other. After all, any theory that tries to explain something has at least two elements, that what 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, and 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 you're looking at concurrent or time-lagged relationships (see chapter 9) between variables.

In concurrent within-person relationships, you're examining how the fluctuations in one variable coincide with the fluctuations in another variable. For instance, van der Steen and colleagues (2017) examine the concurrent associations between momentary stress, affective and psychotic symptoms in populations with varying risk for psychosis. As another example, Hermans and colleagues (2020) study the relation between activity pleasantness and positive affect. When one has data from other channels than ESM, one can also examine relations between experiential, physiological, and behavioral systems. As an example, Vaessen and colleagues (2018) described a curvilinear relationship between self-reported activity-related stress and cortisol fluctuations. In principle, the variables do not even have to be from the same individual. 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. For instance, Sels and colleagues (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 a 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. Here, the researchers were interested in whether low levels of self-esteem or high levels of anxiety were able to predict the fact that soon after a paranoid episode could occur, and therefore tested lagged relations between the former two and the latter (the lagged refers to the fact that 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 the subsequent engagement in non-suicidal self-injury in individuals diagnosed with borderline personality disorder. A special, but often studied type of lagged relation asks the question 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 behavior or 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 use 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). The large benefit of this type of analyses over concurrent relations, is that they give an indication of the temporal directionality of the relation between variables, hinting at what comes first and what comes second and therefore 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 directional. This network approach has in recent years become very popular to study for instance the organization of symptoms in psychopathology (e.g., Borsboom & Cramer, 2013; Robinaugh et al., 2020), and has been applied to ESM data (e.g., Klippel et al., 2018; Wigman et al., 2015).

All of the above is in principle strictly concerned with understanding the behavior of one or more variables, how they unfold, relate, and interact over time within one or more individuals (or between individuals as 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 relate 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 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.

A first question pertains to charting the observed variation and trying to understand 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 (just like 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 underling the dynamics of their emotional lives.

A second question then pertains to trying to understand the observed variation by directly relating it to third (person-level) variables. The list of individual differences variables that typify and/or shape people's feelings, thoughts, and behavior 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, much of the concrete research examples 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. For examples with more biologically oriented individual differences, for instance 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 also, it is 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 are needed for causal interpretations, which calls for deliberately manipulating factors so to change people's feelings, thoughts and behaviors in daily life.

## **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 their object of study during the normal daily routines that make up our lives. With as much as can be learned from that, it is limited in the sense that sometimes researchers may be interested in the effects of particular type of events on people's feelings, thoughts, and behaviors, and in daily life these may be difficult to predict, let alone control. It is also limited in the sense that normal daily life may more often than not be rather ordinary, and may not always elicit very much of the feelings, thoughts, or behaviors we are interested in. Moreover, when studying the effect of interventions, there is an increasing interest in studying their effects not only in the lab or in the clinical center, but also how interventions eventually impact people's functioning in daily life. For these reasons, ESM is being used to also study different 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 the literature where ESM was used to study people's feelings, thoughts and behavior around well-specified type 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. A famous  $n=1$  example of where researchers examined changes in the wake of an important event is the study by Wichers and colleagues (2016) where they

examined symptom fluctuations before and after decreasing anti-depressant medication in a single patient. These kinds of designs are very meaningful to study 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 a manipulation in 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 before and after they were subjected to a classic stress manipulation (having to give a public speech). While the manipulation itself is classic in laboratory research, the interesting element is that here in this type of research, researchers can examine the effects beyond the lab, and see how the effect of such type of manipulation is anticipated and lives on in daily life.

Finally, researchers may want to test the effect of interventions meant to benefit people's everyday feelings, thoughts, and behaviors. In a way, whether or not people's real lives are eventually affected by certain type of intervention or treatment should be the ultimate criterion to evaluate them on (see chapter 12). Increasingly researchers are therefore 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 the ability to differentiate negative emotions in the context of daily life. A special example of intervention is when the ESM data itself is being used as an element of the intervention, for instance when the idiographic data is used to provide feedback and lifestyle advice (see, e.g., Kramer et al., 2014). Finally, the use of smartphones in contemporary ESM research carries in it the possibility to deliver interventions using the same smartphone and evaluating them at the same time. This is the domain of so-called ecological momentary interventions (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**

Egon Dejonckheere & Yasemin Erbas



In this chapter, we will discuss how you can set up your own study using experience sampling methodology (ESM; Csikszentmihalyi & Larson, 1987). The aim of this chapter is 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 these study settings as a function of the research question you wish to answer.

The chapter consists of three parts. In the first part, we will start by describing five ESM studies that differ in the study design that was adopted. These studies will give a first impression of what types of study designs are possible. They will also serve as examples to illustrate which decisions need to be made with regard to the different design parameters, and we will refer to these examples throughout the chapter. For illustrative purposes, we selected five ESM protocols that considerably differ from each other in various ways. In the second part of this chapter, we will introduce a number of design parameters. We will provide a descriptive overview of the parameters and discuss a number of different options. 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. Finally, in the third part, we will introduce a generic theoretical framework that will help you to fine-tune these parameters to the needs of your individual study.

Before we start, however, we would like to make two general disclaimers. First, not all recommendations we present here are entirely evidence-based. While ESM is certainly not a new data collection method (its first use dates back from the eighties; see chapter 1), methodological studies that evaluate the effect of some parameter specifications are still largely lacking. Therefore, there is still relatively little empirical knowledge on how to select the optimal parameter specification for your ESM protocol. Consequently, we will provide you with some guidelines based on our personal experience, and back up our rationale with findings from methodological research where possible. 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 highly depends on your specific research question, your study population, your research infrastructure, and so on. Therefore, this chapter cannot tell you what the best design is for your particular ESM study. Instead, its aim is to promote careful attention to certain aspects that need to be considered when launching your own

ESM study. As such, it will provide you with a general framework to carefully design an ESM study that is optimal for your research question.

### 3.1 A selection of five ESM studies as an example

Here, we provide a short discussion of five different ESM studies. While all studies investigate how certain aspects of our emotions unfold in daily life, they differ drastically from each other with respect to their design. We have selected these studies because they will give you a flavor of the different types of study designs that you can implement to answer different types of research questions. In Table 1 at the end of this paragraph, you will find an overview of the most important characteristics of the five studies.

#### 3.1.1 Study 1: The ‘typical’ ESM study - example 1

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 included three groups: patients with psychotic illness (high vulnerability), their first-degree relatives (intermediate vulnerability), and control subjects (low vulnerability), and each group consisted of 50 participants.

The ESM part of this study lasted for 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 signaled ten times a day at random moments during the waking hours (from 7.30 a.m. until 10.30 p.m.). At each signal, participants were instructed to complete an ESM assessment form. This form consisted of 46 items, among which items to measure positive and negative emotions, and 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 seven-point Likert scales (1-7), indicating “not at all” to “very”. For stress, four different measures were computed which assessed event-related, activity-related, thought-related and social stress. To measure event-related stress, participants were first asked to report and rate the most important event that occurred between the current and the previous measurement occasion on a seven-point bipolar scale ranging from -3 (very unpleasant) to 3 (very

pleasant). Activity-related stress was measured with three items (all on seven-point Likert scales) assessing participants' current activity (e.g., "I am not skilled to do this activity"). Thought-related stress was measured with a single item (i.e., "my current thought is unpleasant"), rated on a seven-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 time-frame, the time at which participants indicated they completed the ESM questions was compared with the actual time the digital watch signaled the participant. All ESM questionnaires that were filled in 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 Study 2: The 'typical' ESM study - example 2***

The second study that we discuss is by Pe and colleagues (2013). Here, the researchers were interested in whether the ability to update positive and negative stimuli in working memory (as evaluated with an affective n-back task in the lab) related to the cognitive and affective components of subjective well-being in daily life and at the trait level (assessed with traditional self-report questionnaires). Because the research question revolved around the person-level working memory ability, its main focus was to investigate between-person differences, and 95 participants took part in this study. In a baseline session in the lab, participants completed the affective n-back task among other lab experiments, and a battery of self-report questionnaires that comprised the Satisfaction With Life scale (Diener et al., 1985) to measure the cognitive component of subjective well-being, and the positive and negative affect scales of the Positive and Negative Affect Schedule (Watson et al., 1988) to compute an index for affect balance, the affective component of subjective well-being.



At the end of the lab session, the ESM part of the study started. Participants received a palmtop computer, and instructions on how to use it. The ESM part lasted seven consecutive days. The waking hours of each day (from 10 a.m. until 10 p.m.) were divided into ten equal time-intervals, and at a random time within each interval participants received a momentary assessment. Thus, in total, participants received ten momentary assessments per day for seven days. At each assessment occasion, participants filled in a questionnaire that consisted of 23 items, including six items (with a randomized order at each measurement occasion) measuring their current emotions (happiness, relaxation, sadness, anxiety, dysphoria, and anger). All emotions were measured on a continuous slider scale from 1 (not at all) to 100 (very much). These emotion ratings were used to compute a daily life index for affect balance, the affective component of subjective well-being.

Participants received up to €70 for completing the baseline measures and the ESM protocol. On average, participants completed 92% of the momentary ESM surveys.

### **3.1.3 Study 3: The ‘measurement burst’ ESM study**

The third study that we discuss 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). We refer to this ESM protocol as ‘measurement burst’ because it involves a design that consists of bursts of intensive repeated measurements within a shorter time-period, and these bursts are repeated longitudinally, over larger time-intervals (Stawski et al., 2015). A measurement burst design is ideal for the study of short-term variability, long-term change, and the individual differences therein, and it is therefore an ideal design for the current study that was interested in both between-person and within-person effects. Here, the study consisted of three measurement waves or ‘bursts’ of seven consecutive days of ESM, and each ESM protocol started with a baseline session in the lab. The second wave took place four months after the first wave, and the third wave took place 10 months after the first wave. The study had a sample size of 202 participants, which is larger than the first study’s sample size.

At each baseline session, participants completed lab-tasks and self-report questionnaires, including the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977), which was used to determine their depressive symptom severity. The lab sessions also included structured clinical interviews. At the end of the lab session, participants received a Motorola Defy Plus Smartphone along with instructions for its use, after which the ESM phase began.

Similar to the ‘typical’ ESM study, waking hours of each day (from 10 a.m. until 10 p.m.) were divided into ten equal time-intervals, and at a random time within each interval, participants received a momentary assessment. Thus, in total, participants received ten momentary assessments per day for seven days in each of the three waves. At each assessment occasion, participants filled in a questionnaire that consisted of 24 items, including six items (with a randomized order at each measurement occasion) measuring their current positive (happy, relaxed, cheerful) and negative emotions (sad, anxious, depressed, angry, stressed). All emotions were measured on a continuous slider scale from 1 (not at all) to 100 (very much). For each participant, the means of the positive and negative emotions were then computed for each measurement occasion, and correlated across the measurement occasions, which gave an indication of the participant’s level of affective bipolarity.

As an incentive, participants were paid up to €60 per wave, and 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 wave, respectively.

#### ***3.1.4 Study 4: The ‘single case’ ESM study***

The fourth study is by Wichers and Groot (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 study consisted of only one participant who was a clinical patient: The participant was 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 more details].

During the course of the study, his antidepressant was gradually discontinued. The participant and researchers were blind to the dose reduction scheme, and the participant was obviously also blind to the research question. Because the study focused on within-person change over a longer time period for a single participant, we refer to this study as a ‘single case’ or N=1 ESM study.

In total, the participant received 1,474 momentary questionnaires over the course of 239 days. Similar to the ‘typical’ and ‘measurement burst’ ESM studies, the waking hours were divided into ten equal intervals, and at a random time within each interval, the participant received a momentary survey on a palmtop computer. Each momentary survey consisted of 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 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 on a daily basis for an extended period of time. As such, there was no additional monetary incentive, and the participant complied to 62% of the momentary surveys.

### **3.1.5 Study 5: The ‘short & intensive’ ESM study**

The fifth example is the second 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 with the DynAffect model. According to this model, there are three major processes that underlie individual differences in how our emotional experiences change over time [i.e., the affective home base, variability, and attractor state; see Kuppens and colleagues (Kuppens, Oravecz, et al., 2010) for more information regarding the model]. To evaluate this model, the researchers looked at data from two studies. The first study was rather similar to the ‘typical’ ESM study we discussed earlier. The second study was different from previous studies because it focused on micro within-person changes (i.e., changes within individuals on a very short time-frame). Therefore, we refer to this study as the ‘short & intensive’ ESM study.

Participants were 60 students from KU Leuven university. The study started with a lab session where participants completed self-report questionnaires and received a Tungsten E2 palmtop computer along with instructions for its use. After this initial session, the ESM period started, and only lasted for four consecutive days. The waking hours of each day were divided into 50 equal time intervals, and the palmtop was programmed to prompt a survey once within each interval. As a result, the average time between two momentary surveys was 17 minutes. Consequently, compared to the 10 momentary surveys a day in the previously discussed ESM studies, this study had a very intense sampling scheme and was more intrusive. Therefore, at each sampling occasion, the participant received only seven items. One of these items was an affect grid, which is a single-item measure designed to simultaneously assess subjectively felt valence and arousal (Russell et al., 1989), and the affect grid was used as an indication of participants' core affect. Participants were paid €50 for their participation and they completed on average 87% of the momentary surveys.

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. 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 density	Device
‘ typical’ -1	150 (patients, relatives, controls)	6 days	10/day	Semi-random	46 items	Wristwatch/ paper-pencil
‘ typical’ -2	95 students	7 days	10/day	Semi-random	23 items	Palmtop
‘ measurement burst’	200 students	3 x 7 days	10/day	Semi-random	24 items	Smartphone
‘ single case’	1 MDD patient	239 days	10/day	Semi-random	13 items	Dedicated device
‘ short & intensive’	50 students	4 days	50/day	Semi-random	7 items	Palmtop

## 3.2 The most focal design parameters of an ESM protocol

In the second part of this chapter, we will introduce a number of 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 density and study device. Like we already mentioned, this is not a conclusive list. Rather, these are the most basic, and some of the most important parameters that structure your ESM protocol. Consequently, these are the factors that you definitely need to consider first when you start designing your ESM study, regardless of what type of study you are planning.

### 3.2.1 *Study duration*

As the name gives away, 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 again the specification of this parameter mainly depends on your research question. Are you interested in the relationship between certain variables between individuals (for instance, in how updating of positive stimuli in working memory is associated with subjective well-being, the research question in the ‘typical’ ESM study)? Or are you interested in processes within individuals (for instance, whether changes in the level of emotional inertia can predict the onset of a depressive episode, the research question in the ‘single case’ ESM study)? And do you expect these changes 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 ‘short & intensive’ ESM study)?

As we have seen, the two ‘typical’ ESM studies lasted six and seven days. But this is not the gold standard, a few days more or less is not a big difference. A recent meta-analysis based on 79 datasets showed that the duration of the studies ranged from one to 150 days, the mean duration across the studies was 11.2 days, and most studies (68%) lasted between two and ten days (Vachon et al., 2019).

Importantly, typically, when running an ESM study, not all participants start at 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 get you to sample both week-days and weekend-days in all participants, eliminating potential differences due to week-days (when most people work or study) versus weekend-days (when most people have more time for leisure activities and social contact), is recommended. But perhaps you are only interested in week-days (e.g., when you study people's behavior in the office), or only in weekend-days (e.g., when you study how people divide their free time, or how working parents interact as a family with their children). In that case, a protocol of seven consecutive days would not be very useful, and you can decide to sample only during office hours or during the weekends (depending on the focus of your study), for a number of weeks.

An ESM study can also be interrupted by measurement-free phases. While the second 'typical' ESM study had a duration of seven consecutive days, the 'measurement bursts' study also had an ESM protocol of seven days, but the protocol was repeated three times (i.e., three waves over the course of a one-year period). 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 'short & intensive' study was quite a bit shorter than the other studies and lasted only four days, because the researchers were interested in micro-level changes (and when we discuss the parameter 'assessment frequency', you will understand that this study was shorter because the assessment frequency was very high, making it a very demanding study protocol for the participants; see also Part 3 of 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, beeps or prompts) per day, and mainly depends on what a meaningful frequency is for the topic of your study (see 3.3 of this chapter for more information). For instance, if you are interested

in sleep quality, it is enough to only assess this 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 you are interested in people's current emotional states, which is a highly variable construct, a higher sampling frequency is likely more informative (Kuppens, Oravecz, et al., 2010).

Another important question with regard to assessment frequency, is whether you are interested in micro-level or macro-level changes. If we look at our examples, we see that the 'short & intensive' study, which was looking at micro-level changes in affect, had a sampling frequency of 50 momentary assessments per day. The other four studies were designed to look at changes on a slower time scale, and had ten momentary assessments per day (note that ten is not the gold standard, although it is an often-used number). Finally, as a general rule, you want to be sure that the study duration and the assessment frequency together make for enough data points to do the required analyses (see Part 3 of this chapter for more information).

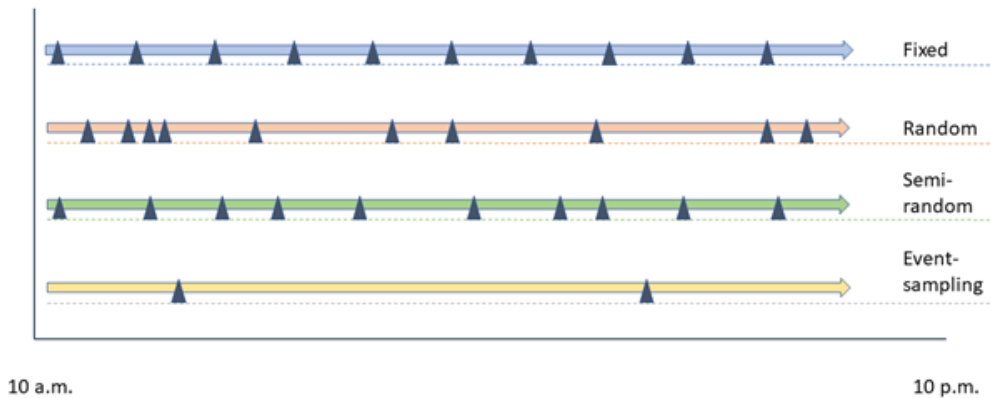
There are two parameters that are related to the assessment frequency: the start- and end-point of the day, and the between-prompt interval. The start- and end-point of the day refers to the time of the first assessment and the time of the last assessment, respectively. Here, you need to find out what the typical waking-hours are of your population and make sure that in normal circumstances, all assessments fall within these hours. A recent re-analysis of ten 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). Second, the between-prompt interval refers to the time in between consecutive assessments within a day. This is of course highly dependent on the number of assessment points: 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. There are different ways in which



assessments can be generated. Here, we will discuss the fixed scheme, the random scheme, the semi-random scheme and 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, between the start- and end-point of the ESM protocol. There are quite some advantages to a fixed scheme. When the assessments are scheduled at a fixed time, they are predictable. This will likely result in higher compliance to 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 will not be violated. However, there are also disadvantages to such a fixed scheme. The fact that the prompts are predictable, decreases the ecological validity of the study and increases reactivity to the method as people may change their behavior to answer the questionnaire (e.g., not starting an argument with your spouse, because you anticipate an assessment in two minutes). This would be unfortunate, because the high ecological validity is one of the main assets of ESM. Finally, because assessments always take place at the same time, assessment selection bias may occur, meaning that certain data can be over- 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 to be missing, 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 doesn't include evening assessments, then this data will be structurally missing and it will affect the generalizability of your 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 to the fixed sampling scheme. The main advantages 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 fact that participants cannot know when to expect an assessment can also result in higher participant burden and a decrease in compliance to the ESM protocol. Moreover, this type of sampling can, in extreme cases, also result in a very unequal distribution of the assessments. For instance, all assessments can take place in the morning, and information about the rest of the day will be missing. In such a case, the data will not be representative of the entire day. Additionally, such an uneven distribution of assessments will also lead to unequal distances between the time-points, resulting in a violation of the underlying assumption of many statistical models. To our knowledge, this type of sampling is not used very often, because the advantages 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 in between the start- and end-point of the day can be divided into a number of equal time intervals (and this number 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), a relatively low participant burden, and relatively small negative consequences for 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 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 event you are interested in is rare, or when you are interested in very specific situations or behaviors. If you are interested in such rare events or specific situations or behaviors, it is possible that the other sampling schemes will miss them. By explicitly asking participants to respond to a questionnaire after such an event, you can ensure that the event will not be missed. The main advantage of this event-based sampling scheme is that the assessments are tailored to study specific, potentially rare 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 (Himmelstein et al., 2019) investigating interpersonal behavior 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. 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 have to decide which events to include. For instance, when participants are asked to report every stressful event, it is possible that the type of events that are selected differ between participants. Additionally, because the assessment will only be initiated *after* the event has taken place, the questions with regard to 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 there will be no information available 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.

### 3.2.4 *Questionnaire density*

Questionnaire density refers to the number of questions assessed at each measurement occasion. Again, there is no gold standard for the optimal number of items. Similar to most other parameters, the number of items depends on important study factors (see 3.3 for more information). However, as also discussed in the fourth chapter on questionnaire development, there is research showing that generally, a higher density of a 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 higher questionnaire density negatively affected both data quality and compliance. In our example studies, the questionnaire in the first ‘typical’ ESM study consisted of 46 items, the questionnaire in the second ‘typical’ ESM study had 23 items and the questionnaire in the ‘measurement bursts’ ESM study had 24 items. The ‘single case’ ESM study had less items, only 13, but the study also lasted much longer than the other two studies. Finally, the ‘short & intensive’ ESM study lasted only four days. However, the sampling frequency was very high: participants had to fill in 50 questionnaires a day. As such, the questionnaire in this study consisted of only seven items. Thus, you see how the decision for this parameter already interacts with other parameters (see also 3.3 for interdependencies between ESM parameters).

There are two parameters that are relevant for the measurement occasions: the amount of time it takes a participant to initiate an ESM questionnaire, also referred to as the response delay, and 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 et al., 2021), showed that studies immensely differ from each other with regard to 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 and colleagues (2021) show that larger 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). With regard to retrospective questions, it is possible that larger response delays increase recall bias. In our example studies, the allowed response time was 15 minutes in the first ‘typical’ ESM study, 90 seconds in the second ‘typical’ ESM study and in the ‘measurement bursts’ study, 15 minutes in the ‘single case’ study, and 90 seconds in the ‘short and intensive’ 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 apply a maximum completion time. For instance, the second ‘typical’ ESM study and the ‘measurement bursts’ study both 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 responding, and it may be needed to account for that. For this reason, some researchers also recommend setting a limit for the minimum completion time (e.g., McCabe et al., 2012).

### 3.2.5 *Study device*

A device refers to the instruments that are used for data collection. There are many different ways through which ESM data can be collected. While the paper-and-pencil method (in combination with a programmed wrist watch that would prompt participants to fill in a momentary survey) traditionally used to be the most common way, nowadays, **electronic devices such as smartphones**

seem to be the preferred method for data collection (Trull & Ebner-Priemer, 2013). These electronic devices are now easily available, and indeed many people already own a mobile phone themselves. While researchers can rely on *research-dedicated devices* to conduct an ESM study, these can be quite expensive, especially if you only aim to run a single ESM study. Therefore, alternatively, you can use the smartphone of your participant, and directly download an ESM app to his or her device (e.g., m-Path; [www.m-path.io](http://www.m-path.io)). 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 (for instance prisoners) may not always have access to smartphones, and using research-dedicated devices may be necessary.

Apart from the paper-and-pencil method and electronic devices, there are also 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 where researchers called the participants once a day for seven days to collect their answers; Ryff et al., 2017). Another option is by using **online interfaces** (such as Qualtrics or Survey Monkey) where the participant answers to 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, through platforms such as [www.surveysignal.com](http://www.surveysignal.com)). In our example studies, data from the ‘measurement bursts’ ESM study was collected with research-dedicated smartphones, while the data from the single case study used a dedicated device and two other studies used palmtops.

All of these sampling methods have their advantages and disadvantages. The paper-and-pencil method has the important advantage that it is free to use. Moreover, most people will be familiar with this method. However, there are also some important disadvantages to consider. If a paper diary gets lost (e.g., it is easy to ‘forget’ on the bus), all data will be lost. Additionally, the data preparation process is long, and the data needs to be entered manually, which makes it susceptible for 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 the validity of the data. It must be noted though that another study by Jacobs and colleagues (2005) reported most people to comply with the ESM protocol when using paper-and-pencil approaches. In addition, a study comparing paper-and-pencil with a digital approach 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 smaller risk for errors. 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' own phones, and no extra devices will need to be carried. Another advantage is that with electronic devices, each questionnaire entry will receive a time-stamp. This way, the researchers will know exactly when a questionnaire was filled in. Furthermore, the researcher can program the questionnaires in such a way that the participant is required to answer them within a given time-frame (for instance, within 10 minutes of the actual prompt), and it is not possible to access the questionnaire after this period. As such, the backfilling issue associated with the paper-and-pencil method does not have to be an issue when using electronic devices. 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 when the electronic device does get lost, usually only a small amount of data will get lost with it. Finally, with the 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.

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. First, there may be issues with usability/accessibility of the electronic devices. Not all participant populations may be familiar with electronic devices (see 3.3 for more information), there may be ergonomic problems, or problems with the interface. 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). 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 device. 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. 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.

Finally, as mentioned before, there are also other methods 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 paper-and-pencil (Stone, Shiffman, et al., 2003). Therefore, based on these findings, 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 refrained to a purely descriptive level, covering each of the criteria in an isolated way as if they were to 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, like we mentioned in the introduction of this book chapter, each research hypothesis will call for its own individual study design, and there are no clear-cut rules or fixed regulations that you should follow. Instead, this section aims to provide a framework that may guide you in your decision making, and has the ultimate goal to promote careful consideration about the optimal set-up for your ESM study's parameters in light of the research question that you wish to answer.



Contrasting the five different ESM studies we reviewed earlier, it becomes easily evident that there is considerable heterogeneity in the protocols that 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 they tried to answer was the most important factor in determining their design choices for a daily life study (Janssens et al., 2018). But how do you tailor the design of your ESM study around your specific research question? In a general way, you could argue that the configuration of your study settings is the result of an 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' life at every given moment, but are limited in the study 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 study feasibility (i.e., reality; Hektner et al., 2007). In the next paragraphs, we will discuss these two opposing forces in a bit 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 validly 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 the fact that you would receive an uninterrupted stream of information about people's daily lives, with the level of detail similar to a movie. Additionally, as participants move in their personal ecologies, measurement would (b) not interfere with their daily routine, providing researchers with the most authentic insight into people's lives. As such, measurement practices would not unwillingly alter subjects' behavior (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), nor would important life events affect their responding (e.g., even though participants were scared to death, they would still complete a momentary survey about their

feelings; Sun et al., 2019). 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 behavior, would always result in honest responses; Fisher, 1993), and would easily eliminate reactivity issues (e.g., repeatedly prompting 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' life in all its facets, (d) not having to make an a priori selection of the constructs they wish to track, nor would they have to specify a predefined study period. Similarly, in terms of number of participants, drawing a particular sub-sample from the population they wish to make inferences about would not be required. Taken together, in an ideal world, maximizing the parameters of your 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 dream protocol is impossible, due to hindering laws and practical constraints (Elsschot, 1934). The crucial reason here is that, to gain insight into the dynamics of participants' internal world, contemporary ESM studies are still largely confined to self-report methods (although attempts to infer people's internal states via passive sensing are well underway; Mehl et al., 2021). Having your daily routine repeatedly interrupted to reflect and report on your mood, thoughts or symptoms is quite a challenging and burdensome task for participants to carry out on a regular basis (e.g., Fuller-Tyszkiewicz et al., 2013). Consequently, in terms of time, for example, ESM studies cannot last forever, pushing researchers to carefully consider the duration of their study. Similarly, researchers cannot harass participants constantly, inviting them to deliberately think about the timing and number of measurement occasions they want to impose (Silvia et al., 2014). In terms of constructs, researchers are limited in the number of items they can assess per construct, and how many constructs they can evaluate per momentary survey (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 sample size planning (Lafit et al., 2021). In sum, lowering the parameters of an ESM study has the advantage that it is less burdensome for participants and hence more feasible to carry out.

The previous discussion of these two forces illustrates that they work in an opposing direction with respect to determining the optimal value of your ESM parameters. Theoretically, maximizing the intensity of your ESM parameters provides you with more information about people's life, and hence would give you the most valid answer to the research question at hand. However, in reality this boost comes with 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. 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 intensity of your ESM parameters and the validity of collected data (see Figure 3.2). Although more information is better to a certain degree, further increasing the intensity of your ESM protocol will not automatically lead to an increase in validity. To the contrary, at a certain point on the real versus ideal continuum, a further increase in the value of your ESM parameters will be associated with a decrease in the validity of your data. Consequently, when tailoring your ESM study around a specific research question, your implicit aim is to pursue the saddle point in this curve-linear relation (i.e., where the relation between intensity in parameters and validity flips from positive to negative). Around this value, you may accomplish the most valid answer for your research question, because the parameters of your ESM protocol allow you to collect a maximum amount of information, with a minimum amount of burden or resources. As such, this sweet spot resembles the perfect trade-off between data quantity and quality, and constitutes the value of your assessment frequency, study duration, questionnaire density, etc. that yields maximum validity. Importantly, and we will continue to reiterate this point, the position of this optimal value will be different for different research questions (see Figure 3.2).

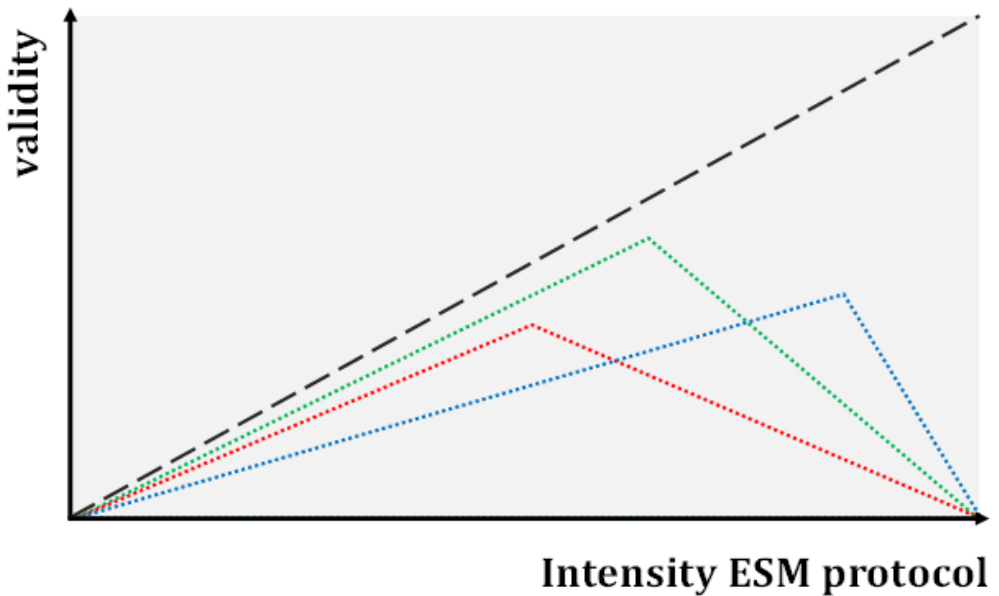


Fig. 3.2. A graphical representation of the inherent tension between the validity of ESM data (i.e., data quality) and the intensity of the protocol (i.e., data quantity). The black long-dashed line represents the ideal scenario, where maximizing the parameters of an ESM protocol leads to a more valid answer to 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 validity of the data. The implicit aim is to find the saddle-point in this relation, where validity is maximized and 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).

### 3.3.3 Defining validity in an ESM context

But how should we understand validity in the context of daily life studies? Like any other data collection method, and very broadly speaking, the validity of an ESM protocol refers to the extent to which its data, information or conclusions accurately correspond to the real world (Cureton, 1951). However, how this general definition specifically manifests in the context of ESM, and what particular barometers reveal adequate or poor validity of a protocol, is quite unique for this discipline. In the next paragraphs, we make a distinction between quantitative and qualitative validity aspects of ESM data.

### 3.3.3.1 *Quantitative indicators of validity in ESM*

The most important quantitative indicators of validity are compliance and attrition, and they are referred to as ‘quantitative’ because they can be numerically summarized. 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 was missed by a participant. Missingness is inevitable in daily life studies, and a small number of missed measurement occasions is generally not problematic. Indeed, in naturalistic settings answering to a measurement prompt may not always be feasible, either because the participant did 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, missingness can undermine validity, specifically when the pattern of missed occasions is no longer completely at random, but consistently depends 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, in the case of specific situations or events, missingness is problematic when it depends on the phenomenon that you are investigating (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 missing nature of these prompts, it is very difficult to determine what people are actually doing at that time, but the recent development of unobtrusive measurement devices allows researchers to by-pass this obstacle. In an inventive ESM study, for example, Sun and colleagues (2019) used Electronically Activated Recorders (EAR; Mehl, 2017) that logged short audio snippets of participants’ real-world behavior 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 missingness is not entirely dependent on specific instances that happen in the lives of participants.

Second, attrition (also called retention or drop-out) refers to the proportion of participants that does not reach the end of the ESM protocol, but prematurely wishes to abort the study (Vachon et al., 2019). Usually, attrition shows strong relations with compliance, because the participants that quit the study early are typically those who found that the protocol was too burdensome and interfered too much with their everyday lives (Delespaul, 1995). For this reason, it is common practice that these cases are not included into 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 at random, the danger exists that the study sample may not accurately represent the population you try 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). In case person-level information was collected during a baseline session prior to the ESM study, it may be worthwhile to statistically check whether drop-outs 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). 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 to understand why participants found the study too demanding (e.g., Eisele et al., 2020; the prompts were sent too early in the day, taking part in the study was not compatible with participants' jobs, carrying around another mobile device next to their personal smartphone was irritating, etc.).

### *3.3.3.2 Qualitative indicators of validity in ESM*

With respect to the most crucial qualitative indicators of validity, issues of interference and reactivity deserve special attention. 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 ones). Interference refers to the degree with which taking part in an ESM study hinders the occurrence of naturalistic or authentic behavior (e.g., playing sports, driving a car, taking a nap, etc.; Scollon et al., 2009). Once 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 hindering, potentially altering participants' behavior (Hormuth, 1986). However, whether interference becomes problematic because it substantially invalidates the data you collected, depends on both the participants' characteristics, as well as the ESM protocol that was adopted. With respect to participants, interference may be more of an issue when subjects already have a lot on their plate next to taking part in a study (e.g., parents with a new-born, students taking finals, couples organizing their wedding day, etc.). With respect to features of the study itself, the density of the protocol, together with the exact timing of the measurement prompts may affect the overall study interference. Specifically, having to complete more surveys on a daily basis, 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 with 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 behavior (Scollon et al., 2009). The mechanisms via which repeated self-assessments may initiate change in responding 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 so down like 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 behavior; Buu et al., 2020) or depressive symptoms (e.g., alleviation of depressed feelings; Broderick & Vikingstad, 2008; Kramer et al., 2014), than the rating of 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 behavior (Munsch et al., 2009). Subject susceptibility may evidently also play a role in reactivity. Conner and Reid (2012), for instance, found that reactivity in happiness was particularly evident in participants who showed low levels of trait negative affect. 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).

### ***3.3.4 Study factors that determine the optimal value of your ESM parameters***

Now that the fundamentals of our overarching framework are clear, and we know which quantitative and qualitative indicators to consider when evaluating the validity of ESM data, we provide some general guidelines about which aspects to pay attention to when trying to determine the optimal value for the parameters of your ESM protocol. Grossly, these aspects can be divided into three categories: (a) sample characteristics, (b) construct features, and (c) the statistical analyses you will rely on to answer your research question. The summary we provide is not meant to be an exhaustive overview, but rather reflects some basic considerations to take into account.



### 3.3.4.1 *Sample characteristics*

A first important point of attention pertains to the question 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 cornucopia of daily life research with patients suffering from depression (e.g., Heininga et al., 2019), borderline (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 start- and end-point of the day in light of 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 miss important information about their sleeping behavior, invalidating the representativeness of the data for their daily life. Individually adapting the start- and end-point of an ESM day to the sleep-wake cycle of a participant, or the assistance of passive sleep sensors that allows 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 density of your ESM protocol. Although compliance and retention rates are still acceptable in patient groups (i.e., typical compliance around 70-80% depending on the protocol), clinical populations tend to miss more measurement prompts and drop out from an ESM study more frequently, particularly 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., 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 (e.g., Cordier et al., 2016; Rah et al., 2006; Schmiedek et al., 2010), but particular 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 your protocol when studying younger age groups. Second, age may have implications for the device you will 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., not during classes, only after they finished homework, etc.), elderly may not be very familiar with these recent technologies, making a smartphone-based ESM study impractical. Relatedly, old participants' vision abilities may be impaired, favoring old-school paper-and-pencil techniques over answering survey items on more recent, yet smaller 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). With respect to the duration, assessment frequency and questionnaire density of your ESM protocol, attenuating these parameters may be justified.

Finally, the specific location of your participants may be something worthwhile to consider when designing your ESM study. Particularly, the device that you will use to collect subjects' momentary responses is partly determined by the presence of a (stable) mobile network or internet connection. Several kinds of ESM software regularly connect with a server to download the most recent information about the sampling scheme or to upload completed surveys to the database (to prevent complete data loss). In other cases, the entire sampling scheme is programmed on the mobile devices in advance, which requires no active internet connection at the time of the study, and allows participants to move freely in offline and remote contexts. When studying hard-to-reach participant populations (e.g., refugees, homebound, etc.; Gould et al., 2020) or exploring the impact of unusual circumstances (e.g., lockdown due to a pandemic, civil war, etc.; Stieger et al., 2020), written or telephone-assisted instructions to program the ESM software on their own personal phones (compared to handing out research-dedicated devices in real life), gives researchers the possibility to set up the study entirely from a distance, not needing to interact with participants directly.

### 3.3.4.2 *Construct features*

With respect to the specific feelings, symptoms or behaviors under study, you should think carefully about the prevalence of the phenomena that you are investigating. Among other parameters, this will have important consequences for the duration of your ESM study, because the regularity with which your 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 behaviors 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 your survey items can affect the degree to which participants will endorse momentary statements, and that altering the phrasing may change a constructs' 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 depressed patient groups, momentary endorsements of euphoria versus happiness differ in terms of frequency due to differences in item intensity (Heininga et al., 2019).

The prevalence of your constructs will also have implications for the sampling scheme that is adopted in your ESM study (Ram et al., 2017). With time-contingent ESM designs, the real-time occurrence of very rare feelings, symptoms or behaviors may easily be overlooked (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 occurred, but this sampling scheme has the drawback that it typically cannot give real-time information about the antecedents, as participants only initiated the questionnaire after the phenomenon of interest took place. Alternatively, modifying the phrasing 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 behavior 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 its experience; Colombo et al., 2020).

Closely related to the prevalence of psychological constructs is the time scale on which your phenomenon of interest operates. It pertains to the question how quickly a particular emotion, symptom or behavior is subject to change (Boker et al., 2009), and is critical to evaluate in light of the assessment frequency of your ESM design. On the one hand, oversampling the construct when temporal changes can hardly take place 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 tracking 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 this emotional recovery (Mestdagh & Dejonckheere, 2021). Nevertheless, determining the appropriate time scale on which psychological constructs change is far from trivial, and relying on different time frames will produce different conclusions (Neubauer & Schmiedek, 2020). As a guiding principle, the time interval between consecutive measurement occasions should be shorter than the rate of change of the phenomenon under study, in order to adequately pick up on the serial dependency between discrete assessments (Ram et al., 2017).

### *3.3.4.3 Statistical analyses*

Finally, it is vital to tailor your ESM protocol in function of the statistical analyses you will perform to obtain the answer to your research question in mind. Similar to other data collection methods, it is advisable to think about the

statistical models or techniques you will rely on to test your hypotheses before actually collecting the data. Modeling 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 prescreening 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 enroll 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 main 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 make valid and ecologically valid inferences. Because it is rather unlikely that participants will complete all measurement occasions due to interference (Vachon et al., 2019), it is advisable to anticipate a certain degree of missingness by introducing slightly more assessment occasions than that are strictly needed.

A specific type of research question related to the within-person level is the one that 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 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, urging you to increase the measurement frequency of your protocol.

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

### ***3.3.5 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 it is preferable to take into account 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 ultimate goal remains unchanged: Designing an ESM study that is as less obtrusive as possible, but allows you to answer your research question in the best possible way.

To illustrate the importance of taking into account the relative interdependencies between ESM parameters, let's consider and compare the protocols of these two real-life ESM studies, the 'single-case' ESM study and the 'short & intensive' 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 blend of the parameters study duration, questionnaire density and assessment frequency. In the 'single-case' ESM study, the researchers were interested in predicting the onset of a MDE based on changes in emotion dynamics. Because depressive relapse does not take place that 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) in order 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 density of the protocol. In order to not compromise validity, it is likely that the optimal value of these parameters is lower than the prototypical ESM study. Indeed, to minimize interference, poor compliance or drop-out, the selection of these study settings was tailored to the study duration of the protocol, leading to smaller values than usual. In contrast, in the ‘short & intensive’ 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 time scale 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 the average ESM study (i.e., 50 times per day). Once more, the selection of this value also had implications for the other study settings. To ensure the protocol would not be too burdensome, the researchers decreased the duration of the study and the questionnaire density.

It is of note, however, that the current examples are rather extreme in the selection of their ESM parameters (and therefore excellent for educational purposes). Within ‘normal’ ranges, the few methodological ESM studies that were conducted (e.g., Eisele et al., 2020) find little evidence for these complex interactions, suggesting that daily life researchers should not be too worried about these interdependencies under normal circumstances.

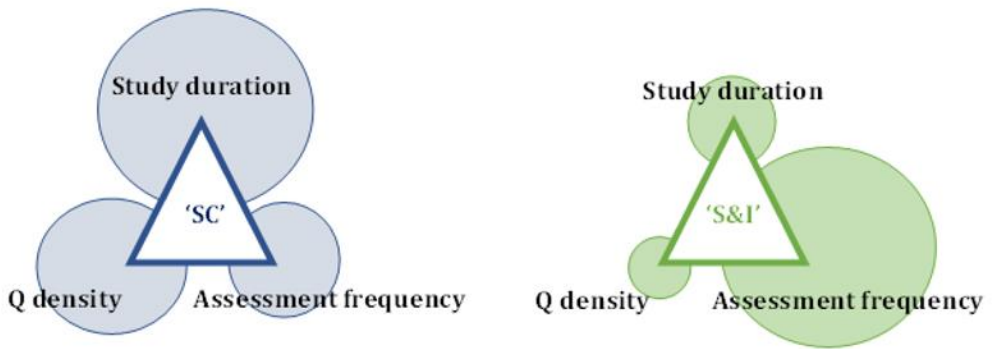


Fig. 3.3 A graphical representation of the inherent interdependencies between ESM parameters. Larger circles indicate higher parameter values. SC = Single-Case ESM study; S&I = Short & Intensive ESM study.

### 3.3.6 Conclusion: Pilot your study

As emphasized throughout this chapter, the presented guidelines do not have an imperative or exhaustive status. Rather, the overarching and general framework was designed to guide you in your decision-making about the specification of your ESM parameters, and to point you to specific considerations you should take into account when pursuing the optimal study settings. This is in part due to the fact that systematic inquiry about the implications of ESM design choices is still running behind (Himmelstein et al., 2019). Answering this type 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, this excludes a possible 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 that shape your ESM study that you did not previously think of. It is therefore crucial to design an initial plan of your ESM protocol and to try it out on a pilot sample. When severe validity issues arise, it may be worthwhile to redesign and adapt your protocol in function of the quantitative and qualitative feedback you received. Quantitatively, problematic compliance or severe drop-out rates may inform you that the protocol was too burdensome. Qualitatively, interviews with participants that took part in your study may highlight particular aspects of your design that undermine the validity of your ESM protocol (e.g., reactivity due to



phrasing of certain questions, sampling schemes that start too early, interference with natural activities, etc.). In sum, designing a good ESM study will be a process of trial-and-error, similar to any other data collection method in psychological science.

## **CHAPTER 4**

# **QUESTIONNAIRE DESIGN AND EVALUATION**

Gudrun Eisele, Zuzana Kasanova, 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. Since this field is still developing, this chapter is not intending to be prescriptive nor definitive, but instead provide a general overview of issues and considerations that need to be addressed at different stages of questionnaire design. First, we discuss how to formulate individual ESM items to assess a specific construct, and list some guidelines for the construction of optimal ESM items that capture dynamic daily-life experiences. Second, we discuss principles and considerations that are essential to constructing a rational ESM questionnaire, consisting of several (sets of) ESM items. Third, we give an overview of tools to assess the quality of individual items and scales. Finally, we introduce an ESM item repository initiative and discuss assessments that can complement self-reports.

## 4.1 Constructing individual ESM items

ESM can be used to assess a broad range of different concepts in daily life; 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 behaviors (“Since the last beep, have you consumed alcohol?”), physical experience (“How hungry do you feel right now?”), and cognitive states (“Since the last beep, have you been ruminating?”), and many more. However, some basic principles and considerations are essential in order to pose a meaningful ESM question that can easily be answered by all participants under any circumstances.

### 4.1.1 *Capture dynamic phenomena*

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 is in contrast to static phenomena that are expected to remain stable for a relatively longer period of time, 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 favorite food?" is expected to be more stable. Food preferences can change, but unlikely on a time scale of hours to days. Therefore, repeatedly assessing this throughout the day is redundant and not meaningful. A meaningful ESM item should capture state-like features that assess something that is specific to the present moment and hence shows sufficient within-person variability. A very general formulation tends to measure more trait-like features, rather than momentary states. To illustrate, "Do you feel the urge to do things impulsively right now?" tends to capture momentary impulsivity, while "Are you an impulsive person?" captures impulsivity as a more stable trait-like feature. Note that simply adding "at the moment" to a question is not sufficient to make it a momentary question. Questions such as "At the moment, are you an impulsive person?" can be confusing as they simply mix trait-like assessments using momentary time references.

Note that not all phenomena are strictly momentary states or stable traits, such as a depressive episode. From a clinical perspective, a depressive episode typically refers to a time period of at least two weeks during which certain specific depressive features were present, implying (limited) stability over time. However, it is also thought to express itself in changes in the momentary levels and fluctuations of depressive features over time. While these momentary levels and more subtle fluctuations 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"). As a general rule, the reference period of phenomena that can be captured through an ESM assessment range from minutes to hours.

Note that sometimes concepts are assumed to be static from a theoretical perspective, but in reality can fluctuate across time within individuals, such as self-esteem (Oosterwegel et al., 2001; P. S. 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 the section 4.3.2 describing the assessment of the quality of one or more items and chapter 3).

#### **4.1.2 Different timeframes**

When we formulate individual ESM items, a choice has to 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?”. This time reference reduces recall bias to a minimum, and is therefore preferred in many instances. Note that this advantage often outweighs the advantages of capturing every single aspect of every moment in daily life, using more retrospective questions assessed in a past time interval (see next paragraph). Keep in mind that also in case of momentary assessments, explicitly adding the time reference will improve 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 in a specific time interval, rather than momentarily. The most commonly used time interval is the time period between the moment the question is being asked and the previous ESM questionnaire, 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?”.

Time intervals can be used as time reference for several reasons. First, time intervals are suited to capture behavior, thoughts, events etc. that occur less frequently. If such concepts would be assessed momentarily (“right now”), the risk exists that many, if not all, occurrences are missed. Asking about the occurrence in 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 would repeatedly ask participants whether they are currently engaged in binge eating, they will miss many or most occurrences of such behavior, unless the assessment schedule and

the occurrence of the behavior line up perfectly. However, this is very unlikely, so assessing whether one or more binge-eating episodes occurred since the previous assessment will result in a better overview of the overall occurrence. Similarly, if a researcher wants to assess the occurrence of events that happen relatively infrequently, such as a dispute, assessing its occurrence in a past time interval will be more informative. Second, time intervals might be more suited to assess phenomena that are ongoing, requiring assessment over a longer period of time. For example, the use of emotion regulation skills typically occurs over the course of minutes to hours. Often, 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 choose a specific time scale on which a phenomenon is expected to change for conceptual reasons. To illustrate, if sleep disturbances or medication use are only expected to occur once per day or once per night, it would be most appropriate to assess them 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, in case it is desirable to model associations between variables, the time reference that was used to assess each variable has implications for the nature of the relationship you are modelling. 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, this association will be of a 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 behavior or thoughts are assessed in a specific time interval, it is unknown when exactly they happened in this time interval, unless it is specifically assessed. For example, a reported event 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 actual 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 and also 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. Above all, 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. In case 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 to know whether people feel cheerful while using alcohol, both current alcohol use and positive affect should be assessed momentarily. However, if a person wants to obtain information regarding the extent to which a person consumes alcohol in daily life, one could consider assessing the use of alcohol since the previous assessment. This would provide a better estimate of overall alcohol use than when current alcohol use is repeatedly assessed. If the research question is whether the use of alcohol is followed by positive affect, one could associate alcohol 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 alcohol consumption on mood.

#### ***4.1.3 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. 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 paranoia can all occur in daily life of some people



and can be captured through ESM, these are likely to have limited variability in the majority of the participants. Most people likely score on the lowest range of the response scale, creating a floor effect. Irritation, enthusiasm and suspiciousness, on the other hand, are experiences that most people can endorse from mild to high levels. Third, explicit assessment of a concept is often, but not always preferred. Sometimes, explicit assessment can actually induce thoughts or behavior. 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 persons 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 model 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 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, 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. Instead, use terms and words that participants would use themselves to describe a concept. In case of 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 make sure to pilot test all items. Sixth, avoid reflective questions, in which participants are asked to reflect on behavior, 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, it is more likely to trigger socially desirable answers or answers that reflect global self-views rather than actual momentary states. Therefore, a more optimal option would be to ask participants to rate their current sadness and next to indicate whether they are currently alone or not. Next, one can easily model the association between sadness and being alone. Last, 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 centers on social contacts, asking participants “Did you initiate the contact?” can be very meaningful in some social contexts, but definitely not in all contexts. Instead, questions should be formulated in such a way that they are relevant in a large variety of contexts.

*Box 4.1. Checklist for optimally formulated ESM questions*

	Make them short and to the point
	Avoid extreme wording
	Consider implicit instead of explicit assessments
	Avoid excessive use of negative wording
	Avoid jargon
	Avoid reflective questions
	Questions should be relevant in all or most contexts

#### 4.1.4 *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 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. For categorical variables, such as current activity type or the binary assessment of the occurrence of behavior, multiple or single choice response options would be most appropriate. Next, bivariate data can be collected using a 2-D grid, in which two theoretically related variables are assessed simultaneously. Examples are dyadic behavioral or emotional data, and the valence and arousal dimension of core affect. Last, text-data can be collected, allowing open-ended responses. This can be used to collect certain thoughts, remarks, or specifications of previously reported answers.

The choice of response scales has been suspected to influence the 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. Findings on differences between scales in non-ESM studies have been mixed. Most studies have found largely equivalent reliability and validity for discrete and continuous rating scales (Gries et al., 2017; Kuhlmann et al., 2017; Lukacz et al., 2004). However, it has been argued that Visual Analog Scales (VAS) reach interval measurement level (i.e., the distance between responses on the scale can be interpreted), while discrete scales only reach ordinal level (Reips & Funke, 2008). In addition, responses on visual analogue scales have been found to show different distributional properties, suggesting that participants are more likely to endorse extreme responses than in Likert scales (Studer, 2012). However, there has also been research indicating that discrete scales may be more user-friendly, because they were found to lead to shorter response times and lower dropout rates than continuous scales (Couper et al., 2006). Even within one category of scale, small variations of 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 (Matejka et al., 2016) and continuous scales where a slider is used lead to different responses than scales where participants clicked to select a response (Funke, 2016). It is important to highlight that, to our

knowledge, no ESM studies have compared different scales with each other. Yet, the effects of different scales may be different on small phone screens. More research is therefore needed before judging the suitability of these different scales for ESM research. In any case, if possible, response scales should be used consistently throughout the questionnaire to avoid confusion for participants.

## **4.2 Constructing a questionnaire**

Once the different 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

### **4.2.1 Order of questions**

For 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). If participants are expected to 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 in 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 a possibility that questions influence answers on subsequent questions. If a participant for instance perceives questions about symptoms as confronting, this 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 up 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 keep questions with the same time frame after each other 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 time frames but to keep items using the same time frame 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 making scoring errors. For instance, while switching between a unipolar Likert scale (e.g., 1 to 7) to a multiple-choice list should take minimal effort, frequent switching between unipolar and bipolar (e.g., -3 to +3) Likert scale might slow participants down or confuse them.

#### **4.2.2 *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 the literature, questionnaire lengths range from 1 to over 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 3 minutes to fill in (Kimhy et al., 2012), and preferably less. When designing the questionnaire, a balance needs to be found between gathering sufficient information while 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 signs of careless responding than a 30-item version of the same questionnaire (Eisele et al., 2020). In line with this, 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 (Ono et al., 2019; Vachon et al., 2019). 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 adjusted content depending on

previous input and adjusted frequency testing, can also be useful. To illustrate the first, only if a participant indicates strong negative feelings, extra follow-up questions can be triggered to 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 4 items measuring positive affect. Instead of assessing all these items every time, we could measure a randomly chosen subset of 3 items per assessment moment. This can allow to reduce the burden for participants, while we would still assess all the items we were interested in 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 avoid that participants start to focus more on their symptoms as a result of the ESM questionnaire and become 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.

### 4.2.3 *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 enough 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 accuracy 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 of 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.

### 4.3 Assessing measurement quality

Once the items and questionnaires are developed, it is important to evaluate them. Broadly speaking, measurement quality refers to the ability of the item(s) to serve as proxies for the underlying construct that the researcher is aiming to measure. Using the example of measuring 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 relevant that the items are carefully designed and selected than in other forms of assessments.

#### 4.3.1 *Pilot testing*

An 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 as being alone when sitting in their office with colleagues (while not interacting), others would classify this same situation as not being alone. Being at home and alone in a room while others are 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. 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.

#### **4.3.2 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.

#### **4.3.3 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 true and how much variance 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 assessing 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 in the construct and the reliability of assessing within-person change. Different ways of calculating the latter type of reliability have been proposed in the literature. They require multiple items measuring the same construct at every assessment. Nezlek (2017) describes how estimates of the between-person and within-person, assessment moment-level reliability can be obtained. In this approach, a 3-level model is fitted with individual items nested in measurement occasions, and occasions 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. However, this approach assumes 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 remaining error variance. 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 level. In another approach, the omegas for each level are estimated with help of structural equation modeling, which allows fitting a multilevel CFA (Bolger & Laurenceau, 2013; Muthén & Asparouhov, 2011). Alternatively, 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). Methods also differ in the extent to which they take the auto-correlated nature of ESM data into account. A detailed description of the individual methods is beyond the scope of this chapter, for more information we refer the reader to the references above and to chapter 9. Up to now, no gold standard for calculating reliability in ESM studies has been established. Further, all methods described above rely on the inclusion of multiple items measuring the same construct. However, many concepts in ESM are assessed with single items and there is currently no consensus on how to determine the reliability for these single items in an ESM study.

#### 4.3.4 *Different forms of validity*

Besides being reliable, a good measure also needs to be valid (see also chapter 3). This means that it measures the construct that we are intending to measure. Because we usually have no way of directly assessing the latent construct (e.g., “momentary anger”), we need to rely on other indices to judge the validity of our measure. One such aspect of validity that can be evaluated is the claim that items from a scale indeed measure a single underlying construct. Multilevel applications of factor analysis can be used to examine this unidimensionality of a construct. If we do for instance include 10 items that are all supposed to assess the underlying construct of “momentary anger”, 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, we need to reconsider our conceptualization of “momentary anger”. Additionally, if we find that our construct is not unidimensional, the calculation of a sum score of the items might not be sensible.

Further, 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). Another way to assess validity is to look at the relationships of our measure with

other constructs. A good measure needs to relate to other constructs in predictable ways. Factor analysis can also be helpful to investigate the relationships between different constructs in the same instrument (when we assume that a construct such as anger is made up of related, yet distinct sub constructs, for instance “experience of anger/hostility” and “arousal of anger”). In ESM, an important consideration is that relationships between different constructs can be different at the within and between-person level. For example, individuals without psychopathology 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 without psychopathology who frequently experience fear, also frequently experience anger, meaning that the two emotions are positively related at the between-person/trait level. Such a discrepancy between within and between-person structure was for instance observed by Borah and colleagues (2018) for a measure of aggression. In this example, a two-factor structure leads to the best fit at the within-person level, while a single factor was sufficient to explain variance at the between-person level. Wilhelm and Schoebi (2007) give another example of different factor structures depending on the level of analysis. They identified a three-factor structure of mood at the within-person level and a two-factor structure at the between-person level. This highlights 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 with 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. Validity can also be assessed by investigating 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 the relationships expected with other constructs can help researchers to decide on the validity and on what information one is most interested in. Further, 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 The ESM Item Repository**

There is currently no consensus on how even very commonly investigated constructs should be measured in ESM studies. This lack of consensus manifests itself in the use of widely varying questions for single constructs (e.g., May et al., 2018; Singh & Bjorling, 2019). The quality and comparability of these different measures are largely unknown, which makes it difficult to compare findings obtained in different studies. To address these problems, Kirtley and colleagues have recently launched the ESM item repository (<https://osf.io/kg376/wiki/home/>). The aim of this project is to collect items used in ESM studies and to evaluate their psychometric properties. This is supposed to facilitate the selection of good measures and the exchange of researchers' experiences with measures. We recommend consulting the item repository before setting up a new ESM study and contributing your items if you have already collected data.

#### **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 behaviors 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 different alternatives to self-report have been appearing in recent years. Examples include the use of actigraphy to assess sleep and movement, sound snippets to assess social interaction, 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. However, the quality of data from these sensors can be low. These types of passive assessments can also reduce the burden placed on participants.

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 comes 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 recent consensus statement on ethical and safety practices for digital monitoring studies of suicidal behavior 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 a number of 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 the 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.

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**

Whilst 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 around 81% of Americans owning a smartphone (Pew Research Center, 2019). Furthermore, people who are frequently underrepresented in research - including Black and Hispanic



individuals, and those on low incomes - are among some of the most “smartphone dependent” individuals, meaning that they use their smartphone as their primary means of contact and internet access (Pew Research Center, 2019). Adolescents are also a group where smartphone ownership is high (van Roekel et al., 2019) and smartphone ownership is increasing among individuals with mental health problems (Torous et al., 2018). Good access to smartphones and the internet is, however, not the case everywhere or for everyone. In low-income countries, for example, men are disproportionately more likely to own smartphones than women (Steele, 2019), creating 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 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.

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 et al., 2021) 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 their 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 participant’s smartphone. Lending participants smartphones does of course increase the cost of conducting the research, so 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 behaviors that are heavily stigmatized, e.g., suicidal behavior, or even illegal, such as substance use, or underage drinking. Anonymity and confidentiality are often key in studies aiming to elicit information regarding 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, as well as adolescents themselves. It is important that researchers are aware of their legal duty of care with regard to confidentiality of information disclosed to them during the study, especially by children and adolescents, and that this 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 when ESM involves collecting audio or photo/video samples. This may occur if data are being collected using an Electronically Activated Recorder (EAR), which captures short snippets of audio during participants' everyday life. Whilst informed consent is necessarily obtained from study participants, individuals around the participant have not 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 on the topic, and makes a number of practical suggestions to ensure bystanders are informed they may be recorded, e.g., by participants wearing a pin badge saying "this conversation may be recorded!"

The previous two considerations have concerned 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 this 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 text 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](https://osf.io/g4jfv)).

### 5.3 Real-time data, real-time responsibility?

Whilst 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 behaviors that may put them at risk, for example suicidal behavior, eating disorders or substance use. In these cases, the likely purpose of the study is to attempt to capture these behaviors. 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 also 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. For a detailed discussion of ethical issues involved in ESM research on Non-Suicidal Self-Injury, see Kiekens and colleagues (2021).

First, the researcher must decide whether or not 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, 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 behavioral 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. Recent 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, as well as the threshold for and nature of the resulting intervention should be clearly explained to participants in the participant information documents, as well as during briefing. Equally, if responses will not be monitored, this should also 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 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. For an excellent discussion of different intervention and safety monitoring protocols for ESM studies of suicidal behavior, see Bai and colleagues (2020).

Third, having decided upon a more intensive intervention, who will be responsible for delivering this? 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. 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 responsible clinician for the study, should a participant become highly distressed. In this case, if a participant's responses give cause for concern, the responsible clinician attached to the study should be the person to contact the participant. 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 issues of research interrupting participants' continuity of care, by researchers intervening in a crisis, which the participants' clinician is better placed to treat, but may be unaware of. 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 behaviors. 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 in balance 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).

How then can researchers minimize participant burden in ESM research? Recent 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, whilst 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 behaviors may in fact induce those thoughts and behaviors (e.g., Myin-Germeys et al., 2009) or may cause participants to alter their behavior (Palmier-Claus et al., 2011). This is also a frequently recurring concern of ethical committees. Low compliance rates may speak to 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, this was unrelated to study duration (Husky et al., 2014). Law and colleagues (2015) also found no significant effect of repeated daily questioning about suicide on individuals’ suicidal thoughts and behaviors. Recently, an elegant study by Coppersmith and colleagues (2021) 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. 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., 2021). This study provides the strongest evidence to date that repeatedly asking about suicidal ideation has no significant iatrogenic effects.

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 behaviors, including increased introspection and awareness of both positive and negative behaviors (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, whilst researchers and ethics committees are often concerned about reactivity to ESM in terms of iatrogenic effects, the existing literature suggests these concerns are unfounded. 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, which 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. 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 this. 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 behaviors, existing research suggests this worry is unfounded. Of course, this does not remove the need for a

thorough safety protocol to ensure participants' well-being during ESM research participation.



# **DURING: CONDUCTING AN ESM STUDY**

## **CHAPTER 6**

# **EXPERIENCE SAMPLING PLATFORMS**

Jeroen Dennis Merlijn Weermeijer, Glenn Kiekens & Martien Wampers



Up and till now, you have learned what types of research questions we can answer with the experience sampling method (ESM) and how to design a study that is methodologically and ethically sound. 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 different hardware (e.g., smartphone, wearable) and software components elements (e.g., dashboard, app) involved. Numerous platforms already exist that allow researchers to operationalize their protocol in the flow of everyday life, which makes 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 important 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. Rather, it aims to provide researchers a starting point by highlighting communalities and meaningful differences between platforms.

## **6.1 The online dashboard**

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

### ***6.1.1 ESM questionnaires***

ESM-questionnaires measure thoughts, experiences, and behavior in real-time (see chapter 4). While the construction 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, while some questions require a particular response scale option (e.g., multiple choice or slider) this also requires its availability 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. Modifiable slider questions concern questions in which the dashboard ideally allows for the

anchors and range to be freely adjusted. This, so that 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). Next, 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.

When considering more complex questionnaires that make use of branching (see chapter 4) or audio-visual stimuli, two important considerations are to be noted. When making use of branching, it is essential to check whether branching can actually be applied. This is important because 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 three, 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 most often needs to 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 to save time, but also to ensure 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). Yet, other settings may benefit from (semi)random sampling schedules that differ per participant. In this case, it is hence advised to use a dashboard that offers individualized (semi)random sampling. What this entails, is a feature in which a researcher needs to, instead of individual time points, indicate the length of the time interval in addition to how many notifications should be presented randomly, and different, 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. For questionnaires that are self-initiated, the implementation 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 device data related to bodily (e.g., increased heart rate) or environmental conditions (e.g., sound or GPS location) is more complicated (see also chapter 12). The challenge here is for the ESM platform to integrate the passively collected sensor data, analyze 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 to 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 which may be tedious to develop. For example, considering heart-rate alone would not be sufficient to provide a questionnaire in those moments that leads up to a panic attack as heart-rate may increase by intensive movement.

### **6.1.3 *Enrollment of participants***

Ideally, the ESM dashboard will generate a single study code, scannable QR code, or a web link that allows participants to enroll 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 enrollment of participants, is whether the dashboard allows for a flexible, instead of fixed, start of the ESM schedule. The fixed start would concern a starting date that will be the same for all participants, regardless of when they enroll (e.g., first beep on Monday the 15th of June). A

flexible start concerns a starting point relative to the enrollment date (e.g., first beep on the first Monday following enrollment, 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 visualization of data make it possible to stimulate compliance by facilitating the provision of visual feedback. Similarly, it allows for 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. Requiring additional development of new onboard analysis tools or visualizations unique to for instance a clinical study is furthermore 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). Yet, 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 recommend researchers to test this process out prior to the start of the study.

### **6.2 ESM apps**

So far, we have 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. Afterward, 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 apps. While the former app is developed to function and work with only one type of system software (e.g., android- vs. iOS-only app), the latter is based on web technologies and works cross-platform (i.e., a single app 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 it is developed for (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 has to do with native apps requiring a unique and more difficult codebase for each system software it is run on, whereas hybrid apps share a single codebase that is generally easier to implement. While performance may favor native apps for those interested in mobile sensing, experts suggest that hybrid apps may eventually be equal to 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 by default due to hardware- and software-based fragmentation (Han et al., 2012). Smartphones that run on iOS and Android phones cover about 99% of the current market share (Karthick & Binu, 2017) and for both of them there are different operating systems versions in circulation. Similarly, phones run on different hardware (i.e., processors, sensors, graphic cards, etc.). This issue makes 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 furthermore holds 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 have an effect on 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., allowing an app to send push notifications). Additionally, updates of 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 (L. Wang et al., 2017). 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 does not apply, 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. 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. This, as the computation of stress may be hidden and under the protection of intellectual property rights by the developer. Similarly, raw data may not necessarily be available from commercially developed devices.

## **6.4 Legal considerations**

Related to ESM software and hardware, there are laws that need to be taken into consideration. This concerns laws on data privacy and use of electronic devices, as well as the use of ESM in the context of clinical settings.

### ***6.4.1 Data privacy and electronic devices***

Chapter 5 addressed the highly personal and often sensitive nature of ESM data, which brings responsibility regarding data privacy and protection. ESM platforms based in the EU will by default be required to be in line with the European Union’s General Data Protection Regulations (GDPR, <https://gdpr-info.eu/>) that 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 that they need to adhere to (Greenleaf, 2017). This implies that the choice of a 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 is in line with GDPR regulation.

As we use electronic devices in ESM research, be it wearables or smartphones, they are subject to laws related to parameters on health, safety, and environmental protection standards. For example, electronic devices sold within the EU need to have a CE marking, which indicates conformity with EU legislation surrounding the parameters mentioned above. This CE marking is not to be confused with the China Export marking, which is highly similar (figure 6.1).

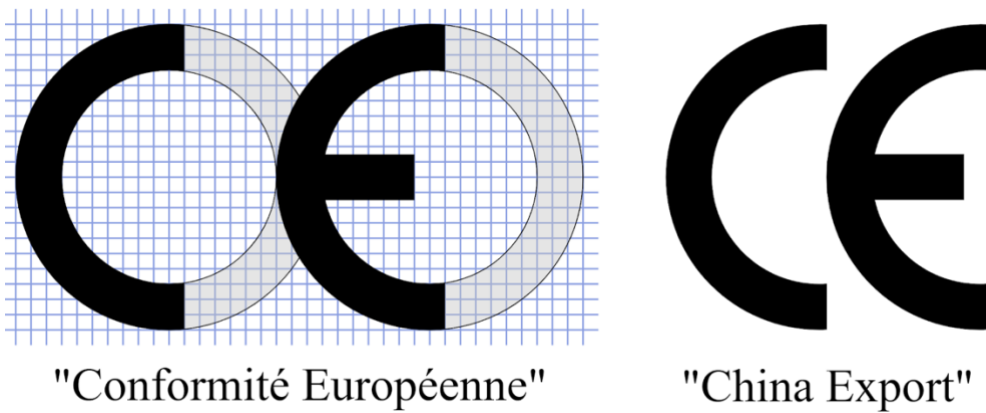


Figure 6.1. CE vs. China Export markings. Image retrieved from [https://upload.wikimedia.org/wikipedia/commons/2/2b/Comparison\\_of\\_two\\_used\\_CE\\_marks.svg](https://upload.wikimedia.org/wikipedia/commons/2/2b/Comparison_of_two_used_CE_marks.svg)

#### 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:

EU [Regulation (EU) 2017/745 on Medical Devices (MDR), article 2(1)] (OJ L 117): “medical 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)] “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 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, and as with many laws, exemptions are possible. For example, the federal drug administration (legal body in the US) states that whilst 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 ‘intends’ and ‘low risk’ are vague. However, they do provide examples under which we also find diagnosis and treatment of psychiatric conditions (US Food and Drug Administration, 2019, p22). It is currently unclear whether or not the same exemption intention applies to the clinical use of ESM software within the EU. Yet, app stores are full of mental health apps (other than ESM) that fit the medical device definition. None of these, to the best of our knowledge, are classified as medical device software. Hence, it seems to suggest, similar to within 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 per se. The updating of application software in order to remain compatible with updated system software is a vital element of sustainability. When application software is managed by a single individual, this 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 in order 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 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 - January 2021). When interested in an alternative ESM platform, you can use this table to see how that platform would compare against the five platforms considered here.

*Table 6.1. Overview of presented ESM platforms <sup>a</sup>*

	m-Path	Movisens	RADAR	SEMA3	Expiwell
<b>Online dashboard</b>					
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.	I.D.	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	I.D.	Yes	I.D.	No	No
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
<b>ESM app</b>					
Native/Hybrid	Native	Native	Native	Hybrid	Hybrid
Operating system compatibility	Android/iOS	Android	Android/iOS	Android/iOS	Android/iOS
Adjustable notification sound & durations	Yes	Yes	No	Yes	No
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	I.D.	Yes	Yes	No	No

Table 6.1, continued

	m-Path	Movisens	RADAR	SEMA3	Expiwell
<b>Wearable</b>					
Integrated data	No	Yes	Yes	No	No
<b>Legal</b>					
GDPR compliant	Yes	Yes	Yes	P.	Yes
CE marking	No	P.	No	No	No
MDR compliant	No	No	No	No	No
510(k)	No	No	No	No	No
Other	No	No	No	Legal framework Australia	No
<b>Profile</b>					
Founding date	2019	2009	2016	2013	2015
Country	Belgium	Germany	UK	AUS	US
Number of paid employees	4	16	U.	5	U.
Number of active users	175+	850+	U.	250+	2100+
<b>Cost<sup>b</sup></b>					
Free	Yes	Yes	Yes	Yes	Yes
Premium	Yes	Yes	No	No	yes

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

<sup>a</sup> For additional ESM platforms, please see:

[https://docs.google.com/spreadsheets/d/18R9x9Qbl9tADJGpJBjID\\_T9EWZeQ\\_4W3OFdn3iKRU7U/edit#gid=204277638](https://docs.google.com/spreadsheets/d/18R9x9Qbl9tADJGpJBjID_T9EWZeQ_4W3OFdn3iKRU7U/edit#gid=204277638)

<sup>b</sup> 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, using 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. Whilst these additional features make each platform different, it is here 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 then that whilst the platforms share similar features for basic application of ESM, the user interface of different platform components (e.g., dashboard, app, wearable) differs 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 budget and envisioned research design, it is hence recommended to pilot the platforms 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 which was used to provide 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 aspired needs perfectly.





## **CHAPTER 7**

# **BRIEFING AND DEBRIEFING IN AN EXPERIENCE SAMPLING STUDY**

Aki Rintala, Silke Apers, Gudrun Eisele, and Davinia Verhoeven



This chapter aims to inform you about the importance of implementing a briefing and debriefing session with your participant within your ESM study. If you are designing a study using ESM, a briefing session will be one of the most important parts of your study. This is mainly because your 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. Absence of the researcher or the lack of a proper briefing session might increase the risks of mistakes or violation of the study protocol. Therefore, to minimize these risks, participants must be properly informed about the procedures of the study. An effective briefing is also crucial to ensure compliance and data quality of your study (Palmier-Claus et al., 2011; Rintala et al., 2019). The aim of a successful briefing session is to motivate your participant to follow your 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 that you are planning to brief to 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. Overall, it is important to practice your briefing session before meeting your participants, in order 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 would always recommend briefing your participant individually, as that would give 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 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 for example using real-life meetings, video 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 is asking 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 give important new insights on 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’*.

You can mention to your participant that the study might increase insights or knowledge about certain feelings or activities that will help to understand participants’ condition better in the future.

Inform your participant about the number of study days and the number of assessments that will be delivered during one day. Go through different scenarios with your participant. It is important that your participant knows what to expect and understands 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 up in real life. We are interested to know how you feel, what you think, what activities you are involved in, 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 6 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’.* For example, if you are conducting a study with a population with depressive symptoms, 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 time frame in which they can expect this. 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. 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 do not have the possibility to 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, that is normal when studying 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'.*

If you are using a time window for answering the questionnaire in your study protocol, do not mention this exact time frame to your participant (unless your study has a very short response window, for example less than 2 minutes). For example, *'One ESM questionnaire will only be available for a certain time, and then it will disappear'.*

You can explain to your participant that the assessments will continue regardless of whether the previous assessment was filled in or not. It is okay to miss an assessment, the purpose is not to change daily routines. For example, *'If you are not able to fill in the ESM questionnaire immediately, you can answer it for a short period of time 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'.*

Repeat several times the most important issues that your participant must know. Remember, your participant will be on his/her own for the entire ESM study period, so they need to fully understand and remember everything they need to do. For example, *'It is important to answer all the questions just before the ESM notification went off. When you hear a notification, take your phone and answer the questions immediately'.*

Explain to your participant that it is important to think of the moment just before the notification went off, because otherwise your participant might fill in the ESM questionnaire based on the feelings experienced while filling in the questions. This is not what we are looking for in the responses. You can also explain this directly to your participant. Also, it is important that your participant carries the phone with them at all times and that they keep the volume turned on. You can also mention to your participant to check the phone from time to time if your participant is in a situation where the phone needs to be switched on mute (e.g., in the theater or a concert). One option is also to keep the phone on vibrate mode to increase the chances to hear the notification. Explain to your participant

not to linger over a question. Encourage your participant to fill in the first thing that comes in mind. However, the questions have to be entered in a calm way. If your ESM application offers the possibility to change a given answer, you can mention that to your participant. Emphasize that it is very important to keep their normal daily or nightly routines. For example, *‘Keep your normal daily/nightly routines. If you want to go to sleep, take a nap, or if you do not want to be disturbed, then you can put the smartphone somewhere where you cannot hear it (such as in the living room or kitchen), or switch the sound settings off’*.

Talk to your participant on how to manage situations in which somebody asks what he/she is doing when filling in an ESM questionnaire. This helps to decrease social pressure or disturbance in these situations. It is good that your participant would have an answer prepared for situations in daily life explaining 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. You need to check if all questions are clear, if the sound is sufficient, or if any technical issues occur’*.

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 there may be for a situation to count as an *‘event’*, for example if you only want participants to report social interaction of a particular length (e.g., it must have lasted at least 5 min) that 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 in order to avoid the participant to scroll too fast through the questions before your instructions are finished. When you hold the phone, you can use all the time you need to explain every question without going too fast 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, feeling, 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 (something that you would normally explain later on), explain that first before continuing with the rest. 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 him/her: *‘What do you think it would stand for?’*.

Go through your ESM items one by one. If you decide to include an item where a participant can answer more than one option (e.g., 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. It is important that your participant understands all the items correctly, and also understands how to answer questions. 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 understands the anchor points. Repeat what your participant has entered. For example in a 7-point Likert scale: *”You answered a 1 here, so this means that the statement ‘I feel cheerful’ is not true, is that correct?”*

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

### **7.1.4 Additional information**

#### Technical instructions

If participants receive a phone or another device for their participation in the study, they need to understand how to use the device properly. For example, it is important that participants know how often and how to charge the device, how to switch the device on and off, how to lock and unlock the screen, and how to contact the researcher if 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.

#### Talk about the upcoming period during which the study runs

Does your participant expect the next few days to represent their normal routine, or are there any special events planned? Will there be moments in which ESM may not be possible to fill in? Discuss these moments (perhaps it is possible to fill in ESM) and negotiate if necessary. Make sure that the phone is not left at home, and that your participant takes it with him/her. You can also explain that the phone can be switched to a vibrating mode, if necessary.

#### Do's and don'ts!

- Make sure you explain your ESM questionnaire in a more general way. For example, we want to learn more about your feelings, your everyday activities, your social interactions and so on. But avoid referring to a participant's disorder or disease. When you are conducting a study within a population that has psychotic experience, you don't want to emphasize that you want to monitor their hallucinations and delusional thoughts or behavior.
- It is not allowed 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 2 minutes)
- Explain the total number of assessments that can be expected to be received randomly throughout the day, but don't mention the time-intervals (e.g., 90-minutes time blocks). If your participant knows the time-intervals, they might start changing their daily routines.

- Ask your participant for real life examples during the ESM briefing. Let him/her respond to the questions realistically, applied to that exact moment. Do not let your participant use a fictitious example.
- Repeat several times that the participant needs to think about the moment right before the notification when they fill out a questionnaire.

*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, you are recommended to ask how many assessments they have filled out that day and the day before, just to get an idea on 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 less assessments than expected, you can encourage them to fill out more. Inquire about reasons why they were not able to fill out more assessments and try to find ways of how they can 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 given one). 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, 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 (of course without creating an unsafe situation).
- Mention that the ESM has a time period of answering without mentioning the exact time.

### 7.1.5 FAQs

#### *How do I brief a depressive patient?*

Avoid using disease-specific terms such as ‘mood’ or ‘negative emotions’ etc. Emphasize that these diary entry questions are general questions that are used for the general population as well. Try to use general wording without focusing specifically on the feeling item. Do remember to say that we are interested in all experiences of the day such as activity, social interactions, and sleep.

#### *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 try to explain the reasoning behind the questions that they perceived as stressful.

#### *When should I contact the participant?*

We recommend that you contact your participant on the 2nd day of the ESM period. If possible, you could monitor your data collection before the contact. If there is a poor compliance (< 30 %) for the first study days, you have the possibility to discuss that in the contact call. It also gives you the opportunity to check for technical issues.

If compliance of participants is monitored in real time, researchers also need to decide whether or not 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 that might further interfere with their daily routines.

#### *Can I tell my participant how many assessments have been filled in if it is asked?*

We do not recommend to give 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.

*What if my participant says that he/she cannot fill in the ESM during the evenings or mornings or during work?*

It is common that your participant may question the relevance of filling in the assessments in situations where your participant may be a bit more reluctant to use it. In these situations, try to approach your participant in a way that increases his/her awareness towards the method. For example, your participant could say that there is a 3-hour time window every evening when ESM cannot be filled in due to hobbies. In this example, you as a researcher could discuss that the notification might come during the time when your participant is travelling to the location, or just before starting the hobby. Although it may be more likely in this situation that the notification is missed more frequently, try to motivate your participant to still check the phone from time to time. Explain to your participant that it is also important to receive information about moments when people are for instance outside, with others, or relaxing. The goal is to get an overview of all kinds of situations and contexts. Therefore, it is important that they also fill it out in these circumstances.

### ***7.1.6 ESM questionnaire for researcher***

In order 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 thinking about something constantly that worries you, 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

The main objective of the debriefing session is to gather information on how the study went from your participant's point of view. The debriefing session is recommended to start with 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 to be 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 the use, or 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 use your chosen feedback questionnaires or interview that is related to your study success. 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
  - a) What kind of events (open question)
- 3) Were you able to express your experiences via the app? 1-7
- 4) Did the ESM period influence your mood? 1-7
- 5) Did the ESM period influence your daily routine? 1-7
- 6) 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**

<b>GENERAL</b>	Check
Explain the purpose of the electronic diary and why it is useful to fill in	
Take your time to explain every item one-by-one	
Repeat the most important things (f.e., keeping the app with you at all times)	
Keep the instructions positive, explain what you expect from the participant	
<b>BRIEFING PARTICIPANT</b>	
Participant is asked to fill in the questionnaire immediately after receiving the notification on how he/she is feeling, doing, etc. right in the moment (i.e., just before the notification)	
Keep normal daily or nightly routines	
Explain 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 (f.e., questions from other people)	



---

## CONCLUSION

---

What are your participants' thoughts on using the electronic diary?

---

Discuss upcoming period: normal routines vs special events

---

**AFTER:  
THE ANALYSIS OF ESM  
DATA**



## **CHAPTER 8**

# **STRUCTURING, CHECKING, AND PREPARING ESM 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 6 days using random time sampling. For variables that are measured at each assessment moment, we will therefore obtain 6000 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 to choose 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	Time	...	Time	Time	Time	...	Time	
			1	2	...	60	1	2	...	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 1800 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 6000 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. 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	
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 in 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 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 post-treatment phase).

## 8.2 Example

For illustration purposes, we will make use of a subset of the data from an ESM study including 328 participants that were asked to fill in a questionnaire assessing their mood and several contextual variables 10 times per day over the course of 6 days. The participants fell into three groups based on their mental health status: 112 participants were healthy controls, 109 participants had a life-time 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 between-prompt interval was around 90 minutes with a standard deviation of approximately 30 minutes. The prompts ('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, [www.real-leuven.be](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 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)<sup>1</sup>, 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 1104 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 / satisfied), 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 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 whether participants would rather do something else than whatever activity they were engaged in at the time of the beep, whether they had consumed coffee / 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, then 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.

---

<sup>1</sup> To be precise, since this question was asked at the end of the questionnaire, this variable indicates the time when a subject had finished filling in the questionnaire.

### 8.3 Software and Code

In addition to the actual dataset, the interested reader will find on the book website ([www.real-leuven.be](http://www.real-leuven.be)) R code corresponding to all of the following data checks, data preparation steps, and the actual analyses to be described further below. 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 analyzing ESM and other types of data. Resources focused on other statistical software packages will be provided at the end. However, it should be noted that the reader will 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 analyzing 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 (as is necessary when the responses to the questionnaires are collected via paper booklets), various data checks should still routinely be 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 not 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, then the values should either be fully missing for all rows corresponding to a particular subject or none of them. Moreover, if the values are not missing, then they should really be time-invariant (i.e., constant) within the subject.<sup>2</sup> Checking variables age, sex, and mental health status reveals no such inconsistencies in the illustrative dataset. Moreover, 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 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 ascertain 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 analyzed 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 end up 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).<sup>3</sup>

---

<sup>2</sup> 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 when this information was collected (e.g., when the baseline questionnaire was administered).

<sup>3</sup> To be precise, for all first beeps, the beep times should be  $\geq 450$  and  $< 540$ , for all second beeps, the beep times should be  $\geq 540$  and  $< 630$ , and so on.

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) belong 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 in (i.e., a pleasantness rating was provided). One could then consider recoding the 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 barplot 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 a priori or it can be enforced once the data have been collected.

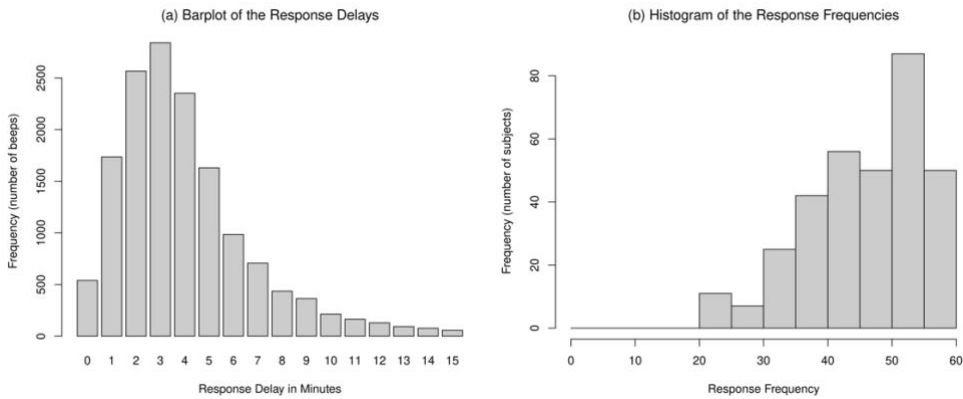


Figure 8.1. (a) Barplot of the response delays and (b) Histogram of the response frequencies

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 for 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., mark 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. Leaving this issue aside, based on the response frequencies, we can compute the compliance rates (i.e., response

frequencies /  $60 \times 100\%$ ) 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%).<sup>4</sup>

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’ the dataset to the subject level when reporting summary statistics about time-invariant variables (otherwise, the repeated values 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 (i.e., 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 3 items used to measure ‘positive affect’ and similarly the mean of the 6 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 or do so but 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 2

---

<sup>4</sup> 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 the 3 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 on, 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 centered’ 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 on, we may not want to lag values across the night (e.g., 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 6 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 gray and white shading within each subfigure correspond to the 6 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 decreasing pattern (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 6 study days.

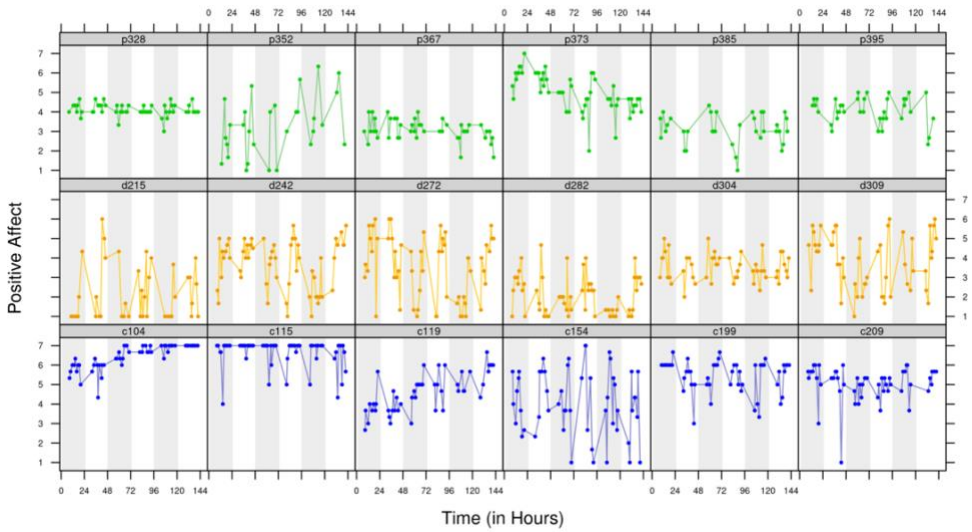


Figure 8.2. Positive affect over time for 6 randomly selected subjects from each of the three mental health status groups (*c* = control, *d* = depressed, *p* = psychotic)

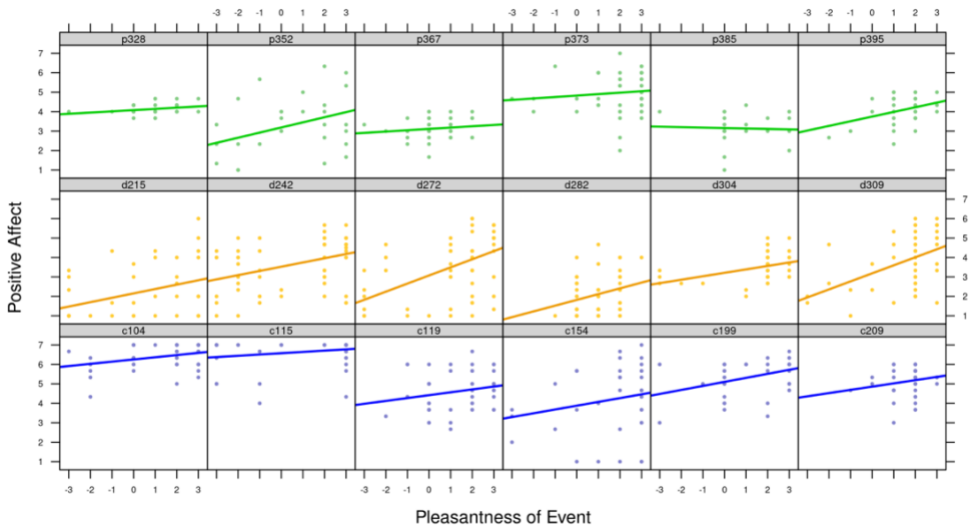


Figure 8.3. Positive affect as a function of the pleasantness of the most important event since the previous beep for 6 randomly selected subjects from each of the three mental health status groups (*c* = control, *d* = depressed, *p* = psychotic) with per-subject regression lines superimposed

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 helpful. 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 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.

## **CHAPTER 9**

# **STATISTICAL METHODS FOR ESM DATA**

Wolfgang Viechtbauer



Before elaborating 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 (not surprisingly!) can 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 of 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 troubles of repeatedly collecting information about an individual's level of affect across varying levels of

stress, we are able to 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 analyze 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- (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 in this direction, 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  $i$ th subject at the  $j$ th assessment moment (hence, in the illustrative dataset,  $i = 1, \dots, 328$  and  $j = 1, \dots, 60$ ). Then a two-level model to analyze the data is given by

$$y_{ij} = \mu + u_i + \epsilon_{ij}, \quad (1)$$

where  $\mu$  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  $\epsilon_{ij}$  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  $i$ th subject and  $\mu$  as the average of these subject-specific averages (so in essence,  $\mu$  is the average of averages!).

This idea is illustrated in Figure 9.1, showing 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 the 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  $\hat{\mu}$ . The faint lines extending from the subject-specific averages to the individual observations represent the observed errors (also often referred to as the ‘residuals’).



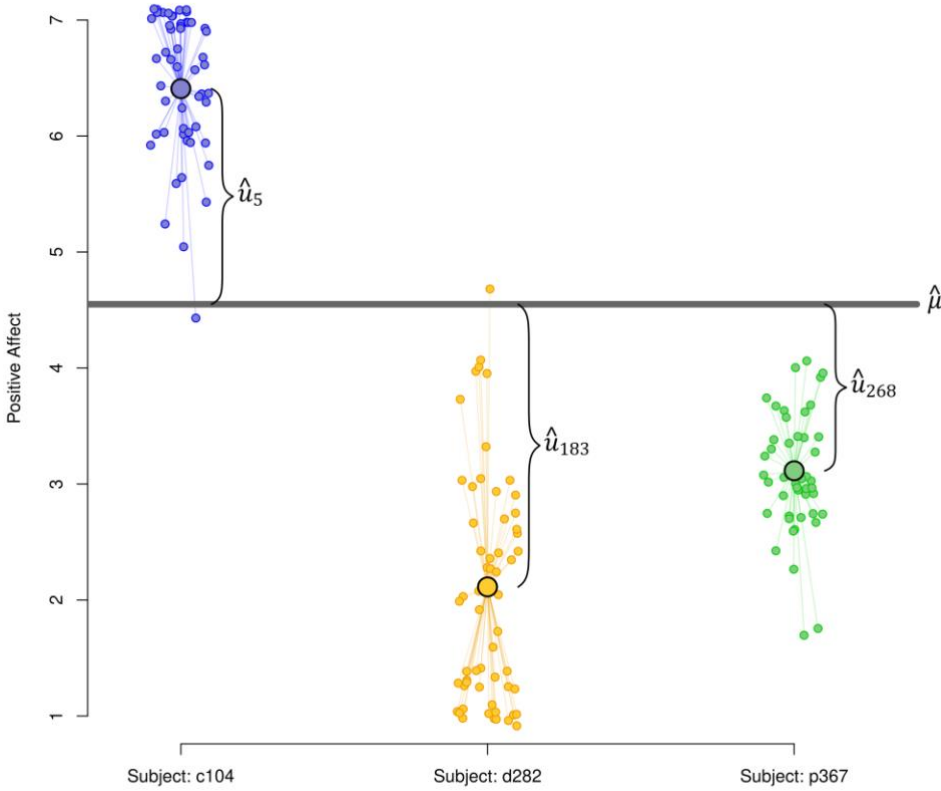


Figure 9.1. The values of positive affect for three subjects from the illustrative dataset, their estimated subject-specific averages, and the estimated overall average

In the context of the model above, we are (usually) not so interested in the subject-specific averages themselves, but 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  $\mu$  and the subject-specific averages are normally distributed (i.e.,  $u_i \sim N(0, \tau^2)$ ) and hence  $\tau^2$  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.,  $\epsilon_{ij} \sim N(0, \sigma^2)$ ) and hence  $\sigma^2$  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  $\mu$  is given by  $\hat{\mu} \pm 1.96SE[\hat{\mu}]$ , 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.<sup>1</sup>

The confidence interval above reflects the uncertainty in our estimate of  $\mu$ , but it does not tell us anything about the distribution of the subject-specific averages themselves (except for where, under the assumptions of the model, we estimate the center of this distribution to be). 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 particular 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.

---

<sup>1</sup> Before somebody sends angry emails about this interpretation of the confidence interval: Yes, this particular interval either covers the unknown but fixed value of  $\mu$  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  $\mu$ .

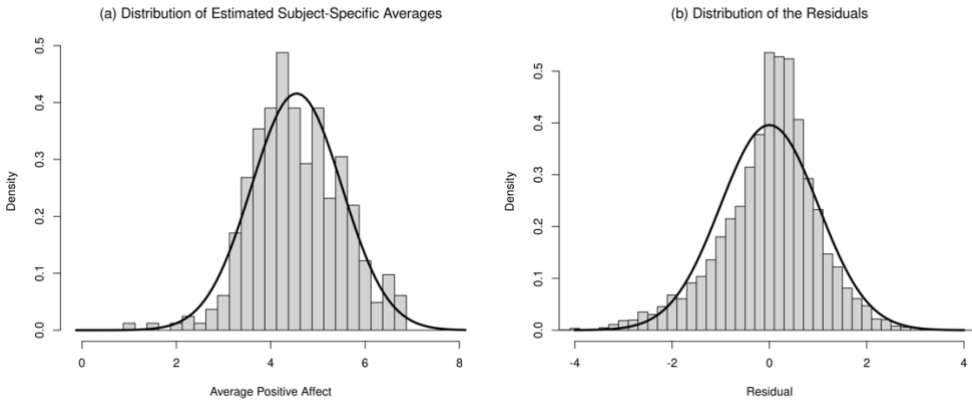


Figure 9.2. Histogram of (a) the estimated subject-specific averages and (b) the residuals, with normal distributions superimposed on the histograms

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  $\bar{y}_i$ . 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 ‘shrunk’) 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  $\bar{y}_i$  (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  $\bar{y}_i$  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  $\bar{y}_i$  (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.<sup>2</sup> Figure 5(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  $\pm 2$ . 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.<sup>3</sup>

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  $\text{ICC} = 0.920 / (0.920 + 1.1015) \approx 0.48$ . Hence, approximately half of the variability in positive affect is due to between-person differences and the other half due to momentary fluctuations.

---

<sup>2</sup> 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.

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

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 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  $H_0: \text{ICC} = 0$  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.<sup>4</sup>

### 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 type of participants tend to report lower versus higher positive affect on average.<sup>5</sup> 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 between those in the three different mental health status groups, then a corresponding model would be

$$y_{ij} = \beta_0 + \beta_1 \text{male}_i + \beta_2 (\text{age}_i - 35) + \beta_3 \text{dep}_i + \beta_4 \text{psy}_i + u_i + \epsilon_{ij}, \quad (2)$$

where  $\text{male}_i$  is a ‘dummy variable’ coded 0 for female and 1 for male participants,  $\text{age}_i$  is the age of the  $i$ th participant,  $\text{dep}_i$  is coded 1 for those in the depression group and 0 otherwise, and  $\text{psy}_i$  is coded 1 for those in the psychosis group and 0 otherwise (hence,  $\text{dep}_i = \text{psy}_i = 0$  for those in the control group). Note that the constant 35 is subtracted from the participants’ age values, which makes the

---

<sup>4</sup> Not surprisingly, a test of  $H_0: \text{ICC} = 0$  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 6 randomly selected participants, we would still have more than 99% power to reject the null hypothesis.

<sup>5</sup> 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.

model intercept (i.e.,  $\beta_0$ ) more interpretable.<sup>6</sup> Therefore,  $\beta_0$  denotes the average level of positive affect for female participants that are 35 years of age and that are in the healthy control group,  $\beta_1$  reflects the difference in average positive affect between male and female participants,  $\beta_2$  denotes how the average level of positive affect differs for participants that are one year apart in their age,  $\beta_3$  denotes the difference between those in the depression group versus those in the control group, and  $\beta_4$  denotes the difference between those in the psychosis group versus those in the control group. Assumptions about  $u_i$  and  $\epsilon_{ij}$  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.,  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_4$ ), the corresponding standard errors (SE), the z-values (i.e.,  $z = \text{estimate} / \text{SE}$ ), 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  $\pm 1.96$ , then  $p < .05$  and hence we reject the null hypothesis that the true coefficient is equal to 0). 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 their age, the coefficient is not significantly different from 0 ( $p = .32$ ) and even participants 20 years apart would be estimated to differ only by 0.1 points. However, participants from the depression group have, on average, a 1.2 points lower level of positive affect compared to those from the control group ( $p < .001$ ), a sizable difference. Similarly, the results indicate a 0.8 points lower average positive affect level for those in the psychosis group compared to the healthy controls ( $p < .001$ ).

---

<sup>6</sup> A common practice is to subtract the mean age of the participants (i.e., 36.5), but this is not a requirement. In fact,  $\beta_0$  would then not be interpretable unless the mean age is also known/reported. Instead, we suggest to manually choose a meaningful value for ‘centering’ the age variable, which does not change the fit of the model but makes the interpretation of the intercept explicitly clear.

*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)*

<b>coefficient</b>	<b>estimate</b>	<b>SE</b>	<b>z-value</b>	<b>p-value</b>
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  $\epsilon_{ij}$  values, namely  $\hat{\tau}^2 = 0.704$  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 denoted 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.<sup>7</sup> 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 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

<sup>7</sup> 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 can just set the value to 0.

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

where *eventpl*<sub>*ij*</sub> denotes the reported event pleasantness value of the *i*th subject at the *j*th assessment and, as before, *y*<sub>*ij*</sub> 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 this 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*<sub>0*i*</sub> and *u*<sub>1*i*</sub>, which allow the intercept and slope of the *i*th subject to differ from  $\beta_0$  and  $\beta_1$ , 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 gray line corresponds to the line defined by the estimated values of  $\beta_0$  and  $\beta_1$ . The estimated deviation between the intercept of the 5th subject from the average intercept is also indicated (i.e.,  $\hat{u}_{0,5}$ ), as well as 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 in their variability. For the random effects, we assume  $u_{0i} \sim N(0, \tau_0^2)$  and  $u_{1i} \sim N(0, \tau_1^2)$  and hence  $\tau_0^2$  denotes the between-subject variance in the intercepts and  $\tau_1^2$  the between-subject variance in the slopes. Moreover, the random effects might be correlated with each other and we use  $\rho$  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.

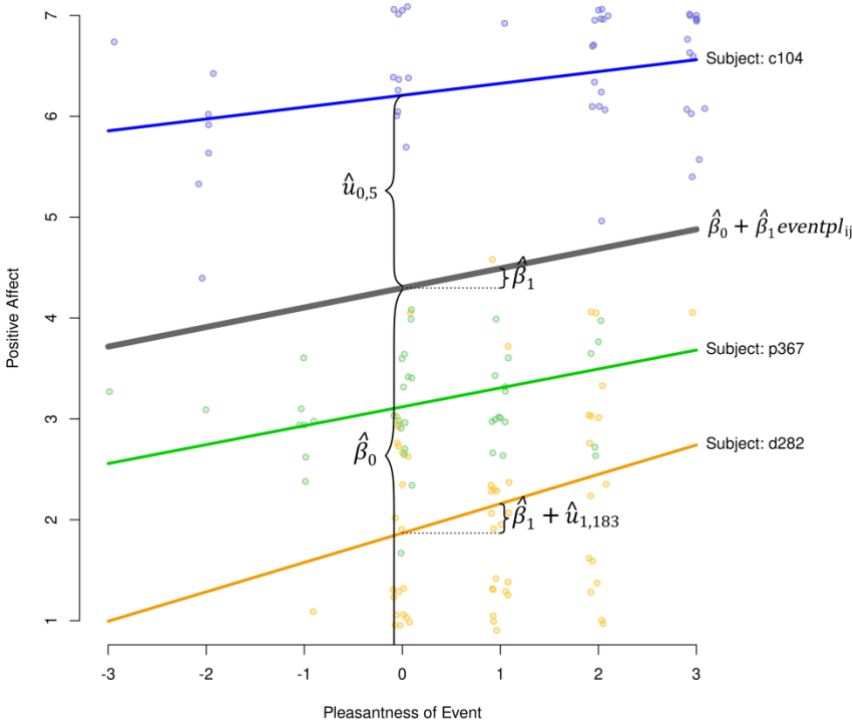


Figure 9.3. The values of positive affect as a function of event pleasantness for three subjects from the illustrative dataset, the estimated subject-specific regression lines, and the overall regression line

Fitting this model yields  $\hat{\beta}_0 = 4.298$  ( $SE[\hat{\beta}_0] = 0.054$ ),  $\hat{\beta}_1 = 0.194$  ( $SE[\hat{\beta}_1] = 0.008$ ),  $\hat{\tau}_0^2 = 0.908$ ,  $\hat{\tau}_1^2 = 0.0111$ ,  $\hat{\rho} = -0.44$ , and  $\hat{\sigma}^2 = 0.893$ . 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 gray line in Figure 9.3), a difference of  $6 \times 0.2 = 1.2$  points.

Note that  $\hat{\beta}_1$  is the estimate of the average slope. However, for particular participants, the relationship between event pleasantness and positive affect could be stronger or weaker. As we did earlier, we can compute an interval, namely  $\hat{\beta}_1 \pm 1.96\hat{\tau}_1$ , 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.,  $6 \times 0.4$ ).

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  $\hat{\beta}_0$  and  $\hat{\beta}_1$  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 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) show the distribution of the residuals, which are again a bit too peaked for a normal distribution, but otherwise fairly symmetrically distributed around 0.

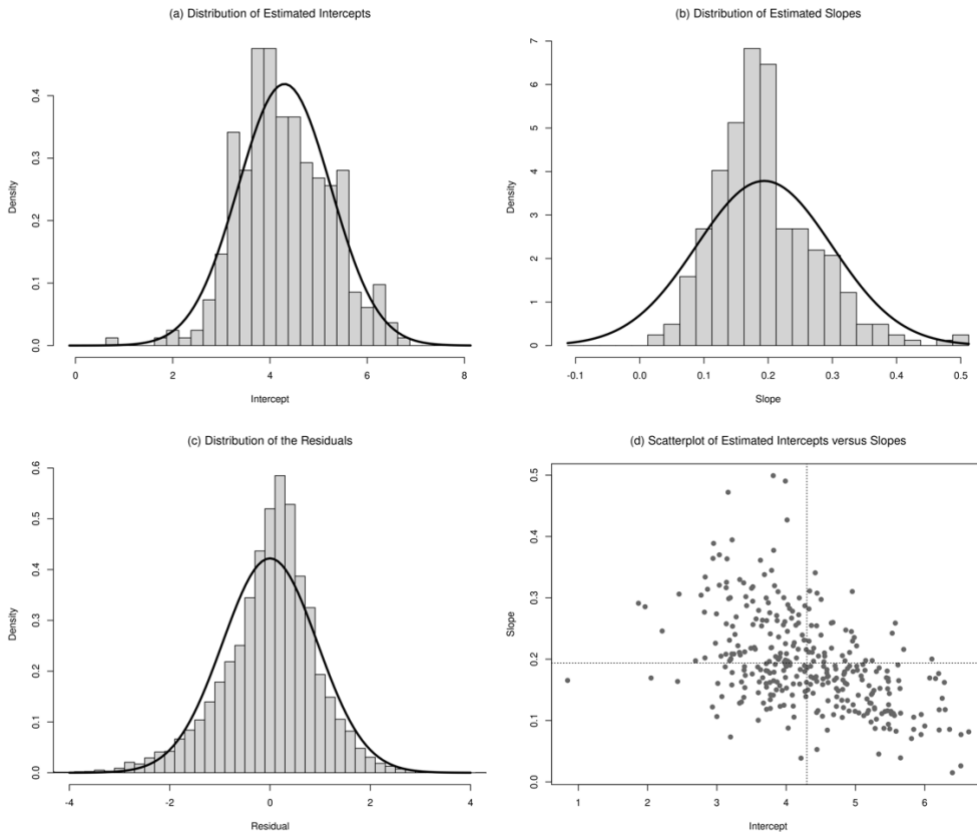


Figure 9.4. Histogram of (a) the estimated subject-specific intercepts, (b) the estimated subject-specific slopes, (c) the residuals, and (d) a scatterplot of the estimated intercepts versus the slopes

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 of  $\hat{\rho} = -0.44$  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 robustness of individuals with high positive affect to the occurrence of unpleasantness/stressful events, but might also be an artifact due to a ceiling effect in the scale used to measure positive affect.

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  $\sigma^2$  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 is 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

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

where  $dep_i$  and  $psy_i$  are dummy variables coded as described earlier. Hence,  $\beta_0$  denotes the average positive affect for neutral events for those in the control group,  $\beta_1$  and  $\beta_2$  indicate how the average level differs for those in the depression and psychosis groups compared to the control group,  $\beta_3$  is the average slope of the association for those in the control group, and  $\beta_4$  and  $\beta_5$  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.,

Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). 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  $\hat{\tau}_0^2 = 0.646$ ,  $\hat{\tau}_1^2 = 0.0085$ ,  $\hat{\rho} = -0.26$ , 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 $\times$ event pleasantness	0.124	0.018	6.920	<.001
psychotic $\times$ event pleasantness	0.025	0.019	1.295	.195

Hence, the results indicate significant differences in the average intercepts of the control and the depression and psychosis groups, with those in the latter two group showing significantly lower average levels of positive affect for neutral events (both  $p < .001$ ). This is most pronounced in the depression group, with an almost 1.3 points lower average intercept 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 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, so 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 is able to 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  $\hat{\tau}_1^2 = 0.0111$  and  $\hat{\tau}_1^2 = 0.0085$ , we find

$$R^2 = \frac{0.0111 - 0.0085}{0.0111} = 0.23$$

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.

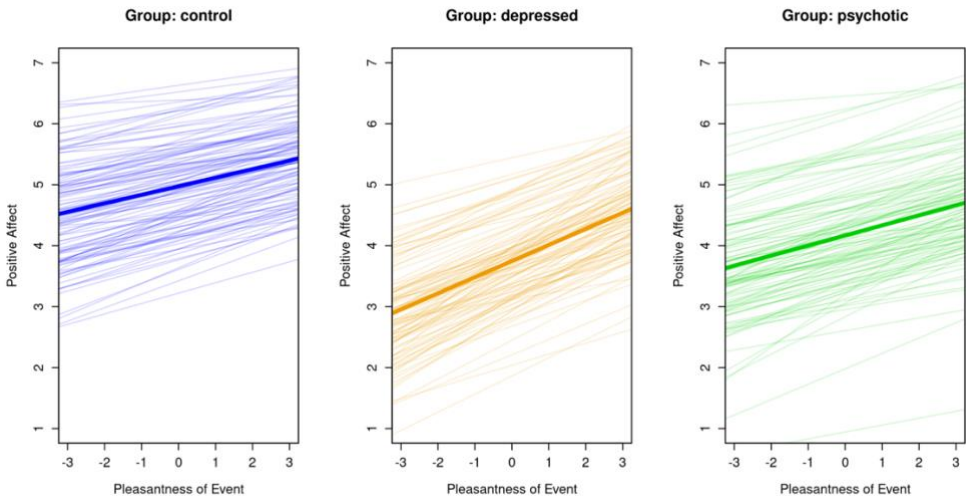


Figure 9.5. Estimated average and subject-specific regression lines for the association between event pleasantness and positive affect within the three mental health status groups

## 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 suggests 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; L. P. Wang & Maxwell, 2015). To do so, we need to fit the model

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

where  $\overline{eventpl}_i$  denotes the average of the event pleasantness ratings of the  $i$ th 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 this model, the intercept  $\beta_0$  represents the average level of positive affect for participants whose average event pleasantness rating is equal to 0 and when the momentary event pleasantness rating is also equal to 0,  $\beta_1$  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  $\beta_2$  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  $\hat{\beta}_0 = 3.701$  ( $SE[\hat{\beta}_0] = 0.101$ ),  $\hat{\beta}_1 = 0.626$  ( $SE[\hat{\beta}_1] = 0.065$ ),  $\hat{\beta}_2 = 0.192$  ( $SE[\hat{\beta}_2] = 0.008$ ),  $\hat{\tau}_0^2 = 0.710$ ,  $\hat{\tau}_1^2 = 0.0111$ ,  $\hat{\rho} = -0.33$ , and  $\hat{\sigma}^2 = 0.893$ .

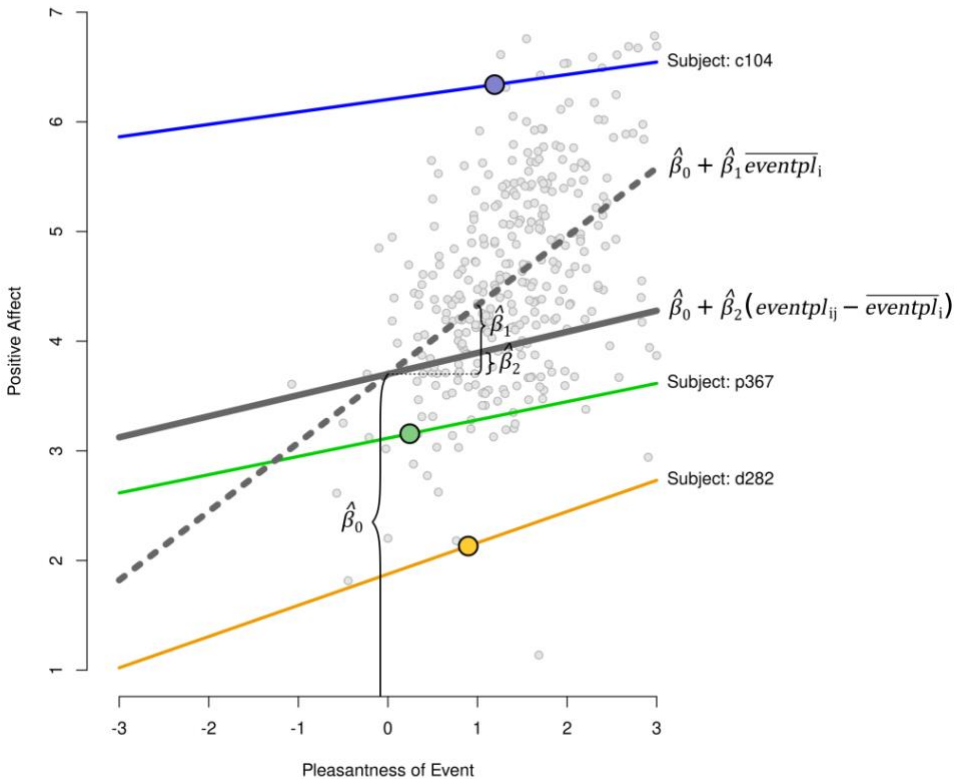


Figure 9.6. Scatterplot of the per-person average event pleasantness values versus the estimated average positive affect values, the between- and average within-person associations, and the subject-specific regression lines for the within-person association for three subjects

In this model  $\beta_1$  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 gray points). The gray dashed line corresponds to this between-person relationship, which in essence models the association between these two sets of averages.

On the other hand,  $\beta_2$  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 gray line. To be precise, we should note that  $\beta_2$  really corresponds to the average within-person association, since the model again allows for the slope of this coefficient to vary across participants (and also their intercepts). 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.

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 to differentiate 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  $\beta_1$  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  $\hat{\beta}_1 = 0.194$  based on model (3) and  $\hat{\beta}_2 = 0.192$  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  $\beta_1$  in model (3) mostly reflects the within-person association. We can also see this reflected in the standard errors of  $\hat{\beta}_1$  and  $\hat{\beta}_2$  in model (5). The so-called ‘relative efficiency’ with

which these two coefficients can be estimated is given by  $SE[\hat{\beta}_1]^2 / SE[\hat{\beta}_2]^2$ , 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 is also close to  $19680 / 328 = 60$ , 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), 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.,  $\hat{\sigma}^2 = 0.893$ ) with that of model (1) (i.e.,  $\hat{\sigma}^2 = 1.015$ ), 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 the 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 as 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 of 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, we have tried to avoid any kind of phrasing with respect to the within-person association that would attribute causality to this finding, 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 to 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

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

where  $eventpl_{ij-1}$  denotes the 'lagged' version of the event pleasantness variable that we created during the data preparation step. For this model, we find  $\hat{\beta}_0 = 4.464$  ( $SE[\hat{\beta}_0] = 0.055$ ),  $\hat{\beta}_1 = 0.080$  ( $SE[\hat{\beta}_1] = 0.008$ ),  $\hat{\tau}_0^2 = 0.933$ ,  $\hat{\tau}_1^2 =$

0.0063,  $\hat{\rho} = -0.37$ , and  $\hat{\sigma}^2 = 0.981$ . 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  $\hat{\beta}_1 = 0.194$  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.,  $\hat{\sigma}^2 = 1.015$ ) 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 the event would on average have occurred about 45 minutes before the beep.<sup>8</sup> Ignoring the directionality 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 the participants' mood might have dissipated to some extent in the meantime. Given this, it is actually quite remarkable to find that such effects may still linger on more than 2 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

---

<sup>8</sup> Quite plausibly, participants may 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.

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, Kuiper, and Hamaker (2018) and the references therein for further details on this approach.

To reduce concerns with 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,  $\hat{\beta}_1$  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 1178 rows were lost as a consequence of setting the lagged event pleasantness to missing for the first beep within each day, the remaining 2018 rows were lost due to intermittent missingness. 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 at least to 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<sub>i</sub>* and *psy<sub>i</sub>* in the model and allowing these dummy variables to interact with *eventpl<sub>ij-1</sub>*). 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 again 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 centered values of the lagged event pleasantness variable as predictors), although for the same reasons as 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 also 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), but 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

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

$\beta_1$  denotes the average autocorrelation while  $\beta_2$  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, that is, 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 of 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.,  $\tau_0^2$ ,  $\tau_1^2$ , and  $\tau_2^2$  are the variances of  $u_{0i}$ ,  $u_{1i}$ , and  $u_{2i}$ , respectively) and are allowed to be correlated (i.e.,  $\rho_{01}$  is the correlation between  $u_{0i}$  and  $u_{1i}$ ,  $\rho_{02}$  is the correlation between  $u_{0i}$  and  $u_{2i}$ , and  $\rho_{12}$  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 undersampling, 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

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

where  $\beta_3$  denotes the average autocorrelation of the control group,  $\beta_4$  and  $\beta_5$  allow the average autocorrelation to differ for the depression and psychosis groups from that of the control group,  $\beta_6$  is denotes the average lagged relationship of the control group, and  $\beta_7$  and  $\beta_8$  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,  $\beta_1$  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  $\beta_1$ . 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; L. P. Wang & Maxwell, 2015). For example, if people tend to experience more unpleasant events early during 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

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

where  $\beta_9$  now denotes the average linear trend in positive affect in the control group and  $\beta_{10}$  and  $\beta_{11}$  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,  $\hat{\tau}_0^2 = 0.739$ ,  $\hat{\tau}_1^2 = 0.0298$ ,  $\hat{\tau}_2^2 = 0.0032$ ,  $\hat{\tau}_3^2 = 0.0003$ , 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	24.135	<.001
depressed	-1.056	0.187	-5.639	<.001
psychotic	-0.556	0.204	-2.723	.007
positive affect <sub>t-1</sub>	0.325	0.025	12.821	<.001
event pleasantness <sub>t-1</sub>	-0.004	0.012	-0.336	.737
response time	0.001	0.004	0.302	.763
depressed × positive affect <sub>t-1</sub>	0.060	0.034	1.787	.074
psychotic × positive affect <sub>t-1</sub>	-0.003	0.037	-0.089	.929
depressed × event pleasantness <sub>t-1</sub>	0.031	0.017	1.883	.060
psychotic × event pleasantness <sub>t-1</sub>	0.019	0.018	1.025	.305
depressed × response time	0.005	0.006	0.746	.456
psychotic × response time	0.000	0.007	0.039	.969

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 to the control group just fails to be significant ( $p = .06$ ). These results therefore call into question to some extent whether there really is, on average, an association between event pleasantness and positive affect, at least when using 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 modeling 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 in the context of the models discussed above, but this 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 chapters serve as an initial introduction to some of the main issues and approaches to consider when analyzing 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 analyzing longitudinal data. For those interested in the use of R for mixed-effects modelling, we recommend Pinheiro and Bates (2000) and Galecki and Burzykowski (2003), 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 analyze 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 analyze 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  $t=1,\dots,T$ . 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 analyze a binary outcome we can use the logistic mixed-effects regression model (see, Demidenko, 2013; Hedeker & Gibbons, 2006; McCulloch & Neuhaus, 2014; H. 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

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

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  $(2 \times 2)$  covariance matrix  $\Sigma_v$ . The diagonal elements of the covariance matrix of the random effects are given by the variances  $\sigma_{v0}^2, \sigma_{v1}^2$ , meanwhile, the off-diagonal elements by the covariance between the random effects denoted by  $\sigma_{v01}$ .

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 ten 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), and their positive affect which is 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* with 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.

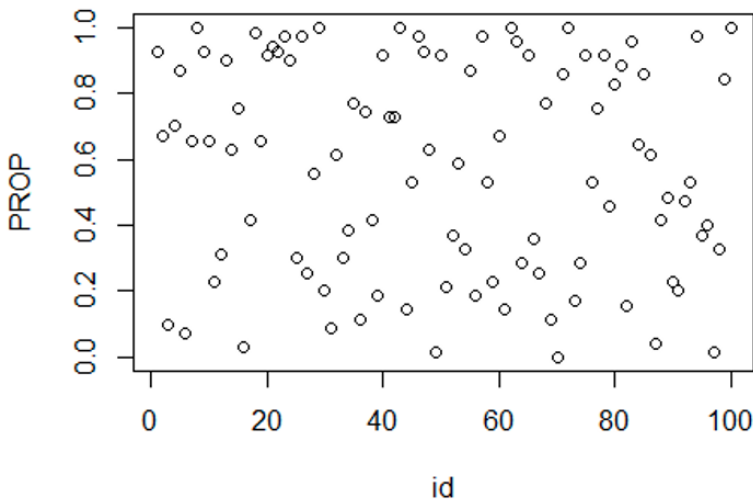
```
summary(data_company$Company)
```

##	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 that 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 centered 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 ( $\sigma_{v_0}$ ) equal to 4.545. To estimate the intra-class correlation coefficient (ICC) for generalized mixed-effects models, we use the package sjstats (Lüdtke, 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 goodness-of-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 ( $\text{Chisq}=31.732$ ); the degrees of freedom for the test (i.e., the difference in the number of parameters); and the p-value of the test ( $\text{Pr(>Chisq)}=1.287\text{e-}07$ ). If  $\text{Pr(>Chisq)}$  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- $R^2$  (fixed effects) = 0.01

Pseudo- $R^2$  (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

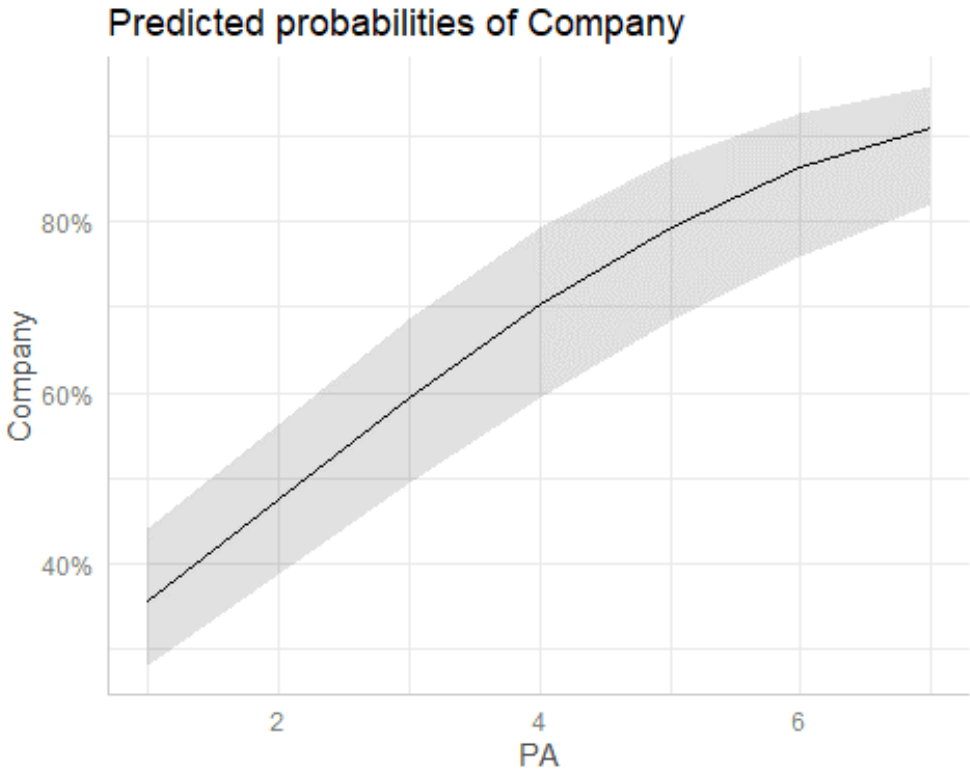
Grouping variables:

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 analyze count data we can use the Poisson mixed-effects regression model (see, Gibbons et al., 2008; Hedeker & Gibbons, 2006). This approach allows modeling the conditional mean of the outcome variable  $y_{it}$  as a function of a predictor  $x_{it}$  assuming the following form:

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

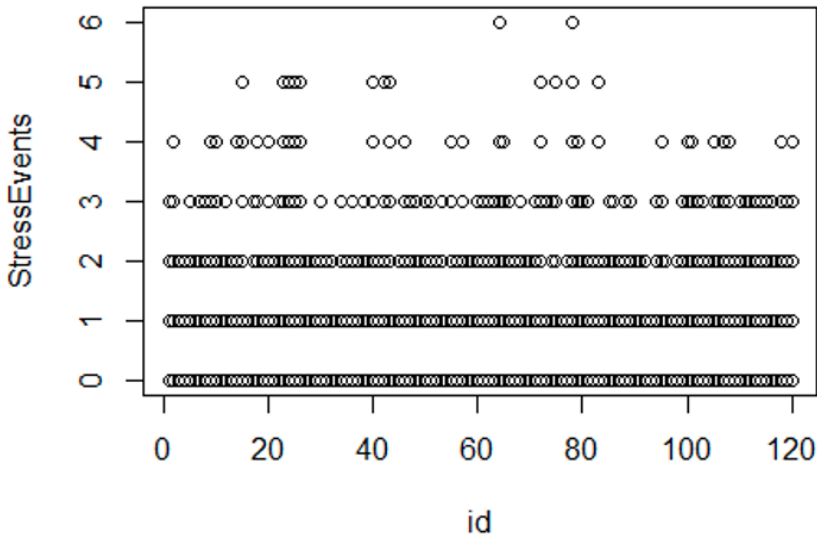


where  $E[\cdot]$  denotes the mean,  $\exp(\cdot)$  is the exponential function,  $\nu_{0i}$  and  $\nu_{1i}$  are the random effects of the random intercept and slope. The random effects are assumed to be multivariate normally distributed with mean zero and (2x2) covariance matrix  $\Sigma_v$ .

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* that 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 centered 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")`.

```
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

```

The output exhibits the estimated variance and correlations of the random effects, 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 be decreased by -0.20 units.

### 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 modeling a positive skewed outcome using the Gamma distribution with a log link function, and it can be modeled as follows

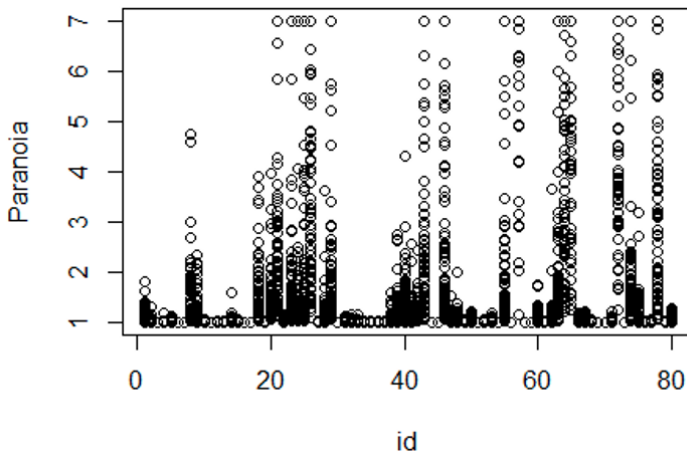
$$E[y_{it}|x_{it}, v_{0i}, v_{1i}] = \exp(\beta_0 + \beta_1 x_{it} + v_{0i} + v_{1i} x_{it})$$

where  $E[\cdot]$  denotes the mean,  $\exp(\cdot)$  is the exponential function,  $v_{0i}$  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* that includes data from 80 participants of an ESM study with ten 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 seven-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 1.066802 4.453393
## 2  1  1  2    2  2 1.121378 2.982072
## 3  1  1  3    3  3 1.008293 2.834387
## 4  1  1  4    4  4 1.028296 5.985223
## 5  1  1  5    5  5 1.067525 3.424695
## 6  1  1  6    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 centered 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 increases their negative affect 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).

Mixed-effects modeling is a flexible approach because it allows to capture and quantify the variance at different clustering levels. In chapter 9, we presented

the two-level multilevel 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 modeling 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 ten 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

$$\text{Level 1: } \text{Valence}_{ijt} = \gamma_{0ji} + \varepsilon_{ijt}$$

$$\text{Level 2: } \gamma_{0ji} = \beta_{00i} + v_{0ji}$$

$$\text{Level 3: } \beta_{00i} = \beta_{000} + v_{0i}$$

where  $\gamma_{0ji}$  represents the mean of individual  $i$  on day  $j$  and the error  $\varepsilon_{ijt}$  represents the deviation from the participant's affective valence on day  $j$  at beep  $t$ . The participant's mean level is denoted by  $\beta_{00i}$  and  $\nu_{0ji}$  is the deviation of the day mean level for this participant from the trait level.  $\beta_{000}$  denotes the grand mean of valence for the population, and  $\nu_{0i}$  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  $\sigma_{\varepsilon}^2$ . The random effects  $\nu_{0ji}$  are  $\nu_{0i}$  are normally distributed with mean zero and variance  $\sigma_{\nu_0}^2$  and  $\sigma_{\nu_0}^2$ , 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 participant clustering structure:

$$\rho_2 = \frac{\sigma_{\nu_0}^2}{\sigma_{\nu_0}^2 + \sigma_{\nu_0}^2 + \sigma_{\varepsilon}^2}$$

The Level 3 ICC describes the proportion of the total variance of the outcome that is accounted for by three-level structure of the data in which days are clustered within participants:

$$\rho_3 = \frac{\sigma_{\nu_0}^2}{\sigma_{\nu_0}^2 + \sigma_{\nu_0}^2 + \sigma_{\varepsilon}^2}$$

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. 2019). 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      M
ax
## -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 ( $\sigma_{\epsilon}$ ) is 0.75. The estimated standard deviation of the day level ( $\sigma_{v_0}$ ) is 0.17, and the estimated standard deviation that accounts for the variability at the person level ( $\sigma_{v_0}$ ) is 0.42. The Level 2 ICC of 0.230 means that 23% of the variability in the outcome can be accounted for by the between-person differences. The Level 3 ICC of 0.038 means that 3.8% of the outcome variance can be accounted for by the three-level structure of the data. 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  $(1 | \text{id}/\text{day})$ . 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 to 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), or 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 difference 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 1 shows a path diagram of a bivariate VAR(1) model. The autoregressive effects (solid arrows) are considered to be measures of emotional inertia (Kuppens, Allen, et al., 2010), whereas the cross-regressive effects (dashed arrows) reflect the effect of past affect on current PA and NA.

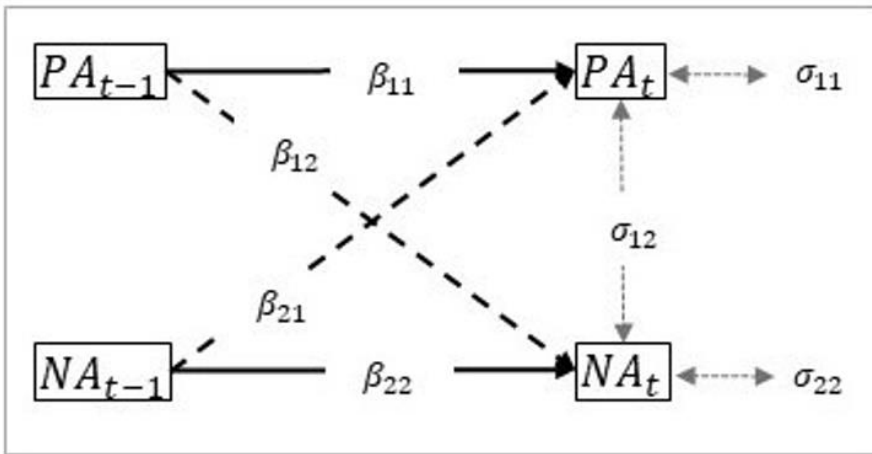


Figure 10.1 Bivariate VAR(1) model of PA and NA.

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 analyze 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 this model, 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 ten beeps per day over six days. The data include the variables positive affect (PA) and negative affect (NA). The variables are person-mean centered (centered 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  2      1      1  4.617290 1.771382 3.608206
2.545103
## 3  3  3      1      1  2.696894 1.991890 4.617290
1.771382
## 4  4  4      1      1  2.492036 1.467974 2.696894
1.991890
## 5  5  5      1      1  3.319702 1.888963 2.492036
1.467974
## 6  6  6      1      1  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. First, we fit a linear mixed-effect model to obtain the estimates of  $\beta_{11}$  and  $\beta_{12}$ . The first argument in `lmer()` is a formula that defines the structure of the fixed effects  $PA \sim 1 + PA\_lag + NA\_Lag$ . 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`).

```
# 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

```

The estimation output includes the estimated value of the fixed autoregressive effect  $\beta_{11}$  equal to 0.38, and the fixed cross-regressive effect  $\beta_{12}$  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  $\beta_{22}$  and cross-regressive  $\beta_{21}$  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  $\sigma_{12}$ .

```
# 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, whereas 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.

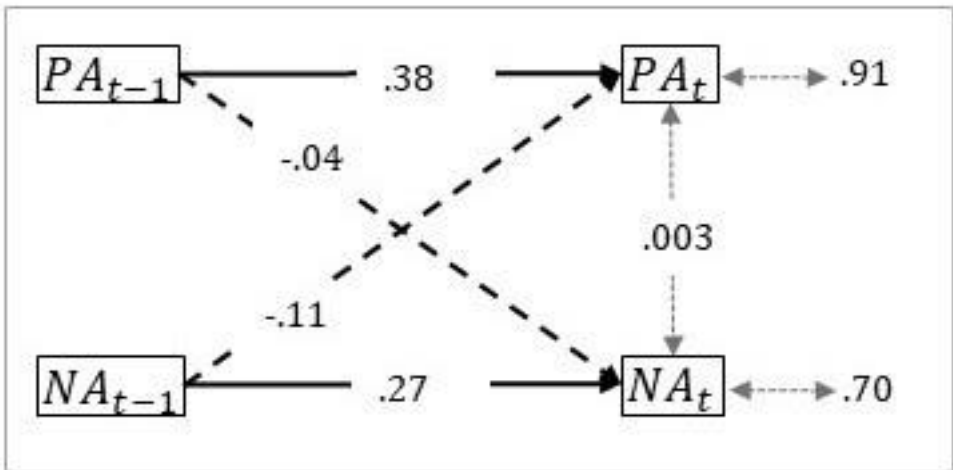


Figure 10.2 Estimated VAR(1) model of PA and NA.

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, a question that arises is 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 and showed 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 analyze three types of non-normal responses. Specifically, we illustrate how to fit a logistic mixed-effects model to analyze dichotomous or binary outcomes. We also present the Poisson mixed-effects model for modeling count data, and the inverse Gaussian mixed-effects model for analyzing 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 overdispersed count data: negative binomial, zero-inflated Poisson and negative binomial models, and Hurdle Poisson and negative binomial models. Good general resources on overdispersed count data include Brooks and colleagues (2017), Hall (2000), Molenberghs & Verbeke (2006), Zhang & Yi (2020), Zuur and colleagues (2009). Finally, we note that ESM data often includes semicontinuous outcomes. In this case, a proportion of responses are equal to a single value. This value often represents whether an individual engaged in a behavior. The remaining values follow a continuous and skewed distribution. In this case, two-part mixed effect models have been proposed for modeling 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 taking into account 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 was missing 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 the methodological tools for conducting power analyses for a number 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. Even though there are still additional considerations and more complex research questions, the set of methodological tools introduced in this chapter can be extended to more complex scenarios.

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, behavior, or thoughts over time in daily life, or how these themselves are related to time-varying predictors (see also chapter 2). All these research questions have in common that the goal 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 one. 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, Gibbons, and Waternaux (1999); Moerbeek, Breukelen, and Berger (2000); Moerbeek, Van Breukelen, and Berger (2001); Moerbeek and Maas (2005); Raudenbush (1997); Raudenbush and Liu (2001); Snijders and Bosker (1993); Wang, Hall, and Kim (2015).

The main drawback of the analytical approach is that currently existing formulas are limited to simple research questions (see e.g., 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 e.g., 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; Z. Zhang, 2014; Z. Zhang & Wang, 2009). Even though these methodological tools are useful, they do not consider a feature that characterizes ESM data, where 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 has been 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 that can be assessed using multilevel regression models (see chapter 2), and 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	<i>Group differences in mean level:</i> estimates differences between two groups of individuals in the mean of the outcome variable.
Model 2	<i>Effect of a continuous time-invariant predictor on the mean level:</i> estimates whether an individual-specific time-invariant variable predicts individual differences in the mean level of the outcome variable.
Model 3	<i>Effect of a Level 1 continuous predictor (random slope):</i> 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:* estimates whether the time-invariant predictors moderates 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 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. Secondly, 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 *PowerAnalysisIL* 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 one for individuals diagnosed with MDD and zero 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.

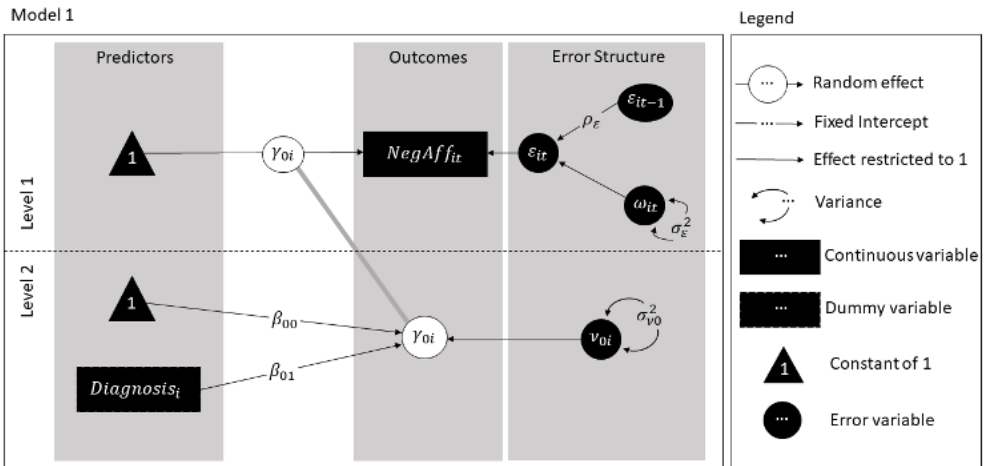


Figure 11.1 Graphical representation of the multilevel model to estimate group differences in the mean level of negative affect between individuals with MDD and healthy control participants.

The multilevel regression model is written as follow:

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

$$\text{Level 2: } \gamma_{0i} = \beta_0 + \beta_1 \text{Diagnosis}_i + v_{0i}$$

Inter-individual differences in negative affect are modeled by the random intercept  $\gamma_{0i}$ . 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  $\sigma_{v_0}$ . The model can be interpreted as follow: for participants in the reference group (controls), the mean level of negative affect equals  $\beta_0$  for individuals diagnosed with MDD, the mean level of negative affect is given by  $\beta_0 + \beta_1$ .

To account for the temporal dependencies in ESM data, we allow for serially correlated errors. We assume that the level-1 errors  $\varepsilon_{it}$  follow a first-order autoregressive (AR(1)) process (Goldstein et al., 1994), where the correlation between two consecutive errors is denoted by  $\rho_\varepsilon$  and  $\sigma_\varepsilon$  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<sup>1</sup>. The dataset includes 40 individuals that have been diagnosed with MDD and 60 control subjects. They all participated in an ESM study with ten 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* with the participants' identification number, *day* denoting the study day, *beep* 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 (J. Pinheiro et al., 2017). We set REML as the estimation method<sup>2</sup>, and we specify

<sup>1</sup> 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.

<sup>2</sup> 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

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
```

---

biased when the sample size is small. REML estimates unbiased variance components by taking into account the degrees of freedom of the fixed fixed-effects estimates. As a result, when the number of participants is small, it is recommended to use REML.

The model provides the estimates of the population parameters that subsequently will be used to perform the power analysis. The estimated fixed intercept  $\beta_{00}$  is 2.06; the differences in the mean level of negative affect between the two groups  $\beta_{01}$  is 0.37. The standard deviation  $\sigma_\epsilon$  and autocorrelation  $\rho_\epsilon$  of the level-1 errors are 1.05 and 0.27, respectively. And the standard deviation of the random intercept  $\sigma_\nu$  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 we set the number of participants for the healthy controls (Group 0) and MDD group (Group 1) respectively 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.*

**Choose a model (more information in panel About the Method):**

Model 1: Group differences in mean level

Model 1: Group differences in mean level

Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$

Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01}Z_i + \nu_{0i}$

$Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise

AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$

Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

**Number of participants in Group 0 (reference group)**

20, 30, 40, 60, 80, 100

**Number of participants in Group 1**

20, 30, 40, 60, 80, 100

**Number of time points**

70

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  $\beta_{00}$  is set to 2.06, and the effect of the level-2 dummy variable  $\beta_{01}$  is set to 0.37. We set the standard deviation  $\sigma_{\varepsilon}$  and autocorrelation of the Level 1 errors  $\rho_{\varepsilon}$  to 1.05 and 0.27, respectively. The standard deviation of the random intercept  $\sigma_{v_0}$  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 1000. 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.



Fixed intercept:  $\beta_{00}$

2.06

Effect of the level-2 dummy variable on the intercept:  $\beta_{01}$

0.37

Standard deviation of level-1 errors:  $\sigma_{\epsilon}$

1.05

Autocorrelation of level-1 errors:  $\rho_{\epsilon}$

0.27

Standard deviation of random intercept:  $\sigma_{\nu_0}$

0.52

☒ Estimate AR(1) correlated errors  $\epsilon_{it}$

Type I error:  $\alpha$

0.05

Monte Carlo Replicates

1000

Choose the method to fit linear mixed-effects model

Maximizing the restricted log-likelihood

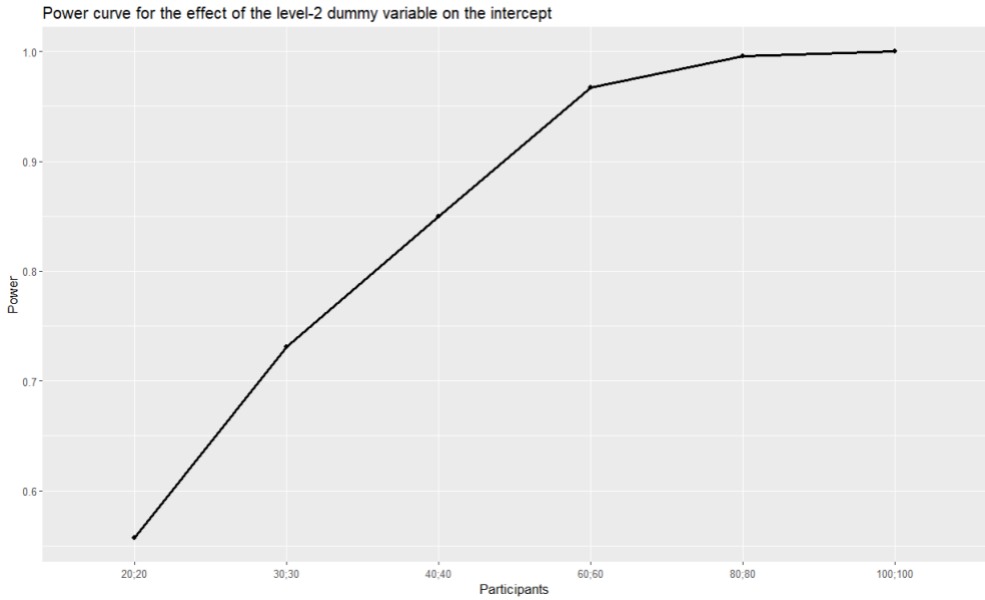
Estimate Computational Time

Compute Power

Reset Page

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 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.,  $\beta_{01}$ ) is 55.4%. This result implies that in only 554 out of the 1000 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.



*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.*

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.

*Table 11.2 Summary of the fixed effects across 1000 Monte Carlo replicates to estimate mean differences in negative affect between participants with MDD and healthy control participants.*

Fixed Effects	N <sub>0</sub>	N <sub>1</sub>	True Value	Mean	Std. error	Bias	(1- $\alpha$ ) Coverage	Power
Fixed intercept ( $\beta_{00}$ )	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 ( $\beta_{01}$ )	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 we have kept the number and spacing of time points fixed. However, researchers might not only be interested 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, and for each simulation we specify the number of time points to 20, 40, 70, 100, and 140. Figure 4 shows the screenshots of the *PowerAnalysisIL* app. In each window, the 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).

### 1. Simulation for $T = 20$

Choose a model (more information in panel About the Method):

Model 1: Group differences in mean level

Model 1: Group differences in mean level

Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$

Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01} Z_i + \mu_{0i}$

$Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise

AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$

Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

Number of participants in Group 0 (reference group)

20,40

Number of participants in Group 1

20,40

Number of time points

20

### 2. Simulation for $T = 40$

Choose a model (more information in panel About the Method):

Model 1: Group differences in mean level

Model 1: Group differences in mean level

Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$

Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01} Z_i + \mu_{0i}$

$Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise

AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$

Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

Number of participants in Group 0 (reference group)

20,40

Number of participants in Group 1

20,40

Number of time points

40

Figure 11.3 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

4. Simulation for  $T = 100$

Choose a model (more information in panel About the Method):  
Model 1: Group differences in mean level

Model 1: Group differences in mean level  
Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$   
Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01} Z_i + \mu_{0i}$   
 $Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise  
AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$   
Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

Number of participants in Group 0 (reference group)  
20,40

Number of participants in Group 1  
20,40

Number of time points  
100

3. Simulation for  $T = 70$

Choose a model (more information in panel About the Method):  
Model 1: Group differences in mean level

Model 1: Group differences in mean level  
Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$   
Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01} Z_i + \mu_{0i}$   
 $Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise  
AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$   
Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

Number of participants in Group 0 (reference group)  
20,40

Number of participants in Group 1  
20,40

Number of time points  
70

5. Simulation for  $T = 140$

Choose a model (more information in panel About the Method):  
Model 1: Group differences in mean level

Model 1: Group differences in mean level  
Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$   
Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01} Z_i + \mu_{0i}$   
 $Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise  
AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$   
Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

Number of participants in Group 0 (reference group)  
20,40

Number of participants in Group 1  
20,40

Number of time points  
140

6. Simulation for  $T = 200$

Choose a model (more information in panel About the Method):  
Model 1: Group differences in mean level

Model 1: Group differences in mean level  
Level 1:  $Y_{it} = \gamma_{0i} + \epsilon_{it}$   
Level 2:  $\gamma_{0i} = \beta_{00} + \beta_{01} Z_i + \mu_{0i}$   
 $Z_i$  is a dummy variable equal to one if participant is in Group 1 and 0 otherwise  
AR(1) errors  $\epsilon_{it}$  with autocorrelation  $\rho_\epsilon$  and variance  $\sigma_\epsilon^2$   
Number of participants: introduce an increasing sequence of positive integers (comma-separated). The length of the sequence must be the same in the two groups.

Number of participants in Group 0 (reference group)  
20,40

Number of participants in Group 1  
20,40

Number of time points  
200

Figure 11.3 continued

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<sub>i</sub>*). 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 interests. For a broader discussion on the impact of the sample size at the different levels, we refer to Snijders (2005).

*Table 3. Summary of the fixed effect 01 representing the Level 2 dummy variable on the intercept across 1000 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- $\alpha$ ) 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 by simulating data sets 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 for sure include missing assessments (e.g., Fuller-Tyszkiewicz et al., 2013; P. Santangelo et al., 2014; Stone, Broderick, et al., 2003). Therefore, when using the app, we recommend specifying the expected number of completed measurement occasions. Besides, 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 conduct 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 this chapter, we introduced a methodological approach to perform 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 e.g., 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.

## 11.6 Conclusions

In this chapter, we have briefly reviewed current approaches for conducting power analysis in multilevel models. In particular, we have 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.

We note that power analysis requires that the investigator determines 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.

To sum up, we highlight that the importance of conducting power analysis to justify the selection of the sample size in ESM studies is not only related to the reliability and replicability of psychological research (Munafò et al., 2017). Additionally, 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**

# **ECOLOGICAL MOMENTARY INTERVENTIONS: FROM RESEARCH TO CLINICAL PRACTICE**

Ana Teixeira



In the last years, we have witnessed an exponential growth of Ecological Momentary Interventions (EMI), which uses electronic devices (such as smartphone apps) to implement individually tailored interventions in real-time at moments when it is most needed and in the real world (Myin-Germeys et al., 2016). This type of intervention has the advantage of helping patients to extend skills and behavioral strategies learned in the therapy office into their real-life (Luxton et al., 2011; Miralles et al., 2020; Myin-Germeys et al., 2016). Advances in technology and wearable biosensors have opened the door to the development of innovative EMIs and their implementation in clinical practice (Colombo et al., 2019).

This chapter comprises four parts that provide an overview of this field of research and implementation in clinical practice. In the first part, we aim to give a general introduction to EMI and how complex it can be to tailor interventions to individuals or groups. In the second part of the chapter, we discuss how to monitor and evaluate EMIs considering the full life cycle of this type of m-health interventions and shortly present how this assessment is currently being done in research. In the third part, we focus on the advantages of incorporating EMI in clinical practice and the importance of including clinicians' and patients' perspectives in EMI development. Moreover, in the fourth and final part of the chapter, we will discuss barriers and additional steps towards clinical implementation. In this chapter, patients are sometimes referred to as end-users.

## 12.1 Ecological Momentary Interventions

EMIs can be defined as momentary psychological interventions delivered via mobile technologies (e.g., smartphones) in the flow of daily life (Heron & Smyth, 2010). They are cost-effective ways of providing psychotherapy and making it available to larger groups of individuals (Heron & Smyth, 2010; Myin-Germeys et al., 2016). EMIs can be offered as a stand-alone intervention, which means that they are employed on their own as a preventive (self-help) tool. An example of a stand-alone EMI is MoodPrism, a CBT-based and transdiagnostic smartphone app (Bakker et al., 2018). This app includes self-monitoring of mood and offers psychoeducational information on mental health (Bakker et al., 2018). In addition, EMIs can also be offered as a blended care intervention, which means that they are implemented as an add-on to other face-to-face psychological interventions. The advantage of blended care is the potential of EMIs to boost

the effects of therapy and reduce the amount of contact between clinicians and patients (Gee et al., 2016). One example of a blended care EMI is the Acceptance and Commitment Therapy in Daily Life (ACT-DL), which was developed to increase psychological flexibility in everyday life as an add-on to ACT therapy (Myin-Germeys et al., 2021).

Some EMIs merge the Experience Sampling Method (ESM) with intervention/treatment strategies delivered outside the therapy office (Myin-Germeys et al., 2016). The addition of ESM assessments to EMIs has the advantage of offering relevant information that can tailor the intervention to specific moments of need for that individual and evaluate the effect of the intervention itself (Heron & Smyth, 2010; Luxton et al., 2011). For example, by monitoring social contexts, thoughts, mood, behaviors, and symptoms in real-time, EMIs can offer help when clients need it the most (for instance, when they feel distressed). This is done by offering a specific exercise or strategy in real-time to help the client deal with stressful situations (just-in-time interventions). Interventions can also be content-tailored to individuals based on information known at a pre-intervention stage, as it is the case of EMIs that were developed explicitly to a particular mental health problem or diagnosis (e.g., depression, anxiety, stress) (Luxton et al., 2011). Like ESM, EMIs are considered ecologically valid since they are delivered in the natural flow of daily life in real-time and at specific moments (Heron & Smyth, 2010).

According to Carter and colleagues (2007), EMIs can be differentiated into three levels of complexity depending on the interaction between the intervention tool and the end-user: low interaction, interactive and integrative.

The low interaction level refers to EMIs that are delivered equally to all end-users and only offer prompting messages (such as reminders, notifications, motivational messages or instructional exercises), which does not require much (inter)action from the end-user (Miralles et al., 2020; Schueller et al., 2017). An example is an app that only includes psychoeducation information, or offers tips or learning materials in written, audio, or video formats (Carter et al., 2007; Schueller et al., 2017). The end-users access this information anytime they wish or need. Self-help apps are a good example of low complexity.

The *interaction level* refers to EMIs that provide interventions based on the information collected from the end-user, for instance, receiving a specific

intervention or tip based on a particular score obtained in a questionnaire, such as a reported high level of subjective stress (Carter et al., 2007; Schueller et al., 2017). It is considered interactive because different scores on a variable of interest (such as stress levels) will prompt different interventions or tips. This type of EMI is the same for all end-users, and no real personalization is possible. Most current m-health apps are of low interaction or interactive complexity (Miralles et al., 2020).

Contrary to the other levels of EMIs, the integrative level is more complex because the intervention is completely tailored to a specific end-user and delivered at the right moment. In this type of EMI, a learning system keeps evolving by continually improving users' interventions based on their pattern of responses and interactions with the system (Carter et al., 2007; Schueller et al., 2017). For instance, apps that provide tailored interventions based on software algorithms from machine learning analysis and information collected passively through sensors incorporated in the smartphones or wearables (Schueller et al., 2017). A concrete example is the Just-In-Time Adaptive Interventions (JITAIs). JITAIs can detect the right moment when an intervention should be provided by collecting past information from the individual and their environment in real-time (Nahum-Shani et al., 2018; Schueller et al., 2017).

When developing an EMI, researchers need to consider when an intervention should be given (for instance, an intervention can be triggered at fixed times, or at random times along the day or after a particular event), what kind of data will be collected, and the level of complexity of the EMI (Carter et al., 2007; Luxton et al., 2011). All these decisions will depend on the goal of a particular EMI.

## **12.2 Research and Assessment of EMI's**

Several mobile apps have been developed over the last years, including in the field of mental health (Gee et al., 2016; Miralles et al., 2020; Versluis et al., 2016). The World Health Organization developed a practical guide for conducting research and assessment of digital health interventions (World Health Organization., 2016). This guide describes the several maturity life cycle phases of digital health interventions (such as EMIs), namely: pre-prototype/prototype, pilot, demonstration, scale-up, and integration/sustainability. During those

phases, monitoring and evaluating processes occur to assess the digital health intervention.

The monitoring process refers to collecting and analyzing data continuously to assess the technical functionality (the ability of the system to operate as planned) and stability (the ability of the system to operate consistently) and assess fidelity (functionality and stability are maintained during implementation) and quality (content and delivery) of the intervention. This process will provide information about how well the intervention is being implemented, identify necessary adjustments, and what needs to be optimized (World Health Organization., 2016).

The evaluation process refers to the systematic and objective assessment of the intervention on the following outcomes: usability (the intervention is user-friendly), feasibility (the ability of the system to work as planned in a specific context), efficacy and effectiveness (health impact of the intervention in research and non-research environment, respectively), and economic/financial evaluations (affordability of the intervention) (World Health Organization., 2016). In short, “while monitoring asks: is the project doing things right? evaluation asks: is the project doing the right things?” (Pritchett et al., 2013).

These monitoring and evaluation outcomes are assessed in specific phases of the life cycle. For instance, the technical functionality and stability outcomes are assessed in the pre-prototype/prototype phase, while fidelity and quality are assessed in all other phases of the maturity cycle. Regarding the evaluation outcomes, usability and feasibility are assessed in the pre-prototype/prototype phase, efficacy in the pilot phase, and effectiveness in the demonstration phase. The two later stages of the life cycle (scale-up and integration/sustainability) focus on scaling up the intervention to a national level; assessing costs of the intervention and cost-effectiveness, and assessing the integration and sustainability of evidence-based interventions in the health system (including policies and practices, interoperability, human resources)<sup>1</sup>.

Regarding research on EMIs for mental health, most studies are still in the initial phases of the life cycle (pre-prototype/prototype and pilot), assessing mostly usability and efficacy outcomes (Miralles et al., 2020). Nevertheless, only a minority of studies conduct Randomized Controlled Trials (RCT) or pilot RCTs, and from those, only a few assess efficacy (Miralles et al., 2020). This shows that researchers recognize the importance of guaranteeing that their interventions are user-friendly, but more efforts are necessary to monitor and evaluate digital interventions robustly. Studies with follow-up data that are fundamental to check behavior change and implementation of skills in daily life usually do not extend the six months (Heron & Smyth, 2010; Rathbone & Prescott, 2017).

Several meta-analyses and systematic reviews studies have shown a small to medium effect of EMIs on reducing depression, anxiety, and stress compared to control conditions (Firth et al., 2017; Gee et al., 2016; Versluis et al., 2016). However, the quality of those studies' quality is considered low (Gee et al., 2016; Versluis et al., 2016). Studies assess mostly stand-alone EMIs, while blended EMIs are considered to have larger effects (McDevitt-Murphy et al., 2018; Versluis et al., 2016). Thus, there is the need for high-quality RCT designs and clinical trials that include longer follow-ups and larger sample sizes to robustly assess efficacy outcomes of EMIs (Gee et al., 2016; Rathbone & Prescott, 2017).

---

<sup>1</sup> For more detailed information on this topic, we recommend the reading of the WHO guide on monitoring and evaluating digital health intervention available for download: <https://www.who.int/reproductivehealth/publications/mhealth/digital-health-interventions/en/>



RCTs are especially necessary for mental health problems other than depression and anxiety disorders (Miralles et al., 2020).

An interesting alternative to RCT is the micro RCT. The latter has been considered a better alternative to assess the efficacy of EMI because its design allows assessing the immediate or long-term effects of specific components of the intervention, how they change over time, determining when these components are more efficacious and what (psychological or contextual) variables moderate their effects (Klasnja et al., 2015). This method follows a sequential factorial design, where specific intervention components or no intervention are randomly assigned to each participant at particular time points (Klasnja et al., 2015).

In the initial stages of developing and evaluating EMIs, qualitative research is crucial to collect feedback from clinicians and patients who will implement or receive EMIs (Heron & Smyth, 2010; Luxton et al., 2011). In the literature, we see a growing interest in qualitative research and research involving clinicians and patients. Clinicians' involvement allows identifying challenges and opportunities for EMI in healthcare settings. In contrast, patients' involvement allows assessing the acceptability and usability of EMIs from their perspective (Heron & Smyth, 2010; Luxton et al., 2011). After few uses, most individuals discontinue using mental health apps, making it relevant to study the end-user's perspective on app content and preferences to increase engagement (Nicholas et al., 2017). This pilot testing phase that includes qualitative information should precede an RCT or a micro RCT studies to identify problems and correct them in time (Heron & Smyth, 2010).

### **12.3 Benefits of EMI for clinical practice**

The development and study of EMI's are, of course, intended to be implemented in clinical practice to help support the evolution of mental health care. One important avenue by which EMIs realize this is by taking a person-centered approach (Myin-Germeys, 2020). EMIs that incorporate ESM assessments are an excellent way to accomplish that because they allow patients and clinicians to assess and monitor subjective mood, social environment, behaviors, and symptoms in real-time and real-world (Myin-Germeys et al., 2016). These assessments can be shared and visualized in the therapy room, providing

clinicians and patients with a better picture and understanding of problems and potential personalized targets for treatment (Myin-Germeys et al., 2016). In this context, patients are considered active actors of their treatment and recovery process and not merely passive receivers of care (Morley & Floridi, 2020; Myin-Germeys, 2020). Clinicians and patients have different kinds of expertise: therapists have clinical and theoretical knowledge on mental health problems and treatment options, while patients know their own qualitative and experiential aspects (Charles et al., 1999; Morley & Floridi, 2020). EMIs allows patients to quantify their experiences and bring them to the table. This makes patients more active partners in the process, which helps to improve mutual understanding, share responsibility, and make decisions together (Economou et al., 2019; Myin-Germeys, 2020; Slade, 2017).

Another way in which EMIs can contribute to clinical practice is by helping to bridge what is known as the therapy-real world gap that traditional psychotherapy faces by extending therapy from the therapy office to individuals' everyday lives (Miralles et al., 2020; Myin-Germeys et al., 2016; Schueller et al., 2017). They offer patients the opportunity to practice new skills and behaviors learnt in the therapy office in their day-to-day life and several life contexts. This is especially relevant because research shows that skills learned in therapy do not always translate to patients' life contexts (Heron & Smyth, 2010; Versluis et al., 2016). Patients often struggle to generalize and implement these skills in their personal life (Versluis et al., 2016).

Therefore, adding technology to traditional care would facilitate patients' involvement in their healthcare process, giving them the necessary tools for self-management and self-monitoring (Hollis et al., 2015; Miralles et al., 2020). It also allows them to stay committed to therapy, feel empowered and actively contribute to their recovery process (Hollis et al., 2015; Luxton et al., 2011). There are already mental health online platforms that facilitate EMI implementation in clinical practice, such as m-Path (<https://m-path.io/>). However, it also requires time for researchers to set up ESM questionnaires and the interventions and teach clinicians and patients how to use a dashboard and an app (Heron & Smyth, 2010).

Research on m-health apps for mental health problems shows that the most represented targets are depression, stress, anxiety, substance-related and

addictive disorders, schizophrenia spectrum/psychotic disorders, trauma and stressor-related disorders obsessive-compulsive disorder (Miralles et al., 2020; Rathbone & Prescott, 2017). These intervention targets seem to be consistent with the most prevalent disorders worldwide, but some are still underrepresented (Miralles et al., 2020). Regarding the type of psychological intervention techniques offered or studied through EMIs, research shows that they mainly involve CBT, Acceptance and Commitment Therapy, mindfulness, behavioral activation, and relaxation (Versluis et al., 2016). Although there is a consistent growth of EMIs, there still is a research-to-practice gap, which means that EMI has not yet been adequately integrated into clinical practice and healthcare systems (Myin-Germeys, 2020).

## 12.4 Needs and barriers for clinical implementation

Despite the apparent clinical potential of EMIs, few EMIs are actually implemented in routine mental health care. The translation from research findings and evidence-based practices into (mental) healthcare – implementation science – has progressed slowly (Hollis et al., 2015; Myin-Germeys, 2020; Wensing & Grol, 2019). The scientific evaluation of the entire life cycle of digital interventions (usability, feasibility, efficacy, and effectiveness) can take several years (Rathbone & Prescott, 2017). It is also hard to keep up with rapid technological development (Hollis et al., 2015; Rathbone & Prescott, 2017).

More than developing new EMIs, there is a need to provide scientific evidence to those already available, especially those developed for people with mental health problems. It is critical to offer clinicians evidence-based EMIs that they can safely use in their clinical work and with their patients. This is especially relevant because non-validated apps might provide potentially harmful or inaccurate content and worsen individuals' symptoms (Larsen et al., 2016; Luxton et al., 2011; Miralles et al., 2020).

In addition, there are concerns regarding data security and privacy. Most apps collect sensitive information beyond what is relevant for intervention goals; do not present privacy policy information; and share user information with third parties for storage and analysis (e.g., google analytics) or marketing intentions without disclosing this information in their privacy policy (Huckvale et al., 2019; Miralles et al., 2020; O'Loughlin et al., 2019; Parker et al., 2019). Recently, we

have seen significant efforts from the European Commission to regulate (mental) health apps. Apps with a medical purpose need to comply with Medical Device Regulation and obtain CE marking (EUR-Lex. Regulation (EU). 2017) and comply with EU General Data Protection Regulation (EUR-Lex. Regulation (EU). 2016).

There is also the need to involve end-users in creating, designing, and assessing EMIs to increase long-term use and provide person-centered care (Torous et al., 2018). The very few studies that have involved patients in those phases reported higher treatment adherence and engagement (Biagianti et al., 2017; Goodwin et al., 2016; Killikelly et al., 2017). Thus, it is necessary to address and overcome the described needs and barriers to successfully disseminate and implement EMI in clinical practice and bridge the gap between research and practice (Kristensen et al., 2016; Myin-Germeys, 2020; Wensing & Grol, 2019).

## **12.5 Conclusion**

Technological advances set the stage for person-centered care in psychiatry healthcare, providing the necessary tools to identify and personalize treatment targets (Myin-Germeys, 2020). These advances increase the pressure to provide evidence-based EMIs in the clinical market that complies with medical device regulations to ensure the best patient care (Myin-Germeys, 2020). The integration of EMIs in clinical practice and healthcare systems have the advantage of reducing intervention costs by improving treatment efficiency and decreasing treatment time (Heron & Smyth, 2010; Rathbone & Prescott, 2017). Nevertheless, continuing efforts need to be made to support clinicians in terms of knowledge, training and implementation of an EMI in their clinical practice (Kerst et al., 2020).



## **CHAPTER 13**

# **PASSIVE SENSING**

Joana De Calheiros Velozo, Koen Niemeijer & Thomas Vaessen



Over the past decade, there has been a steep increase in the use of wearables and other means of passive sensing (data collection without immediate active involvement of the user) to capture aspects of real-life behavior and environment. For instance, according to the Web of Science database, there has been over a tenfold increase in the number of published scientific reports annually on wearables (i.e., body-worn devices that measure bodily or environmental signals) since 2010. No doubt, this rise has to do with the technological advances in the fields of microelectronics and software development, but also improvements in esthetic design during the past decade. Likewise, the development of smartphones has taken flight, adding even more possibilities for the passive ecological investigation of daily-life dynamics. The results of these developments on crucial aspects for both general and research purposes, such as measurement accuracy and reliability, battery life, comfort, esthetics, and user friendliness, have led to the utilization of wearables in both science and personal use on an unprecedented scale. These wearables have become particularly popular with eESM researchers who use passive sensing to add to and further develop their traditional protocols.

Commercial wearables such as sports watches and fitness trackers have become mainstream products. In 2019, there were a projected total of 722 million wearables connected worldwide (Statista, 2019); twenty-one percent of American adults owned a smartwatch or fitness tracker, with projections for the next years indicating an increasing trend (PewResearch, 2020). Most commercial wearables can now continually measure physiological variables for days. Whereas even basic wearables are able to estimate heart and respiration rate through photoplethysmography (PPG) and accelerometer-based movement, higher-end wearables complement these measures with electrocardiography (ECG), galvanic skin response (GSR), and skin temperature sensors. The widespread personal use of these wearables has undoubtedly lowered the threshold for scientific data collection – not only in terms of availability, but also of user acceptance – 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 passive collection of daily life data, due to its many applications and its global availability. An estimated 3.2 billion people worldwide own a smartphone (Newzoo, 2019), with percentages of >76 for populations of developed countries (PewResearch, 2019). 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 using an accelerometer, gyroscope, GPS, and Bluetooth – features that are available in most smartphones today. Social interactions can be monitored through Bluetooth, microphone, or app usage, calls, and text messages. Screen on/off time, typing speed and patterns, and Wi-Fi connectivity are only some of the other possibilities smartphones offer to capture behavioral and environmental markers of daily living. 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. Many features of daily life, such as an individual's environment or behavior, can be assessed using passive sensors. Combining passive sensing with ESM potentially holds several important benefits. First, it could add information to ESM measures. Although passive sensing may never be able to assess an individual's personal experience, it can enrich ESM data by adding biological, behavioral, and contextual variables of interest. Second, it could substitute certain ESM measures. In some cases, passive sensing may be able to replace ESM items asking for a description of a phenomenon that can be directly measured using passive sensing (e.g., location, ambient noise, etc.). Third, it could improve the precision of ESM, increasing the sampling of moments of interest by triggering ESM questionnaires when pre-set passive sensor-based thresholds are reached. However, additional sensors and protocols may also increase the burden of participating in ESM research, potentially influencing compliance and the measurements themselves. In this chapter, we will discuss the benefits and potential challenges of passive sensing for ESM research.

### **13.1 Using passive sensing to enhance ESM research**

ESM has garnered a significant following in part because of its unique ability to account for context. Passive sensors are particularly useful to amplify ESM by collecting a wide array of contextual factors that would have not been possible otherwise. Passive sensing encompasses two main branches of data collection: data that is passively collected directly from the participant's smartphone (e.g., microphone, bluetooth, swiping activity, GPS, message and call logs), and data collected from an additional wearable device be that a wristband,

an earpiece, a chest patch, or anything additional that is provided to the participant to wear.

### ***13.1.1 Smartphone sensing***

Smartphones have effectively become an extension of the self (C. S. Park & Kaye, 2019) and as such provide a means that may be particularly useful to supplement ESM research, providing more fine-grained information on several well-studied fields. Social behavior, for instance who the participant is with and whether they are actively engaged in a social interaction or not, is a construct that may greatly benefit from the addition of passive sensing as it is oftentimes difficult to capture precisely with ESM items alone. In a standard ESM protocol, participants may be asked whether they are in the company of others and if so, to appraise this company, but little is known about the specifics of their company. This is where passive sensing can come in to augment our knowledge of the social context; while ESM collects information on subjective experience, passive sensing can gather information on behavior and surroundings (Lathia et al., 2013). Together, data from the microphone, Bluetooth, calls and messages, as well as various apps can help identify if the participant is in company or not, with how many people, how close or far, and whether they are actively interacting or not (Bachmann, 2015). In fact, data collected from the smartphone has been used as a complement to ESM in order to estimate not only if the participant is having or is near a conversation but also more specific information such as how long the conversation lasts (Morshed et al., 2019).

Naturally, social behavior is no longer limited to the physical realm but extends to the virtual world as well. Online social behavior is difficult to capture via questionnaires as it comprises a wide complex range of behaviors starting from more direct active forms of online interaction (e.g., messaging, calling) to more passive forms such as liking, sharing, and commenting. Smartphone usage can easily track every aspect of these behaviors without adding any additional burden to the participant and further contributing to the ESM (Morshed et al., 2019).

Smartphone data such as Bluetooth and GPS further work to amplify ESM by providing near continuous measurement on various aspects surrounding the individual that are scarcely captured through ESM unless it is the specific focus of the study. The specifics of an individual's daily trips for example are

generally omitted from ESM even though it has shown to have a direct effect on their affect. A study combining GPS with ESM for example found that it may not be where people are going that affects their mood but how they get there. That is, regardless of where they are going, people generally report a more positive mood when engaging in an active mode of transportation such as walking or biking rather than driving or taking a bus. Likewise, a commute that includes greenery and large water bodies was found to be better for mood than commuting through dense urban areas (Glasgow et al., 2019).

GPS not only informs on the participant's travel patterns (Glasgow et al., 2019; Reinau et al., 2015), but also on situational factors such as weather, daylight, location type, proximity to certain structures (i.e., entertainment, nature, social buildings, offices, shopping centers, etc.) (Mackerron & Mourato, 2013), to name a few examples. Given that people report higher levels of well-being when completing a questionnaire outdoors, and/or on sunny days versus indoors, and/or on cold dark days (Feddersen et al., 2016; Kööts et al., 2011; Messner & Wanke, 2011), access to this type of data offers the opportunity to assess their role in daily life, as well as control for them in more complex statistical models.

These are but a few examples of the possibilities passive smartphone sensors bring. Some researchers have even claimed that solely by tracking patterns in typing or swiping pressure, it is possible to identify when an individual is stressed or not (Exposito et al., 2018). And while promising, we must note that these applications come with several new challenges. For instance, Bluetooth interactions can only be established when nearby devices have this function activated, and it may be difficult to identify the type of interacting device (mistaking a printer for another smartphone may lead to very erroneous conclusions!). Users must have their smartphone on them for the sensors to collect meaningful information at all. At the same time, we can but acknowledge the youngness of this field and even though passive sensing will not allow us to measure everything, it allows us to measure much more than ESM alone.

### ***13.1.2 Wearable sensing***

Like smartphones, developments and access to wearable technology has proven a great asset for ambulatory research. For instance, accelerometry derived from body-worn sensors, ideally placed on the torso, provides much more reliable information about bodily movements than do smartphone sensors.

Accelerometry and/or physiological markers such as heart rate and breathing can further inform on the intensity of the participant's physical activity (Treuth et al., 2004). More and more studies investigate the possibility of predicting different experiential states such as affect types through variables acquired through passive sensing. For instance, several studies have tried to predict the experience of acute stress using machine learning models based on physiological and motion parameters that have been collected using passive sensing (Hovsepian et al., 2015; Smets et al., 2018). Although these models as of yet do not reach desired performance levels, in time they may provide possibilities to estimate experiential states, at least to a certain extent, without active involvement of the participant. However, as these models inevitably are built upon subjective experience as ground truth, the question can always be posed whether passive sensing is a better approach than self-report.

The same applies to sleeping behavior. While ESM studies may include a few items inquiring about sleeping patterns such as length, and quality of sleep (i.e., waking up multiple times, ease falling asleep, etc.), passive sensing can further inform on this by directly tracking the physiology and movement associated with the participant's sleep. By adding passive sensing to ESM, it is possible to investigate the effects of the more detailed aspects of sleep quality, such as the exact moment the participant fell asleep and woke up, how much did the participant move in their sleep (Staples et al., 2017), to name a few.

In short, passive sensing can significantly enhance ESM by expanding or in some cases replacing ESM within the scope of interest. Although ESM already covers a broad range of measures, passive sensing can further amplify them by providing more comprehensive data to complement it.

### **13.2 Using passive sensing to substantiate ESM research**

Passive sensing is not limited to enhancing ESM data, in some cases it can also be used to corroborate or substantiate it. That is, passive sensing can work as a means to check (or confirm) the subjective data reported. While it is not yet widely used in this context it presents enormous potential to better study certain populations or behaviors that are particularly difficult to track precisely otherwise. For example, it is notably challenging to study individuals who suffer from substance abuse as they may be either unable to or unwilling to honestly respond to successive self-reported questionnaires on their substance use. For that reason,

rather than relying solely on their ESM reports, GPS data combined with physiological data can provide further evidence to substantiate their reports. Using GPS allows the researcher to check whether the participant is in close proximity to places where substance is likely to be present such as bars for example, and physiology can help identify substance use by confirming whether there is a physiological response associated with taking said substance or not. In addition, certain tools can further corroborate whether a substance has or not been used, for example in the case of a breathalyzer to check for alcohol intake (Bertz et al., 2018). Needless to say, there may also be ethical objections to certain usage of passive monitoring. Some of these considerations are discussed in chapter 5.

Likewise, young children are also difficult to study accurately as they may struggle to complete electronic surveys. In a study investigating physical activity in children between the ages of five and seven, accelerometers tracking their activity was combined with an ESM completed by the parents (Engelen et al., 2015). The addition of passive sensing worked to confirm children's actual physical activity.

In some cases, it is not the population that presents the challenge but the behavior itself. In fact, there are certain behaviors that we are simply unable to report correctly as we grossly under- or over- estimate them. This is the case of phone or internet usage that are commonly heavily underestimated (Yuan et al., 2019). In these cases, passive data may be more reliable than ESM self-reports, providing a more accurate and timely depiction of the participants' phone usage.

Smoking is another behavior that is difficult to measure accurately with self-reported measures. In fact, smokers frequently underestimate how often and how much they actually smoke. Using Bluetooth e-cigarettes in parallel with ESM to measure the puff count and duration, Pearson and colleagues (2017) found that smokers underestimated how much they smoked 87% of the time, with an average difference of 33 puffs per day.

These are mere examples where passive sensing has and can be used to add value to ESM by either enhancing and/or corroborating certain measures that may not be entirely reliable on their own due to either the demographic or the measure itself. However, there are other potential applications and no doubt these will be further explored and applied in future research.

### 13.3 Triggering ESM questionnaires based on passive sensors

An important challenge with ESM is when to trigger the questionnaires (cf. event-contingent sampling) so that it is able to capture a wide range of events that are representative of the participant's daily life while also collecting enough of the event of interest (i.e., fleeting experiences or situations). Setting-up a trigger schedule that can fulfil those two requirements whilst also keeping participant burden to a minimum and compliance to a maximum is a delicate balancing act that can be substantially improved with passive sensing technology (Lathia et al., 2013).

One way where passive sensing can facilitate ESM is by triggering beeps in the moments of interest to the study, so a more event-based approach. In fact, certain research questions do not require data representative of every aspect of an individual's life, but more focused data on specific events as is the case for research on eating behavior for example. By using a smartwatch that can trigger a beep when it detects a meal episode it limits probing to the relevant moments only (Morshed et al., 2020). Passive sensing can also help increase compliance by triggering beeps only at opportune times such as when the participant is using the phone, after ending a phone-call or finishing a text (Fischer et al., 2011), or based on activity transitions by use of accelerometers (Pejovic & Musolesi, 2014), rather than in the middle of a sports activity or when driving, where answering a questionnaire would be difficult, and in some cases dangerous. Adjusting the trigger schedule to the participant's lifestyle via passive sensing has shown to not only significantly improve compliance rate, but more importantly effectively improve data quality by as much as 96% and reduce the time between trigger and survey completion (Gosh et al., 2019; Lathia et al., 2013).

In addition to triggering individual beeps, the whole day schedule could be adjusted based on passive sensor data, for instance, to match the participant's sleep-wake rhythm. Based on either smartphone use data or physiological data, it could be inferred when a participant is awake and thus could respond to the beep schedule. Going one step further, the same could be done with the triggering of an entire ESM period. Predetermined passive sensor data could here trigger the initiation of an ESM period, for instance when physiology or smartphone sensors detect a change in functioning (e.g., changes in stress levels, sleeping patterns, social media use, movement, etc.). Although these applications are still

speculative, they have the potential of becoming essential additions to standard ESM studies in the not-too-distant future.

### 13.4 Participant burden

An important aspect to consider before deciding to implement passive sensing in a study is the impact on participant burden. That is, whether or not adding passive sensing will increase or decrease participant burden and if it does, whether its use can be justified. As mentioned above, passive sensing can significantly reduce the burden on participants in two key ways: first, by replacing ESM items that can easily be monitored passively (i.e., without participant involvement) such as social interactions (Doryab et al., 2019). Second, passive sensing may be able to guide the trigger schedule so that questionnaires appear only at informative moments (e.g., Bertz et al., 2018; Halem et al., 2020). Nevertheless, there are instances where passive sensing can actually increase participant burden.

There are several technical issues that may increase perceived participant burden. Namely, the added battery-drain if a participant's device is used, and/or the burden of having to carry an additional device in the case of research phones or wearables (e.g., chest-patch, wristband, earpiece). The latter case can be especially burdensome as participants need to remember to carry this additional device and, in some cases, occasionally charge it. Some devices such as chest-patches or earpieces may also be cumbersome to use causing an added discomfort to the participant. As mentioned before, privacy issues are another important factor to consider. Depending on what kind of passive sensing is used, it can have an impact on the participants' privacy and may even compromise their anonymity. For instance, tracking GPS information may potentially reveal the participant's home address and consequently their identity. While there are ways to address and thus avoid such ethical concerns, participants may still feel that they are being monitored which can cause uneasiness. If as a result of this discomfort participants either avoid using their phone or adjust their behavior, then this may in turn compromise ecological validity. Finally, the burden on the researcher is also of note. Passive sensing requires an active involvement of the researcher to ensure devices and or software function properly, collect the appropriate data, and abide by GDPR regulations. Subsequent pre-processing and data analysis will also involve more time and energy. The use of passive sensing should therefore

always be carefully deliberated prior to initiating a study so that any challenges do not overshadow its benefits.

### 13.5 Conclusions

Passive sensing is a new and developing field providing a myriad of opportunities for ESM research. Technological advances will keep driving new developments in hardware, whereas improvement of computational models will determine the possibilities of application.

With this world of new possibilities also come new deliberations. First, we mentioned potential ethical issues related to the monitoring of sensitive or personally identifiable information. In some cases, data may be immediately processed such that the information that is left has lost identifiable features. In the case of GPS data, instead of using an individual's exact coordinates, the data can be processed into a measure of distance to a pre-specified location (e.g., home, work). It may also be tempting to collect data on as many measures as possible when data collection is as easy as some forms of passive monitoring. However, that alone is naturally not a justification. A thorough evaluation of the potential ethical issues of using passive sensing is, however, necessary to determine if its use is warranted. A second point is related to the participant's burden. Although we have advocated that, when used adequately, the addition of passive sensing to ESM may decrease this burden, very often it means that the participant has to perform more actions (e.g., recharge wearables, install apps) and experiences more discomfort (e.g., psychological discomfort from the invasion of their privacy, physical discomfort from the wearables). The researcher therefore must at all times consider the costs and benefits of adding passive sensing to their ESM protocol, not in the last place because an increase in perceived burden may have consequences for data quality and participant compliance. A final deliberation is of more theoretical nature: as stated in this chapter, with passive sensing one can never directly measure an individual's experience and therefore, when the aim is to measure experience, the measurement error will almost always be greater than in self-report measurements. There will be further developments in the field of predicting experiences based on passive sensing data, which will keep improving model performance. However, in the field of *experience* sampling, the most straightforward method of assessment will always remain self-report. For each



variable of interest, a thoughtful consideration between these two assessment methods will be necessary to determine the optimal approach. Passive sensing can be an invaluable addition in the study of daily life by substituting ESM items, enriching ESM data, or triggering ESM sampling schemes or individual beeps. As long as the issues that come with it are taken into consideration, it has the potential to forever change ESM research for the better.

# INDEX

	4E-approaches	1.2.3, p16
A		
	Accelerometer	6.3, p110; 8.1, p140; 13, p253; 13.1.2, p256; 13.2, p258; 13.3, p259
	Acceptability	5.4, p98; 12.2, p246
	Adverse event	5.3, p96
	Anonymity	4.5, p89; 5.2, p95; 5.3, p98; 13.4, p260
	Application software	6.5, p113
	Assessment frequency	3.2.2, p44
	Attrition	3.3.3.1, p59
	Autocorrelation	2.1.1, p27; 3.3.4.3, p66; 9.8, p179; 11.3.1, p228
	Autoregressive effects	10.3, p208
B		
	Backfilling	3.2.5, p51
	Best linear unbiased predictions	9.2, p160; 9.4, p167
	Between-person/trait level	4.3.4, p88
	Between-person variation / relationships	1.2.4, p17; 2.1, p24; 2.1.1, p25; 2.1.1, p26, 3.1.2, p37 4.3.2, p85; 9, p156; 9.2, p158; 9.6, p172; 10.1.1, p192; 10.2, p206; 11.1, p220
	Between-prompt interval	3.2.2, p45; 8.2, p141
	Branching questions	4.2.2, p82; 6.1.1, p106; 8.4, p145
	Briefing	5.3, p97; 7, p121
C		
	CE marking	6.4.1, p111; 12.4, p249
	Checkbox questions	6.1.1, p106
	Clinical application	1.2.4, p18
	Co-designing	5.3, p97
	Compatibility	5.1, p94
	Completion time	3.2.4, p49; 8.4, p146
	Compliance	3.2.2, p45; 3.2.4, p49; 3.3.3.1, p58; 4.2.2, p82; 5.5, p99; 6.1.4, p108; 7.1.1, p122; 7.1.5, p129; 8.4, p146; 11.5, p236; 13, p254; 13.3, p259
	Concurrent relationships	2.1.2, p28
	Confidentiality	5.2, p95
	Contextual science	1.2.3, p17
	Continuous-time models	9.7, p178

	Control questions	4.2.3, p83
	Count outcome	10, p187; 10.1.2, p197
	Cross-level interactions	9.5, p169
	Cross-regressive effects	10.3, p208
	Cross-sectional associations	4.3.3, p86; 9, p155
D		
	Database	6.1.5, p108
	Debriefing	7, p121
	Device-initiated	6.1.2, p107
	Dichotomous outcome	10, p187; 10.1.1, p188
	Diurnal patterns	3.3.4.2, p65; 9.9, p183
E		
	Eating disorders	3.2.3, p48; 3.3.1, p55; 3.3.4.2, p64; 4.1.2, p75; 5.3, p96
	Ecological momentary interventions	1.2.4, p18; 2.2, p32; chapter 12, p239
	Ecological psychology	1.2.1, p14
	Ecological validity	1.2.1, p15; 3.2.3, p46; 13.4, p260
	Electronic devices	3.2.5, p50; 6.4.1, p111; 8.4, p145; 12, p241
	Empty model	9.3, p162
	Enrollment	6.1.3, p107
	ESM Item Repository	4.4, p89
	Ethical aspects	1.1, p12; chapter 5, p91
	Ethical committees	5.5, p99
	Ethical objections	chapter 5, p91; 13.2, p258
	Event-contingent sampling	1.1, p12; 3.2.3, p46; 6.1.2, p106; 8.2, p140; 13.3, p259
	Experimental and intervention paradigms	2, p23
F		
	Feasibility	3.3, p54; 5.4, p98; 7.2, p132; 11, p220; 11.5, p236; 12.2, p244; 12.4, p248
	Fixed sampling	3.2.3, p46; 6.1.2, p106; 7.1.2, p123
	Fixed start	6.1.3, p107
	Flexible start	6.1.3, p108
	Formulation	4.1.3, p77
G		
	Gamma distribution	10.1.3, p200
	GDPR	5, p93; 6.4.1, p111; 13.4, p260
	Generalized mixed effects model	10, p187; 10.1, p187
	GPS	4.5, p89; 6.1.2, p107; 8.1, p140; 13.1.1, p255; 13.2, p258; 13.4, p260
H		
	Hardware	6, p105; 6.5, p113

	Hardware- and software-based fragmentation	6.2.2, p109
	Hierarchical data	1.2.4, p18; 11.1, p220
I		
	Iatrogenic effects	5.5, p100
	Inclusivity	5.1, p93
	Inertia	9.8, p180
	Informed consent	5, p93; 5.2, p95
	Initial elevation bias	2.1.1, p26
	Intercept-slope correlation	9.4, p165
	Interference	3.3.3.2, p60
	Internal consistency	4.3.3, p86
	Interventions	2.2, p32; 5.3, p96; chapter 12, p239
	Intra-class correlation coefficient	2.1.1, p25; 4.1.1, p74; 4.3.2, p85; 9.2, p161
	Intra-class correlation coefficient for generalized mixed-effects models	10.1.1, p191
	Inverse Gaussian	10.1.3, p200
L		
	Lagged variable	8.4, p150; 9.7, p176; 10.3, p209
	Length of questionnaires	4.2.2, p82
	Likelihood ratio test	10.1.1, p194
	Likert Scale	4.1.4, p80; 4.2.1, p82; 6.1.1, p106; 7.1.3, p126; 10.1.1, p188; 10.1.2, p198; 10.1.3, p201; 10.2, p204
	Linear mixed-effects model	10, p187
	Logistic mixed-effects regression model	10.1.1, p188
	Long format	8.1, p139
M		
	Measurement-burst design	3.1.3, p38; 8.1, p141; 8.4, p143
	Measurement quality	4.3, p84
	Missing data	3.3.3.1, p58; 8.4, p143; 9.1, p156
	Mixed-effects models	9.1, p156; chapter 10, p185
	Monitoring of ESM responses	5.3, p97
	Multilevel data	9.1, p156
	Multilevel model	4.3.3, p86; 9.1, p156; 10.1.1, p192; 11.1, p220; 11.2, p221; 11.3.1, p224
	Multilevel vector autoregressive model of order one (VAR(1))	10, p187; 10.3, p208
N		
	Non-natural variations	2.2, p31
	Non-normal errors	10.1, p187

	Notification	6.1.2, p106; 6.2.2, p109; 7.1.2, p123; 7.1.3, p125; 7.1.4, p127; 7.1.5, p129; 7.2, p131; 12.1, p242
O	Observational ESM research	2.1, p23
	Online dashboard	6.1, p105
	Online interfaces	3.2.5, p51
	Open data	5.3, p95
	Open Science Framework	5.2, p96
P	Paper-and-pencil method	3.2.5, p50
	Participant burden	3.2.3, p47; 5, p93; 5.4, p98; 13.3, p259; 13.4, p260
	Passive data	4.5, p89; 5, p93; 6.1.2, p107; 6.3, p110; 8.1, p140; chapter 13, p251
	Personalized approaches	1.2.4, p17; 12.3, p247
	Personalized data	1.2.4, p18
	Pilot testing	3.3.6, p69; 4.3.1, p84; 12.2, p246
	Poisson mixed-effects regression model	10.1.2, p197
	Platform	Chapter 6, p103; 12.3, p247
	Pop-up window	5.3, p97
	Power analysis	11.1, p220
	<i>PowerAnalysis</i> IL shiny app	11.3.1, p224
	Predictive power	2.1.1, p25
	Privacy	5, p93; 6.4.1, p111; 12.4, p248; 13.4, p260
	Privileged observer	1.2.1, p14
Q	Quantified self	1.2.2, p15
	Questionnaire density	3.2.4, p49
	Questionnaire development	1.1, p11; chapter 4, p71
R	Radio questions	6.1.1, p106
	Random effects	9.1, p156; 9.4, p165; 9.5, p170; 9.8, p180; 10.1.1, p188; 10.1.2, p198; 10.1.3, p201; 10.2, p205
	Random sampling	3.2.3, p46; 6.1.2, p106; 7.1.2, p123
	Reactivity	1.1, p12; 3.2.3, p46; 3.3.1, p55; 3.3.3.2, p59; 3.3.6, p69; 4.2.3, p83; 5, p93; 5.5, p99
	Reliability	4.3.3, p85; 13.2, p258
	Remote ESM data collection	5.1, p94
	Research-dedicated devices	3.2.5, p51
	Research questions	2, p23
	Response delay	3.2.4, p49; 8.4, p145
	Response frequencies	8.4, p146

	Response scale	4.1.4, p80; 6.1.1, p105
	Retrospective bias	1.1, p11; 4.1.2, p76
S		
	Safety monitoring protocols	5.3, p97
	Sample size planning	11, p219
	Sampling scheme	3.2.3, p45
	Schedule template	6.1.2, p106
	Self-harm	5.3, p96
	Self-initiated	6.1.2, p107
	Self-predictability	2.1.1, p27
	Self-report diary techniques	1.1, p10
	Self-tracking	1.2.2, p15
	Semi-random sampling	3.2.3, p46; 6.1.2, p106; 8.2, p141
	Shrinkage	9.2, p160
	Signal-contingent	6.1.2, p106
	Simulation-based approach to conduct power analysis	11.4, p236
	Single-case studies	1.2.4, p18
	Skins	6.2.2, p109
	Slider questions	6.1.1, p105
	Software	6, p105
	Start- and end-point of the day	3.2.2, p45
	State-like features	4.1.1, p74
	Statistical power	11, p219
	Structured diary method	1.1, p12
	Study device	3.2.5, p50
	Study duration	3.2.1, p43
	Subjective experience	1.1, p11; 1.2.2, p16; 13.1.2, p257
	Subject-level means	8.4, p148
	Subject-specific average	9.2, p157
	Substance use	3.3.3.2, p60; 5.3, p96; 12.3, p247; 13.2, p257
	Suicidal behavior	5.3, p96
	System software	6.5, p113
T		
	Three-level model	9.1, p156; 10.2, p203
	Time-dependency	2.1.1, p27
	Timeframes	4.1.2, p74
	Time-lagged relationships	2.1.2, p28
	Time trends	9.9, p181
	Time-varying variables	2.1, p24; 2.1.1, p24; 2.1.2, p28; 8.1, p140; 10.1.1, p190; 10.1.2, p199; 10.1.3, p202; 11.1, p220; 11.2, p222
	Two-level model	9.1, p156; 9.2, p157; 9.3, p162; 11.3.1, p224

U		
	Usability	3.2.5, p52; 5.1, p94; 12.2, p244; 12.4, p248
V		
	Validity	2.1.1, p25; 3.2.5, p52; 3.3.3, p57; 4, p73; 4.3.4, p87
	Variance accounted for ( $R^2$ )	9.3, p164; 9.4, p169; 9.5, p171
	Vector-autoregressive model	2.1.2, p29
	Visual Analog Scale	4.1.4, p80; 7.1.3, p126
W		
	Wearable technology	1.2.2, p16; 6.1.2, p107; 6.3, p110; 12, p241; 13, p253; 13.1.2, p256
	Wide format	8.1, p139
	Within-person dynamics	10.3, p208
	Within-person mean centering	8.4, p148; 9.6, p172
	Within-person/state level	4.3.4, p88
	Within-person variation / relationships	1.2.4, p17; 2.1, p; 2.1.1, p25; 2.1.1, p26; 2.1.1, p27; 4.1.1, p74; 4.1.3, p77; 4.3.2, p85; 4.3.3, p86; 9, p155; 9.2, p158; 9.4, p164; 9.6, p172; 11.1, p220

# REFERENCES

- 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.
- Arean, P. A., Hoa Ly, K., & Andersson, G. (2016). Mobile technology for mental health assessment. *Dialogues Clin Neurosci*, 18(2), 163-169. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/27489456>
- Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychol Methods*, 24(1), 1-19. doi:10.1037/met0000195
- 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.
- Bachmann, A. (2015). Towards smartphone-based sensing of social interaction for ambulatory assessment. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (pp. 423-428).
- 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. *J Clin Child Adolesc Psychol*, 1-13. doi: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. doi:10.1371/journal.pone.0165762
- Bakker, D., Kazantzis, N., Rickwood, D., & Rickard, N. (2018). A randomized controlled trial of three smartphone apps for enhancing public mental health. *Behav Res Ther*, 109, 75-83. doi:10.1016/j.brat.2018.08.003
- Barker, R. G. (1975). *Ecological Psychology: Concepts and Methods for Studying the Environment of Human Behavior*. Stanford, California: Stanford University Press.
- Barta, W. 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. doi:10.2196/19852
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.
- Belisario, J. S. M., Doherty, K., O'Donoghue, J., Ramchandani, P., Majeed, A., Doherty, G., . . . 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.
- 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.
- Bertz, J. W., Epstein, D. H., & Preston, K. L. (2018). Combining ecological momentary assessment with objective, ambulatory measures of behavior and physiology in substance-use research. *Addict Behav*, 83, 5-17. doi:10.1016/j.addbeh.2017.11.027
- Biagiante, B., Hidalgo-Mazzei, D., & Meyer, N. (2017). Developing digital interventions for people living with serious mental illness: perspectives from three mHealth studies. *Evid Based Ment Health*, 20(4), 98-101. doi:10.1136/eb-2017-102765
- Blozis, S. A., McTernan, M., Harring, J. R., & Zheng, Q. (2020). Two-part mixed-effects location scale models. *Behav Res Methods*, 52(5), 1836-1847. doi:10.3758/s13428-020-01359-7
- Boker, S. M., Molenaar, P. C., & Nesselroade, J. R. (2009). Issues in intraindividual variability: individual differences in equilibria and dynamics over multiple time scales. *Psychol Aging*, 24(4), 858-862. doi: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). New York: Guilford.
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annu Rev Psychol*, 54, 579-616. doi: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)*. New York, NY: Guilford Press.
- Boon, B., Stroebe, W., Schut, H., & Ijntema, R. (2002). Ironic processes in the eating behaviour of restrained eaters. *Br J Health Psychol*, 7(Pt 1), 1-10. doi: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. *J Interpers Violence*, 886260518812068. doi:10.1177/0886260518812068

- Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol*, 9, 91-121. doi: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. *Psychiatr Serv*, 70(11), 983-991. doi:10.1176/appi.ps.201900050
- Bouisson, J., & Swendsen, J. (2003). Routinization and emotional well-being: an experience sampling investigation in an elderly French sample. *J Gerontol B Psychol Sci Soc Sci*, 58(5), P280-282. doi:10.1093/geronb/58.5.p280
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71, 791-799. doi: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. *Front Psychol*, 6, 272. doi: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. doi:10.1037/a0032400
- Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., . . . Snippe, E. (2019). What do centrality measures measure in psychological networks? *J Abnorm Psychol*, 128(8), 892-903. doi:10.1037/abn0000446
- Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., . . . Kuppens, P. (2016). Assessing Temporal Emotion Dynamics Using Networks. *Assessment*, 23(4), 425-435. doi:10.1177/1073191116645909
- Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., . . . Tuerlinckx, F. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PLoS One*, 8(4), e60188. doi:10.1371/journal.pone.0060188
- Broderick, J. E., & Vikingstad, G. (2008). Frequent assessment of negative symptoms does not induce depressed mood. *J Clin Psychol Med Settings*, 15(4), 296-300. doi:10.1007/s10880-008-9127-6
- Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., . . . Bolker, B. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2), 378-400.
- 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*. Bristol, UK: University of Bristol.

- Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR(1) based models do not always outpredict AR(1) models in typical psychological applications. *Psychol Methods*, 23(4), 740-756. doi: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. *Behav Res Ther*, 137, 103775. doi:10.1016/j.brat.2020.103775
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*, 14(5), 365-376. doi: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. *Addict Behav*, 102, 106198. doi:10.1016/j.addbeh.2019.106198
- Capon, H., Hall, W., Fry, C., & Carter, A. (2016). Realising the technological promise of smartphones in addiction research and treatment: An ethical review. *Int J Drug Policy*, 36, 47-57. doi:10.1016/j.drugpo.2016.05.013
- Carter, B. L., Day, S. X., Cinciripini, P. M., & Wetter, D. W. (2007). Momentary health interventions: Where are we and where are we going. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Eds.), *Science of Real-Time Data Capture Self-Reports in Health Research* (pp. 289-307). New York: Oxford University Press.
- Chang, W., Cheng, J., Allaire, J. J., Xie, Y., McPherson, J., & and others. (2019). Shiny: Web Application Framework for R. Retrieved from <https://CRAN.R-project.org/package=shiny>
- Charles, C., Gafni, A., & Whelan, T. (1999). Decision-making in the physician-patient encounter: revisiting the shared treatment decision-making model. *Soc Sci Med*, 49(5), 651-661. doi:10.1016/s0277-9536(99)00145-8
- 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. doi:10.1037/emo0000232
- Clark, L. A., Watson, D., & Leeka, J. (1989). Diurnal variation in the positive affects. *Motivation and Emotion*, 13(3), 205-234.
- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*: Academic press.
- Collip, D., van Winkel, R., Peerbooms, O., Lataster, T., Thewissen, V., Lardinois, M., . . . Myin-Germeys, I. (2011). COMT Val158Met-stress interaction in psychosis: role of background psychosis risk. *CNS Neurosci Ther*, 17(6), 612-619. doi:10.1111/j.1755-5949.2010.00213.x
- Colombo, D., Fernandez-Alvarez, J., Garcia Palacios, A., Cipresso, P., Botella, C., & Riva, G. (2019). New Technologies for the Understanding,

- Assessment, and Intervention of Emotion Regulation. *Front Psychol*, 10, 1261. doi:10.3389/fpsyg.2019.01261
- Colombo, D., Suso-Ribera, C., Fernández-Alvarez, 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. doi:https://doi.org/10.1007/s10608-020-10122-3
- Conner, T., & Reid, K. (2012). Effects of intensive mobile happiness reporting in daily life. *Social Psychological and Personality Science*, 3, 315-323.
- Conner, T. S., & Mehl, M. R. (2015). Ambulatory Assessment – Methods for studying everyday life. In R. Scott, S. Kosslyn, & N. Pinkerton (Eds.), *Emerging Trends in the Social and Behavioral Sciences*. Hoboken, NJ: Wiley.
- Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: a program for designing efficient multilevel studies. *Behav Res Methods*, 40(1), 236-249. doi:10.3758/brm.40.1.236
- Coppersmith, D. D., Fortgang, R., Kleiman, E., Millner, A., Yeager, A., Mair, P., & Nock, M. (2021). Effect of frequent assessment of suicidal thinking on its incidence and severity: High-resolution real-time monitoring study. *The British Journal of Psychiatry*, 1-3. doi: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. *Dev Neurorehabil*, 19(2), 103-110. doi:10.3109/17518423.2014.915244
- Couper, M., Tourangeau, R., Conrad, F., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24(2), 227-245.
- 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. doi:10.1016/0304-3959(96)03125-9
- Csikszentmihalyi, M., & Larson, R. (1984). *Being Adolescent: Conflict and Growth in the Teenage Years*. s.l.: BasicBooks.
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the Experience-Sampling Method. *J Nerv Ment Dis*, 175(9), 526-536. doi:10.1097/00005053-198709000-00004
- Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (Vol. 621-694): American Council on Education.
- Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu Rev Psychol*, 62, 583-619. doi:10.1146/annurev.psych.093008.100356
- de Bruin, L., Newen, A., & Gallagher, S. (2018). *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.
- de Haan, S. (2020). *Enactive Psychiatry*. New York, NY: Cambridge University Press.

- 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. *Front Psychol*, 7, 891. doi: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.
- 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. doi: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. *Depress Anxiety*, 34(9), 836-844. doi:10.1002/da.22653
- 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. doi: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*: Springer Nature.
- Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., . . . Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *J Pers Soc Psychol*, 114(2), 323-341. doi: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. *Nat Hum Behav*, 3(5), 478-491. doi:10.1038/s41562-019-0555-0
- Delespaul, P. (1995). *Assessing schizophrenia in daily life: the Experience Sampling Method*. Maastricht: Universitaire Pers Maastricht.
- Demidenko, E. (2004). *Mixed models: Theory and applications*. Hoboken, NJ: Wiley.
- Demidenko, E. (2013). *Mixed models: theory and applications with R*: John Wiley & Sons.
- 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. doi: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.

- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *J Pers Assess*, 49(1), 71-75.  
doi:10.1207/s15327752jpa4901\_13
- Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*: CRC press.
- Doryab, A., Villalba, D. K., Chikersal, P., Dutcher, J. M., Tumminia, M., Liu, X., . . . Dey, A. K. (2019). Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data. *JMIR Mhealth Uhealth*, 7(7), e13209. doi:10.2196/13209
- 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. *Psychol Med*, 48(2), 208-228. doi:10.1017/S0033291717001659
- Economou, M., Souliotis, K., Peppou, L. E., & Dimopoulos, Y. (2019). Shared decision-making in mental health care: have we overlooked the collective level? *Eur Arch Psychiatry Clin Neurosci*, 269(4), 481-482. doi:10.1007/s00406-018-0954-7
- 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*, 1073191120957102. doi: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*. doi:10.3389/fpsyg.2021.580684
- Elsschot, W. (1934). *Het humelijk*.
- Engelen, L., Bundy, A. C., Lau, J., Naughton, G., Wyver, S., Bauman, A., & Baur, L. (2015). Understanding Patterns of Young Children's Physical Activity After School--It's all About Context: A Cross-Sectional Study. *J Phys Act Health*, 12(3), 335-339. doi:10.1123/jpah.2013-0153
- Epskamp, S. (2020). graphicalVAR: Graphical VAR for Experience Sampling Data. R package version 0.2.4. Retrieved from <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. Retrieved from <https://CRAN.R-project.org/package=mlVAR>
- EUR-Lex. Regulation (EU). (2016). 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. Retrieved from

- <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>
- EUR-Lex. Regulation (EU). (2017). 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. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745>
- Exposito, M., Hernandez, J., & Picard, R. W. (2018). Affective keys: towards unobtrusive stress sensing of smartphone users. In *proceedings of the 20th international conference on human-computer interaction with Mobile devices and services adjunct* (pp. 139-145).
- Farewell, V. T., Long, D. L., Tom, B. D. M., Yiu, S., & Su, L. (2017). Two-Part and Related Regression Models for Longitudinal Data. *Annu Rev Stat Appl*, 4, 283-315. doi:10.1146/annurev-statistics-060116-054131
- Feddersen, J., Metcalfe, R., & Wooden, M. (2016). Subjective wellbeing: why weather matters. *J.R. Stat. Soc.*, 179, 203-228.
- Federal Food, Drug, and Cosmetic Act.* (s.201(h)).
- Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry*, 16(3), 287-298. doi:10.1002/wps.20472
- 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* (pp. 181).
- Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proc Natl Acad Sci U S A*, 115(27), E6106-E6115. doi:10.1073/pnas.1711978115
- Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, 20, 303-315. doi:<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. *J Pers Soc Psychol*, 109(6), 1090-1104. doi:10.1037/a0039517
- 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 assessment of suicidal ideation and its proximal risk factors using ecological momentary assessments. *J Abnorm Psychol*, 127(8), 758-769. doi:10.1037/abn0000381

- Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., . . . Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychol Bull*, 143(2), 187-232. doi: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.
- Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *J Pers Soc Psychol*, 65(1), 45-55. doi:10.1037//0022-3514.65.1.45
- Frijters, P., & Beaton, T. (2012). The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization*, 82(2-3), 525-542.
- Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomin, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychol Assess*, 29(9), 1120-1128. doi: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. doi:10.1016/j.bodyim.2013.06.003
- 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.
- Galecki, A., & Burzykowski, T. (2003). *Linear mixed-effects models using R*. New York: Springer.
- Gee, B. L., Griffiths, K. L., & Gulliver, A. (2016). Effectiveness of mobile technologies delivering ecological momentary interventions for stress and anxiety: A systematic review. *Journal of the American Medical Informatics Association*, 23(1), 221-229. doi:doi:10.1093/jamia/ocv043
- Gibbons, R. D., Segawa, E., Karabatsos, G., Amatya, A. K., Bhaumik, D. K., Brown, C. H., . . . Mann, J. J. (2008). Mixed-effects Poisson regression analysis of adverse event reports: the relationship between antidepressants and suicide. *Stat Med*, 27(11), 1814-1833. doi:10.1002/sim.3241
- Gibson, J. J. (2015). *The Ecological Approach to Visual Perception*. New York: London: Psychology Press.
- Glasgow, T. E., Le, H. T., Geller, E. S., 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).
- 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. *J Clin Child Adolesc Psychol*, 1-17. doi:10.1080/15374416.2020.1741377
- Goldstein, H. (2011). *Multilevel statistical models*. Chichester, UK: John Wiley & Sons.
- Goldstein, H., Healy, M. J., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Stat Med*, 13(16), 1643-1655. doi:10.1002/sim.4780131605
- Goodwin, J., Cummins, J., Behan, L., & O'Brien, S. M. (2016). Development of a mental health smartphone app: perspectives of mental health service users. *J Ment Health*, 25(5), 434-440. doi:10.3109/09638237.2015.1124392
- Gosh, S., Ganguly, N., Mitra, B., & De, P. (2019). Designing an experience sampling method for smartphone based emotion detection. *IEEE Transactions on Affective Computing*. doi:10.1109/TAFFC.2019.2905561
- 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.
- Green, A. S., Rafaeli, E., Bolger, N., Shrout, P. E., & Reis, H. T. (2006). Paper or plastic? Data equivalence in paper and electronic diaries. *Psychol Methods*, 11(1), 87-105. doi: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.
- 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.
- Gries, K., Berry, P., Harrington, M., Crescioni, M., Patel, M., Rudell, K., . . . Vernon, M. (2017). Literature review to assemble the evidence for response scales used in patient-reported outcome measures. *J Patient Rep Outcomes*, 2, 41. doi:10.1186/s41687-018-0056-3
- Groot, P. C. (2010). Patients can diagnose too: How continuous self-assessment aids diagnosis of, and recovery from, depression. *J Ment Health*, 19(4), 352-362. doi:10.3109/09638237.2010.494188
- Halem, S., van Roekel, E., Kroencke, L., Kuper, N., & Denissen, J. (2020). Moments That Matter? On the Complexity of Using Triggers Based on Skin Conductance to Sample Arousing Events Within an Experience Sampling Framework. *European Journal of Personality*, 34(5), 794-807.
- Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4), 1030-1039. doi:10.1111/j.0006-341x.2000.01030.x

- 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.
- 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. *IEEE* (pp. 83-92).
- Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358), 320-338.
- Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. New York: Wiley.
- 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 Behavioral Statistics* 24(1), 70-93.
- 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.
- Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., . . . Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry*, 19(1), 59. doi: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.
- 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.
- Hermans, K., Achterhof, R., Myin-Germeys, I., Kasanova, Z., Kirtley, O., & Schneider, M. (2019). Improving Ecological Validity in Research on Social Cognition. In *Social Cognition in Psychosis* (pp. 249-268): Academic Press.
- Hermans, K., Myin-Germeys, I., Gayer-Anderson, C., Kempton, M. J., Valmaggia, L., McGuire, P., . . . Reininghaus, U. (2020). Elucidating negative symptoms in the daily life of individuals in the early stages of psychosis. *Psychol Med*, 1-11. doi:10.1017/S0033291720001154
- Heron, K. E., & Smyth, J. M. (2010). Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *Br J Health Psychol*, 15(Pt 1), 1-39. doi: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.  
doi: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. *Arch Sex Behav*, 23(4), 453-463.  
doi: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 behavior and affect in social situations. *Psychol Assess*, 31(7), 952-960.  
doi:10.1037/pas0000718
- Hoemann, K., Khan, Z., Feldman, M. J., Nielson, C., Devlin, M., Dy, J., . . . Quigley, K. S. (2020). Context-aware experience sampling reveals the scale of variation in affective experience. *Sci Rep*, 10(1), 12459.  
doi:10.1038/s41598-020-69180-y
- 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.
- Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, 345(6202), 1340-1343.  
doi:10.1126/science.1251560
- Hollis, C., Morriss, R., Martin, J., Amani, S., Cotton, R., Denis, M., & Lewis, S. (2015). Technological innovations in mental healthcare: harnessing the digital revolution. *Br J Psychiatry*, 206(4), 263-265.  
doi:10.1192/bjp.bp.113.142612
- Hormuth, S. E. (1986). The sampling of experiences in situ. *Journal of Personality*, 54, 262-293. doi: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. *J Abnorm Psychol*, 126(1), 89-95.  
doi: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. doi:doi: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*.  
doi: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. *Psychol Bull*, 141(4), 901-930. doi:10.1037/a0038822

- Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., & Kumar, S. (2015). cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *Proceedings of the ACM International Conference on Ubiquitous Computing UbiComp (Conference)* (pp. 493-504).
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3 ed.). New York: Routledge.
- Huckvale, K., Torous, J., & Larsen, M. E. (2019). Assessment of the Data Sharing and Privacy Practices of Smartphone Apps for Depression and Smoking Cessation. *JAMA Netw Open*, 2(4), e192542. doi:10.1001/jamanetworkopen.2019.2542
- Hurlburt, R. T. (1993). *Sampling Inner Experience in Disturbed Affect*. Springer US.
- 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 Res*, 220(1-2), 564-570. doi: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.
- Impett, E. A., Strachman, A., Finkel, E. J., & Gable, S. L. (2008). Maintaining sexual desire in intimate relationships: the importance of approach goals. *J Pers Soc Psychol*, 94(5), 808-823. doi: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.
- 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 Sci*, 76(21), 2431-2443. doi:10.1016/j.lfs.2004.10.045
- Jacobson, N. C., Bentley, K. H., Walton, A., Wang, S. B., Fortgang, R. G., Millner, A. J., . . . Coppersmith, D. D. L. (2020). Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bull World Health Organ*, 98(4), 270-276. doi: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. *Psychol Methods*, 13(4), 354-375. doi:10.1037/a0014173
- 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 Med Res Methodol*, 18(1), 140. doi:10.1186/s12874-018-0579-6
- Ji, L., Chow, S. M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling Missing Data in the Modeling of Intensive

- Longitudinal Data. *Struct Equ Modeling*, 25(5), 715-736.  
doi:10.1080/10705511.2017.1417046
- 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 Behav Res*, 50(3), 334-349.  
doi: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.  
doi: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. *Psychol Sci*, 30(6), 863-879.  
doi:10.1177/0956797619838763
- Kalokerinos, E. K., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., . . . Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proc Natl Acad Sci U S A*, 117(17), 9270-9276.  
doi: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. *Psychol Sci*, 18(7), 614-621. doi:10.1111/j.1467-9280.2007.01948.x
- Karthick, S., & Binu, S. (2017). Android security issues and solutions. In *International Conference on Innovative Mechanisms for Industry Applications (ICIMLA)* (pp. 686-689).
- 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. *J Abnorm Psychol*, 129(1), 122-130. doi:10.1037/abn0000453
- Kasanova, Z., Oorschot, M., & Myin-Germeys, I. (2018). Social anhedonia and asociality in psychosis revisited. An experience sampling study. *Psychiatry Res*, 270, 375-381. doi:10.1016/j.psychres.2018.09.057
- 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. *Eur Arch Psychiatry Clin Neurosci*, 270(2), 139-152. doi:10.1007/s00406-018-0974-3
- Kidd, C., & Hayden, B. Y. (2015). The Psychology and Neuroscience of Curiosity. *Neuron*, 88(3), 449-460. doi:10.1016/j.neuron.2015.09.010
- Kiekens, G., Hasking, P., Nock, M. K., Boyes, M., Kirtley, O., Bruffaerts, R., . . . Claes, L. (2020). Fluctuations in Affective States and Self-Efficacy to Resist Non-Suicidal Self-Injury as Real-Time Predictors of Non-Suicidal

- Self-Injurious Thoughts and Behaviors. *Front Psychiatry*, 11, 214. doi:10.3389/fpsy.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*. doi:10.31234/osf.io/yp86x
- Killikelly, C., He, Z., Reeder, C., & Wykes, T. (2017). Improving Adherence to Web-Based and Mobile Technologies for People With Psychosis: Systematic Review of New Potential Predictors of Adherence. *JMIR Mhealth Uhealth*, 5(7), e94. doi:10.2196/mhealth.7088
- Kimhy, D., Myin-Germeys, I., Palmier-Claus, J., & Swendsen, J. (2012). Mobile assessment guide for research in schizophrenia and severe mental disorders. *Schizophr Bull*, 38(3), 386-395. doi:10.1093/schbul/sbr186
- Kirtley, O., Achterhof, R., Hagemann, N., Hermans, K. S. F. M., Hiekkaranta, A. P., Lecei, A., & 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*. doi:10.31234/osf.io/jp2fk
- Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., & Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34S, 1220-1228. doi:10.1037/hea0000305
- 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. *J Abnorm Psychol*, 126(6), 726-738. doi: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. *Depress Anxiety*, 35(7), 601-608. doi:10.1002/da.22730
- Klippel, A., Viechtbauer, W., Reininghaus, U., Wigman, J., van Borkulo, C., Merge, . . . Wichers, M. (2018). The Cascade of Stress: A Network Approach to Explore Differential Dynamics in Populations Varying in Risk for Psychosis. *Schizophr Bull*, 44(2), 328-337. doi:10.1093/schbul/sbx037
- Kööts, L., Realo, A., & Allik, J. (2011). The influence of the weather on affective experience. *Journal of individual differences*, 32(2), 74-84.
- Koval, P., Kalokerinos, E. K., Greenaway, K., Medland, H., Kuppens, P., Nezlek, J. B., . . . Gross, J. J. (submitted). Emotion Regulation in Everyday Life: What Can We Learn from Global Self-Reports?
- 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. doi:10.1037/a0024756

- 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. doi: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. *Front Psychol*, 6, 1997. doi:10.3389/fpsyg.2015.01997
- Kramer, I., Simons, C. J., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., . . . 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. doi: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. *J Pain*, 18(3), 319-331. doi:10.1016/j.jpain.2016.11.006
- Kristensen, N., Nymann, C., & Konradsen, H. (2016). Implementing research results in clinical practice- the experiences of healthcare professionals. *BMC Health Serv Res*, 16, 48. doi:10.1186/s12913-016-1292-y
- Kuhlmann, T., Dantlgraber, M., & Reips, U. D. (2017). Investigating measurement equivalence of visual analogue scales and Likert-type scales in Internet-based personality questionnaires. *Behav Res Methods*, 49(6), 2173-2181. doi:10.3758/s13428-016-0850-x
- Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychol Sci*, 21(7), 984-991. doi:10.1177/0956797610372634
- Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: accounting for individual differences in the temporal dynamics of affect. *J Pers Soc Psychol*, 99(6), 1042-1060. doi:10.1037/a0020962
- Kuppens, P., & Verduyn, P. (2017). Emotion dynamics. *Curr Opin Psychol*, 17, 22-26. doi:10.1016/j.copsyc.2017.06.004
- 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, Advanced online publication*. doi:10.31234/osf.io/dq6ky
- Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963-974. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/7168798>
- Lamont, A. (2008). Young children's musical worlds: Musical engagement in 3.5-year-olds. *Journal of Early Childhood Research*, 6, 247-261. doi:<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. *Stat Methods Med Res*, 22(3), 324-345. doi:10.1177/0962280212439578
- 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.
- Larsen, M. E., Nicholas, J., & Christensen, H. (2016). A Systematic Assessment of Smartphone Tools for Suicide Prevention. *PLoS One*, 11(4), e0152285. doi:10.1371/journal.pone.0152285
- Lathia, N., Rachuri, K. K., Mascolo, C., & Rentfrow, P. J. (2013). Contextual dissonance: Design bias in sensor-based experience sampling methods. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 183-192).
- 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. *Psychol Assess*, 27(4), 1171-1181. doi: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. *Behav Ther*, 38(2), 132-143. doi:10.1016/j.beth.2006.06.004
- Levine, L. J., & Safer, M. A. (2002). Sources of Bias in Memory for Emotions. *Current Directions in Psychological Science*, 11(5), 169-173.
- 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.
- Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Front Psychol*, 6, 1171. doi:10.3389/fpsyg.2015.01171
- Lobo, L., Heras-Escribano, M., & Travieso, D. (2018). The History and Philosophy of Ecological Psychology. *Front Psychol*, 9, 2228. doi:10.3389/fpsyg.2018.02228
- Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116, 75.
- Long, J. A. (2020). jtools: Analysis and Presentation of Social Scientific Data. R package version 2.1.0. Retrieved from <https://cran.r-project.org/package=jtools>
- Lüdtke, D. (2021). sjstats: Statistical Functions for Regression Models (Version 0.18.1). Retrieved from <https://CRAN.R-project.org/package=sjstats>
- 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.



- Lukacz, E. S., Lawrence, J. M., Burchette, R. J., Lubert, K. M., Nager, C. W., & Buckwalter, J. G. (2004). The use of Visual Analog Scale in urogynecologic research: a psychometric evaluation. *Am J Obstet Gynecol*, 191(1), 165-170. doi:10.1016/j.ajog.2004.04.047
- Luxton, D. D., McCann, R. A., Bush, N. E., Mishkind, M. C., & Reger, G. M. (2011). mHealth for mental health: Integrating smartphone technology in behavioral healthcare. *Professional Psychology: Research and Practice*, 42(6), 505-512. doi:10.1037/a0024485
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 1(3), 86-92.
- Mackerron, G., & Mourato, S. (2013). Happiness is greater in natural environments. *Global environmental change*, 23(5), 992-1000.
- 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.
- 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 Behavior. *Front Psychol*, 9, 1485. doi:10.3389/fpsyg.2018.01485
- Matejka, J., Glueck, M., Grossman, T., & Fitzmaurice, G. (2016). *The effect of visual appearance on the performance of continuous sliders and visual analogue scales*. Paper presented at the 2016 CHI Conference on Human Factors in Computing Systems.
- 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. *J Appl Psychol*, 97(5), 951-966. doi:10.1037/a0028380
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annu Rev Psychol*, 59, 537-563. doi:10.1146/annurev.psych.59.103006.093735
- May, M., Junghaenel, D. U., Ono, M., Stone, A. A., & Schneider, S. (2018). Ecological Momentary Assessment Methodology in Chronic Pain Research: A Systematic Review. *J Pain*, 19(7), 699-716. doi:10.1016/j.jpain.2018.01.006
- 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.
- McDevitt-Murphy, M. E., Luciano, M. T., & Zakarian, R. J. (2018). Use of Ecological Momentary Assessment and Intervention in Treatment With

- Adults. *Focus (Am Psychiatr Publ)*, 16(4), 370-375.  
doi: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.  
doi:https://doi.org/10.1177/1948550617708015
- Meers, K., Dejonckheere, E., Kalokerinos, E. K., Rummens, K., & Kuppens, P. (2020). mobileQ: A free user-friendly application for collecting experience sampling data. *Behav Res Methods*, 52(4), 1510-1515.  
doi:10.3758/s13428-019-01330-1
- Mehl, M. R. (2017). The Electronically Activated Recorder (EAR): A Method for the Naturalistic Observation of Daily Social Behavior. *Curr Dir Psychol Sci*, 26(2), 184-190. doi:10.1177/0963721416680611
- Mehl, M. R., Eid, M., Wrzus, C., Harari, G. M., & Ebner-Priemer, U. (2021). *Mobile Sensing in Psychology: Methods and Applications*: Guilford Press.
- Messner, C., & Wanke, M. (2011). Good weather for Schwarz and Clore. *Emotion*, 11(2), 436-437. doi:10.1037/a0022821
- Mestdagh, M., & Dejonckheere, E. (2021). Ambulatory assessment in psychopathology research: Current achievements and future ambitions. *Curr Opin Psychol*, 41, 1-8. doi:10.1016/j.copsyc.2021.01.004
- Miralles, I., Granell, C., Diaz-Sanahuja, L., Van Woensel, W., Breton-Lopez, J., Mira, A., . . . Casteleyn, S. (2020). Smartphone Apps for the Treatment of Mental Disorders: Systematic Review. *JMIR Mhealth Uhealth*, 8(4), e14897. doi: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 Behavioral and Social Sciences* 7(1), 11-24.
- 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.
- Moerbeek, M., & van Breukelen, G. J. P. (2000). Design Issues for Experiments in Multilevel Populations. *Journal of Educational and Behavioral Statistics* 25(3), 271-284.
- 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.
- Mohr, D. C., Shilton, K., & Hotopf, M. (2020). Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *NPJ Digit Med*, 3, 45. doi:10.1038/s41746-020-0251-5
- Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, 2(4), 201-218.

- Molenberghs, G., & Verbeke, G. (2006). *Models for discrete longitudinal data*: Springer Science & Business Media.
- Monk, T. H., Houck, P. R., & Shear, M. K. (2006). The daily life of complicated grief patients--what gets missed, what gets added? *Death Stud*, 30(1), 77-85. doi:10.1080/07481180500348860
- Morley, J., & Floridi, L. (2020). The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. *Sci Eng Ethics*, 26(3), 1159-1183. doi:10.1007/s11948-019-00115-1
- Morren, M., van Dulmen, S., Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: a systematic review. *Eur J Pain*, 13(4), 354-365. doi:10.1016/j.ejpain.2008.05.010
- Morshed, M. B., Kulkarni, S. S., Li, R., Saha, K., Roper, L. G., Nachman, L., . . . Abowd, G. D. (2020). A Real-Time Eating Detection System for Capturing Eating Moments and Triggering Ecological Momentary Assessments to Obtain Further Context: System Development and Validation Study. *JMIR mHealth and uHealth*, 8(12), e20625.
- Morshed, M. B., Saha, K., Li, R., D'Mello, S. K., De Choudhury, M., Abowd, G. D., & Plötz, T. (2019). Prediction of mood instability with passive sensing. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (Vol. 3, pp. 1-21).
- Munafò, M., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Du Sert, N. P., . . . Ioannidis, J. P. (2017). A Manifesto for Reproducible Science. *Nature Human Behaviour*, 1(1), 0021.
- Munsch, S., Meyer, A. H., Milenkovic, N., Schlup, B., Margraf, J., & Wilhelm, F. H. (2009). Ecological momentary assessment to evaluate cognitive-behavioral treatment for binge eating disorder. *Int J Eat Disord*, 42(7), 648-657. doi:10.1002/eat.20657
- 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). New York, NY: Taylor and Francis.
- Myin-Germeys, I. (2020). Digital technology in psychiatry: towards the implementation of a true person-centered care in psychiatry? *Eur Arch Psychiatry Clin Neurosci*, 270(4), 401-402. doi:10.1007/s00406-020-01130-1
- Myin-Germeys, I., Henquet, C., & Myin, E. (n.d.). *Embodied, Embedded Cognition and Mental Health: The Missing E in Psychiatry*.
- 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. doi:10.1002/wps.20513

- Myin-Germeys, I., Klippel, A., Steinhart, H., & Reininghaus, U. (2016). Ecological momentary interventions in psychiatry. *Curr Opin Psychiatry*, 29(4), 258-263. doi: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. *Psychol Med*, 39(9), 1533-1547. doi:10.1017/S0033291708004947
- Myin-Germeys, I., van Aubele, E., Vaessen, T., Steinhart, H., Klippel, A., Lafit, G., . . . Reininghaus, U. (2021). Efficacy of Acceptance and Commitment Therapy in Daily Life (ACT-DL) in early psychosis: Results from the multi-centre INTERACT randomized controlled trial. *medRxiv*. doi:10.1101/2021.05.28.21257986
- Myin-Germeys, I., van Os, J., Schwartz, J. E., Stone, A. A., & Delespaul, P. A. (2001). Emotional reactivity to daily life stress in psychosis. *Arch Gen Psychiatry*, 58(12), 1137-1144. doi: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 (JITAI) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Ann Behav Med*, 52(6), 446-462. doi:10.1007/s12160-016-9830-8
- 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. doi:10.1159/000507993
- Newzoo. (2019). Newzoo Global Mobile Market Report 2019. Retrieved from <https://www.newzoo.com>
- Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, 69, 149-155.
- 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. *J Med Internet Res*, 19(4), e105. doi:10.2196/jmir.7273
- Nock, M., Kleiman, E., Abraham, M., Bentley, K. H., Brent, D. A., Buonopane, R. J., . . . Pearson, J. L. (2021). Consensus Statement on Ethical & Safety Practices for Conducting Digital Monitoring Studies with People at Risk of Suicide and Related Behaviors. *Psychiatric Research & Clinical Practice*. doi:doi:10.1176/appi
- Nusbaum, E. C., Silvia, P. J., Beaty, R. E., Burgin, C. J., Hodges, D. A., & Kwapił, T. R. (2014). Listening between the notes: Aesthetic chills in everyday music listening. *Psychology of Aesthetics, Creativity, and the Arts*, 8(104-109).

- Nutt, D., Wilson, S., & Paterson, L. (2008). Sleep disorders as core symptoms of depression. *Dialogues Clin Neurosci*, 10(3), 329-336. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18979946>
- O'Loughlin, K., Neary, M., Adkins, E. C., & Schueller, S. M. (2019). Reviewing the data security and privacy policies of mobile apps for depression. *Internet Interv*, 15, 110-115. doi:10.1016/j.invent.2018.12.001
- 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. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745>
- 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. *J Med Internet Res*, 21(2), e11398. doi: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. *J Pers*, 69(5), 689-708. doi:10.1111/1467-6494.695160
- Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A., . . . Dunn, G. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatr Scand*, 123(1), 12-20. doi:10.1111/j.1600-0447.2010.01596.x
- Park, C. S., & Kaye, B. K. (2019). Smartphone and self-extension: Functionally, anthropomorphically, and ontologically extending self via the smartphone. *Mobile Media & Communication*, 7(2), 215-231.
- Park, M., Thom, J., Mennicken, S., Cramer, H., & Macy, M. (2019). Global music streaming data reveal diurnal and seasonal patterns of affective preference. *Nat Hum Behav*, 3(3), 230-236. doi:10.1038/s41562-018-0508-z
- Parker, L., Halter, V., Karliychuk, T., & Grundy, Q. (2019). How private is your mental health app data? An empirical study of mental health app privacy policies and practices. *Int J Law Psychiatry*, 64, 198-204. doi:10.1016/j.ijlp.2019.04.002
- Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., . . . Gotlib, I. H. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, 3(2), 292-300.
- 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. doi:10.1016/j.cognition.2012.10.002
- Pearson, J. L., Elmasry, H., Das, B., Smiley, S. L., Rubin, L. F., DeAtley, T., . . . Abrams, D. B. (2017). Comparison of Ecological Momentary

- Assessment Versus Direct Measurement of E-Cigarette Use With a Bluetooth-Enabled E-Cigarette: A Pilot Study. *JMIR Res Protoc*, 6(5), e84. doi:10.2196/resprot.6501
- Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., & Nicolson, N. A. (2006). Diurnal mood variation in major depressive disorder. *Emotion*, 6(3), 383-391. doi:10.1037/1528-3542.6.3.383
- Pejovic, V., & Musolesi, M. (2014). InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 897-908).
- Pew Research Center. (2019). Mobile Fact Sheet. Retrieved from <https://www.pewresearch.org/internet/fact-sheet/mobile/>
- PewResearch. (2019). Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally. Retrieved from <https://www.pewresearch.org/>
- PewResearch. (2020). About one-in-five Americans use a smart watch or fitness tracker. Retrieved from <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. Retrieved from <https://CRAN.R-project.org/package=nlme>
- Pinheiro, J. C., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.
- Poulton, A., Pan, J., Bruns, L. R., Jr., Sinnott, R. O., & Hester, R. (2019). A Smartphone App to Assess Alcohol Consumption Behavior: Development, Compliance, and Reactivity. *JMIR Mhealth Uhealth*, 7(3), e11157. doi:10.2196/11157
- Pritchett, L., Samji, S., & Hammer, J. (2013). It’s all about MeE: using structured experiential learning (“e”) to crawl the design space. Working Paper 322. Washington (DC): Center for Global Development. Retrieved from [http://www.cgdev.org/sites/default/files/its-all-about-mee\\_1.pdf](http://www.cgdev.org/sites/default/files/its-all-about-mee_1.pdf)
- 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. *Front Hum Neurosci*, 12, 501. doi:10.3389/fnhum.2018.00501
- Quinlivan, L., Cooper, J., Meehan, D., Longson, D., Potokar, J., Hulme, T., . . . Kapur, N. (2017). Predictive accuracy of risk scales following self-harm: multicentre, prospective cohort study. *Br J Psychiatry*, 210(6), 429-436. doi: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. Retrieved from <https://www.R-project.org/>

- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.
- Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385-401. doi:http://dx.doi.org/10.1177/014662167700100306
- 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 Physiol Opt*, 26(5), 483-489. doi: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. *Res Hum Dev*, 14(3), 253-270. doi:10.1080/15427609.2017.1340052
- Rathbone, A. L., & Prescott, J. (2017). The use of mobile apps and SMS messaging as physical and mental health interventions: Systematic review. *Journal of Medical Internet Research*, 19(8), e295. doi:10.2196/jmir.7740
- Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. *Psychological Methods* 2(2), 173-185.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2 ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., & Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychol Methods*, 6(4), 387-401. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11778679
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.
- 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 behavior. *Biol Psychol*, 131, 54-62. doi:10.1016/j.biopsycho.2016.09.002
- Reinau, K. H., Harder, H., & Weber, M. (2015). The SMS-GPS-Trip method: a new method for collecting trip information in travel behavior research. *Telecommunications Policy*, 39(3-4), 363-373.
- Reips, U. D., & Funke, F. (2008). Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behav Res Methods*, 40(3), 699-704. doi:10.3758/brm.40.3.699
- Rintala, A., Wampers, M., Myin-Germeyns, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychol Assess*, 31(2), 226-235. doi: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. doi:doi: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. *Psychol Bull*, 132(1), 1-25. doi: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. *Psychol Med*, 50(3), 353-366. doi:10.1017/S0033291719003404
- Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychol Bull*, 128(6), 934-960. doi:10.1037/0033-2909.128.6.934
- 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. doi:10.21037/mhealth.2017.10.01
- RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. Boston, MA: 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. doi:doi: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 behavioral and related sciences* (pp. 27-54). Cham, Switzerland: Springer.
- Ryff, C., Almeida, D. M., Ayanian, J., Carr, D. S., Cleary, P. D., Coe, C., . . . Williams, D. (2017). *Midlife in the United States (MIDUS 2), 2004-2006*.
- Safer, M. A., & Keuler, D. J. (2002). Individual differences in misremembering pre-psychotherapy distress: personality and memory distortion. *Emotion*, 2(2), 162-178. doi:10.1037/1528-3542.2.2.162
- Santangelo, P., Bohus, M., & Ebner-Priemer, U. W. (2014). Ecological momentary assessment in borderline personality disorder: a review of recent findings and methodological challenges. *J Pers Disord*, 28(4), 555-576. doi: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. *J Abnorm Psychol*, 126(8), 1057-1065. doi:10.1037/abn0000288
- Schembre, S. M., Liao, Y., O'Connor, S. G., Hingle, M. D., Shen, S. E., Hamoy, K. G., . . . Boushey, C. J. (2018). Mobile Ecological Momentary Diet



- Assessment Methods for Behavioral Research: Systematic Review. *JMIR Mhealth Uhealth*, 6(11), e11170. doi:10.2196/11170
- Schiepek, G., Aichhorn, W., Gruber, M., Strunk, G., Bachler, E., & Aas, B. (2016). Real-Time Monitoring of Psychotherapeutic Processes: Concept and Compliance. *Front Psychol*, 7, 604. doi:10.3389/fpsyg.2016.00604
- Schmiedek, F., Lovden, M., & Lindenberger, U. (2010). Hundred Days of Cognitive Training Enhance Broad Cognitive Abilities in Adulthood: Findings from the COGITO Study. *Front Aging Neurosci*, 2. doi:10.3389/fnagi.2010.00027
- Schoevers, R. A., van Borkulo, C. D., Lamers, F., Servaas, M. N., Bastiaansen, J. A., Beekman, A. T. F., . . . Riese, H. (2020). Affect fluctuations examined with ecological momentary assessment in patients with current or remitted depression and anxiety disorders. *Psychol Med*, 1-10. doi:10.1017/S0033291720000689
- Schueller, S. M., Aguilera, A., & Mohr, D. C. (2017). Ecological momentary interventions for depression and anxiety. *Depress Anxiety*, 34(6), 540-545. doi:10.1002/da.22649
- Schwartz, S., Schultz, S., Reider, A., & Saunders, E. F. (2016). Daily mood monitoring of symptoms using smartphones in bipolar disorder: A pilot study assessing the feasibility of ecological momentary assessment. *J Affect Disord*, 191, 88-93. doi:10.1016/j.jad.2015.11.013
- Scollon, N. C., 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.
- Selig, J. P., Preacher, K. J., & Little, T. D. (2012). Modeling Time-Dependent Association in Longitudinal Data: A Lag as Moderator Approach. *Multivariate Behav Res*, 47(5), 697-716. doi: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. *J Pers Soc Psychol*, 119(1), 136-158. doi:10.1037/pspi0000212
- 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. *J Consult Clin Psychol*, 65(2), 292-300. doi:10.1037/0022-006x.65.2.292.a
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annu Rev Clin Psychol*, 4, 1-32. doi: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). New York, NY: Guilford Press.

- Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavel, F. D., . . . Bolger, N. (2018). Initial elevation bias in subjective reports. *Proc Natl Acad Sci U S A*, 115(1), E15-E23. doi:10.1073/pnas.1712277115
- Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, 31(4), 471-481.
- Silvia, P. J., Kwapil, T. R., & Walsh, M. A. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior research methods*, 46(1), 41-54.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Singh, N. B., & Bjorling, E. A. (2019). A review of EMA assessment period reporting for mood variables in substance use research: Expanding existing EMA guidelines. *Addict Behav*, 94, 133-146. doi:10.1016/j.addbeh.2019.01.033
- Slade, M. (2017). Implementing shared decision making in routine mental health care. *World Psychiatry*, 16(2), 146-153. doi:10.1002/wps.20412
- Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass*, 2(1), 245-261.
- Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., . . . Van Hoof, C. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ Digit Med*, 1, 67. doi:10.1038/s41746-018-0074-9
- 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. *Psychother Psychosom*, 88(4), 249-251. doi: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 Behavioral Science*: John Wiley & Sons.
- 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.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- 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 Disord*, 6(3), 267-277. doi: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. Retrieved from <https://osf.io/g4jfv/>

- 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. *NPJ Schizophr*, 3(1), 37. doi:10.1038/s41537-017-0038-0
- Statista. (2019). Number of connected wearable devices worldwide from 2016 to 2022. Retrieved from <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).
- Steeg, S., Quinlivan, L., Nowland, R., Carroll, R., Casey, D., Clements, C., . . . 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. doi:10.1186/s12888-018-1693-z
- Steele, C. (2019). What is the Digital Divide? Retrieved from <http://www.digitaldividecouncil.com/what-is-the-digital-divide/>
- Stein, K. F., & Corte, C. M. (2003). Ecologic momentary assessment of eating-disordered behaviors. *Int J Eat Disord*, 34(3), 349-360. doi:10.1002/eat.10194
- Stieger, S., Lewetz, D., & Swami, V. (2020). Psychological well-being under conditions of lockdown: An experience sampling study in Austria during the covid-19 pandemic. *International Journal of Environmental Health Research*. doi:<https://doi.org/10.31234/osf.io/qjhfp>
- 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. doi: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. doi: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.
- Stone, A. A., & Shiffman, S. (1994). Ecological Momentary Assessment (EMA) in Behavioral Medicine. *Annals of Behavioral Medicine*, 16(3), 199-202.
- Stone, A. A., Shiffman, S., Schwartz, J. E., Broderick, J. E., & Hufford, M. R. (2003). Patient compliance with paper and electronic diaries. *Control Clin Trials*, 24(2), 182-199. doi:10.1016/s0197-2456(02)00320-3
- Studer, R. (2012). Does it matter how happiness is measured? Evidence from a randomized controlled experiment. *Journal of Economic and Social Measurement*, 37(4), 317-336.

- Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, 24, 127-136.
- Sun, J., Rhemtulla, M., & Vazire, S. (2019). Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? doi:<https://doi.org/10.31234/osf.io/5tcwd>
- 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. *J Abnorm Psychol*, 117(1), 143-153. doi:10.1037/0021-843X.117.1.143
- Thewissen, V., Bentall, R. P., Oorschot, M., J. A. C., van Lierop, T., van Os, J., & Myin-Germeys, I. (2011). Emotions, self-esteem, and paranoid episodes: an experience sampling study. *Br J Clin Psychol*, 50(2), 178-195. doi:10.1348/014466510X508677
- Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., . . . Marcus, G. M. (2018). Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. *JAMA Cardiol*, 3(5), 409-416. doi:10.1001/jamacardio.2018.0136
- Tooez, J. A., Grunwald, G. K., & Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Stat Methods Med Res*, 11(4), 341-355. doi: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. *JMIR Ment Health*, 3(2), e16. doi:10.2196/mental.5165
- Torous, J., Wisniewski, H., Liu, G., & Keshavan, M. (2018). Mental Health Mobile Phone App Usage, Concerns, and Benefits Among Psychiatric Outpatients: Comparative Survey Study. *JMIR Ment Health*, 5(4), e11715. doi:10.2196/11715
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychol Bull*, 133(5), 859-883. doi:10.1037/0033-2909.133.5.859
- Treuth, M. S., Schmitz, K., Catellier, D. J., McMurray, R. G., Murray, D. M., Almeida, M. J., . . . Pate, R. (2004). Defining accelerometer thresholds for activity intensities in adolescent girls. *Medicine and science in sports and exercise*, 36(7), 1259.
- Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annu Rev Clin Psychol*, 9, 151-176. doi: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. *Psychol Assess*, 21(4), 457-462. doi: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. *J Abnorm Psychol*, 129(1), 56-63.  
doi:10.1037/abn0000473
- 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*. Retrieved from  
<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 Res*, 245, 99-104. doi: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. *J Med Internet Res*, 21(12), e14475. doi: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. doi:10.1016/j.psyneuen.2018.06.007
- 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.  
doi: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.  
doi:doi: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* (pp. 1-12).
- Van den Bergh, O., & Walentynowicz, M. (2016). Accuracy and bias in retrospective symptom reporting. *Curr Opin Psychiatry*, 29(5), 302-308.  
doi:10.1097/YCO.0000000000000267
- Van der Gucht, K., Dejonckheere, E., Erbas, Y., Takano, K., Vandemoortele, M., Maex, E., . . . Kuppens, P. (2019). An experience sampling study examining the potential impact of a mindfulness-based intervention on emotion differentiation. *Emotion*, 19(1), 123-131.  
doi:10.1037/emo0000406
- van der Steen, Y., Gimpel-Drees, J., Lataster, T., Viechtbauer, W., Simons, C. J. P., Lardinois, M., . . . Myin-Germeys, I. (2017). Clinical high risk for psychosis: the association between momentary stress, affective and

- psychotic symptoms. *Acta Psychiatr Scand*, 136(1), 63-73.  
doi:10.1111/acps.12714
- 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. *Am J Geriatr Psychiatry*, 26(12), 1231-1243. doi:10.1016/j.jagp.2018.06.004
- van Roekel, E., Keijsers, L., & Chung, J. M. (2019). A Review of Current Ambulatory Assessment Studies in Adolescent Samples and Practical Recommendations. *J Res Adolesc*, 29(3), 560-577. doi:10.1111/jora.12471
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.
- Verhagen, S. J., Hasmi, L., Drukker, M., van Os, J., & Delespaul, P. A. (2016). Use of the experience sampling method in the context of clinical trials. *Evid Based Ment Health*, 19(3), 86-89. doi:10.1136/ebmental-2016-102418
- 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. *J Med Internet Res*, 18(6), e152. doi: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 Behav*, 49(4), 1157-1177. doi: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). Luxembourg City: University of Luxembourg.
- von Baeyer, C. (1994). Reactive effects of measurement of pain. *The Clinical Journal of Pain*, 10, 18-21. doi:https://doi.org/10.1097/00002508-199403000-00004
- 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.
- Wang, L., Lo, B. P. L., & Yang, G. Z. (2017). Multichannel reflective PPG earpiece sensor with passive motion cancellation. *IEEE Transactions on Biomedical Circuits and Systems*, 1(4), 235-241.
- Wang, L. P., & Maxwell, S. E. (2015). On disaggregating between-person and within-person effects with longitudinal data using multilevel models. *Psychol Methods*, 20(1), 63-83. doi:10.1037/met0000030
- Wang, C., Hall, C.B. & Kim, M. (2015). *Stat Methods Med Res*, 24(6), 1009-29. doi: 10.1177/0962280212437452

- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol*, 54(6), 1063-1070. doi:10.1037//0022-3514.54.6.1063
- Wensing, M., & Grol, R. (2019). Knowledge translation in health: how implementation science could contribute more. *BMC Med*, 17(1), 88. doi:10.1186/s12916-019-1322-9
- Wichers, M., Groot, P. C., & Psychosystems, E. S. M. G. E. W. S. G. (2016). Critical Slowing Down as a Personalized Early Warning Signal for Depression. *Psychother Psychosom*, 85(2), 114-116. doi:10.1159/000441458
- Wickham, H., Averick, M., Bryan, J., Chang, W., MacGowan, L. D. A., François, R., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686.
- Wigman, J. T., van Os, J., Borsboom, D., Wardenaar, K. J., Epskamp, S., Klippel, A., . . . 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. *Psychol Med*, 45(11), 2375-2387. doi:10.1017/S0033291715000331
- Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life. *European Journal of Psychological Assessment*, 23(4), 258-267.
- Wolf, G. I., & De Groot, M. (2020). A Conceptual Framework for Personal Science. *Frontiers in Computer Science*, 2, 21.
- World Health Organization. (2016). Monitoring and evaluating digital health interventions: a practical guide to conducting research and assessment. Licence: CC BY-NC-SA 3.0 IGO.
- 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. *Contemp Clin Trials*, 33(5), 869-880. doi:10.1016/j.cct.2012.05.004
- Yuan, N., Weeks, H. M., Ball, R., Newman, M. W., Chang, Y. J., & Radesky, J. S. (2019). How much do parents actually use their smartphones? Pilot study comparing self-report to passive sensing. *Pediatr Res*, 86(4), 416-418. doi:10.1038/s41390-019-0452-2
- Zettle, R. D. (2016). *The Wiley Handbook of Contextual Behavioral Science*.
- 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. *Stat Med*, 30(20), 2562-2572. doi: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. doi:10.1186/s12859-020-03803-z
- Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: methods and software. *Behav Res Methods*, 46(4), 1184-1198. doi:10.3758/s13428-013-0424-0

- Zhang, Z., & Wang, L. (2009). Statistical power analysis for growth curve models using SAS. *Behav Res Methods*, 41(4), 1083-1094.  
doi:10.3758/BRM.41.4.1083
- Zuur, A., 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 center for **R**esearch on **E**xperience sampling and **A**mbulatory methods **L**euvven (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 to study emotional, behavioral, 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, both in the Research group of quantitative psychology and individual differences (<https://ppw.kuleuven.be/okp/home/>) and in the Center for Contextual Psychiatry (<http://www.ccp-leuven.be>) and they join forces in REAL. The aim of the center is to be a leading center for research into ecological and ambulatory methods for the study of daily life, advancing both fundamental science and clinical practice, and offer 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 health care. She likes rebuilding houses (and gradually gets better at it) as well as being on the road with her bike - both for short and long trips.

**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 modeling 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 has an unexplained interest in the vanitas theme in medieval art.

**Egon Dejonckheere** is a clinical psychologist and post-doctoral researcher at KU Leuven – University of Leuven in Belgium. 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 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 championship hip-hop, 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 the 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 traveling and reading. To the dismay of her neighbours, she also loves playing the violin.

**Gudrun Eisele** is a clinical psychologist and doctoral student 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.

**Zuzana Kasanova** was, at the time of writing this book, a post-doctoral researcher in the Center for Contextual Psychiatry of KU Leuven, Belgium and in the Department of Psychology of the Comenius University in Bratislava, Slovakia. Her expertise lies in the intersection between human neurobiological processes and daily-life behavior, and how it relates to motivation and apathy. In her current role as a Translational Program Manager at KU Leuven Research & Development, her mission is to help bring scientific innovation to the healthcare market and create impact on patients' lives. Due to her putative attachment issues, she continues to collaborate with every academic group she has ever been part of.

**Marlies Houben** is a postdoctoral researcher at Mind-Body Research and Center for Contextual Psychiatry, KU Leuven. Her research focuses on emotional functioning in daily life in relation to psychopathology, both actively measured using self-report and passively monitored using wearables and smartphone sensors. 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.

**Olivia J. Kirtley** is a Senior Postdoctoral Research Fellow at 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 pain. She also leads a number of 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, wild swimming, and buying books she never has time to read.

**Jeroen Weermeijer** is a research psychologist at the Center for Contextual Psychiatry at KU Leuven. His research focuses on the clinical implementation of the Experience Sampling Method. He has a particular interest in user-experience perspectives, statistics, and software development. He has a profound hatred of COVID-19, as it has prevented him from surfing for almost two years.

**Glenn Kiekens** is a clinical psychologist with a double-degree Ph.D. in Psychology (Curtin University, Australia) and Biomedical Sciences (KU Leuven, Belgium). He is currently working as a postdoctoral researcher at KU Leuven. His research focuses on advancing scientific knowledge about why non-suicidal self-injurious thoughts and behaviors (e.g., cutting, hitting oneself) emerge and how to predict and prevent their occurrence. More broadly, he is interested in research on mental health in adolescents and emerging adults and the clinical implementation of digital mental health tools and applications. Glenn loves to travel and explore new places and cultures, and is the proud dog owner of an adorable Great Dane.

**Martien Wampers** obtained a PhD in Psychology and works as a research psychologist in UPC KU Leuven and as database manager in the Research Group Psychiatry at KU Leuven. In that capacity, she is involved in most studies using the Experience Sampling Method 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 research coordinator at the Center for Contextual Psychiatry at KU Leuven – University of Leuven in Belgium. She coordinates all

CCP projects that examine real-time and real-world person-environment interactions in the field of mental health, by using the Experience Sampling Methodology (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 a (mostly) wonderful daughter.

**Davinia Verhoeven** is a research assistant at the Center for Contextual Psychiatry, with over six years of experience in the field of mental health. She did a lot of data collection for studies that used ESM and did probably more than 1000 ESM briefings. She loves the simple things; good laughs, good food and the smell of new cars.

**Wolfgang Viechtbauer** is associate professor of methodology and statistics in the Department of Psychiatry and Neuropsychology and the School for Mental Health and Neuroscience 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 used 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, she is a postdoctoral researcher at the Research Group of Quantitative Psychology and Individual Differences and the Center of Contextual Psychiatry 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, and 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 theater and dance, and she would like to have a second life to create a

performance based on the history of feminism in Latin America and the latest album of Nathy Peluso.

**Ana Teixeira** is a clinical psychologist and a postdoctoral researcher at the Center for Contextual Psychiatry, at KU Leuven. Her PhD was on developing and assessing an Optimal Functioning Therapy for adolescent depression based on Positive Psychology. In the last two years, she has been focused on developing, optimizing, and implementing EMIs in clinical healthcare settings and individuals' daily lives. Ana is a proud mother of Le Chouchou (an adopted dog) and Lindinho (a cat that moved into her house). She has been a vegetarian for almost 10 years, and enjoys reading and talking about the Bible to others. She is originally from the Azores, a group of islands that you should visit one day.

**Joana De Calheiros Velozo** is a doctoral student at the Center for Contextual Psychiatry, at KU Leuven in Belgium. Her PhD focuses on the affective and physiological response to daily-life stressors and its role in the development of mental illness. She is also interested in the potential of wearables for diagnosing and treating mental health concerns. Joana is vegan, and in her free time enjoys learning to skateboard with her husband and taking their adorable basset hound for long walks.

**Koen Niemeijer** is a doctoral student 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. He began his chapter with two hands and finished with one. He also enjoys going for a run without being dragged to the hospital afterwards.

**Thomas Vaessen** is an assistant professor at the Centre for eHealth and Well-being Research at the University of Twente, the Netherlands, and a postdoctoral 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 behavior 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.

In an earth-shattering discovery, scientists recently revealed that all of people's thoughts, feelings, behaviors, and social interactions take place in daily life.

Experience sampling methods (ESM) are the gold standard in studying people's thoughts, feelings, and behavior in the context of daily life. In this open handbook, experts from the center for Research on Experience sampling and Ambulatory methods Leuven (REAL) join forces to provide up-to-date and empirically-based advice for designing, conducting, and analyzing data from ESM studies. The book is designed as a step-by-step guide that walks the reader through the different steps of an ESM study: starting from the type of research questions that can be formulated and how this translates into the design of the study; ethical issues; considerations that come into play when actually conducting the study, such as calculating sample size; choice of software; and the importance of briefing/debriefing. This guide also provides a thorough introduction to preprocessing and analyzing the resulting intensive longitudinal data, ending with some future developments in the areas of mobile intervention and passive sensing. As such, this handbook represents an open, complete, and up-to-date introduction to ESM research for novice and expert researchers alike.

---

*"The advent of smartphones, apps, and wearables has ultimately integrated real-life and real-time methodology, such as the Experience Sampling Methodology (ESM), into the toolbox of psychological science. Whereas books on how to analyze intensive longitudinal data are numerous, wisdom on designing ESM studies was for a long time only transferred by word-of-mouth from senior experts to their PhD students. The Open Handbook of Experience Sampling Methodology, edited by Inez and Peter, closes this gap entirely. This handbook will be an abundant resource in educating upcoming generations."*

- Prof. Dr. Ulrich Ebner-Priemer, Central Institute of Mental Health, Mannheim

*"This book delivers perfectly on its title. The Open Handbook is an immensely practical step-by-step guide to running studies using experience sampling methods (ESM). Combining foundational knowledge of ESM with the latest evidence-based developments, this handbook sympathetically walks readers through a decisional journey of what to do "before", "during" and "after" conducting their ESM study. Deftly edited by leading ESM experts Inez Myin-Germeys and Peter Kuppens, The Open Handbook is a must-read for graduate and postgraduate students, early career researchers, and seasoned researchers wanting to innovate their methods and quell their ESM FOMO. I highly recommend this resource."*

- Tamlin S. Conner, Ph.D., Department of Psychology, University of Otago