New Approaches to Methodological ESM Research: Integrating Smartphone Sensing, Experimentation, and Predictive Modeling

Thomas Reiter

München, 2025

New Approaches to Methodological ESM Research: Integrating Smartphone Sensing, Experimentation, and Predictive Modeling

Inauguraldissertation

zur Erlangung des Doktorgrades der Philosophie

an der Ludwig-Maximilians-Universität München

vorgelegt von

Thomas Reiter

aus Fürstenfeldbruck

München, 2025

Erstgutachter: Prof. Dr. Markus Bühner

Zweitgutachterin: Prof. Dr. Ramona Schödel

Tag der mündlichen Prüfung: 17.10.2025

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich in den letzten Jahren auf meinem Weg begleitet haben.

Zuerst möchte ich mich bei meinem Doktorvater Prof. Dr. Markus Bühner bedanken. Er hat mich während meiner gesamten Promotionszeit immer unterstützt, mir den Rücken freigehalten und die Möglichkeit gegeben, mich weiterzuentwickeln – sowohl innerhalb als auch außerhalb von Academia.

Besonderer Dank gilt zudem meiner Betreuerin und Zweitprüferin Prof. Dr. Ramona Schödel, ohne die dieses Kapitel meines Lebens vermutlich nie begonnen hätte. Vielen Dank für die herausragende Betreuung und Zusammenarbeit, die ich mit dir erleben durfte und die für mich stets von Vertrauen, Wertschätzung und Verlässlichkeit geprägt war.

Als Nächstes danke ich dem gesamten Coping-with-Corona (CoCo) Studienteam – den PIs Mitja, Maarten und Markus, allen studentischen Hilfskräften im Projekt und natürlich den anderen PhDs Julian, Julian und Sophia. Danke für die gute und spaßige Zusammenarbeit im Rahmen zahlreicher Zoom-Termine, CoCo-Treffen und Marketing-Aktionen. Ein mit Sicherheit nicht immer leichtes Projekt wurde erst durch euch zum Erfolg!

Zudem möchte ich mich beim gesamten Phonestudy-Team bedanken. Sei es durch fachlichen Input, das Teilen von Erfahrungen oder wiederholte last-minute Veränderungen an der App kurz vor Studienstart – auf euch war immer Verlass! Besonderer Dank gilt hierbei Flo, Anil, Oli, Tobi, Sarina, Fiona und Larissa.

Außerdem danke ich dem gesamten (aktuellen oder ehemaligen) Lehrstuhl für die familiäre Teamatmosphäre, die lustigen Weinabende, die gute Zusammenarbeit in der Lehre und dafür, dass ich auf euch alle immer zählen konnte, wenn ich bei einem Problem Unterstützung gebraucht habe. Danke für die schöne Zeit Flo, Philipp, David, Yannick, Nensy, Lukas, Philipp, Larissa, Felix, Caro, Lena, Flo, Ai, Ningzhe, Lena-Marei, Lilia, Ramona und Markus.
An dieser Stelle noch einmal besonders erwähnen möchte ich meinen Freund und Weggefährten in Promotion und Studium Philipp sowie Nensy und Lukas aus Büro

3323 (oder wie auch immer unser Büro sonst genannt wurde). Egal ob beim fachlichen Austausch oder anderen (mal mehr mal weniger weltbewegenden) Diskussionen in oder nach der Arbeit – durch euch kam auch an stressigen oder nervigen Arbeitstagen der Spaß im Büro nie zu kurz.

Bei Philipp, Sophia und Flo möchte ich mich zudem für das Lektorat dieser Arbeit bedanken.

Danke auch an meine Freunde Philipp, Moni, Oli, Sarah, Basti und alle, die hier nicht namentlich genannt sind, dass ihr während der kompletten Zeit immer hinter mir standet und zu einer sehr guten Work-Life-Balance beigetragen habt. Schließlich danke ich meinen Eltern Gabi und Peppi, meiner Schwester Susi, meinem Schwager Simon und meinem Opa, dass sie mir auf meinem Weg immer Rückhalt gegeben und ihn ermöglicht haben. Zuletzt danke ich Sophia, die in den letzten Jahren eng an meine Seite gewachsen ist, dafür dass sie mich auf dieser Reise unterstützt und in schwierigen Momenten ausgehalten hat.

Contents

List of Figures

List of Tables

List of Abbreviations

AA	Ambulatory Assessment
API	Application Programming Interface
AUC	Area Under the (Receiver Operating Characteristic) Curve
BFSI	Big Five Structure Inventory
CoCo	Coping with Corona [Research Project]
CV	Cross-Validation
DAG	Directed Acyclic Graph
EAR	Electronically Activated Recorder
EMA	Ecological Momentary Assessment
ESM	Experience Sampling Method
EU-GDPR	European Union General Data Protection Regulation
GLMM	Generalized Linear Mixed Model
GPS	Global Positioning System
HCI	Human-Computer Interaction
ILD	Intensive Longitudinal Data
MAR	Missing At Random
MCAR	Missing Completely At Random
MCC	Matthews Correlation Coefficient
MNAR	Missing Not At Random
OR	Odds Ratio
OSF	Open Science Framework
OSM	Online Supplemental Material
PANAS	Positive and Negative Affect Schedule
PANAS-X	Positive and Negative Affect Schedule - Expanded Form
PHQ-9	Patient Health Questionnaire-9
PWB	Psychological Well-Being (Scale)
RQ	Research Question
SWLS	Satisfaction With Life Scale
TA-EG	Fragebogen zur Erfassung der Technikaffinität als Umgang mit und Einstellung zu elektronischen Geräten [German Scale Measuring Technology Affinity]
WEIRD	Western, Educated, Industrialized, Rich, and Democratic

Abstract

Over the past decades, Ambulatory Assessment (AA) methods have gained significant traction in psychological research due to their potential to capture data on individuals' behavior, experiences, and physiology in real-life settings. Among these, the Experience Sampling Method (ESM) has emerged as a particularly influential approach, involving repeated in-situ self-reports across multiple days. Compared to traditional one-time assessments, ESM offers enhanced temporal resolution and ecological validity, while reducing recall bias. However, ESM studies also introduce specific methodological challenges such as a large number of researcher degrees of freedom in study design, high participant burden and increased rates of missing data—or non-compliance—which may systematically affect or bias study results.

This dissertation aims to advance the methodological understanding of ESM, focusing on two central issues: (1) the phenomenon of non-compliance and its potentially systematic nature, and (2) the effects of specific design decisions within ESM protocols on data quality and study outcomes. To address these topics, the research specifically expands existing work along two dimensions: incorporating passive data collection via smartphone sensing and combining observational and experimental data using both predictive and explanatory modeling.

Smartphone sensing refers to the unobtrusive collection of behavioral and contextual data through sensors embedded in mobile phones (e.g., GPS, accelerometer, microphone, and usage logs). This technique can provide objective measures of participants' behaviors or physical and social environments, supplementing or replacing self-reports and potentially mitigating participant burden. Moreover, smartphone sensing enables novel methodological investigations by offering rich, temporally detailed data streams without requiring active engagement from participants.

The dissertation comprises two empirical studies. Study 1 investigates non-compliance in ESM using over 25,000 questionnaire responses from 592 participants. Predictive models based on person-level traits, contextual information, and smartphone sensing data were developed to predict whether participants would miss a given ESM survey. Both linear and non-linear machine learning models achieved above-chance predictive performance, suggesting that non-compliance is not random but systematically influ-

enced by behavioral and contextual factors. For example, participants were more likely to respond when located at home, work, or on public transport—contexts inferred from sensing data. These findings have implications regarding the handling of systematic non-response and the development of ESM study designs aimed at increasing participant compliance.

Study 2 experimentally manipulated two design dimensions in a four-week ESM study with 395 participants: timing (fixed vs. random) and contingency (direct vs. indirect triggering of notifications). Results indicated that indirect protocols led to higher response rates, but also increased response latencies, potentially affecting data quality. Although self-reported momentary well-being did not differ across conditions, the timing protocol influenced other outcomes, including patterns of smartphone use and their associations with well-being. These findings emphasize the trade-offs associated with different ESM design decisions and the importance of empirical evaluation of ESM design aspects.

Together, the two studies contribute to a better understanding of how study design and participant behavior interact in ESM research. By integrating smartphone sensing and diverse analytical approaches, this work provides actionable insights for both applied and methodological ESM researchers. Moreover, the present dissertation serves as an example of how psychological research in general and methodological ESM research in particular can benefit from the integration of smartphone sensing and the interplay of explanatory and predictive modeling approaches. Furthermore, it highlights the need for updated researcher training, ethical standards, and data handling procedures to accommodate emerging technologies and complex data streams in psychological science.

Zusammenfassung

Der Einsatz von Forschungsmethoden aus dem Bereich *Ambulatory Assessment* hat in der psychologischen Forschung über die letzten Jahrzehnte immer weiter an Bedeutung gewonnen. Unter dem Überbegriff Ambulatory Assessment werden verschiedene Methoden zusammengefasst, die alle zum Ziel haben, Daten über Verhalten, Physiologie oder Erlebnisse von Menschen in ihrem Alltag zu erheben. Einer der prominentesten Vertreter der Familie der Ambulatory Assessment Methoden ist das *Experience Sampling* (engl.: *Experience Sampling Method, ESM*), das inzwischen auch in anderen Forschungsdisziplinen wie der Medizin, den Wirtschaftswissenschaften oder der Medieninformatik eingesetzt wird. In ESM-Studien werden die Studienteilnehmenden typischerweise zu verschiedenen Zeitpunkten über mehrere Tage hinweg wiederholt zu Themen wie aktuellen Erlebnissen, Verhalten oder Gefühlen befragt. Hierfür kommen kurze Fragebögen zum Einsatz, die früher häufig mit Stift und Papier und heute meist mittels anderer Hilfsmittel (z.B., E-Mail, PDAs, Smartphone-Apps) beantwortet werden. Im Vergleich mit globaleren, einmaligen Befragungen werden ESM verschiedene Vorteile zugeschrieben: Eine höhere zeitliche Auflösung, die geringe Störung und bessere Erfassung des alltäglichen Lebens und eine geringere Gefahr der Verzerrung der erhobenen Daten durch Erinnerungsfehler seitens der Teilnehmenden.

Neben diesen Vorteilen ergeben sich aus dem Einsatz von ESM auch spezifische Herausforderungen für Teilnehmende und Forschende. Zum Beispiel geht die in der Regel hohe Zahl an Fragebögen in ESM-Studien normalerweise auch mit einer gewissen Anzahl an verpassten oder nicht beantworteten Fragebögen einher. Dies ist nicht zuletzt auf die hohe zeitliche Auflösung als auch die damit verbundene höhere Belastung der Teilnehmenden zurückzuführen, die durch die wiederholten Befragungen entsteht. Hierbei ist es möglich, dass dieses Nicht-Beantworten – auch *Non-Compliance* genannt – und das daraus resultierende Fehlen von Daten nicht rein zufällig ist, sondern systematischer Natur. Genau wie bei einmaligen Befragungen kann das systematische Fehlen von Daten zu Problemen wie Verfälschungen und Fehlschlüssen in der darauffolgenden Datenanalyse führen.

Außerdem müssen Forschende während der Planung und Durchführung von ESM-Studien eine Vielzahl an Entscheidungen hinsichtlich Studiendesign und Auswertung

treffen. Allein der Faktor, wann und wie die Teilnehmenden befragt werden, kann die Interpretation der Studienergebnisse stark beeinflussen. Häufig stützen sich die Forschenden beim Treffen dieser Designentscheidungen auf theoretische oder pragmatische Überlegungen und weniger auf die Ergebnisse empirischer Studien und Evaluation.

Das übergreifende Ziel dieser Dissertation ist daher, zu einem besseren methodologischen Verständnis von ESM beizutragen und Forschenden somit Anhaltspunkte für ihre Designentscheidungen zu bieten. Hierbei soll besonders auf die zwei eingangs beschriebenen Probleme von Non-Compliance in ESM-Studien und möglichen Auswirkungen verschiedener Designaspekte bzw. ESM-Studienprotokolle eingegangen werden. Bestehende Forschung wurde hierfür besonders hinsichtlich zweier Aspekte erweitert.

Erstens wurde zusätzlich zur bisher meist ausschließlichen Nutzung von Selbstberichten auch auf *Smartphone Sensing* zurückgegriffen. Smartphone Sensing stellt eine vergleichsweise neue Methode aus der Ambulatory Assessment Methodenfamilie dar, bei der Daten passiv und unaufdringlich mit Hilfe von Smartphones gesammelt werden. Hierbei werden die zahlreichen in heutigen Smartphones verbauten Sensoren genutzt, um beispielsweise Daten über den Aufenthaltsort (GPS), Umgebungsgeräusche (Mikrofon), Bewegung (Accelerometer) oder Handynutzung (Smartphone Logs) der Teilnehmenden zu sammeln. Damit lassen sich objektive Informationen über verschiedene Aspekte des Lebens der Teilnehmenden ermitteln, die sonst möglicherweise nur ungenau, lückenhaft oder verzerrt mit Hilfe von Selbstberichten erfasst werden können. Ein weiterer Vorteil von Smartphone Sensing ist, dass die Teilnehmenden – abgesehen von der meist notwendigen Installation einer Sensing App – nicht aktiv zur Datensammlung beitragen müssen. Zudem bietet Smartphone Sensing die Möglichkeit auch dann Daten zu sammeln, wenn andere Methoden der Datenerhebung an ihre Grenzen stoßen, beispielsweise wenn ein Fragebogen nicht beantwortet wird, weil Teilnehmende ihn nicht bemerken oder aktiv ignorieren. Die vielen Vorteile von Smartphone Sensing können nicht nur für inhaltliche Forschung, sondern auch für methodologische Forschung genutzt werden.

Zweitens wurden in der vorliegenden Dissertation sowohl *Beobachtungsdaten* mittels *prädiktiver Modellierung* als auch *experimentelle Daten* mittels modellbasierter inferenzstatistischer Verfahren analysiert. Hierdurch werden sowohl der Erklärungsaspekt als auch der Vorhersageaspekt der Psychologie bedient, die sich als Wissenschaft der Be-

schreibung, Erklärung und Vorhersage menschlichen Erlebens und Verhaltens versteht. Das Zusammenspiel verschiedener Methoden der Datenerhebung und Modellierung zeigt, wie die psychologische Forschung davon profitieren kann, ihre Forschungsfragen mit verschiedenen Ansätzen zu untersuchen.

Die vorliegende Dissertation setzt sich dabei aus zwei Studien zusammen, die zusammen im aktuellen Forschungskontext eingeordnet und diskutiert werden:
Studie 1 befasst sich genauer mit fehlenden Antworten oder Non-Compliance im Rahmen von ESM-Studien. Hierfür wurde auf Basis von über 25.000 Beobachtungen von 592 Teilnehmenden vorhergesagt, ob Teilnehmende einen bestimmten ESM-Fragebogen beantworten oder nicht. Die hierfür verwendeten Daten beinhalteten sowohl Personen-, Kontext- als auch Verhaltensvariablen, die mit Hilfe von globalen Selbstberichten, ESM-Selbstberichten und Smartphone Sensing gesammelt wurden. Im Rahmen eines Vergleichs verschiedener linearer und nicht-linearer Machine Learning Modelle konnte das Nichtbeantworten von ESM-Fragebögen überzufällig gut vorhergesagt werden. Dies kann als Indiz einer systematischen Nichtbeantwortung von ESM-Fragebögen durch die Studienteilnehmenden gedeutet werden. Wenn dieser Umstand in der Analyse oder Modellierung außer Acht gelassen wird, kann dies zu einem gewissen Maß an Verzerrung der Ergebnisse – compliance bias — führen. Im Detail zeigte sich, dass das Antwortverhalten unter anderem durch vergangenes Verhalten sowie den physischen Kontext der Teilnehmenden vorhergesagt werden kann. So antworteten Teilnehmende beispielsweise mit höherer Wahrscheinlichkeit auf ESM-Fragebögen, wenn sie schneller auf vorangegangene Fragebögen reagiert hatten oder sich zum Zeitpunkt der Befragung gerade zu Hause, in der Arbeit oder im Zug befanden. Letztere Informationen wurden hierbei aus Smartphone Sensing Daten abgeleitet. Die Ergebnisse der Studie gehen mit Implikationen für das Design von ESM-Studien als auch mit Lösungsansätzen zum Umgang mit fehlenden Daten und zur Optimierung des Antwortverhaltens in ESM-Studien einher.

In **Studie 2** wurden spezifische Designaspekte von ESM-Protokollen und deren Einfluss auf Quantität und Qualität der erhobenen Daten, sowie mögliche Verzerrungen daraus folgender Studienergebnisse untersucht. In einer präregistrierten, vierwöchigen Studie mit 395 Teilnehmenden wurden das Timing (fixe vs. variable Zeitpunkte) und die

Kontingenz (direktes Auslösen zum gegebenen Zeitpunkt vs. indirektes Auslösen bei der nächsten Handynutzung nach dem gegebenen Zeitpunkt) der ESM-Fragebögen experimentell manipuliert. Anschließend wurde die Auswirkung dieser Aspekte hinsichtlich verschiedener Kriterien untersucht. Es zeigte sich, dass indirekte ESM-Protokolle die Antwortraten der Teilnehmenden verbessern, jedoch auch die Datenqualität beeinträchtigen können, da beispielsweise die Latenz der Antworten zunimmt. Das selbstberichtete momentane Wohlbefinden der Teilnehmenden wies keine Unterschiede zwischen den ESM-Protokollen auf. Das Timing der Fragebögen wirkte sich jedoch auf andere Studienergebnisse, wie die Smartphone-Nutzung der Teilnehmenden sowie deren Zusammenhang mit anderen Variablen zum Wohlbefinden aus. Die Studie unterstreicht damit die Notwendigkeit, Designentscheidungen in ESM-Studien sorgfältig zu prüfen und deren Vorteile und Nachteile gegeneinander abzuwägen, um eine valide und zuverlässige Datenerhebung zu gewährleisten.

Insgesamt leisten beide Studien einen direkten Beitrag zum besseren Verständnis des Antwortverhaltens und potenzieller Nebeneffekte bestimmter Studienprotokolle in ESM-Studien. Unter anderem konnte gezeigt werden, dass Smartphone Sensing Daten einen wertvollen Beitrag für die prädiktive Modellierung von (Non-)Compliance in ESM-Studien leisten können, der über die ausschließliche Nutzung von Selbstberichtdaten hinausgeht. Zudem wurde Smartphone Sensing in das Design und die Evaluation verschiedener ESM-Protokolle eingebaut, um mögliche Nebeneffekte der jeweiligen Protokolleigenschaften besser zu verstehen. Es zeigt sich, dass der Einsatz von Smartphone Sensing und die Anwendung von erklärender und prädiktiver Modellierung inkrementelle Beiträge zur methodischen Erforschung von ESM leisten können. Ferner kann die vorliegende Dissertation als allgemeines Beispiel für die Integration neuartiger Ansätze und Methoden in die psychologische Grundlagenforschung gesehen werden. Die gewonnenen Erkenntnisse können dabei sowohl als Anhaltspunkte für angewandt Forschende bei der Planung und Durchführung inhaltlicher Studien verwendet werden, als auch zukünftige methodologisch orientiertere Studien zum besseren Verständnis von ESM motivieren und inspirieren. Neben dem Hervorheben der Stärken und Vorteile befassen sich die zwei Studien und die vorliegende Arbeit auch mit den Limitationen und Herausforderungen im derzeitigen und zukünftigen Einsatz der verwendeten Methoden. So stellt der Einsatz von Smartphone Sensing und die Verwendung von erklärender

und prädiktiver Modellierung neue Anforderungen an die Ausbildung psychologisch Forschender und macht die Entwicklung neuer Standards zur Verarbeitung und Veröffentlichung der daraus entstehenden Daten notwendig. Darüber hinaus ergeben sich neue Fragestellungen hinsichtlich Ethik und Sicherheit der gesammelten Daten und eingesetzten Modelle sowie potenzielle Einsatzgebiete und Synergien interdisziplinärer Zusammenarbeit.

1 General Introduction

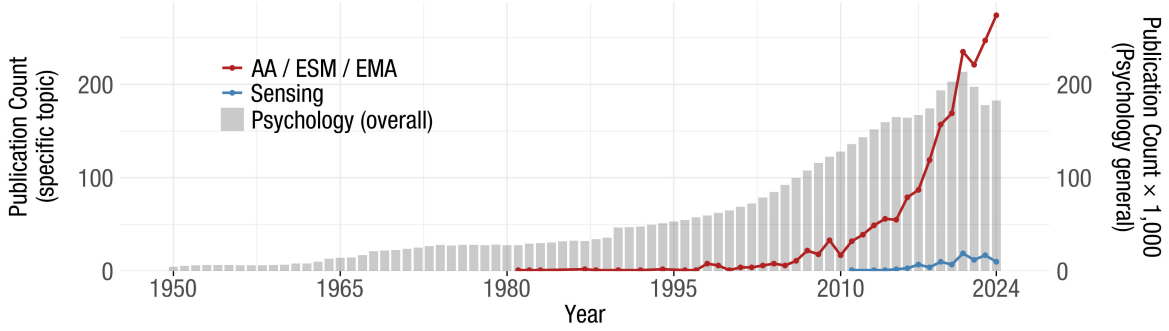
1.1 The Rise of Ambulatory Assessment in Psychological Research

In recent decades, the use of *Ambulatory Assessment* (AA) to investigate individuals' experiences, states, and behaviors has continuously increased (Seizer et al., 2024; Trull & Ebner-Priemer, 2014). Although the term AA is often used interchangeably with related concepts for data collection such as experience sampling, ecological momentary assessment, or the daily diary method, it can rather be considered an umbrella term for all of these different assessment methods (Trull & Ebner-Priemer, 2014). The different data collection methods of AA have in common that they can be applied in the field to capture self-reports, physiological, or behavioral data in unrestrained life settings (Fahrenberg et al., 2007). Often, they are computer-assisted (e.g., automated calls, e-mails) or rely on smartphones (e.g., using specific survey apps) and other electronic devices (e.g., actigraphs, fitness trackers, etc.). The rise in AA is accompanied by a rise in studies and methods concerned with intensive longitudinal data (ILD) which AA often results in (Hamaker & Wichers, 2017). Arguably one of the most common methods from the spectrum of AA which has also seen a relatively greater increase in associated publications compared to the average in psychology (cf. Figure 1.1) is the *Experience Sampling Method* (ESM; Fritz et al., 2024; Seizer et al., 2024).

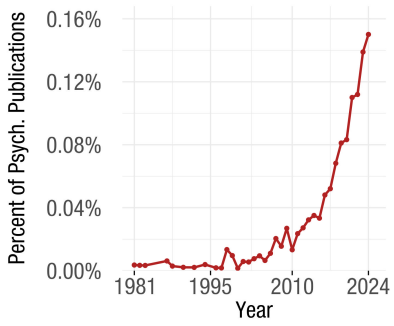
Initially introduced and applied in psychology by Csikszentmihalyi, Larson, and Prescott in the late 70s to early 80s for studying everyday life (Csikszentmihalyi et al., 1977; Larson & Csikszentmihalyi, 1983; Prescott & Csikszentmihalyi, 1981), ESM is now used across different disciplines including medicine, economics, computer science, or the behavioral sciences in general and is often combined with other data collection methods (Scollon et al., 2003; van Berkel et al., 2017). In ESM studies, participants are typically asked to answer several short questionnaires multiple times during the day or over multiple days (Eisele et al., 2022). Like AA, the term ESM is often used interchangeably with similar methods, one of them being Ecological Momentary Assessment (EMA). However, although having its roots in ESM, the concept of EMA has expanded to also include the monitoring of physiological processes or behaviors via technical devices rather than relying solely on self-reports of subjective states or behaviors as is the case for ESM (Trull & Ebner-Priemer, 2009). The repeated measurements of ESM are

Figure 1.1
Trends in ESM- and Smartphone Sensing-related Publications in Psychology

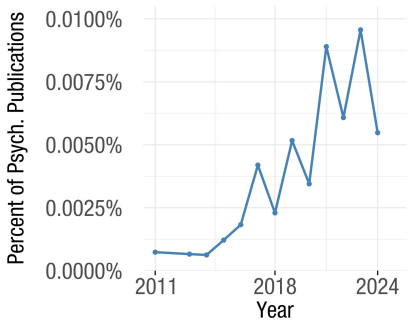
(a) *Absolute Number of Publications per Year*



(b) *Proportion of ESM Studies*



(c) *Proportion of Sensing Studies*



Note. Lines in panel (a) display the absolute number of articles containing "Ambulatory Assessment", "Experience Sampling Method", or "Ecological Momentary Assessment" (red) and "Smartphone Sensing" or "Mobile Sensing" (blue) in their title or abstract (left y-axis). Grey bars represent the number of overall articles in the research area *Psychology* during the same period (right y-axis). Lines in panels (b) and (c) display the corresponding relative proportion of ESM and Sensing related studies among the overall number of psychological publications.
Data source: Clarivate *Web of Science*.

associated with different advantages. First, due to the repeated nature of the data collection, ESM allows researchers to study dynamic within-person processes as well as individual differences across them (Hamaker & Wichers, 2017). This is often considered one of the strongest benefits of ESM when compared with global reports (Scollon et al., 2003). Second, instead of being conducted in laboratory settings, ESM assessment takes place in the real world, catching participants in real-life situations (Myin-Germeys et al., 2009). This promises advantages for ecological validity and generalizability (Scollon et al., 2003). Third, ESM assessments happen in real-time in (or very close to) the moment an experience, behavior, or event of interest occurred. This helps reduce different data quality issues often associated with one-time, retrospective, self-report surveys, such as recall biases (Conner & Mehl, 2015; Myin-Germeys et al., 2009; Scollon et al., 2003).

1.2 Challenges in ESM Research: (Non-)Compliance and Other Stories

Despite the various proposed advantages, newly adopted methods such as ESM often come with new methodological challenges for researchers planning to apply them. The following sections further elaborate on two key issues often discussed in the context of ESM.

1.2.1 (Non-)Compliance, Data Quantity, and Missing Responses in ESM Studies

First, in ESM, the advantage of more granular data can come at the price of higher participant burden compared to "traditional" one-time surveys (Fritz et al., 2024). Throughout the course of a prototypical two-week ESM study, with 5 assessments per day each containing 15 items (which can be considered a standard ESM protocol with short to medium sized survey length; cf. Eisele et al., 2022; Vachon et al., 2019; Wrzus & Neubauer, 2023), participants are required to answer more than 1000 items in total which can be burdensome. Accordingly, different ESM design aspects have already been evaluated regarding their effects on participant burden. For example, longer surveys have been found to increase participants' perceived burden (Eisele et al., 2022). This increased burden may impact participants' general motivation to participate in ESM studies, increase dropout rates, or (at least) reduce participants' compliance with the ESM protocol resulting in higher non-response rates (Eisele et al., 2022; Ottenstein & Werner, 2021).

Apart from ESM design characteristics, non-compliance or non-response to ESM surveys can also be related to participant or context characteristics. Besides merely decreasing data quantity, participants selectively answering only specific ESM surveys (e.g., at specific times or during specific activities) can also introduce potentially systematically missing data. If this data missingness (often classified into Missing Completely at Random [MCAR], Missing at Random [MAR], and Not Missing at Random [NMAR]; Little & Rubin, 1987) is not handled appropriately during analysis, it can lead to different inference problems such as decreased power and biased parameter and standard error estimates (Ji et al., 2018).

1.2.2 Specific ESM Protocols and How They Affect Data and Study Outcomes

In addition to these general data quantity and compliance related issues that need to be considered when conducting an ESM study, there are more specific methodological issues related to study design that researchers have to keep in mind. Already the traditional trichotomy of interval-, signal-, and event-contingent ESM protocols (Wheeler & Reis, 1991) gives a first hint about the many facets and decisions regarding the optimal way of designing an ESM study. However, often these design decisions are made based on theoretical, pragmatic, or heuristic considerations and circumstances, rather than on systematic inquiry (Fritz et al., 2024; Himmelstein et al., 2019). Moreover, even if based on proper theoretical considerations, the decision on whether to use an event-contingent protocol (i.e., trigger an ESM questionnaire upon occurrence of a specific event), an interval-contingent protocol (e.g., trigger an ESM questionnaire every 2 hours), or versus a signal-contingent protocol (i.e., triggering ESM signals at randomly scheduled times) may affect the collected data. For example, participants could get used to receiving their survey at 9 a.m. and start to habitually always answer it in the same manner or with lower effort or motivation possibly decreasing data quality (Eisele et al., 2023). Moreover, selection of a specific ESM protocol might even impact study findings as the data collected via one protocol could systematically differ from data collected via another protocol. For example, always asking a participant at 8 a.m. and 2 p.m. could lead to reaching a usually energetic participant in their sleepiest states throughout the day, namely in the morning and after lunch. Accordingly, the design of ESM studies can affect the data collected and should be based on thorough considerations and, at best, systematic evaluation.

1.3 Smartphone Sensing: A New Member to the AA Family

One comparably new data collection method from the spectrum of AA which has so far not been addressed in detail is *smartphone sensing* (Conner & Mehl, 2015). It makes use of the ubiquity of smartphones in today's everyday lives for passive data collection (Harari et al., 2016). Similar to ESM (although smaller in magnitude and with a less steep trajectory; cf. Figure 1.1), the number of studies incorporating smartphone sensing has increased over the last years (e.g., Krämer et al., 2024; Müller et al.,

2020; Schoedel et al., 2023). Just as the number of researchers applying smartphone sensing increased, so did the number of providers of (commercial) smartphone sensing solutions to researchers (e.g., Beierle et al., 2018; Niemeijer et al., 2023). This suggests that smartphone sensing studies will become easier to conduct for future researchers and further continue to be integrated in psychological research. The reasons for including smartphone sensing in psychological research are manifold: First, given the high adoption rates of smartphones and the variety of sensors included, smartphone sensing offers the possibility to collect diverse and rich data sets from large samples (Harari et al., 2016; Miller, 2012). Second, smartphone sensing happens unobtrusively, removing the need for participants to interact with additional devices or actively answer questions on topics such as smartphone usage or mobility (Harari et al., 2016). In addition, the unobtrusive nature of data collection in smartphone sensing may also reduce the likelihood of participants showing reactive behaviors (Conner & Mehl, 2015). Third, with smartphone sensing, individuals' everyday behaviors can be assessed objectively, mitigating known issues of self-reports, such as biases and distortions because participants no longer have to recall and describe past behaviors, social interactions, or other activities (Harari et al., 2017). In many cases, researchers use smartphone sensing data together with other AA approaches such as ESM or daily diaries allowing for subsequent combination of the data (Meegahapola & Gatica-Perez, 2020). By integrating self-reports and smartphone sensing, researchers can not only make use of the individual advantages of both methods but even profit from additional synergies. According to Keusch and Conrad (2022), the integration allows for verification and contextualization of collected data (e.g., Kern et al., 2021; Lathia et al., 2017), quantification of relationships between measures (e.g., Wang et al., 2014), the definition and measurement of composite metrics (e.g., Reiter et al., 2024), or the triggering of measurements contingent on sensed events (e.g., Roos et al., 2023).

1.4 Leveraging Smartphone Sensing for Methodological ESM Research

Just as for applied research settings, smartphone sensing has the potential for advances in methodological research and can be used to improve the understanding and design of ESM studies. The objective and unobtrusive collection of additional data types paves the way for addressing gaps left by research that lacks smartphone

sensing integration, and even enables the investigation and evaluation of ESM from a completely new perspective. For example, smartphone sensing measures have been compared to self-reports to assess measurement validity in the context of mobile phone usage or physical activity representing a first methodological use-case of integrating smartphone sensing (Lathia et al., 2017; Yuan et al., 2019).

Additionally, smartphone sensing offers an additional way to advance the understanding of ESM which becomes most apparent when compared with traditional ESM. Contrary to ESM, smartphone sensing allows for data collection even if participants miss or actively decide not to answer ESM surveys. In these cases, smartphone sensing offers a unique way to gather data when other data collection methods reach their limits. This mitigates one problem inherent to and naturally complicating the investigation of missing data mechanisms — namely, that we usually lack any information beyond the fact that the data is missing (although there are some noteworthy and clever research designs, collecting data around the missing data for example by using electronically activated recorders (EAR) in addition to ESM surveys as done in Sun et al. (2021)). Thus, smartphone sensing stands out as one candidate approach for better understanding missing data mechanisms in ESM settings.

Apart from this, smartphone sensing can also be used to motivate, design, and evaluate new ESM protocols. For example, it was already used to trigger surveys based on participants' locations (Kreuter et al., 2020) or phone calls (Sugie, 2018). This eliminates the need for participants to "become aware" of specific events on their own and thus may be beneficial to validity in event-contingent ESM designs. Moreover, it can be used to trigger survey notifications in opportune moments, possibly leading to decreased feelings of interruption or increased response rates (Pielot et al., 2017). Accordingly, the role of the smartphone is changing from being a mere tool for data recording to a tool directly associated with study design (e.g., when smartphone sensed measures are used to trigger surveys; cf. Roos et al., 2023). However, this makes it necessary to also consider smartphone sensing metrics during the evaluation of study designs relying on smartphone sensing. This becomes even more critical given that smartphone-derived metrics themselves increasingly become variables of interest in psychological studies (Christensen et al., 2016; große Deters & Schoedel, 2024).

1.5 The Present Dissertation

1.5.1 Rationale

The present dissertation contributes to a better methodological understanding of ESM by exploiting the advantages of smartphone-based data collection and of different analysis approaches. Herein, a special focus is placed on the aspects of general (non-)compliance in ESM and side effects of specific ESM design decisions such as the choice of certain ESM protocols. In particular, two strategies and approaches are applied and combined.

1.5.1.1 Combination of Multiple Data Sources to Address ESM Challenges. The present work relies on a combination of multiple data sources. More precisely, the studies included will not only use traditional one-time survey data, but also integrate intensive longitudinal data from ESM surveys and smartphone sensing data collected passively via the PhoneStudy research app (https://www.phonestudy.org/en). This makes it possible to benefit from the unique and diverse advantages inherent to the different data collection approaches. Specific momentary or person-related information such as momentary mood, which often requires introspection and is therefore difficult to infer from smartphone sensing data, can be gathered directly via participant self-reports. Other measures including behaviors or contextual information can conveniently be derived from smartphone sensing data collected unobtrusively without requiring potentially annoying or burdensome self-reports from participants. Thus, this dissertation highlights the benefits of combining different data sources to more comprehensively address methodological questions. Indeed, it is among the first to exploit the advantages of smartphone sensing (e.g., unobtrusiveness, no selective missingness, objectivity, and high granularity) in the context of methodological ESM research.

1.5.1.2 Combination of Multiple Analysis Methods to Address ESM Challenges. In addition to the mere combination of different data sources, the present dissertation draws upon the two cultures of statistical modeling —namely the culture of data (or explanatory) modeling and the culture of algorithmic (or predictive) modeling (Breiman, 2001; Shmueli, 2010)— to address its respective research questions. On the one hand, the variety of variables collectible via smartphone sensing will be used

to predict noncompliance in ESM studies. This goal of predicting an actual behavior (i.e., responding to a specific ESM survey) based on a large set of predictor variables represents a use case well-suited for applying machine learning methods. Moreover, it responds to the call to move psychology towards becoming a more predictive science (Yarkoni & Westfall, 2017). On the other hand, experimental manipulation and methods from descriptive and inferential statistics are used to understand how the data collected via ESM are affected by design aspects of ESM protocols. This interplay highlights that psychological research can greatly benefit from drawing on both cultures of statistical modeling in order to appropriately address its research questions and progress as a science.

1.5.2 Manuscripts of this Dissertation

In detail, this dissertation contributes to these goals through two empirical studies listed in Table 1.1.

Study 1 directly addresses the methodological challenge of compliance in ESM studies. For this, a dataset of over 25,000 observations from 592 participants is used. In a benchmark experiment, different machine learning algorithms are compared in predicting compliance based on a combination of data collected via traditional surveys, ESM, and smartphone sensing. The findings reveal systematic compliance biases, such as the influence of past response behavior and physical context on questionnaire completion. These insights have implications for the design, optimization, and analysis of future ESM studies. They point to the possibility of enhancing data quantity by considering participant behavior and context during study design. Moreover, they may provide reference points for analyzing ESM study data by suggesting variables that should be included in the data analysis. Thus, Study 1 not only highlights the potential of integrating diverse data sources but also advances methodological understanding by providing actionable recommendations to address missing data and improve compliance in ESM research.

Study 2 empirically examines how specific design decisions in ESM protocols influence data quantity, data quality, and potential biases in study outcomes. In a pre-registered four-week ESM study with 395 participants, the ESM protocol characteristics

Table 1.1
Overview of Author Contributions in Co-authored Publications

Study	Publication Reference	CRediT Author Contributions
1	**Reiter, T.**, & Schoedel, R. (2024). Never miss a beep: Using mobile sensing to investigate (non-)compliance in experience sampling studies. *Behavior Research Methods*, *56*, 4038–4060. https://doi.org/10.3758/s13428-023-02252-9	**TR:** Conceptualization, Formal Analysis, Methodology, Visualization, Writing - Original Draft, Writing - Review & Editing RS: Conceptualization, Data Curation, Investigation, Methodology, Supervision, Writing - Review & Editing
2	**Reiter, T.**, Sakel, S., Scharbert, J., ter Horst, J., van Zalk, M., Back, M., Bühner, M., & Schoedel, R. (2025). Side Effects of Experience Sampling Protocols: A Systematic Analysis of How They Affect Data Quality, Data Quantity & Bias in Study Results. *Advances in Methods and Practices in Psychological Science. 8*(3). https://doi.org/10.1177/25152459251347274	**TR:** Conceptualization, Data Curation, Formal Analysis, Methodology, Visualization, Writing - Original Draft, Writing - Review & Editing, Project Administration SS, JS, JtH: Investigation, Data Curation, Writing - Review & Editing, Project Administration MvZ, MBa, MBü: Writing - Review & Editing, Project Administration, Funding Acquisition RS: Conceptualization, Data Curation, Investigation, Methodology, Supervision, Writing - review & editing

Note. Author contributions according to the Contributor Role Taxonomy (CRediT; for more information see Brand et al., 2015). Contributions of the author of this dissertation are in bold.

timing (fixed vs. variable scheduling of surveys) and contingency (direct vs. indirect triggering) were experimentally manipulated and their effects investigated. The findings provide evidence that while indirect notifications improve response rates, they may compromise data quality as reflected in increased response latencies. Furthermore, the timing of surveys was shown to affect participants' smartphone usage behavior and its associations with well-being measures. This highlights how ESM design choices may affect study outcomes. Accordingly, Study 2 demonstrates the need for careful consideration of ESM protocol characteristics to ensure the collection of valid and reliable data. By addressing how protocol decisions affect methodological outcomes, this manuscript advances the methodological understanding of ESM design and its associated trade-offs.

In summary, this dissertation can be considered to be positioned at the intersection of psychology and related disciplines such as human-computer interaction (HCI). It highlights how psychology can both benefit from interdisciplinary collaboration and, in turn, also contribute to other fields. Some of the proposed methods and the app employed draw upon and integrate insights from media informatics and HCI (e.g., Ferreira et al., 2015). Conversely, the findings have methodological implications that can inform and inspire experimental designs in disciplines beyond psychology. This illustrates how interdisciplinary collaboration can enrich psychological science and vice versa.

1.5.3 Open Science Statement

The two studies in this dissertation were guided by and adhere to the principles of open science. However, as a consequence of the complexity of mobile sensing data and the variable extraction procedures, Study 1 adopted a purely exploratory perspective and thus was not preregistered. Study 2, which adopted a more confirmatory perspective, was preregistered prior to data preprocessing and analysis. For both studies, the preprocessing code, analysis code, and (preprocessed) data are provided via the respective Open Science Framework (OSF) repositories referenced in the study sections. For smartphone sensing variables, the data sets only include the preprocessed measures as sharing the raw, time-stamped log data is not possible because of privacy concerns and related data protection regulation and legislation.

1.6 References

Beierle, F., Tran, V. T., Allemand, M., Neff, P., Schlee, W., Probst, T., Pryss, R., & Zimmermann, J. (2018). Tydr: Track your daily routine. android app for tracking smartphone sensor and usage data. *Proceedings of the 5th International Conference on Mobile Software Engineering and Systems*, 72–75. https://doi.org/10.1145/3197231.3197235

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2). https://doi.org/10.1087/20150211

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Christensen, M. A., Bettencourt, L., Kaye, L., Moturu, S. T., Nguyen, K. T., Olgin, J. E., Pletcher, M. J., & Marcus, G. M. (2016). Direct measurements of smartphone screen-time: Relationships with demographics and sleep. *PLoS ONE*, *11*(11), e0165331. https://doi.org/10.1371/journal.pone.0165331

Conner, T. S., & Mehl, M. R. (2015). Ambulatory assessment: Methods for studying everyday life. *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, *2015*, 1–15. https://doi.org/10.1002/9781118900772.etrds0010

Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, *6*(3), 281–294. https://doi.org/10.1007/BF02138940

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

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., Myin-Germeys, I., & Viechtbauer, W. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol

and individual characteristics. *Psychological Assessment*. https://doi.org/10.1037/pas0001177

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment-monitoring behavior in daily life settings. *European Journal of Psychological Assessment*, *23*(4), 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Ferreira, D., Kostakos, V., & Dey, A. K. (2015). Aware: Mobile context instrumentation framework. *Frontiers in ICT*, *2*, 6. https://doi.org/10.3389/fict.2015.00006

Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., et al. (2024). So you want to do esm? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Science*, *7*(3). https://doi.org/10.1177/25152459241267912

große Deters, F., & Schoedel, R. (2024). Keep on scrolling? using intensive longitudinal smartphone sensing data to assess how everyday smartphone usage behaviors are related to well-being. *Computers in Human Behavior*, *150*, 107977. https://doi.org/10.1016/j.chb.2023.107977

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

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

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

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

Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling missing data in the modeling of intensive longitudinal data.

Structural Equation Modeling: A Multidisciplinary Journal, *25*(5), 715–736. https://doi.org/10.1080/10705511.2017.1417046

Kern, C., Höhne, J. K., Schlosser, S., & Revilla, M. (2021). Completion conditions and response behavior in smartphone surveys: A prediction approach using acceleration data. *Social Science Computer Review*, *39*(6), 1253–1271. https://doi.org/10.1177/0894439320971233

Keusch, F., & Conrad, F. G. (2022). Using smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and Methodology*, *10*(4), 863–885. https://doi.org/10.1093/jssam/smab035

Krämer, M. D., Roos, Y., Schoedel, R., Wrzus, C., & Richter, D. (2024). Social dynamics and affect: Investigating within-person associations in daily life using experience sampling and mobile sensing. *Emotion*, *24*(3), 878–893. https://doi.org/10.1037/emo0001309

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, *38*(5), 533–549. https://doi.org/10.1177/0894439318816389

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. Reis (Ed.), *New directions for methodology of social and behavioral sciences* (pp. 41–56, Vol. 15). San Francisco: Jossey-Bass. https://doi.org/10.1007/978-94-017-9088-8_2

Lathia, N., Sandstrom, G. M., Mascolo, C., & Rentfrow, P. J. (2017). Happier people live more active lives: Using smartphones to link happiness and physical activity. *PloS ONE*, *12*(1), e0160589. https://doi.org/10.1371/journal.pone.0160589

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data* (Vol. 1). John Wiley & Sons. https://doi.org/10.1002/9781119013563

Meegahapola, L., & Gatica-Perez, D. (2020). Smartphone sensing for the well-being of young adults: A review. *IEEE Access*, *9*, 3374–3399. https://doi.org/10.1109/ACCESS.2020.3045935

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. https://doi.org/10.1177/1745691612441215

Müller, S. R., Peters, H., Matz, S. C., Wang, W., & Harari, G. M. (2020). Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and gps tracking. *European Journal of Personality*, *34*(5), 714–732. https://doi.org/10.1002/per.2262

Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & Van Os, J. (2009). Experience sampling research in psychopathology: Opening the black box of daily life. *Psychological Medicine*, *39*(9), 1533–1547. https://doi.org/10.1017/S0033291708004947

Niemeijer, K., Mestdagh, M., Verdonck, S., Meers, K., Kuppens, P., et al. (2023). Combining experience sampling and mobile sensing for digital phenotyping with m-path sense: Performance study. *JMIR Formative Research*, *7*(1), e43296. https://doi.org/10.2196/43296

Ottenstein, C., & Werner, L. (2021). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

Pielot, M., Cardoso, B., Katevas, K., Serrà, J., Matic, A., & Oliver, N. (2017). Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–25. https://doi.org/10.1145/3130956

Prescott, S., & Csikszentmihalyi, M. (1981). Environmental effects on cognitive and affective states: The experiential time sampling approach. *Social Behavior and Personality: An International Journal*, *9*(1), 23–32. https://doi.org/10.2224/sbp.1981.9.1.23

Reiter, T., Sakel, S., Scharbert, J., Ter Horst, J., Back, M., Van Zalk, M., Bühner, M., & Schoedel, R. (2024). Investigating phubbing in everyday life: Challenges & lessons for future research. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3613905.3651009

Roos, Y., Krämer, M. D., Richter, D., Schoedel, R., & Wrzus, C. (2023). Does your smartphone "know" your social life? a methodological comparison of day reconstruction, experience sampling, and mobile sensing. *Advances in Methods and Practices in Psychological Science*, *6*(3), 1–12. https://doi.org/10.1177/25152459231178738

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology*, *125*(6), 1442–1471. https://doi.org/10.1037/pspp0000469

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*(1), 5–34. https://doi.org/10.1023/A:1023605205115

Seizer, L., Schiepek, G., Cornelissen, G., & Löchner, J. (2024). A primer on sampling rates of ambulatory assessments. *Psychological Methods*. https://doi.org/10.1037/met0000656

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Sugie, N. F. (2018). Utilizing smartphones to study disadvantaged and hard-to-reach groups. *Sociological Methods & Research*, *47*(3), 458–491. https://doi.org/10.1177/0049124115626176

Sun, J., Rhemtulla, M., & Vazire, S. (2021). Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin*, *47*(11), 1535–1549. https://doi.org/10.1177/0146167220964639

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

Trull, T. J., & Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (esm/ema) in clinical assessment and clinical research: Introduction to the special section [editorial]. *Psychological Assessment*, *21*(4), 457–462. https://doi.org/10.1037/a0017653

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, *50*(6), 1–40. https://doi.org/10.1145/3123988

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14. https://doi.org/10.1145/2632048.2632054

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, *59*(3), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x

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

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

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. *Pediatric Research*, *86*(4), 416–418. https://doi.org/10.1038/s41390-019-0452-2

2 Study 1: Never Miss a Beep: Using Mobile Sensing to Investigate (Non-)Compliance in Experience Sampling Studies

Reference

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

Author Contributions

Thomas Reiter:

Conceptualization, Formal Analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing

Ramona Schoedel:

Conceptualization, Data Curation, Investigation, Methodology, Supervision, Writing – Review & Editing.

2.1 Abstract

Given the increasing number of studies in various disciplines using experience sampling methods, it is important to examine compliance biases because related patterns of missing data could affect the validity of research findings. In the present study, a sample of 592 participants and more than 25,000 observations were used to examine whether participants responded to each specific questionnaire within an experience sampling framework. More than 400 variables from the three categories of person, behavior, and context, collected multi-methodologically via traditional surveys, experience sampling, and mobile sensing, served as predictors. When comparing different linear (logistic and elastic net regression) and non-linear (random forest) machine learning models, we found indication for compliance bias: response behavior was successfully predicted. Follow-up analyses revealed that study-related past behavior, such as previous average experience sampling questionnaire response rate, was most informative for predicting compliance, followed by physical context variables, such as being at home or at work. Based on our findings, we discuss implications for the design of experience sampling studies in applied research and future directions in methodological research addressing experience sampling methodology and missing data.

2.2 Introduction

Trivial as it may sound, what most studies have in common is that they deal with data they have, not data they do not have. Missing data, however, can lead to problems or fallacies if not taken into account in a study's design, data analysis, or interpretation of results (Graham, 2012; Little & Rubin, 1987). For example, individuals suffering from depression might be less willing to participate in a survey because of a lack of energy associated with their illness. The resulting systematic lack of data could lead to biased findings when estimating the prevalence of depression or its association with other variables of interest (Prince, 2012). Biased results, in turn, can have far-reaching consequences, for example, in informing policy makers to ensure adequate mental health care (Shorey et al., 2022).

Non-response bias is a challenge not only for traditional surveys but also for newer data collection approaches such as the experience sampling method (ESM) also often referred to as ecological momentary assessment or ambulatory assessment (Stone, Schneider, & Smyth, 2023). First introduced by Larson and Csikszentmihalyi (1983), ESM has become a data collection tool widely applied across different disciplines, such as medicine, economics, computer science, and behavioral sciences. Its primary idea is to repeatedly assess individuals' behavior, feelings, or thoughts on (pseudo-)random occasions in daily life. As the request to respond is often associated with some kind of audible signal, the repeatedly sent experience sampling (ES) questionnaires have historically often been called *beeps* (Csikszentmihalyi & LeFevre, 1989).

In survey research, the term *non-response (bias)* has become established to refer to missingness. To delineate the terminology used in ESM research, we follow scholars' suggestion and use the term *(non-)compliance (bias)* (van Berkel et al., 2020). In doing so, we aim to highlight the repeated nature of assessments. We thereby also emphasize that we are considering the special case where participants initially committed to participate in a study but then failed to respond for a portion of a study's ESM assessments. In contrast to one-time surveys, natural environments in which ESM studies are conducted come with even more reasons why participants might not answer specific beeps. The frequent need to answer ES questionnaires directly or in a timely manner adds the current context or what a person is doing as another momentary

hurdle in complying with a specific beep (Stone, Schneider, & Smyth, 2023). These daily hurdles are also reflected in the average non-compliance rates in ESM studies ranging between 10% to 30% (Wrzus & Neubauer, 2023). Although ESM has meanwhile established as a data collection tool in psychological research, there are still many open questions regarding the validity of the self-report measures, in particular with respect to potential (non-)missingness of the data. Our study aims to address this gap by using a multi-method approach to explore compliance in an ESM study with a comprehensive set of potential hurdles participants are faced with in their natural daily environments.

2.2.1 Scenarios of Missing Data in ESM Studies

Traditionally, methodological literature distinguishes three types of missing data: *missing completely at random* (MCAR), *missing at random* (MAR), and *not missing at random* (NMAR) (Little & Rubin, 1987; Thoemmes & Mohan, 2015).

MCAR means that the missing observations must be a true random sample of all observations. That is, the probability of missing an observation does not depend on any observed or unobserved variables (Little & Rubin, 1987). An exemplary scenario in ESM studies for this type of missingness would be if the app used for delivering the ES questionnaires sometimes randomly crashes, independent of any participant state (i.e., feelings, behaviors, thoughts) at the moment of crashing.

The less restrictive concept of MAR assumes that missingness depends only on the observed values and not on any unobserved values of the variables (Newman, 2014). An exemplary scenario in ESM studies would be if older participants were more likely not to respond to ES questionnaires but the age of all participants is assessed (e.g., in a pre-questionnaire) and missingness does not depend on any other variables that were not assessed, after accounting for participants' age (e.g., by including it as a control or auxiliary variable [1]).

Finally, the concept of NMAR assumes that missingness depends on the values of the missing variables themselves (Schafer & Graham, 2002). This is the case when the probability of missing an observation depends on variables that were not observed in

[1]Within the context of missing data, variables that only aim to enhance the performance of statistical methods handling missing data even if they are not particularly relevant to the scientific hypotheses of interest, are called auxiliary variables (Collins et al., 2001)

the data set or on the values of the missing variables themselves (van Ginkel et al., 2007). An exemplary scenario in ESM studies would be if participants systematically fail to answer a mood questionnaire when they are in a specific situation, such as being "in the midst of a marital dispute" (Stone, Schneider, & Smyth, 2023, p. 15), waiting for the dentist (which is probably unknown to the researcher), or, more general, whenever they are in a certain mood, for example, in a bad mood.

For the first two types of missingness, the methodological literature proposes easy-to-handle solutions for data analysis: MCAR observations can simply be neglected as listwise deletion will not introduce bias after data exclusion (although statistical power will decrease) (Allison, 2001). For MAR observations, well-known methods such as maximum likelihood estimation or multiple imputation can be used if researchers control for the variables that cause the missingness, even if this procedure does not explicitly model the process of missingness (Gelman & Hill, 2006; Mohan & Pearl, 2021). In contrast, NMAR observations come with certain (statistical) biases in the data analysis if the missingness is not explicitly modeled and, consequently, researchers are in danger of drawing wrong conclusions from their results (Gelman & Hill, 2006). To illustrate, imagine the following thought experiment: We conduct an ES study to examine the relationship between mood and the quality of social interactions. However, we do not consider the characteristics of the contexts our participants encounter when answering the ES questionnaires. Based on our study, we could reach null findings and conclude that mood is not related to the quality of social interaction. However, one explanation for the null findings could simply be that participants systematically did not respond to the beeps in certain contexts associated with bad mood (e.g., in a dentist's waiting room), so we cannot detect a significant association because of the low variance in mood ratings. Thus, the conclusion that mood is unrelated to the quality of social interaction would be biased because we did not take into account that mood depends on contextual characteristics, that is, data were missing systematically. Thus, we cannot consider missing values as MAR if we systematically omitted mood reports in certain contexts, for example, if participants' mood was exceptionally bad or good. Another example for NMAR in an ESM scenario is presented by Stone and Shiffman (2002): Investigating the interplay of chronic pain and psychological well-being, it seems natural due to the nature of the studied constructs that participants are less likely to

answer ES questionnaires while experiencing pain, which could in turn introduce bias to the estimated association between pain and well-being. To conclude, especially data NMAR might be associated with problems and fallacies in ESM research.

In summary, there are different plausible scenarios for missing data but little is known about whether they should be considered MCAR, MAR, or NMAR in the ESM setting. Insight into these mechanisms associated with missing data would help researchers to avoid introducing bias in their statistical analyses. Important countermeasures such as controlling for variables that affect the probability of missingness or, when applying multiple imputation, including these variables in the imputation model are already available in the statistical literature (Gelman & Hill, 2006; Graham, 2009). To use them sensibly, an important next step is to investigate the nature of missingness in ESM studies.

2.2.2 Sampling Biases in ESM Studies

Methodological ESM research has long recognized the importance of understanding missing data and scholars have started to investigate sampling biases in ESM studies from different perspectives.

2.2.2.1 Compliance at the Study Level. Most previous studies have examined participants' overall compliance rate, that is, the percentage of beeps answered, and its association with both study and person characteristics. Study characteristics have ranged from more general design elements such as the overall study duration to more specific ones such as the implementation of particular ESM sampling schemes, the number and duration of ES questionnaires and thus participant burden, or compensation incentives; person characteristics have ranged from participants' socio-demographic to psychological traits such as Big Five personality; (Courvoisier et al., 2012; Eisele et al., 2022; Harari et al., 2017; Hasselhorn et al., 2021; Ottenstein & Werner, 2021; Vachon et al., 2019; van Berkel et al., 2019). In more detail, some studies found negative associations between the overall compliance rate and ES questionnaire duration (Eisele et al., 2022), the number of study days, and the overall number of ES questionnaires (Ottenstein & Werner, 2021). In contrast, the implementation of specific sampling schemes (van Berkel et al., 2019) and monetary incentives (Harari et al., 2017) were found to be associated with higher compliance rates.

Overall the effects of study characteristics on compliance rate seem small or even negligible. Neither Wrzus and Neubauer (2023) nor a recent study by Hasselhorn et al. (2021) which experimentally manipulated different study characteristics found any effects of study characteristics on the overall compliance rate in ESM studies, except for incentivization. Moreover, design choices such as the number or duration of beeps are often determined by the research question at hand and thus cannot be easily adapted. Therefore, in our study, we consider study characteristics fixed and instead focus on person characteristics.

Gender has repeatedly been found to be related to overall compliance rate in previous studies. At the person level, the compliance rate was lower for male participants, and at the sample level, the compliance rate was lower for samples with a higher proportion of male participants (Rintala et al., 2019; Silvia et al., 2013). These effects, however, could not be replicated consistently (Howard & Lamb, 2023).

Depending on the field of ESM research, more specific person characteristics have been investigated in terms of overall compliance rate. For example, psychotic disorders were found to be related to decreased compliance (Sokolovsky et al., 2014; Vachon et al., 2019). In contrast, personality traits, for which an association with non-response has been hinted at by previous survey research (Rogelberg & Luong, 1998; Rogelberg et al., 2003; Satherley et al., 2015), have not been found to be associated to compliance in ESM studies (Courvoisier et al., 2012; Sun et al., 2020).

2.2.2.2 Compliance at the Beep Level. Few previous studies addressed participants' compliance rate at the beep level by modeling the probability of participants answering specific beeps.

Some researchers have explored the association between the compliance rate at the beep level and context-related characteristics. The selection of contextual characteristics under study has ranged from easily accessible smartphone data like battery or charging status (van Berkel et al., 2020) and physical activity features (McLean et al., 2017) parameters of psychological context determined from the responses to the previous beep (e.g., participants' mood or stress at the preceding beep)(Murray, Brown, et al., 2023; Sokolovsky et al., 2014). In addition, electronically activated recorders (EARs) have been used to collect audio snippets of participants' surroundings during

beeps to infer the current context of participants (Sun et al., 2020). The inclusion of contextual characteristics such as physical activity or audio indicators captured via EAR (e.g., whether participants were engaged in social interaction at the time of the beep) increased the overall accuracy for predicting compliance rate at the beep level by 0.5 to 2 percentage points (McLean et al., 2017; Sun et al., 2020).

Apart from these contextual characteristics, some studies have focused on participants' behavioral characteristics (Rintala et al., 2020; Sokolovsky et al., 2014). For example, if participants failed to answer the previous beep, they were more likely to miss the next beep (Rintala et al., 2020). Poly-substance users (i.e., participants using an illicit drug different to cannabis) were also more likely to miss a specific beep (Messiah et al., 2011). No effects on compliance rate at the beep level were, however, found for more general behaviors such as cigarette consumption during the last 30 days (Schüz et al., 2013; Sokolovsky et al., 2014) or aggressive behavior assessed at the previous beep (Murray, Brown, et al., 2023). This partial absence of effects of behavioral characteristics could be a type of methodological artifact: Studies have frequently used only delayed behavioral information from the previous beep as predictors for compliance at the respective beep but not information on behavior at the beep itself because this information is missing if participants fail to answer the respective beep (e.g., Murray, Brown, et al., 2023; Rintala et al., 2020; Silvia et al., 2013). Nonetheless, information about behavior at the moment of the beep response itself could be an additional important source for predicting compliance at the beep level.

2.2.3 How to Collect Data When They are Missing

Previous findings paint a mixed picture of characteristics associated with compliance in ESM studies. Effects and conclusions for person and behavior characteristics are small and often not consistent across studies (Stone, Schneider, & Smyth, 2023; Wrzus & Neubauer, 2023). In addition, the previous literature suggests that context characteristics add little to predicting compliance at the beep level (McLean et al., 2017; Sun et al., 2020). With a good portion of optimism, this could be considered good news for ESM researchers. If no characteristics are found to be systematically related to compliance, in the most optimistic interpretation, this could mean that missing data in ESM studies are simply missing completely at random. However, another explanation

for the unclear pattern of findings could be the limited methodological scope of previous research. For example, some studies have used analogous and rather easy-to-backdate ESM methodology such as paper and pencil or call-based sampling (e.g., Courvoisier et al., 2012; Rintala et al., 2020), specific samples (e.g., 9th and 10th grade smokers, Sokolovsky et al., 2014), or small sample sizes (e.g., n = 57, van Berkel et al., 2019). When trying to detect associations between context or behavior characteristics and compliance, this methodological scope might have led to problems with detecting effects. While person characteristics are often collected via surveys once at the beginning of a study, information on context and behavior is missing – by definition of missing data – when participants do not respond to a specific beep, that is, when they do not provide a self-report on their current context and behavior via the respective ES questionnaire.

To overcome this methodological hurdle, one promising approach is mobile sensing that provides passively collected information (Harari et al., 2016). With smartphones being omnipresent in our daily lives, they are not only perfectly suited to supersede devices previously used for sending beeps in ESM studies such as paper-and-pencil diaries or personal digital assistants (e.g., PALM). They also offer the possibility of continuously collecting a variety of data types without the active engagement or interruption of participants' day-to-day behavior, which in turn can be used to derive contextual and behavioral information, even if participants miss certain beeps (Elmer et al., 2022; Harari et al., 2016; Schoedel et al., 2023). Accordingly, scholars have recently pointed out the huge potential of using mobile sensing as a toolbox to gather further insight into compliance in ESM studies (Murray, Brown, et al., 2023; Sun et al., 2020), for example, by using GPS data instead of self-reported information on locations (Sokolovsky et al., 2014).

2.2.4 The Present Study

In this exploratory study, we adopt a multi-methodological approach and combine smartphone-based ESM with mobile sensing to investigate compliance and to obtain insight into the nature of missing data in ESM studies. In doing so, we address two research questions. In a first step, we investigate whether there are any characteristics at all that are systematically associated with compliance in ESM studies. If so, in a second step, we investigate which characteristics these are.

Therefore, our study focuses on two aspects that we think represent gaps in the current literature on compliance in ESM studies: First, most previous research has investigated overall but not beep level compliance. We think that zooming in on the beep level is an important next step to better understand overall compliance. What exactly makes participants miss a beep? Using information exclusively related to specific beeps might help us discover the (opposing) interplay of both more general participant characteristics and very specific contextual characteristics that might mask each other, finally leading to an inconsistent pattern or null findings at the overall level. Accordingly, our study focuses on compliance at the beep level. For this purpose, we use a large sample with 26,750 beeps sent to 592 participants collected across 10,856 days in total.

Second, many previous studies have focused on a small number of selected variables associated with compliance. That means that only person (Murray, Yang, et al., 2023) *or* only contextual (Boukhechba et al., 2018) *or* only behavioral characteristics (Sun et al., 2020) *or* non-extensive and incomplete combinations thereof (Courvoisier et al., 2012) have been of specific interest. But, in order to better understand compliance in ESM studies, a *comprehensive combination* of these different categories of characteristics is still pending. Thus, in our study, we use an integrative approach combining 402 variables from all three categories (person, context, and behavior) to examine compliance in ESM studies.

2.3 Method

The data for this study were collected in the Smartphone Sensing Panel Study (SSPS), an interdisciplinary research project at LMU Munich in cooperation with the Leibniz Institute for Psychology (ZPID; see Schoedel & Oldemeier, 2020). All procedures adhered to the General Data Protection Regulation (EU-GDPR) and received ethical approval. Not to go beyond the scope of this article, we focus our report on the procedures and measures relevant to our specific research question. A detailed description of the SSPS can be found in Schoedel and Oldemeier (2020).

2.3.1 Transparency and Openness

The study protocol of the SSPS was preregistered [2]. Due to the complexity of mobile sensing data and associated variable extraction procedures, our study adopts a purely exploratory perspective. Accordingly, we did not preregister our study but describe our approach transparently and in detail herein. Due to the privacy-sensitive nature of the mobile sensing data (e.g., timestamped logs in combination with GPS coordinates collected in daily life), we share our data set only as aggregated variables. However, we provide our preprocessing code, analysis code, and further supplemental material in our OSF repository[3] to make our complete data handling pipeline transparent. Data preprocessing and analyses were conducted using the statistical software R (version 4.1.0, R Core Team, 2022). For reproducibility of our analysis, we used the package management tool renv() (Ushey, 2021) and provide a complete list of all R packages used in this paper in the renv.lock file in the OSF repository.

2.3.1.1 Procedures. The initial sample of 850 participants from across Germany was collected with the help of a non-probability online panel provider according to quota representing the German population in terms of gender, age, education, income, religious denomination, and relationship status. In addition, participants had to be between 18 and 65 years of age, fluent in German, and for technical reasons be the sole user of a smartphone with Android version 5 or higher (see Schoedel & Oldemeier, 2020). Participants were compensated dependent on the number of study parts completed. After recruitment, participants were randomly assigned to one of two groups with a study duration of either three months ($n_{group1} = 191$) or six months ($n_{group2} = 659$).

Data collection started for all participants in May 2020. Participants were asked to install our self-developed Android-based mobile sensing app, called PhoneStudy[4], on their private smartphone for the respective study duration. Using the app, various data types (phone usage, Bluetooth connectivity, GPS, etc.) were continuously collected in the background of the device. Each month, participants were sent a link to a 30-minute online survey via the app. These online surveys included questionnaires on

[2] preregistration is available at https://doi.org/10.23668/psycharchives.2901
[3] https://osf.io/jw3bn/
[4] For information on the PhoneStudy project see https://phonestudy.org/en/

socio-demographic and psychological measures (for a complete overview of included measures, see Schoedel & Oldemeier, 2020).

The SSPS also included two 14-day ES waves (in July/August and September/October 2020) during which participants were asked to complete five-minute questionnaires on up to four occasions per day. The ES schedule was pseudo-randomized: Each day (from 7am to 10pm on weekdays and from 9am to 11pm on weekends) was divided into four equally sized time windows and two to four of these time windows were randomly chosen to schedule one ES questionnaire. The timing of a beep within a time window was again randomly chosen while maintaining a minimum interval of 60 minutes between two consecutive beeps. Participants were informed about the ES questionnaire via a notification as soon as they actively used their smartphone for the first time after the scheduled time for the respective beep. Accordingly, the time at which a beep was scheduled did not necessarily match the time at which the notification was presented on the participant's screen. If a participant did not use the smartphone in a time window in which a beep was scheduled, the beep was overwritten and thus the participant did not receive it. This procedure was chosen because our study design required a careful trade-off between sending ES questionnaires randomly but not provoking artificial smartphone usage and thus not distorting or interrupting participants' natural behavior (van Berkel et al., 2019).

2.3.2 Sample

For data quality reasons, we applied several exclusion criteria. For example, we excluded participants who decided to cancel their participation within the first day of the panel study or who had technical problems. As our central research focus was compliance in ESM studies at the beep level, we excluded participants if their study behavior suggested that they were not seriously taking part in the ES waves. This was particularly important to check, as individual study parts such as online surveys, ES, and mobile sensing were compensated independently of each other, and participants were not generally excluded from the panel study if they did not participate in all parts. Thus, we excluded participants, if they canceled their participation in the panel study prior to onset of the first ES wave, did not take part in at least one of the two 14-day ES waves, received fewer than 10 beeps, for example, as a result of participation

withdrawal during the ES waves, or did not react to the beeps (i.e., did not answer more than 5 beeps, or had an answer rate below 20%). These exclusion criteria were much less strict than the compensation criteria of the panel study (at least 14 beeps per ES wave); thus our study results do not apply only to "compliant" participants but can be generalized.

This resulted in a final sample of 592 participants, whose age ranged from 18 to 65 years with a mean of 41.7 years (SD = 12.9) with 55.3% of participants being male (n = 294) and 45.7% (n = 238) female. With respect to educational attainment, 0.6% did not have any degree, 15.4% had lower secondary education, 34.6% had junior secondary education, 29.3% did their A-levels, 19.5% graduated from university, and 0.6% had a PhD.

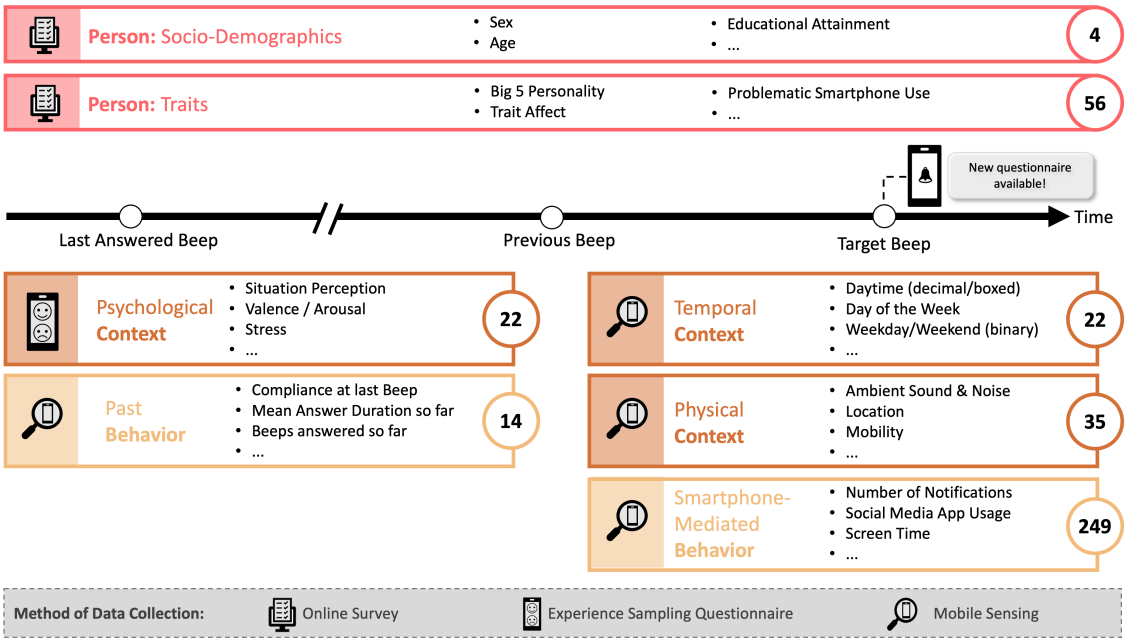
2.3.3 Measures

To handle the vast amount of data assessed via traditional self-reports, mobile sensing, and ES, we applied a predictive modeling approach, using various machine learning (ML) algorithms. Following ML terminology, we refer to the outcome variable as the target variable and the predictor variables as features.

2.3.3.1 Target: Missed Beeps. The response to ES questionnaires (short: beeps) served as the target variable. To avoid artificially provoking smartphone usage, participants only received beeps upon actual usage of their smartphone. A beep was considered as *answered* and therefore coded as 1 if the participant opened the ES questionnaire in the app within 15 minutes after the respective notification and completed the ES questionnaire within further 15 minutes after opening the app. Whenever a participant received a beep but failed to answer it, a beep was considered as *missed* and coded as 0. These included instances in which participants intentionally chose not to respond to a beep, such as by wiping away the notification, or in which participants did not respond to or finish an ES questionnaire within the 15-minute time limit. If however, a participant did not use the smartphone after a scheduled beep within the associated randomly selected time window and therefore did not receive a notification for the ES questionnaire, this case was not considered in our analysis.

2.3.3.2 Features: Person, Context, and Behavior. We extracted a total of *402* features. Not to exceed the scope of this report, we describe categories of extracted features with selected examples. However, a complete list of all features including a short description and the code that was used for feature extraction can be found as supplementary material in our OSF repository.

Figure 2.1
Overview of Features According to Categories and Data Collection Methods



Note. We were interested in compliance at the observational level, that is, how participants responded to a single beep, the target beep, at a given time. To this end, we used information about a participant's response to preceding beeps and information on the current beeps, as represented by the timeline. We assigned this information to different (sub)categories, represented by the colored boxes with selected feature examples. The icons on the left of the boxes depict the respective data collection method. Numbers on the right of the boxes indicate the total number of features for each subcategory. In all, 402 different features were extracted for each beep.

We assigned our features to three main categories: *Person, Context*, and *Behavior* (see Figure 2.1). Person features were assessed via the monthly online surveys. Some of the context features were assessed via ES questionnaires. The major part of the context as well as the behavior features were extracted from mobile sensing data assessed via the PhoneStudy app. Raw sensing data were logged as time-stamped data points stored with data type-specific information (e.g., app name for app usage logs, decibel values for ambient sound logs, longitude and latitude for GPS logs). User-smartphone interactions such as phone or specific app usage, notifications, or screen status were logged *event-based*, i.e., the app recorded data points whenever they occurred. In

comparison, GPS data were logged *interval-based*, that is at fixed time points every 10 to 60 minutes, depending on the respective smartphone model. Ambient sensor data such as sound or light were also logged on an interval basis but only between 6pm and noon so as not to put too much strain on the battery. To obtain an accurate picture of the user's physical context while conserving battery power, GPS and physical activities were additionally logged *change-based*, that is, whenever a change was detected. To do so, GPS and physical activity features were gathered via the Google Fence API[5], the Google Snapshot API[6], and the Google Activity Recognition API [7]. For more details on logging procedures see also Schoedel et al. (2023).

2.3.3.3 Person. The category *Person* comprises features related to participants' self-reported socio-demographics, traits, habits, and preferences. More specifically, we included socio-demographics, personality traits, media usage habits, technology affinity, problematic smartphone usage, and trait affect (see Table 2.1 for an overview of all measures). As problematic smartphone usage was assessed repeatedly in each of our six monthly online surveys, we used those assessments closest to the respective ES wave (i.e., online surveys assessed in the third and fifth study month for the first and second ES wave, respectively).

2.3.3.4 Context. The category *Context* comprises features related to the participant's situation when receiving a beep. The three subcategories *Temporal Context*, *Psychological Context*, and *Physical Context* focus on different levels of abstraction.

2.3.3.4.1 Temporal Context. Features included in this category characterize the time at which participants received beeps. We extracted features from the timestamps automatically recorded when the beeps were sent via the PhoneStudy app. This included the encoding of information on time as both decimal number and daytime category. For example, the timestamp 10:30:00am (Central European Summer Time or UTC + 2) was encoded as numerical (i.e., as 10.5) and categorical (in 2 hour time-boxes). In addition, time was encoded as information on day (weekday/weekend).

[5]https://developers.google.com/awareness/android-api/snapshot-api-overview
[6]https://developers.google.com/awareness/android-api/fence-api-overview
[7]https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionApi

Table 2.1
Overview of Self-Report Measures of the Feature Category Person

Measure	Instrument	Included Features	Reference
Socio-Demographics	Self-Created	4 single items	
Big Five Personality	BFSI	5 factors and 30 facets	Arendasy et al. (2011)
Media Usage Habits	Self-Created	12 single items	
Technology Affinity	TA-EG	4 facets	Karrer et al. (2009)
Problematic Smartphone Usage	Self-Created[a]	3 single items	Lee et al. (2014)
Trait Affect	PANAS[b]	2 facets	Watson et al. (1988)

Note. List of questionnaires used for the assessment of socio-demographics, traits, habits, and preferences. A complete list of single items can be found in the study protocol of the SSPS (Schoedel & Oldemeier, 2020) .
[a] Items were selected and translated from Lee et al. (2014).
[b] German translation by Breyer (2016).

2.3.3.4.2 Psychological Context. Features of this category describe the psychological context in which a beep was sent. As we do not have self-reported psychological context features if a beep was missed, we used the participant's self-report at the last answered beep as a proxy. Measures included in this category were (1) a two-item state affect rating according to the Circumplex Model of Mood (i.e., valence and arousal; Russell, 1980) and (2) a single-item stress rating. Participants were asked to rate all three items on a 6-point Likert scale. In addition, we included (3) a single-item rating for each of the eight situational DIAMONDS (Duty, Intellect, Adversity, Mating, pOsitivity, Negativity, Deception, and Sociality) asking participants how they perceived the current situation (Rauthmann et al., 2014). Participants rated on a binary scale for each dimension of situation perception if it applied or not (Rauthmann & Sherman, 2018).

2.3.3.4.3 Physical Context. Features from this category describe the current physical context, including location, mobility, and environmental cues at a very fine granular level (e.g., ambient light or sound). As these features are enabled by the mobile sensing component of the PhoneStudy app and therefore do not require active logging by participants, they were assessed regardless of whether participants responded to a beep. These features were derived using GPS, activity, light, and sound sensor data. These raw data were preprocessed according to a set of preprocessing pipelines. For example, GPS points were clustered per participant in order to identify each person's home and workplace coordinates (i.e., the center of the cluster in which a

participant was present most frequently between 1am to 5am for home vs. 10am to 4pm on weekdays for workplace). Subsequently, we classified whether a person was at home or at work for a given beep based on actual GPS coordinates and their accuracy logged right before the respective beep. Additionally, we calculated, for example, the distance from home, whether the participant was likely to be traveling (e.g., by car or train), or whether they were in a bright or noisy environment at the time a beep was sent. To get a comprehensive picture of the physical context at the time a beep was answered or missed, we chose a time window of 60 minutes before the respective target beep to aggregate the raw sensing data points for this feature category.

2.3.3.5 Behavior. Features included in the category *Behavior* describe active behaviors in the time window before receiving a beep.

2.3.3.5.1 Smartphone Usage. This category includes features on participants' smartphone-mediated behaviors. Features were extracted based on the timestamped smartphone logs within 60 minutes before the respective target beep. We included general smartphone usage (e.g., time spent on screen, number of incoming calls), smartphone notifications (e.g., number of notifications, latency between receiving notifications and unlocking the smartphone), and app usage (e.g., Communication, Photo, News, or Music). Single apps were categorized to psychologically meaningful categories based on the system proposed by Schoedel et al. (2022). Thereby, we followed the proposed inclusion of app categories with sufficient inter-rater agreement (i.e., Cohen's kappa $> .60$).

2.3.3.5.2 Past Behaviors. This feature category is associated with participant responses to past beeps. Features included in this category were, for example, the number of sent beeps up to the time of the respective target beep or the mean answer latency, that is, the average time between receiving a beep notification and the time when the participant started answering the respective ES questionnaire. It also comprised the mean answer duration (i.e., the average time needed for completing ES questionnaires), the mean answer rate, and whether the previous beep was answered. In these features we only coded information on behaviors that occurred *before* the respective target beep. We did this to design our prediction model (see next section for

more details) to be applicable in real time in future ESM studies. That is, if we would like to apply our prediction model in a new study to predict whether a participant will respond to the next target beep, we would only have information collected up to that specific time point. In this case, features such as the overall answer rate in the study would not be available if a participant is only halfway through.

2.3.4 Data Analysis

2.3.4.1 Machine Learning. We used the previously described features (in total, 402 before and 190 after target-independent preprocessing) to predict whether a specific target beep was answered (or missed) at the observational level. This setting corresponds to a binary classification task. Machine learning predictions were conducted using the mlr3 environment in R (Lang & Schratz, 2023).

2.3.4.2 Preprocessing. We applied both target-independent and target-dependent preprocessing. The first included the replacement of extreme outliers in each feature (±4 standard deviations from the mean) by missing values. We applied this procedure to exclude anomalies in the data most likely caused by technical logging errors, while extreme expressions of features were preserved in the data. Further steps were the removal of features with more than 90% of values missing across all observations and the removal of features with zero or near-zero variance as defined by the default settings of the caret package (i.e., classification of a predictor as having variance near-zero if the percentage of unique values in the samples was less than 10% or if the ratio of frequency of the most common value to the frequency of the second-most common value was greater than the ratio of 95%/5%, Kuhn, 2008). All subsequent target-dependent preprocessing steps, namely scaling and missing data imputation, were integrated into the resampling procedure to avoid overfitting and leakage problems (i.e., information from the test set "leaking" into the training set in the prediction task). As some sensing components were only logged at specific time intervals throughout the day (e.g., ambient sensor data), some features showed a substantive amount of missing values. Imputation was conducted via histogram and tree-based learners, respectively, using the methods implemented in the mlr3pipelines package (Binder et al., 2021).

Regarding preprocessing, we also tested approaches to account for the class imbalance in our target variable such as the assignment of class-dependent weights or oversampling (Sterner et al., 2023) and model-specific hyperparameter tuning (e.g., lambda or mtry for elastic net and random forest, respectively). We reran the models without the described exclusion of extreme outliers (±4 SD of the mean) for which results are provided in the online materials in the OSF repository. However, none of these approaches led to considerable performance improvements but did considerably increase computational costs. Therefore, we report all results based on the default settings in the respected software packages for all hyperparameters.

2.3.4.3 Models. We benchmarked three models for the prediction task, namely (1) standard logistic regression, (2) elastic net regularized logistic regression (hereafter referred to as elastic net; Zou & Hastie, 2005) as implemented in the glmnet package (i.e., cv.glmnet; Friedman et al., 2010), and (3) random forest (Breiman, 2001) as implemented in the ranger package (Wright & Ziegler, 2017). We selected these models as they facilitate the comparison of a familiar approach for classification in the behavioral sciences - ordinary logistic regression - with two more sophisticated, common machine learning algorithms, representing a regularized linear model (i.e., the elastic net) and a non-linear tree-based model (i.e., the random forest). Random forests consisting of many single decision trees automatically take into account interaction effects between variables, because the partitioning within a tree may depend on different predictor variables (for a more detailed introduction to random forest models, see Module 2 in Pargent et al., 2023). Elastic net models, on the other hand, are able to consider interaction effects only if explicitly stated in the model equation. However, we decided against including interaction terms in our analysis, as this would have enormously increased the (already large) number of predictors. Both algorithms are especially well-suited to the modeling problem at hand as they can handle identification or computation issues due to a large number of features and linear dependency among these (e.g., between Big Five factors and facets) (Dormann et al., 2013; Hastie et al., 2009; Pargent et al., 2023). Apart from the described models, we trained a baseline model that served as a reference point to benchmark the other models. This baseline model predicted the most common class of the target variable (i.e., that a given

beep was answered) among all observations in the respective training set by assigning probabilities corresponding to the relative frequency of the class labels in the training set without considering any of the features (Lang et al., 2019).

2.3.4.4 Performance Evaluation. We estimated the prediction performance for the different models by using 10-fold cross-validation with 10 repetitions (10x10 CV) as resampling procedure. Because the basic idea behind the ESM is to collect repeated measurements within individuals, we considered the assumption of independence of residuals to be violated. To account for the nested structure of our data, we applied blocked resampling with participants' unique identifiers as the blocking variable. By using this blocked resampling strategy, we ensured that all observations of one individual completely went into either the test or the training data set but were never split up in order to counteract overoptimistic performance estimates (Dragicevic & Casalicchio, 2020).

To evaluate our models' performances, we used the area under the receiver operating characteristic (ROC) curve. In binary classification tasks, the ROC curve considers both, the true positive rate (sensitivity) and the false positive rate (1 - specificity) of a model to evaluate a model's predictive ability as a function of different discrimination thresholds. Integrating over the ROC curve yields the area under the curve (AUC) metric, which can be thought of as an "integrated measure" between both sensitivity and specificity. The AUC can be interpreted as the probability of the model ranking two randomly selected beeps (one of each class, one answered, one missed) correctly (i.e., the calculated probability of being answered is higher for the answered beep than for to the missed beep) (Viaene & Dedene, 2005). Describing a probability, AUC values can range from 0 to 1. The AUC metric can be considered robust to class imbalance (Boughorbel et al., 2017). A naïve guessing approach, as applied by our baseline model yields an AUC of .50 (Fawcett, 2006). Accordingly, AUC values smaller and larger than .50 represent worse or better prediction performances than our baseline model, respectively. As a more intuitive performance measure for classification, we additionally report Matthew's correlation coefficient (MCC), which is a method of calculating the Pearson product-moment correlation coefficient between actual and predicted values based on the confusion matrix (Chicco & Jurman, 2020). The MCC ranges from -1

to +1, equals zero for the baseline model's predictions, and produces high scores only if good prediction results are obtained in all of the four confusion matrix categories (Chicco et al., 2021). This is why it can be considered a more reliable statistical measure compared to more popular metrics such as accuracy, especially for cases with strong class imbalance (Chicco & Jurman, 2020).

2.3.4.5 Model Interpretation. To gain insights into the prediction models, we performed follow-up analyses by applying two interpretable machine learning tools. As a preview of our results, the elastic net model achieved the highest average prediction performance. Therefore, we decided to focus our interpretable machine learning analyses on this model and to use model-specific techniques exclusively for the elastic net model.

2.3.4.6 Single Feature Importance. We investigated which features were most predictive (i.e., informative) of answered beeps. Accordingly, we estimated standardized beta coefficients and used them as a metric for single feature importance. Due to the large number of features, we did this exclusively for the features that were selected by the elastic net model. In the elastic net's feature selection, the regression coefficients of uninformative features are shrunken to zero. This is done based on shrinkage parameters that are selected using a model-inherent cross-validation. To account for the random component introduced by this cross-validation, we trained 100 separate elastic net models and calculated the rate of inclusion into the final models, the average beta coefficients, and the 10-90 percentile ranges for the average beta coefficients across these 100 iterations.

2.3.4.7 Grouped Feature Importance. As our features can roughly be clustered into the three categories (person, context, and behavior) with several subcategories, we were also interested in whether one of these subcategories was particularly relevant (i.e., informative) for our model's predictions. Accordingly, we conducted a *leave-one-group-out* analysis by comparing the prediction performance of the elastic net model for the full feature set containing all features of all categories with its performance after the features of each of the different subcategories had been excluded. Again, we used a 10x10 CV scheme. If the AUC decreased after excluding a specific feature subcategory, this indicated that the respective feature subcategory was important for the prediction performance.

To anticipate our results, the *past behavior* and *physical context* features led to noteworthy decreases in prediction performance in the *leave-one-group-out* analysis. To further explore the importance of these feature subcategories, we additionally implemented a *leave-one-group-in* analysis. Thus, we trained two models that only used *past behavior* or the combination of *past behavior* and *physical context* feature for prediction.

As the *past behavior* subcategory was by far the most important one, we additionally considered the possible masking effect of this subcategory in relation to all other subcategories by implementing a hierarchical leave-one-group-out approach. Thus, we first excluded features of the subcategory *past behavior* from the feature set and then compared the prediction performance changes after excluding each of the other remaining feature subcategories. For example, one could assume that relevant associations between the personality trait conscientiousness could be related to compliance at a given beep but would then necessarily also be related to compliance at the previous beep. Accordingly, including compliance at the previous beep as a predictor would at least to some extent include, control for, or *mask* effects of the person trait. When, however, excluding the auto-regressive effect of compliance at the previous beep, effects of the personality trait conscientiousness could be detected.

2.4 Results

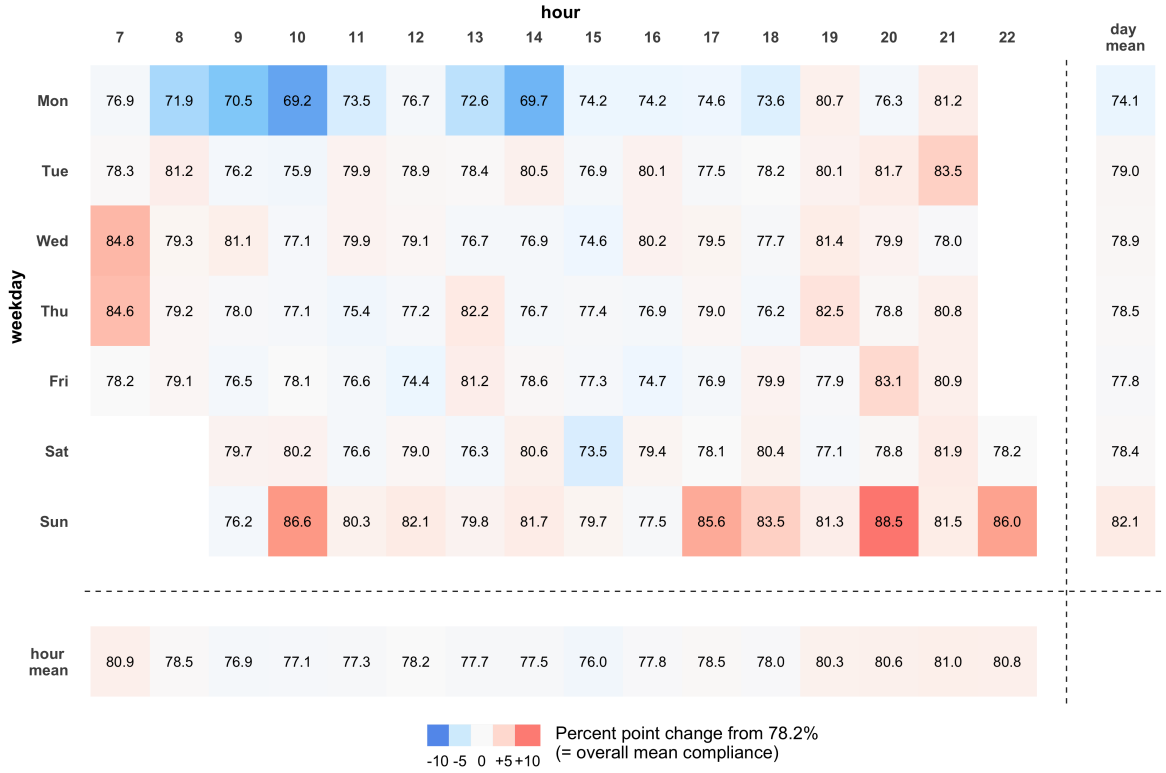
2.4.1 Descriptive Statistics

Overall, participants received 26,750 beeps of which 20,907 were answered, aggregating to an overall compliance rate of 78.2%. At the level of single persons, participants received, on average, 45.2 beeps of which, on average, 35.3 were answered. With a value of 78.6%, participants who reported being male had similar average overall compliance rates as participants who reported being female, at 78.8%. Similarly, average overall compliance rates were comparable for different age groups: 77.8%, 79.0%, and 78.9% for participants aged 18-29, 30-49, and 50+ years, respectively.

In addition, as presented in Figure 2.2, we descriptively investigated the average compliance rates at the beep level and their deviation from the average compliance rate at the overall level (i.e., 78.2%) separately for each combination of weekday and

Figure 2.2

Deviation of Beep Level Compliance Rate from Overall Compliance Rate Depending on Weekday and Daytime



Note. Numbers in the grid cells represent the average compliance rate at the beep level across all participants for the respective weekday and daytime combinations. Daytimes of beeps ranged from 7am/9am to 9pm/10pm on weekdays and weekends, respectively. Right and bottom margin cells represent the beep level compliance rates averaged for the respective weekday and daytime. The degree of coloration represents the degree of deviation from the average overall compliance rate of 78.2%, with reds representing higher and blues representing lower compliance rates at the beep level.

daytime in hourly time bins (see main figure area), for weekdays irrespective of daytime (see right margin), and for daytime irrespective of weekdays (see bottom margin).

On average, the compliance rate at the beep level was, in comparison to the compliance rate at the overall level, lower on Monday mornings (blue cells on the top left) and higher on mornings in the middle of the week and Sunday evenings (red cells in the middle left and on the bottom right). There were no noticeable deviations neither for specific weekdays, irrespective of daytime, nor for specific daytimes, irrespective of weekdays. Please note that due to our ESM design, these descriptive patterns have to be interpreted conditionally, that is, under the condition that participants had to actively use their smartphone in order to be notified of an ES questionnaire and, consequently, to be able to answer (or miss) a beep.

Not to go beyond the scope of this article, further descriptive statistics of our 402 features and their correlation with compliance at the beep level can be found in our OSF repository. Additionally, in the OSF online repository, we provide a descriptive overview of the total number of sent beeps across all participants for each day, daytime, and day × daytime combination.

2.4.2 Prediction of Compliance at the Beep Level

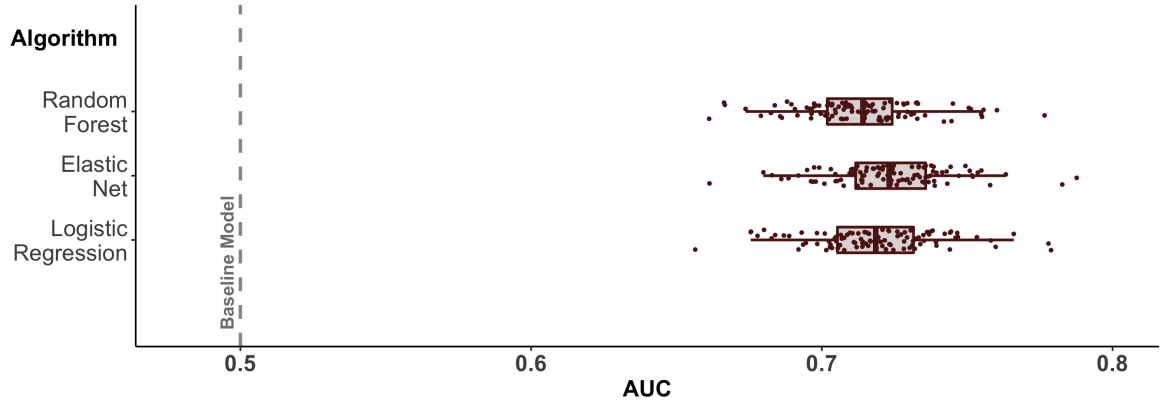
Our central research question was whether there are characteristics that are systematically associated with compliance at the beep level. We applied a machine learning approach to condense the information in our large set of investigated features. To briefly summarize, we did find indications for a compliance bias at the beep level. We compared different models, and they all performed better than our baseline model (AUC = .500). That is, the models were all able to grasp systematic variance in the collection of *person*, *context*, and *behavior* features to make predictions for whether participants answered a beep. In comparison, the standard logistic regression model (M_{AUC} = .719), the elastic net model (M_{AUC} = .723), and the random forest model (M_{AUC} = .713) achieved similar mean prediction performances, but the elastic net model slightly outperformed the other two. The distributions of prediction results across the 100 resampling iterations resulting from our applied 10x10 CV scheme are depicted for all models in Figure 2.3.

When using the MCC as the performance evaluation metric, the linear models (M_{MCC} = .217 for standard logistic regression and M_{MCC} = .194 for elastic net regression) also outperformed the non-linear random forest model (M_{MCC} = .129). Moreover, all three models were better than the baseline model (M_{MCC} = .000).

2.4.3 Interpretation of Compliance Predictions at the Beep Level

Having found indications for a compliance bias at the beep level, we conducted a follow-up analysis to explore which characteristics in particular were predictive of whether participants missed a beep. As mentioned, we only considered the elastic net model, which had the highest AUC in the benchmark.

Figure 2.3
Prediction Performances Across Iterations of Repeated Cross-Validation



Note. Distribution of the area under the operating characteristic curve (AUC) across the resampling iterations of the applied 10x10 CV scheme for random forest, elastic net, and logistic regression models. The gray dotted line at an AUC of .500 represents the prediction performance of the baseline model. AUCs of the single iterations are represented by single dots. The boxes contain all values between the 25% and 75% quantiles. Their middle line indicates the median. For presentation clarity, the AUC scale was cutoff at .500 and .800.

2.4.3.1 Features in Their Individual Role. We trained 100 elastic net models and, for each model, examined which features were most important in terms of their absolute mean standardized beta coefficients. Based on this, we extracted the top 20 features across the 100 models. In doing so, we found that many features (more than 200) were equally important, that is, they had the same mean absolute beta values. Therefore, we applied more strict selection criteria and only included features that had an average standardized beta coefficient > 0.05 or were included in at least 90% of the 100 elastic net models. We present the resulting list of the top 20 features in Table 2.2.

The table shows that features of all categories and subcategories (except *psychological context*) were represented among the 20 most important features. Features of the category *past behavior* particularly stood out, as the *mean answer rate (so far)* had the highest averaged standardized beta coefficient and was by far the most informative feature for the elastic net model's predictions. The two next most important features (*mean answer latency (so far)* and *compliance at last beep*) were also from the category *past behavior*.

Features of the category *physical context* were also represented frequently among the top 20. Accordingly, features such as whether participants were at home or at work, in an environment louder or brighter than a specific decibel or lumen value, or in a

rail or 4-wheel vehicle were consistently included in the elastic net models, with mean standardized beta coefficients ranging up to 0.10. Apart from these, weekday was the only feature of the *temporal context* feature category appearing in the top 20.

Comparing the higher level categories of *behavior*, *context*, and *person*, the latter was the least represented among the top 20 features. *Socio-demographics* and *traits* were, on average, among the features with the lowest standardized beta coefficients and inclusion rates. All person features included in the top 20 (i.e., age, dutifulness, and technology enthusiasm) had standardized beta coefficients below .05.

Table 2.2
Top 20 Important Features in the Elastic Net Models

			Standardized Beta Coefficients		
Variable	Group[a]	% Incl.	*M*	*SD*	10-90% Perc.
Mean Answer Rate (so far)	B2	100	0.50	0.01	[0.49; 0.51]
Mean Answer Latency (so far)	B2	100	-0.14	0.01	[-0.14; -0.13]
Compliance at Last Beep (binary)	B2	100	0.12	0.01	[0.11; 0.13]
Participant at Home (GPS)	C3	100	0.10	0.01	[0.09; 0.11]
Number of Missed Beeps Prior	B2	100	-0.08	0.01	[-0.09; -0.07]
Participant in Rail Vehicle	C3	100	-0.07	0.00	[-0.08; -0.07]
Age	P1	100	0.04	0.01	[0.03; 0.05]
Participant at Work (GPS)	C3	100	0.04	0.01	[0.02; 0.05]
Number of Events Louder Than 55 db	C3	100	-0.04	0.01	[-0.05; -0.03]
Number of Unique Apps Used	B1	100	0.03	0.01	[0.02; 0.04]
Tech.-Enthusiasm (Tech.-Affinity Subfacet)	P2	100	-0.03	0.01	[-0.04; -0.02]
Number of Unique App Categories Used	B1	100	0.03	0.01	[0.02; 0.05]
Duration of Finance Apps Used	B1	100	0.03	0.01	[0.02; 0.04]
Weekday (1=Monday)	C1	100	0.03	0.01	[0.02; 0.03]
Answer Latency at Last Answered Beep	B2	100	-0.02	0.01	[-0.04; -0.01]
Min. Latency of App Notification Usage	B1	96	0.02	0.01	[0.01; 0.04]
Number of App Usages	B1	99	0.02	0.01	[0.01; 0.03]
Dutifulness (Conscientiousness Subfacet)	P2	96	0.02	0.01	[0.01; 0.03]
Number of Events Brighter Than 10 Lumen	C3	99	0.02	0.01	[0.01; 0.03]
Participant in 4-Wheel Vehicle	C3	100	-0.02	0.00	[-0.02; -0.02]

Note. Table of top 20 features as identified from 100 iterations of elastic net model. Features are ordered with respect to their mean standardized beta coefficient across all iterations in which they were included into the model. Some column headings have been abbreviated relative to the published manuscript to fit the layout of the present dissertation. Column '% Incl.' indicates in how many iterations the latter was the case (i.e., coefficient was not shrunken to 0). Criterion for inclusion of features in this table was an inclusion rate of at least 95% (i.e., feature was selected in at least 95 elastic net iterations) or an absolute mean standardized beta coefficient equal to or greater 0.03. Column '10-90% Perc.' contains the 10-90% Percentiles of the standardized beta coefficients.
[a]Group column indicates feature category:
P1 = Person: Socio-Demographics, P2 = Person: Traits,
C1 = Temporal Context, C2 = Psychological Context, C3 = Physical Context,
B1 = Smartphone Usage, B2 = Past Behavior

2.4.3.2 The Role of Features as Groups. Besides considering individual features, we also explored their informativeness in their group constellations. Note that we assigned our features to seven subcategories of *person*, *behavior*, and *context*

characteristics. Figure 2.4 shows the relevance of each subcategory by plotting the average prediction performance (quantified by the AUC across the resampling iterations of the applied 10x10 CV scheme) when each feature category was excluded. We found the largest decrease in prediction performance with the exclusion of the category *past behavior*. The average prediction performance decreased from $M_{AUC} = .723$ (when the category was included for prediction) by .134 to $M_{AUC} = .590$ (when the category was excluded for prediction). The exclusion of other feature categories resulted in smaller changes in the prediction performance, with decreases of .007 for the category *physical context*, followed by .004 for the category *smartphone usage*. Excluding features of the categories person (*socio-demographics* and *traits*), *temporal context*, or *psychological context* resulted in average decreases of below .001.

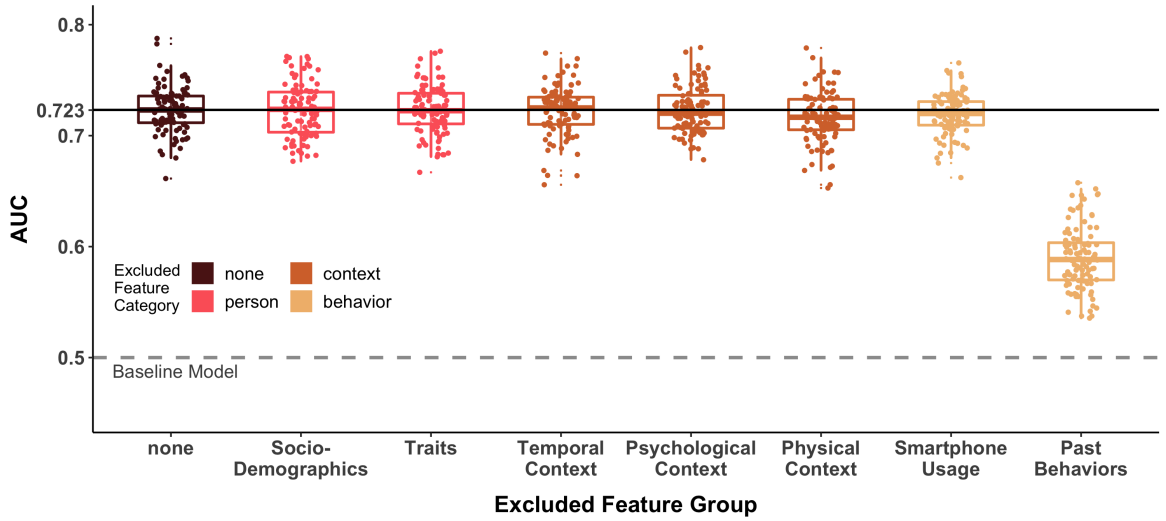
In summary, features from the category *past behavior* were by far most important in predicting compliance at the beep level, followed by features of the category *physical context*. Thus, the results of this leave-one-group-out analysis are in line with both the single feature importance analysis and the leave-one-group-in analysis. In the latter approach, we trained two additional models using *only* the features of the category *past behavior* or a combination of the categories *past behavior* and *physical context*. Both models produced comparable results to the model with features from all categories, with average prediction performance decreasing slightly by .013 and .005, respectively, compared to the full model.

To obtain further insight into which features are relevant for predicting compliance in the ESM setting, we also explored a possible "masking of effects" by the category *past behavior*. That is, we were interested in whether other feature categories are important beyond the dominant category of *past behavior*. To this end, we reran the leave-one-group-out analysis. This time, however, the full model included all features except the *past behavior* category, and in each leave-one-group-out model, one of the other subcategories was additionally excluded. We found that the resulting models achieved lower overall prediction performances compared to the original prediction models, with M_{AUC} ranging from .577 to .590. However, there were no major changes in the rank order of the different feature subcategories. Features of the categories *physical context* and *smartphone usage* were still the two most important feature subcategories

and *traits* remained to carry very low predictive information. We present the results for this additional analysis in the supplemental material in our OSF repository.

Figure 2.4

Prediction Performance of Elastic Net Models after Exclusion of Specific Feature Groups



Note. Distribution of the area under the operating characteristic curve (AUC) across the resampling iterations of the applied 10x10 CV scheme for the elastic net models after exclusion of each single subcategory. The boxplot in dark brown represents prediction performances with the initial full set of features (i.e., no exclusion of categories). For clarity, we include the median performance of this model as a solid black line. The remaining boxplots represent the performance (ordered and colored by feature subcategories) when one of the subcategories was excluded in each case. The gray dotted line at an AUC of .500 represents the prediction performance of the baseline model.

2.5 Discussion

The present study was designed to provide new insights into characteristics related to compliance and consequently the nature of missing data in ESM studies. To this end, we adopted a multi-method approach and combined various *person*, *behavior*, and *context* features collected via surveys, ES questionnaires, and passive mobile sensing to predict compliance at the beep level in an ESM study with 592 participants and more than 25,000 observations collected over several weeks. We used machine learning techniques and found empirical indicators of a compliance bias. Using the interpretable machine learning toolbox, we explored which characteristics were most informative in predicting compliance at the beep level individually and as aggregated feature categories. Features from the subcategory *past behavior* were by far the most relevant, followed by features in the category *physical context*. *Person* and *psychological context* features were of least importance. In the following sections, we discuss our results and how they help shed new light on compliance in ESM studies.

2.5.1 Predictability of Compliance in ESM Studies at the Beep Level

We found that participants' actual behavior - namely whether a beep was responded to or not - was predicted above chance in our ESM study. That is, we found systematic associations between compliance at the beep level and *person*, *behavior*, and *context* features. Each of the three ML models outperformed the baseline model. The linear models (standard logistic regression and elastic net) were not inferior to the non-linear random forest model which has already been observed in previous literature applying machine learning approaches to psychological research questions (Christodoulou et al., 2019; Pargent & Albert-von der Gönna, 2019; Schoedel et al., 2023). One possible explanation for this could be that the true underlying associations are indeed linear and as such could be captured somewhat better by the linear models than by the non-linear random forest model, which can only approximate smooth or linear relationships (Grömping, 2009; Hastie et al., 2009). In addition, non-linear models have problems capturing truly non-linear relationships when measurement error is present in the predictor or outcome variables (Jacobucci & Grimm, 2020). Because some of our features were psychological constructs, this reasoning may also have applied to our study and thus also limited the potential of the non-linear random forest model. Simulation studies showed that this effect is even exacerbated with smaller samples as the linear model is more impervious to sample size (Jacobucci & Grimm, 2020). Our study with 592 participants and more than 25,000 beep level observations would have been among the largest 3% according to a recently published meta-analysis examining compliance in ESM studies (Wrzus & Neubauer, 2023). However, in the machine learning context, sample sizes of several thousand people are not uncommon (Rosenbusch et al., 2021).

To compare the ranking of our models' performance relative to previous studies, we first considered commonly used rules of thumb. With mean performance metrics (AUC) exceeding .700, our models performed at a level which would be considered acceptable (Hosmer Jr. et al., 2013). We additionally inspected the strength of the association between the actual and the predicted response to beeps as a further evaluation metric. The correlations ranged between .129 and .217, so they were low to medium. To summarize, our models were able to predict compliance at the beep level, but the

prediction performance was far from perfect. Thus, despite using a large variety of *person*, *behavioral*, and *context* variables, we found little compliance bias at the beep level. However, given the context - an increasing number of ESM studies across disciplines - even a small compliance bias could be meaningful for the validity of research findings (Götz et al., 2022). Accordingly, the magnitude of research findings biased due to missing data could be considerably decreased if researchers across disciplines explicitly considered compliance bias in ESM studies, for example by including control or auxiliary variables to statistically counteract (Newman, 2014).

Second, we also wanted to compare our results more specifically with effects found in psychological studies addressing similar research questions. This proved to be a challenging task, however, as most previous studies have used an explanatory modeling framework rather than a predictive one (McLean et al., 2017; Sun et al., 2020). They reported in-sample effects, but we evaluated our models out-of-sample, or how they performed on resampled, and thus unseen, observations when predicting compliance at the beep level (Shmueli, 2010). While explanatory modeling is an important strategy to gain a better understanding of psychological processes, psychology as a research discipline has been criticized as strongly focusing on explanation but neglecting prediction (Yarkoni & Westfall, 2017). By combining ideas from explanatory and predictive modeling, psychology has the opportunity to extend its focus and thus increase the generalizability and reproducibility of research results (Hofman et al., 2021). Our study contributes to this debate (Rocca & Yarkoni, 2021; Yarkoni & Westfall, 2017) by applying predictive modeling and aiming at the accurate prediction of actual response behavior. This data-driven approach can help in developing ideas for underlying (causal) mechanisms or generating new hypotheses for explanatory modeling (Shmueli, 2010). Especially for the objective of the present study – the identification of variables linked to participants' missing beeps – predictive modeling was a useful approach because it allowed us to condense information included in a broad set of multi-methodologically collected variables.

2.5.2 Differential Importance of Person, Behavior, and Context Features

Because our models were able to systematically grasp variance in the large set of *person*, *behavior*, and *context* features, we explored which features were related to whether participants missed specific beeps.

2.5.2.1 Past Behavior Predicts Future Behavior. The results of our follow-up analyses provided consistent evidence that study-related past behaviors were most relevant for predicting compliance at the beep level: In particular, participants' average preceding compliance was by far the most informative feature. Considered individually, the top three most relevant features for predicting compliance belonged to the feature category *past behavior*. In the leave-one-group-out analysis, the category *past behavior* was also by far most important. To illustrate, the decrease in prediction performance related to the exclusion of all features of this category was higher than the sum of performance decreases caused by excluding all other feature categories individually. Moreover, the leave-one-group in analysis showed that a model only considering *past behavior* features was able to achieve a prediction performance only slightly inferior to the full model including information of all features from all categories.

The importance of (study-related) past behavior for predicting compliance behavior at the beep level is in line with a "classic" finding in psychology: Past behavior predicts future behavior (Albarracin & Wyer Jr., 2000; Ouellette & Wood, 1998). This has been found consistently in different areas, such as blood donation, physical exercise, or voting to name but a few (Ferguson & Bibby, 2002; Rodrigues et al., 2021; Rogers & Aida, 2011). According to previous literature, "well-practiced behaviors in constant contexts recur because the processing that initiates and controls their performance becomes automatic. Frequency of past behavior then reflects habit strength and has a direct effect on future performance" (Ouellette & Wood, 1998, p. 54). Applied to the ESM setting of our study, this could mean that beep level compliance behavior might have become automated over time in the constant ESM study setting and thus proved to be the most informative predictor.

In contrast to past (study-related) behaviors, smartphone use such as calls or app use immediately before a certain beep was far less relevant for predicting compliance

in ESM studies - both individually and when considered as a feature category. The number of unique apps used (in the past 60 minutes) was the most important feature from this category and could be considered a proxy for diversity of smartphone use. However, single feature effects from this category were very small. This could be related to the fact that there was some asymmetry in the resolution of the target behavior (i.e., snapshot at a specific time point) and the extracted features (i.e., snapshots aggregated over 60 minutes). However, it would also be plausible that digital smartphone use represents a different class of behavior than (analog) study-related behaviors and is therefore less informative for predicting compliance at the beep level.

2.5.2.2 Physical Context Matters (a Bit). *Context features*, particularly those related to *physical context*, played some role – albeit a much smaller one compared to *past behavior*. In line with this, the leave-one-group-in analysis showed that using a combination of *past behavior* and *physical context* features without consideration of all other feature groups achieved a prediction performance that can be considered equivalent to the full model containing information of all features. In more detail, information on whether a participant was at home at the time of receiving a beep was among the most informative features for predicting (non-)compliance. More precisely, being at home was associated with a higher probability of responding to a given beep. Similarly, being at work was associated with a higher probability of answering a beep. Both GPS-based location features have in common a relatively low mobility. That is, participants usually stay at home or at work for relatively long periods. Thus, our results might indicate that features associated with low mobility are associated with a higher probability of responding to a given beep. In line with this interpretation, features of high mobility such as being on a train or in an automobile were associated with a lower probability of responding to a given beep in our study. Overall, this finding is in line with previous studies that found increased compliance when participants stayed at specific locations (e.g., when being at food places or at home; Boukhechba et al., 2018; Rintala et al., 2020) and decreased compliance when participants had a higher level of physical activity (McLean et al., 2017).

Besides features informing about mobility, other contextual features were informative for predicting compliance, especially those enabling a high resolution of physical

surrounding. For example, the number of events louder than 55 decibels or brighter than 10 lumens in the hour before a specific beep were among the top 20 predictors for compliance at the beep level. Thus, *physical context* was related to whether participants reacted to beeps. Note, however, that this finding has to be interpreted with caution, as ambient noise and sound were only measured between 6pm and noon in our study and therefore might have been confounded with temporal information that was assigned to the category *temporal context*. For time features, we found patterns contrasting to previous studies (e.g., Csikszentmihalyi & Hunter, 2003; Rintala et al., 2020). For example, Csikszentmihalyi and Hunter (2003) found decreased compliance rates on Sundays, whereas our study found compliance to increase with progression of the week from Monday to Sunday (indicated by inclusion of the feature *weekday* in the top 20 and its positive standardized beta coefficient). Likewise, on a descriptive basis, Mondays and Sundays were the days with the lowest and highest average compliance respectively in our study. This result, which contrasts with previous literature, may be related to our scheduling and notification approach. In our study, participants were only sent a beep if they actually used their smartphone in the time interval after a scheduled beep. We applied this procedure to capture natural smartphone behavior (van Berkel et al., 2019). Therefore, our participants received beeps only if they had time to use their smartphone, irrespective of day. On free days, such as Sundays, they might have had more time to respond to a beep than on work days. When participants received beeps on Monday mornings, they might have been more likely to dismiss it as they probably used their smartphones, for example, to work through their after-weekend e-mails at work thus experiencing a higher level of stress and therefore responding to fewer beeps (Pindek et al., 2021).

2.5.2.3 The Minor Role of Psychological Features. Finally, the included psychological features contributed little to predicting compliance at the beep level. In more detail, in the *person* category, only age, technology enthusiasm (a subfacet of technology affinity), and dutifulness (a subfacet of the Big Five dimension conscientiousness) were informative, albeit at comparatively low levels. *Psychological context* features such as mood or stress were not among the 20 most important features. Accordingly, removing the categories of *socio-demographics*, *traits*, and *psychological*

context in our grouped feature importance analyses resulted in a negligible reduction in prediction performance. This was the case even when the *past behavior* features were removed first and then additionally the *trait* features, arguing against a masking effect of the past behavior features. Our results for person features are in line with previous research, which has also found no or at most very little systematic non-response bias introduced by person characteristics such as personality traits (Courvoisier et al., 2012; Sun et al., 2020).

Regarding *psychological context* features, our results are also in line with previous studies identifying null findings (e.g., Rintala et al., 2020). However, it should be noted that previous results in this area are inconsistent: Some studies have also found small effects for some psychological context variables (e.g., feeling stressed, upset, or enthusiastic; Murray, Brown, et al., 2023; Silvia et al., 2013). One reason for this ambiguity in previous research could be that the effects for psychological context features might be very small, if present at all, and additionally be methodologically masked. As psychological context features rely on self-reports, this information is missing for a point in time if participants do not respond to the beep. As a workaround, researchers usually use the psychological context information reported in one of the previous beeps to predict compliance (Rintala et al., 2020; Silvia et al., 2013; Sokolovsky et al., 2014). Thus, the included psychological contextual information frequently refers to the participant's psychological state hours before. But as psychological states are highly fluctuating (Fleeson, 2001; Heller et al., 2007), this category of features might be little informative for compliance prediction.

2.5.3 Study Compliance as a Trait?

In summary, a key finding of our study is that *past behavior* features are by far most important for predicting compliance at the beep level. If past compliance behavior predicts future study behavior, this, in turn, leads to the question of whether compliance in ESM studies might be some sort of temporally stable person-level trait. Based on our analyses we cannot rule out the possibility that an actually unobserved (psychological) trait drove our compliance prediction and past behavior is just a kind of observable manifestation of this trait. For example, a person with a high score on the (unobserved) compliance trait, might also be more likely to respond to both the

last and the given beep. Thus, as this (unknown) trait was not explicitly considered as predictor, a direct relation with compliance could, of course, not be observed in our study. Nevertheless, this trait could have effects on compliance, as the *past behavior* features might have carried over its effect. Please note that this is only our post-hoc interpretation and future studies should investigate this assumption, for example, by theory-guided derivation of new constructs or inclusion of known constructs (e.g., specific motivational aspects) in future beep level prediction studies. One additional way to further investigate the assumption of a stable person-trait, would be the use of a measurement burst design. By collecting ESM data during multiple ESM periods (bursts) at different times, stable compliance rates within participants would give some further support to this idea. For the sake of simplicity, we have referred to one single compliance trait in this paragraph. But future research should also investigate if one or maybe even several traits underlie *past behavior* features.

Our study gives a starting point for the search of a compliance trait by limiting the range of eligible constructs. If there is a compliance trait, it seems to be mostly independent of "traditional" psychological traits such as personality or attitudes. Even after excluding the *past behavior* features in our study, we found no considerable decrease in the prediction performance when additionally dropping the *trait* subcategory. Thus, we conclude that the effect of the traits included in our study was not masked by the effect of the past behavior features. The compliance trait might therefore carry different content information or have a less abstract resolution than, for example, established personality traits such as conscientiousness. At the same time, the finding that the included psychological traits (and states) were not related to compliance at the beep level, is rather good news for research disciplines such as personality psychology. The subject of interest such as personality traits or affect states seem not to be strongly and systematically related to missing data in the ESM setting.

2.5.4 Implications for Applied and Methodological Research

Our findings come with implications for both researchers applying ESM in their empirical studies and methodological researchers investigating ESM as their subject of research.

For researchers applying ESM, our results could help to optimize participant compliance at the beep level. For example, if researchers want to know whether a participant took their medication on a particular day (Verhagen et al., 2016), a promising approach to monitor treatment in clinical trials might be to send beeps only in contexts in which participants are most likely to respond, such as when they are at home or at work but not when they are on a train or in a car. A limitation of this compliance optimization approach is, however, that the core idea of ESM (i.e., random sampling across situations, moods, and experiences in everyday life) gets lost. This strategy should therefore be treated with caution, as the randomness of the sampling is arguably restricted when using this compliance-optimized approach. By selecting only the contexts in which participants are most likely to respond to beeps, researchers are likely to introduce a new type of bias, as some specific contexts are already selectively excluded during data collection (Lathia et al., 2013; van Berkel & Kostakos, 2021). Thus, researchers should be aware of the trade-off between optimizing compliance rates on the one hand but also keeping the idea of random sampling in their ESM studies on the other.

Second, our study provides ESM researchers with a guide on which variables to consider as control or auxiliary variables. This could help bring them one step closer to the (desired) goal of missing data at random (MAR) and at the same time one step away from biased study results and errors due to non-compliance bias (Newman, 2014). Based on our results, potential candidates for such control variables are information on participant mobility at the time of receiving a beep (e.g., being at home or at work versus being in a rail vehicle). This information could be operationalized through passive GPS tracking. In this context, developments in smartphone technologies are increasingly facilitating the collection of mobility data for research purposes (Harari et al., 2016; Miller, 2012; Müller et al., 2020).

In addition, scholars have recently highlighted the enormous potential of using mobile sensing for investigating compliance in ESM studies (Murray, Brown, et al., 2023; Sun et al., 2020). As far as we know, our study is one of the first to respond to this call and thus could also serve as a starting point for future methodological research focusing on ESM as a research subject.

First, one possible objective of future research could be to gain a more thorough understanding of the above-mentioned differences between compliance-optimized vs. randomness-optimized approaches applied in ESM studies. One and the same research question could be addressed by collecting data via both approaches and comparing findings, depending on the applied optimization scheme. Moreover, irrespective of the subject of research, effects on compliance could be investigated by experimentally manipulating the type of optimization approach. This comparison, in turn, might help in understanding the possible (intended or unintended) impact of researcher degrees of freedom on findings in ESM studies, such as biases due to study design aspects related to compliance, such as suspending ESM beeps on specific weekdays (Wicherts et al., 2016).

Second, future methodological research could extend our approach of combining ESM and mobile sensing in several important ways to see how robust compliance bias at the beep level in ESM is across different study settings. On the one hand, future studies could apply different ESM designs and investigate if compliance biases depend on the degree of invasiveness of the used ESM schedule (van Berkel et al., 2019). On the other hand, studies could include additional feature categories such as physiological parameters (e.g., heart rate or stress measurements from smartwatches or other wearables). A broader set of included features beyond person, behavior, and context characteristics could further contribute to understanding compliance in ESM studies (Wrzus & Mehl, 2015) and could further increase the prediction performance obtained in our study. Finally, future studies could also compare participants' perception of compliance and their reported reasons for missing beeps with their actual compliance behavior and reasons deduced from objective data to gain further insight into compliance in ESM studies.

2.5.5 Limitations

The present study encountered some limitations. First, our ESM scheme deviated from more "traditional" time-contingent designs reported in the literature. This deviation should be considered when interpreting our results. In most previous ESM studies, participants received a fixed number of notifications at fixed or (quasi-)random times prompting them to respond to a beep (Wrzus & Neubauer, 2023). In contrast, we

used an ES scheme that could be considered a combination of time-contingent and event-contingent sampling (Reis et al., 2014): Beeps were scheduled pseudo-randomly, that is, they were time-contingent in pre-specified intervals across the day but only triggered if participants turned their screen on within a particular time interval. Thus, participants were not proactively notified, for example, via the smartphone's vibration or acoustic signals. Instead, they only received a beep when they used their smartphone of their own accord. Accordingly, our study focused exclusively on the investigation of *active* non-compliance (i.e., participants noticed the beep but actively decided not to respond). In contrast, previous studies with their time-contingent designs did not differentiate between *active* and *passive* (i.e., participants did not notice a beep) non-compliance (Rogelberg et al., 2003). For example, in our study, if participants were doing sports in the morning and therefore did not use the smartphone, they did not receive and thus not miss any beep in this time interval. In contrast, in a study using a standard time-contingent ESM schedule, participants would have been notified to respond to a beep. Non-compliance could then either mean, that they did not notice the beep while doing sports in the morning or actively decided not to respond because, for example, they were enjoying their morning routine. The reason for deviating in the ESM design from previous literature was that we used smartphones not only to deliver beeps but also to collect mobile sensing data. If we had proactively notified participants, we would likely have altered their natural smartphone usage behaviors, which we included as features in our prediction models (van Berkel et al., 2019). Having this trade-off in mind, we decided to put emphasis on collecting naturally occurring (smartphone) behavior. In summary, our results should be interpreted depending on our study procedure, i.e., the times at which notifications were sent can be considered a pseudo-random sample of smartphone usage. For example, we found higher compliance rates on Sundays. However, it is important to keep in mind that fewer beeps than usual were sent on Sundays due to lower smartphone usage. Thus, people were less likely to receive ES notifications on Sundays because they used their smartphone less, but when they did use their smartphones, they were also more likely to respond. To allow more specific conditional interpretations of our results, we provide the distribution of the sent beeps depending on day, daytime, and day $\times$ daytime combination in our online material.

Second, when interpreting our results, we should keep in mind that the lack of some effects, e.g., for the *person* features, might be related to one major challenge of many empirical studies: self-selection or collider-stratification-bias (Bethlehem, 2010; Cinelli et al., 2022). Selection bias arises because our participants may not have entered our sample at random. Rather, the decision to participate in such a time-consuming, intensive longitudinal study is likely influenced by several factors, some of which might overlap with the factors investigated in our study. This impacts how we can interpret our results. This bias can be formalised by means of the directed acyclic graph (DAG) framework. We do not go into detail about the DAG framework at this point, but refer interested readers to Cinelli et al. (2022), Rohrer (2018), or Smith (2020). Nevertheless, we would like to briefly discuss the selection bias and possible consequences for the interpretation of our results in the light of the DAG framework to illustrate a possible scenario of how this bias might arise in ES and mobile sensing studies: On the one hand, someone with a demanding job might be less likely to join the study due to a lack of time. And if they do decide to participate, beep level compliance could be influenced by their job's demands. This creates a situation where the job's demands become a variable that affects both the decision to participate and their beep level compliance (Scollon et al., 2003; Stone, Schneider, Smyth, et al., 2023). Thus, it constitutes a *confounder* variable. On the other hand, it should be considered that some features in our study, and probably especially stable person features, might have also affected the decision to take part in the study. For example, openness in previous research has been found to be related to the willingness to participate in surveys (Marcus & Schütz, 2005). If we then want to investigate how our features are related to beep level compliance, self-selection acts as a *collider* variable because it is affected by both our features (e.g., openness) and (unobserved) other factors (e.g., job demand). According to the DAG framework, to obtain unbiased estimates of the effect of our features on beep level compliance, we should then not condition our estimation upon a collider variable such as self-selecetion (Elwert & Winship, 2014). However, we automatically condition upon self-selection as we only consider data from persons taking part in our study, but not persons deciding against participating. Thus, conditioning on self-selection offers another possible explanation for the effect of our features on compliance at the beep level, at least for those features that also possibly affect self-selection (Cinelli et al.,

2022). We speculate that this described constellation with confounders and colliders could especially apply to features of the *person* category and we might therefore not have found any associations with compliance. The issue of self-selection bias is common in psychological research, but probably especially so in studies like ours that involve intensive data collection. Therefore, future research should address the problem of person variables associated with self-selection in ES or mobile-sensing studies, for example, by contacting non-participants and learning about the "unknown." That is, by examining factors associated with the decision not to participate.

Third, our study and the associated ESM periods took place in July/August and October 2020, which might have led to some distortion of "normality" due to the ongoing COVID-19 pandemic. Accordingly, our results should be interpreted against the background of the COVID-19 pandemic, which may have led to changes in everyday behaviors and contextual characteristics (e.g., time spent at home). However, governmental restrictions in Germany were loosened during the time of data collection. For example, shopping restrictions were suspended, travel restrictions within Germany were loosened, and restaurants and daycares had started re-opening (as can be seen from the data collated by Steinmetz et al., 2020). In line with this, we think that possible pandemic effects on our results are limited in scale. Nevertheless, future research should investigate whether our model generalizes outside of pandemic periods.

Lastly, as we wanted to include a broad range of mobile sensing-based behavioral and contextual features, we designed a research app running only on the Android operating system, as it allows more extensive access to third-party apps (Kreuter et al., 2020). However, only negligible to small differences in key personality traits have been found between users of the two most common smartphone operating systems, Android and iOS, which may be attributed to differences in the socio-demographic composition of the users (Götz et al., 2017; Keusch et al., 2020). Bearing this and the overall sample characteristics (e.g., size, age range, gender distribution) in mind, this rather supports the generalizability of our results (Götz et al., 2017).

2.6 Conclusion

This study used a multitude of features of *person*, *behavior*, and *context* categories to predict compliance at the beep level in an ESM study. Based on a sample of 592 participants and more than 25,000 beeps, we used a combination of more than 400 features collected multi-methodologically via surveys, ES questionnaires, and mobile sensing. Compliance was successfully predicted at the beep level, with both linear and non-linear models investigated in our machine learning benchmark experiment. Using a large variety of person, behavior, and context features, we found indicators of a compliance bias in our ESM study. Our follow-up analyses revealed that study-related past behaviors were most informative in predicting compliance at the beep level, followed by physical context features related to participants' mobility. In contrast, smartphone-mediated behaviors, temporal context, psychological context, and person characteristics played a negligible role in predicting compliance.

Our study has implications for both researchers applying ESM and those doing methodological research on ESM. With ESM being a widely used method across disciplines and smartphones being omnipresent and increasingly used in research, our study contributes a multi-method approach combining traditional and newer data-intensive collection methods to gain insight into compliance bias in ESM studies.

2.7 Declarations

Funding and Acknowledgements:

For the data collection in the context of the Smartphone Sensing Panel Study, our special thanks go to our cooperation partners of the the Leibniz Institute of Psychology (ZPID). The Smartphone Sensing Panel Study, which produced the dataset used in this manuscript, is a joint project of LMU Munich and the ZPID who provided most of the funding for the implementation of the described panel study. In addition, we thank the Schuhfried GmbH for providing the materials to integrate the Big Five Structure Inventory in our panel study. We thank the PhoneStudy team for their work on this comprehensive panel study (special thanks to Fiona Kunz, Florian Bemmann, Larissa Sust, Luzia Heusler, Florian Pargent, and Markus Bühner).

Conflicts of interest/Competing interests:

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval:

Approval was obtained from the ethics committee of LMU Munich. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate:

Informed consent was obtained from all individual participants included in the study.

Consent for publication:

Not applicable as data were anonymized and no identification of participants is possible.

Availability of data and materials:

The pre-processed feature data are available at https://osf.io/jw3bn/

Code availability:

The preprocessing and analysis code for the study is available at https://osf.io/jw3bn/

Authors' contributions:

TR: Conceptualization, Formal Analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing

RS: Conceptualization, Data Curation, Investigation, Methodology, Supervision, Writing – Review & Editing

2.8 References

Albarracin, D., & Wyer Jr., R. S. (2000). The cognitive impact of past behavior: Influences on beliefs, attitudes, and future behavioral decisions. *Journal of Personality and Social Psychology*, *79*(1), 5. https://doi.org/10.1037/0022-3514.79.1.5

Allison, P. D. (2001). *Missing data*. Sage Publications. https://doi.org/10.4135/9781412985079

Arendasy, M., Sommer, M., & Feldhammer, M. (2011). Manual big-five structure inventory bfsi.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*(2), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

Binder, M., Pfisterer, F., Lang, M., Schneider, L., Kotthoff, L., & Bischl, B. (2021). mlr3pipelines - flexible machine learning pipelines in r. *Journal of Machine Learning Research*, *22*(184), 1–7. https://jmlr.org/papers/v22/21-0281.html

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLoS ONE*, *12*(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Boukhechba, M., Cai, L., Chow, P. I., Fua, K., Gerber, M. S., Teachman, B. A., & Barnes, L. E. (2018). Contextual analysis to understand compliance with smartphone-based ecological momentary assessment. *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 232–238. https://doi.org/10.1145/3240925.3240967

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breyer, M., B. Bluemke. (2016). Deutsche version der positive and negative affect schedule panas (gesis panel). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/zis242

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13. https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., Tötsch, N., & Jurman, G. (2021). The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and

markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*(1), 1–22. https://doi.org/10.1186/s13040-021-00244-z

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, 00491241221099552. https://doi.org/10.1177/00491241221099552

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330. https://doi.org/10.1037/1082-989x.6.4.330

Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics. *Psychological Assessment*, *24*(3), 713–720. https://doi.org/10.1037/a0026733

Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in everyday life: The uses of experience sampling. *Journal of Happiness Studies*, *4*(2). https://doi.org/10.1023/A:1024409732742

Csikszentmihalyi, M., & LeFevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, *56*(5), 815. https://doi.org/10.1037//0022-3514.56.5.815

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dragicevic, M., & Casalicchio, G. (2020). *Resampling—stratified, blocked and predefined*. Mlr-Org. Retrieved January 4, 2023, from https://mlr-org.com/gallery/basic/2020-03-30-stratification-blocking/

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire

length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Elmer, T., van Duijn, M. A., Ram, N., & Bringmann, L. (2022). Modeling categorical time-to-event data: The example of social interaction dynamics captured with event-contingent experience sampling methods. https://doi.org/10.1177/02654075221122069

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Ferguson, E., & Bibby, P. A. (2002). Predicting future blood donor returns: Past behavior, intentions, and observer effects. *Health Psychology*, *21*(5), 513. https://doi.org/10.1037/0278-6133.21.5.513

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011. https://doi.org/10.1037/0022-3514.80.6.1011

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Götz, F. M., Stieger, S., & Reips, U.-D. (2017). Users of the main smartphone operating systems (ios, android) differ only little in personality. *PLoS ONE*, *12*(5), e0176921. https://doi.org/10.1371/journal.pone.0176921

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer New York, NY. https://doi.org/10.1007/978-1-4614-4018-5

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*(4), 308–319. https://doi.org/10.1214/18-aoas1157

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harari, G. M., Müller, S. R., Mishra, V., Wang, R., Campbell, A. T., Rentfrow, P. J., & Gosling, S. D. (2017). An evaluation of students' interest in and compliance with self-tracking methods: Recommendations for incentives based on three smartphone sensing studies. *Social Psychological and Personality Science*, *8*(5), 479–492. https://doi.org/10.1177/1948550617712033

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

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-84858-7

Heller, D., Komar, J., & Lee, W. B. (2007). The dynamics of personality states, goals, and well-being. *Personality and Social Psychology Bulletin*, *33*(6), 898–910. https://doi.org/10.1177/0146167207301010

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., et al. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Hosmer Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Howard, A. L., & Lamb, M. (2023). Compliance trends in a 14-week ecological momentary assessment study of undergraduate alcohol drinkers. *Assessment*, *0*(0). https://doi.org/10.1177/10731911231159937

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Karrer, K., Glaser, C., Clemens, C., & Bruder, C. (2009). Technikaffinität erfassen–der fragebogen ta-eg. *Der Mensch im Mittelpunkt technischer Systeme*, *8*, 196–201.

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2020). Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a german national survey. *Sociological Methods & Research*, 0049124120914924. https://doi.org/10.1177/0049124120914924

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, *38*(5), 533–549. https://doi.org/10.1177/0894439318816389

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., & Bischl, B. (2019). Mlr3: A modern object-oriented machine learning framework in r. *Journal of Open Source Software*, *4*(44), 1903. https://doi.org/10.21105/joss.01903

Lang, M., & Schratz, P. (2023). *Mlr3verse: Easily install and load the 'mlr3' package family*. https://github.com/mlr-org/mlr3verse

Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In H. Reis (Ed.), *New directions for methodology of social and behavioral sciences* (pp. 41–56, Vol. 15). San Francisco: Jossey-Bass. https://doi.org/10.1007/978-94-017-9088-8_2

Lathia, N., Rachuri, K. K., Mascolo, C., & Rentfrow, P. J. (2013). Contextual dissonance: Design bias in sensor-based experience sampling methods. *Proceedings of the 2013*

ACM International Joint Conference on Pervasive and Ubiquitous Computing, 183–192. https://doi.org/10.1145/2493432.2493452

Lee, Y.-K., Chang, C.-T., Lin, Y., & Cheng, Z.-H. (2014). The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. *Computers in Human Behavior*, *31*, 373–383. https://doi.org/10.1016/j.chb.2013.10.047

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data* (Vol. 1). John Wiley & Sons. https://doi.org/10.1002/9781119013563

Marcus, B., & Schütz, A. (2005). Who are the people reluctant to participate in research? personality correlates of four different types of nonresponse as inferred from self-and observer ratings. *Journal of Personality*, *73*(4), 959–984. https://doi.org/10.1111/j.1467-6494.2005.00335.x

McLean, D. C., Nakamura, J., & Csikszentmihalyi, M. (2017). Explaining system missing: Missing data and experience sampling method. *Social Psychological and Personality Science*, *8*(4), 434–441. https://doi.org/10.1177/1948550617708015

Messiah, A., Grondin, O., & Encrenaz, G. (2011). Factors associated with missing data in an experience sampling investigation of substance use determinants. *Drug and Alcohol Dependence*, *114*(2-3), 153–158. https://doi.org/10.1016/j.drugalcdep.2010.09.016

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. https://doi.org/10.1177/1745691612441215

Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, *116*(534), 1023–1037. https://doi.org/10.1080/01621459.2021.1874961

Müller, S. R., Peters, H., Matz, S. C., Wang, W., & Harari, G. M. (2020). Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and gps tracking. *European Journal of Personality*, *34*(5), 714–732. https://doi.org/10.1002/per.2262

Murray, A. L., Brown, R., Zhu, X., Speyer, L. G., Yang, Y., Xiao, Z., Ribeaud, D., & Eisner, M. (2023). Prompt-level predictors of compliance in an ecological momentary assessment study of young adults' mental health. *Journal of Affective Disorders*, *322*, 125–131. https://doi.org/10.1016/j.jad.2022.11.014

Murray, A. L., Yang, Y., Zhu, X., Speyer, L., Brown, R., Eisner, M., & Ribeaud, D. (2023). Respondent characteristics associated with adherence in a general population ecological momentary assessment study. *International Journal of Methods in Psychiatric Research*, e1972. https://doi.org/10.1002/mpr.1972

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Ottenstein, C., & Werner, L. (2021). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin*, *124*(1), 54. https://doi.org/10.1037/0033-2909.124.1.54

Pargent, F., & Albert-von der Gönna, J. (2019). Predictive modeling with psychological panel data. *Zeitschrift für Psychologie*. https://doi.org/10.1027/2151-2604/a000343

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231162559

Pindek, S., Zhou, Z. E., Kessler, S. R., Krajcevska, A., & Spector, P. E. (2021). Workdays are not created equal: Job satisfaction and job stressors across the workweek. *Human Relations*, *74*(9), 1447–1472. https://doi.org/10.1177/0018726720924444

Prince, M. (2012). Epidemiology. In P. Wright, J. Stern, & M. Phelan (Eds.), *Core psychiatry (third edition)* (pp. 115–129). Elsevier Health Sciences. https://doi.org/10.1016/B978-0-7020-3397-1.00009-4

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B., & Funder, D. C. (2014). The situational eight diamonds: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677. https://doi.org/10.1037/a0037250

Rauthmann, J. F., & Sherman, R. (2018). S8-i-situational eight-i-deutsche fassung. https://doi.org/10.23668/psycharchives.6568

Reis, H. T., Gable, S. L., & Maniaci, M. R. (2014). Methods for studying everyday experience in its natural context. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 373–403). Cambridge University Press. https://doi.org/10.1017/CBO9780511996481.019

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2020). Momentary predictors of compliance in studies using the experience sampling method. *Psychiatry Research*, *286*, 112896. https://doi.org/10.1016/j.psychres.2020.112896

Rocca, R., & Yarkoni, T. (2021). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211026864. https://doi.org/10.1177/25152459211026864

Rodrigues, F., Teixeira, D. S., Cid, L., & Monteiro, D. (2021). Have you been exercising lately? testing the role of past behavior on exercise adherence. *Journal of Health Psychology*, *26*(10), 1482–1493. https://doi.org/10.1177/1359105319878243

Rogelberg, S. G., Conway, J. M., Sederburg, M. E., Spitzmüller, C., Aziz, S., & Knight, W. E. (2003). Profiling active and passive nonrespondents to an organizational survey. *Journal of Applied Psychology*, *88*(6), 1104. https://doi.org/10.1037/0021-9010.88.6.1104

Rogelberg, S. G., & Luong, A. (1998). Nonresponse to mailed surveys: A review and guide. *Current Directions in Psychological Science*, *7*(2), 60–65. https://doi.org/10.1111/1467-8721.ep13175675

Rogers, T., & Aida, M. (2011). What does 'intending to vote' mean? *HKS Working Paper No. RWP12-056*. https://doi.org/10.2139/ssrn.1971846

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated r code. *Social and Personality Psychology Compass*, *15*(2), e12579. https://doi.org/10.1111/spc3.12579

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161. https://doi.org/10.1037/h0077714

Satherley, N., Milojev, P., Greaves, L. M., Huang, Y., Osborne, D., Bulbulia, J., & Sibley, C. G. (2015). Demographic and psychological predictors of panel attrition: Evidence from the new zealand attitudes and values study. *PLoS ONE*, *10*(3), e0121950. https://doi.org/10.1371/journal.pone.0121950

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147. https://doi.org/10.1037/1082-989x.7.2.147

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology*, *125*(6), 1442–1471. https://doi.org/10.1037/pspp0000469

Schoedel, R., & Oldemeier, M. (2020). Basic protocol: Smartphone sensing panel study. https://doi.org/10.23668/psycharchives.2901

Schoedel, R., Oldemeier, M., Bonauer, L., & Sust, L. (2022). Systematic categorisation of 3,091 smartphone applications from a large-scale smartphone sensing dataset. *Journal of Open Psychology Data*, *10*(1). https://doi.org/10.5334/jopd.59

Schüz, N., Walters, J. A., Frandsen, M., Bower, J., & Ferguson, S. G. (2013). Compliance with an ema monitoring protocol and its relationship with participant and smoking characteristics. *Nicotine & Tobacco Research*, *16*(Suppl_2), S88–S92. https://doi.org/10.1093/ntr/ntt142

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*(1), 5–34. https://doi.org/10.1023/A:1023605205115

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shorey, S., Ng, E. D., & Wong, C. H. (2022). Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. *British Journal of Clinical Psychology*, *61*(2), 287–305. https://doi.org/10.1111/bjc.12333

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, *31*(4), 471–481. https://doi.org/10.1177/0894439313479902

Smith, L. H. (2020). Selection mechanisms and their consequences: Understanding and addressing selection bias. *Current Epidemiology Reports*, *7*, 179–189. https://doi.org/10.1007/s40471-020-00241-6

Sokolovsky, A. W., Mermelstein, R. J., & Hedeker, D. (2014). Factors predicting compliance to ecological momentary assessment among adolescent smokers. *Nicotine & Tobacco Research*, *16*(3), 351–358. https://doi.org/10.1093/ntr/ntt154

Steinmetz, H., Batzdorfer, V., & Bosnjak, M. (2020). The zpid lockdown measures dataset for germany. *ZPID Science Information Online 20 (1)*. https://doi.org/10.23668/psycharchives.6676

Sterner, P., Goretzko, D., & Pargent, F. (2023). Everything has its price: Foundations of cost-sensitive learning and its application in psychology. *Psychological Methods*. https://doi.org/10.1037/met0000586

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, *19*. https://doi.org/10.1146/annurev-clinpsy-080921-083128

Stone, A. A., Schneider, S., Smyth, J. M., Junghaenel, D. U., Couper, M. P., Wen, C., Mendez, M., Velasco, S., & Goldstein, S. (2023). A population-based investigation of participation rate and self-selection bias in momentary data capture and survey studies. *Current Psychology*, 1–17. https://doi.org/10.1007/s12144-023-04426-2

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

Sun, J., Rhemtulla, M., & Vazire, S. (2020). Eavesdropping on missing data: What are university students doing when they miss experience sampling reports? *Personality and Social Psychology Bulletin*, 1535–1549. https://doi.org/10.1177/0146167220964639

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 631–642. https://doi.org/10.1080/10705511.2014.937378

Ushey, K. (2021). *Renv: Project environments* [R package version 0.14.0]. https://CRAN.R-project.org/package=renv

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

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

van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. (2019). Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies*, *125*, 118–128. https://doi.org/10.1016/j.ijhcs.2018.12.002

van Berkel, N., & Kostakos, V. (2021). Recommendations for conducting longitudinal experience sampling studies. In E. Karapanos, J. Gerken, J. Kjeldskov, & M. B. Skov (Eds.), *Advances in longitudinal hci research. human–computer interaction series*. Springer, Cham. https://doi.org/10.1007/978-3-030-67322-2_4

van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research*, *42*(2), 387–414. https://doi.org/10.1080/00273170701360803

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. *BMJ Mental Health*, *19*(3), 86–89. https://doi.org/10.1136/ebmental-2016-102418

Viaene, S., & Dedene, G. (2005). Cost-sensitive learning and decision making revisited. *European Journal of Operational Research*, *166*(1), 212–220. https://doi.org/10.1016/j.ejor.2004.03.031

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, *54*(6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? measuring personality processes and their social consequences. *European Journal of Personality*, *29*(2), 250–271. https://doi.org/10.1002/per.1986

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

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

3 Study 2: Side Effects of Experience Sampling Protocols: A Systematic Analysis of How They Affect Data Quality, Data Quantity & Bias in Study Results

Reference

Reiter, T., Sakel, S., Scharbert, J., ter Horst, J., van Zalk, M., Back, M., Bühner, M., & Schoedel, R. (2025). Side Effects of Experience Sampling Protocols: A Systematic Analysis of How They Affect Data Quality, Data Quantity & Bias in Study Results. *Advances in Methods and Practices in Psychological Science*, *8*(3). https://doi.org/10.1177/25152459251347274

Author Contributions

Thomas Reiter:

Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing, Project Administration

Sophia Sakel, Julian Scharbert, Julian ter Horst:

Investigation, Data Curation, Writing - Review & Editing, Project Administration

Maarten van Zalk, Mitja Back, Markus Bühner:

Investigation, Data Curation, Writing - Review & Editing, Project Administration

Ramona Schoedel:

Conceptualization, Data Curation, Investigation, Methodology, Supervision, Writing – Review & Editing

The article was slightly modified in formatting and notation to align with the style and layout of the present dissertation. Moreover, the manuscript was subject to slight copyediting by the journal introducing slight differences between the original manuscript and the published article.

3.1 Abstract

In studies using the increasingly popular Experience Sampling Method (ESM), design decisions are often guided by theoretical or practical considerations. Yet limited empirical evidence exists on how these choices impact data quantity (e.g., response probabilities), data quality (e.g., response latency), and potential biases in study outcomes (e.g., characteristics of study variables). In a preregistered, four-week study ($N = 395$), we experimentally manipulated two key ESM protocol characteristics for sending ESM surveys: *timing* (fixed versus varying times) and *contingency* (directly versus indirectly after unlocking the smartphone). We evaluated the ESM protocols resulting from the combination of these two characteristics with regard to different criteria: As hypothesized for contingency, indirect protocols resulted in higher response probabilities (increased data quantity). But they also led to higher response latencies (reduced data quality). Contrary to our expectations, the combined effect of contingency and timing did not significantly influence response probability. We did also not observe other effects of timing or contingency on data quality. In exploratory follow-up analyses, we discovered that timing significantly affected response probability and smartphone usage behaviors, as measured by screen logs; however, these effects were likely attributable to time of day effects. Notably, self-reported states showed no differences based on the chosen ESM protocol, and similar trends were found when correlating primary outcomes with external criteria such as trait affect and well-being. Based on the study's findings, we discuss the trade-offs that researchers should consider when choosing their ESM protocols to optimize data quantity, data quality, and biases in study outcomes.

3.2 Introduction

In recent years, the Experience Sampling Method (ESM) has experienced a major upsurge in various psychological disciplines (Wrzus & Neubauer, 2023). ESM has several advantages over traditional data collection methods. Due to its nature of repeatedly assessing feelings and behaviors in everyday life, it reduces recall bias (Lucas et al., 2021; Scharbert, Utesch, et al., 2024), covers real-life situations that are difficult or unethical to induce in the laboratory (Reis, 2012), and enables to study within-person phenomena (Hamaker, 2012).

However, these various advantages of ESM come at a price. Study designs are much more complex and require researchers to make many decisions. One of the most fundamental decisions is the definition of the *ESM protocol*, i.e., when and how the ESM surveys are sent over the course of the study. Thereby, ESM protocols are often selected based on theoretical considerations, such as the frequency with which the psychological states or behaviors of interest occur or change in daily life. Occasionally, they may also result from pragmatic considerations, such as the technical capabilities of the ESM application (app) or tool used. However, with a few exceptions (e.g., van Berkel, Goncalves, Lovén, et al., 2019; van Berkel et al., 2017), a comprehensive methodological investigation of the potential side effects of when and how ESM surveys are sent on ESM study parameters, including the quantity and quality of ESM data collected and biases in resulting study findings, is still lacking. With this study, we aim to help fill this gap.

3.2.1 Overview of the Characteristics of ESM Protocols

Previous studies have used different ESM protocols depending on the research question (Stone & Shiffman, 2002). Traditionally, these protocols were categorized into interval-, signal-, and event-dependent (Wheeler & Reis, 1991). Early ESM studies often used paper-and-pencil questionnaires completed at set intervals, when signaled by devices like pagers, or after certain events (Larsen & Kasimatis, 1991; Moskowitz & Coté, 1995; Wong & Csikszentmihalyi, 1991). With smartphones now commonly used for ESM, participants are typically notified via email or app (e.g., Scharbert, Humberg, et al., 2024; Stieger et al., 2021), blurring the line between interval and signal-contingent

protocols (Horstmann, 2021). Based on current literature, we therefore categorize ESM protocols by two main characteristics: *timing* and *contingency* as well as their combination (see Table 3.1).

Table 3.1
Overview of ESM Protocol Characteristics

		Contingency	
		direct	indirect
Timing	fixed	fixed×direct	fixed×indirect
	varying	varying×direct	varying×indirect

Note. ESM surveys can be scheduled for exactly the same time every study day (fixed timing) or pseudo-randomly (varying timing). In addition, participants can be notified about ESM surveys at the exact time it was scheduled (direct contingency) or at the very next time they use their smartphone after the time it was scheduled (indirect contingency). Timing and contingency conditions can also be combined (see cells).

Timing describes *when* ESM surveys are scheduled and distinguishes between fixed and varying timing. Fixed timing means that ESM surveys are scheduled at exactly the same time every study day (e.g., Gloster et al., 2017; Tripathi et al., 2020). Varying timing, in turn, means that ESM surveys are scheduled at random times throughout the day or pseudo-randomly. That is, they are sent at random times but within fixed intervals (e.g., in the morning, the afternoon, and evening; for examples, see Neubauer et al., 2020; Verhagen et al., 2019).

Contingency describes *how* the ESM surveys are triggered. Direct contingency means that participants are notified about a scheduled ESM survey at the exact time it is scheduled (e.g., Fullagar & Kelloway, 2009; Hartmann et al., 2015). Indirect contingency means that participants receive ESM surveys only after they actively use their smartphone after the scheduled time (van Berkel, Goncalves, Lovén, et al., 2019), for example, when turning on the screen or after answering a call (e.g., Ghosh et al., 2019; Reiter & Schoedel, 2024; Schoedel et al., 2023). Despite the growing number of software frameworks and methodological or applied research on and with context- or interruptibility-aware ESM designs (Bachmann et al., 2015; Fischer et al., 2011; Schoedel et al., 2023; van Berkel, Goncalves, Koval, et al., 2019; Wen et al., 2017), there is still a strong imbalance in favor of direct ESM protocols in the current ESM research landscape. This is likely due to the fact that indirect protocols require at least some sort of passive sensing in order to work and thus tend to require more effort to implement technically.

3.2.2 Side Effects of ESM Protocol Choice on ESM Study Parameters

From a methodological standpoint, ESM researchers should consider that ESM protocol characteristics may affect a variety of ESM study parameters, including the quantity and quality of ESM data and the introduction of bias to study results.

3.2.2.1 ESM Data Quantity. One important aspect of ESM studies is the quantity of ESM data, often referred to as the response probability or compliance of participants (Eisele et al., 2022; Hasselhorn et al., 2022). Previous meta-analytical and systematic reviews have examined the factors underlying the response probability. These include the socio-demographic characteristics of participants, the clinical status of the population under study, as well as study design elements such as study duration, the number of items, and the provision of incentives. (Davanzo et al., 2023; Jones et al., 2019; Wen et al., 2017; Wrzus & Neubauer, 2023). However, many of the primary studies referred to in this work were observational (e.g., Courvoisier et al., 2012; Silvia et al., 2013; Sokolovsky et al., 2014). Studies that experimentally manipulated ESM study characteristics, primarily focused, with few exceptions (e.g., Businelle et al., 2024), on very specific design elements such as the number of items per ESM survey or the number of prompts per day (e.g., Eisele et al., 2022; Hasselhorn et al., 2022; Ottenstein & Werner, 2021).

Regarding the timing and contingency of ESM surveys, and their combination, however, there is only preliminary evidence from one rather small study ($N = 20$) in the field of computer science that investigated whether participants' response probability may depend on these ESM protocol characteristics (van Berkel, Goncalves, Lovén, et al., 2019). We take this pioneering study as a starting point to replicate on a larger scale whether the design of the ESM protocol used is related to response probabilities in ESM studies (*Research Question 1*).

In more detail, we assume that participants are more likely to answer ESM surveys if they are notified the next time they use their smartphone (i.e., indirect mode) than if they are notified directly at the time the ESM survey was scheduled (i.e., direct mode; *Hypothesis 1*). We believe that ESM surveys in the indirect mode are more likely to be noticed since they are sent only during active phone use. In contrast, participants in

the direct mode may miss ESM surveys if they are not at their phones at the scheduled time.

In addition, contingency can also be considered in combination with timing (see Table 3.1). In the direct mode, we expect participants' response probability to be higher in the fixed than in the varying mode. If the ESM surveys are scheduled for the exact same time each day, participants can anticipate when they will occur (Myin-Germeys et al., 2018). This may increase the likelihood of habitually answering ESM surveys in comparison to the varying-direct mode, in which the ESM surveys are sent at slightly varying times each day. In contrast, the indirect mode is not expected to yield a discrepancy in response probabilities between fixed and varying timing. This is because participants only receive ESM surveys upon active usage of their smartphones, which precludes them from anticipating the timing of subsequent surveys, regardless of whether the timing is fixed or varying. Therefore, we hypothesize that the difference in the (increased) probability of responding to ESM surveys in the fixed versus the varying timing protocol is higher when participants are directly notified at the time the ESM survey is scheduled (direct mode) than when they are notified about the ESM survey the next time they use their smartphone (indirect mode; *Hypothesis 2*).

3.2.2.2 ESM Data Quality. Another important aspect of the ESM studies is the quality of the ESM data, which means that the ESM surveys are carefully answered so that they pass certain quality controls and are therefore usable by researchers (for a detailed discussion of careless responding in surveys, see Meade & Craig, 2012). Defining control criteria, however, is not straightforward, as it can involve a variety of different data characteristics (DeSimone & Harms, 2018).

To the best of our knowledge, there is a lack of substantial findings about how ESM protocol characteristics affect data quality indicators. Therefore, we proceed purely exploratory. We follow the recommended best practices recently published by Ward and Meade (2023) for detecting careless responding in online surveys to transfer them to ESM surveys. In particular, we investigate whether participants, depending on different ESM protocols, (a) speed through or do not take enough time for the items, (b) select contradictory statements, or (c) always choose the same answer option, which could be considered indicators of low ESM data quality (*Research Question 2*).

3.2.2.3 Bias in Resulting Study Findings. A final and further aspect of ESM studies is whether the resulting findings are dependent of or biased due to the characteristics of the ESM protocol used. One reason for such bias can be measurement reactivity, which refers to the effect that the instrument or procedure itself systematically distorts the validity of the outcomes collected (Barta et al., 2012). The degree of the bias resulting from measurement reactivity might vary depending on the specific ESM protocol characteristics. For example, participants may be more annoyed by ESM surveys sent at varying times than by those sent at fixed times, because they are less able to adjust to the timing. But they might want to adjust to answer as many ESM surveys as possible in the study in order to receive the compensation. Consequently, they might systematically rate their mood somewhat worse in the varying compared to the fixed mode.

Another reason for this bias can be selective sampling. For example, previous research has shown, that participants reported more social interactions in event-contingent compared to signal-contingent ESM protocols (Himmelstein et al., 2019). Previous research has also shown that the physical context of participants affects whether they respond to ESM surveys. For example, people are more likely to answer surveys when they are at home (Reiter & Schoedel, 2024). Also, people might use their phones more at certain locations, such as at home. As in the indirect mode, ESM surveys are sent only when participants are using their smartphones, responding to ESM surveys might in turn be limited to specific locations such as home. This selective sampling of participants' mood may produce biased estimate that systematically differ from the true target quantity or estimand (e.g., participants' intra-daily mood fluctuations). By contrast, in the direct mode, responses are collected independently of smartphone use, making them less dependent on the participants' location and thus potentially more representative. This could lead to more accurate estimates.

Although we will not be able to empirically identify the specific reasons behind the bias of study findings, we aim to investigate it as a comprehensive phenomenon and how it might depend on ESM protocol characteristics. Accordingly, we explore whether the ESM responses themselves and the association patterns of ESM responses with external constructs (e.g., traits collected via questionnaires) are biased depending on the ESM protocol used (*Research Question 3a*).

As ESM data are increasingly combined with data passively logged on smartphones (e.g., Ebner-Priemer & Santangelo, 2024; Harari et al., 2016; Wright & Zimmermann, 2019), we think that it is important to also include measures derived from smartphone sensing into the investigation of ESM protocols' side effects. Smartphone sensing refers to the approach of continuously collecting different types of data (e.g., screen or app logs, GPS) in the background at high resolution (i.e., several thousand logs per day) in order to derive objective data on daily behaviors such as mobility, physical activity, or sleep (Harari et al., 2016; Miller, 2012; Schoedel & Mehl, 2024). The combination of ESM and sensing is often considered an important new step in psychological research as the two methods can highly benefit from each other offering new research designs and opportunities (Ebner-Priemer & Santangelo, 2024). However, it also changes the role of the smartphone from a mere tool for ESM data collection to a research object itself. To illustrate, smartphone screen time is increasingly being studied as variable of interest (Christensen et al., 2016; Liebherr et al., 2020). For example, researchers investigate associations between objectively assessed smartphone use and psychological well-being (große Deters & Schoedel, 2024; Przybylski & Weinstein, 2017). On the other hand, sending ESM surveys to assess psychological states can artificially provoke smartphone use. Thus, the ESM notifications could trigger an unintended cascade of smartphone usage behaviors, such as quickly checking the weather or briefly replying to a message, that would not have occurred without the initial ESM. In the best case, however, ESM surveys should not elicit measurement reactivity, interrupt participants' naturally occurring behavior, or (actively) provoke smartphone usage. For this, van Berkel, Goncalves, Lovén, et al. (2019) proposed that ESM notifications are sent only upon the next naturally occurring smartphone use after the time the ESM survey was originally scheduled. We take this approach as starting point and explore in our study if the type of ESM protocol has side effects on smartphone usage indicators themselves and also their associations patterns with external constructs (*Research Question 3b*).

3.3 Method

The data for this study were collected as part of the Coping with Corona (CoCo) project conducted by the University of Münster, the University of Osnabrück, and the LMU Munich in Germany from March to July 2023. The project was approved by the Ethics Committee of LMU Munich under the study title „Coping with Corona (CoCo): Understanding individual differences in well-being during the COVID-19 pandemic". This study was preregistered and can be accessed via the project's osf repository (https://osf.io/a5bg4/), which also contains the online supplemental material (OSM), the data pre-processing and analysis code, and an anonymized, pre-processed data set. Raw sensing data cannot be shared publicly due to privacy and related data protection legislation.

3.3.1 Procedures

A diverse set of recruitment strategies was employed to obtain a heterogeneous sample, comprising psychology students and individuals from the general public. To be eligible, participants had to be 18 years or older and for technical reasons use a smartphone with Android operating system (version 7 or higher).

Participants were asked to install the PhoneStudy research app [8] on their private smartphones for four weeks. The app provided them with up to four ESM surveys per day. The exact ESM protocol according to which the surveys were sent was subject to experimental manipulation and is described in the next section. In addition to the daytime ESM surveys, participants received a daily evening survey and the app continuously collected various types of log data in the background of the smartphone. At the start and the end of the study, participants completed an online pre- and post-questionnaire. Participants received up to €75 in compensation depending on the study parts completed (i.e., base compensation for completing the pre- and post-questionnaire and compliance-related bonus for answering more than 50% of ESM surveys; for further details, see Appendix A in the OSM).

[8] https://phonestudy.org/en/

3.3.2 ESM Protocols

We manipulated the ESM protocols, with respect to the two characteristics timing and contingency. Timing was operationalized in two distinct forms. First, in fixed timing, the ESM surveys were scheduled for the same times each day: 7am, 10am, 1pm, and 4pm. Second, varying timing indicated that the ESM surveys were scheduled pseudo-randomly, with one survey occurring in each of the intervals 7am–10am, 10am–1pm, 1pm–4pm, and 4pm–7pm. In order to ensure that there was sufficient time between two consecutive ESM surveys, a minimum of 60 minutes was set.

Contingency was operationalized in two distinct ways. First, direct contingency refers to the situation in which participants were informed about a planned ESM survey at the exact time it was scheduled. Second, indirect contingency refers to the situation in which participants received the ESM surveys as soon as and only if they actively used their smartphone after the scheduled time (e.g., turned on the screen, answered a call). If participants did not use their phone until the time of the next scheduled ESM survey, the corresponding survey was skipped in favor of the next scheduled survey. Consequently, it was possible that participants were not informed about a scheduled ESM survey if they did not use their phone within the requisite time.

The combination of these two parameters resulted in the 2×2 experimental conditions presented in Table 3.1. Each participant was exposed to all four experimentally manipulated ESM protocol conditions, each lasting for seven days. The order of the four ESM protocol conditions in the within-subjects design was randomized across participants. The randomization was accomplished during the app setup right after the participants installed the app on their smartphones. For each participant, each of the four possible experimental conditions were drawn without replacement resulting in a total of 24 possible combinations. We informed our participants in advance that the timing of the ESM surveys would vary and that they might experience a different number of ESM surveys per day over the course of the study. We did this because, during pilot testing, we found that participants thought they would have technical problems switching to another ESM protocol, and were worried about it. In each of the four experimental conditions, ESM surveys timed out (i.e., the notification disappeared) 15 minutes after their initial appearance on the smartphone. After starting the ESM survey, participants had to complete the ESM survey within another 15 minutes.

3.3.3 Participants

In total 510 participants answered the pre-questionnaire and installed the app. After applying our preregistered exclusion criteria (see Appendix A in the OSM for further details), the final sample was comprised of 395 participants, of which 67.6% identified as female, 31.7% as male, and 0.8% as neither male nor female. On average, participants were 27.8 years old (range between 18 and 72 years). 1.2% of participants graduated from lower secondary school, 6.3% graduated from higher secondary school, 57% had finished A-levels, 33.1% graduated from university, and 2.3% held a PhD. We preregistered an a-priori simulation-based sensitivity power analysis for generalized linear mixed models (following Pargent et al., 2024). This sensitivity power analysis was conducted prior to data analysis but after data collection. We took a conservative approach to estimating power, assuming small effect sizes and drawing on previous ESM literature (Wrzus & Neubauer, 2023). For an alpha error level of 5%, our achieved sample size (n = 395) and average number of observations per person (n = 104) allowed us to detect small effects with a power of 80% to 100%.

3.3.4 Measures

3.3.4.1 ESM Data Quantity. As preregistered (see *H1* and *H2*), we defined participants' response probability as the proportion of ESM surveys answered out of all ESM surveys sent to a given participant. Thereby, we only considered those sent ESM surveys in our study for which we could ensure that participants were notified about them. In other words, we excluded cases where participants switched their smartphone off at the scheduled time for the ESM surveys, had set their devices to "do not disturb," or had intentionally not been notified. This latter scenario pertains to the indirect contingency protocols in instances where the participants did not utilize their smartphones until the scheduled time for the subsequent ESM survey leading to the initial survey being 'replaced' by the next survey. We counted ESM surveys as answered if all items of the survey were answered within 30 minutes after participants received the survey notification (i.e., participants started the survey within 15 minutes after notification and finished the survey within 15 minutes after starting). We use the term *response rate* in a descriptive sense and the term *response probability* if we refer to the probability to answer a specific ESM survey as predicted by the specified model

(see section *Data Analysis*). As an additional indicator, we descriptively examined *dropout counts*. To this end, we analyzed the last completed ESM survey of each participant during the study period. If this occurred more than three days before the end of the study, we classified it as an indicator of "silent" study dropout, defined as a participant who did not officially inform us of their withdrawal from the study but ceased answering ESM surveys.

3.3.4.2 ESM Data Quality. We used three different indicators representing common approaches to detect low quality data arising from careless responding in survey research (Gibson & Bowling, 2019; Huang et al., 2012; Scharbert et al., 2023): (1) *Response duration* of single ESM surveys defined as the time difference between opening and finishing an ESM survey; (2) *contradictory response patterns* defined binary as whether participants selected the response options "agree" or "strongly agree" on items that were semantic antonyms of one another within a single ESM survey (i.e., the item pairs feeling "happy" vs. "sad" and "stressed" vs. "relaxed"; see Ward & Meade, 2023); and (3) *repetitive response styles* defined binary as whether the same response option was selected on all eleven subsequently presented state affect items within a single ESM survey. As an additional, not preregistered data quality indicator we examined *response latency*. It was defined as the difference between the time of the originally scheduled ESM survey and the time when participants actually began to respond to a given ESM survey. Besides these items, the ESM surveys also asked about partners and conversational topics of preceding social interactions, co-rumination, and further mood-related states. A full list of the assessed items and further details on the study procedures can be found in Appendix A in the OSM.

3.3.4.3 Bias in Resulting Study Findings. We examined ESM responses and smartphone usage indicators as primary study outcomes and their respective patterns of association with external constructs assessed by self-report questionnaires as secondary study outcomes. We thereby expanded our preregistered analysis plan by additionally including further primary study outcomes (besides smartphone usage) and their associations with external constructs in our exploratory analysis.

The selection of ESM primary study outcomes included *state positive affect* and *state negative affect*, each of which was assessed as the average score of three items (positive affect: "happy", "excited", "relaxed"; negative affect: "angry", "anxious", "sad") of the PANAS-X (Watson & Clark, 1994). Items were rated on a 6-point Likert scale ranging from "strongly disagree" to "strongly agree". In our analyses, we were interested in both the absolute state values of affect and the deviation of people's current affect from their personal mean value. Therefore, we additionally centered the affect state scores around the person-specific mean across the study.

As additional primary study outcomes, we explored smartphone usage measures derived from the smartphone sensing logs. Specifically, we extracted participants' total smartphone usage time and total number of unlock events in the hour around the time of each scheduled ESM survey (for ± 30 minutes; for further details see Reiter & Schoedel, 2024). As we were interested in whether participants used their smartphones more (i.e., longer and more frequently) than usual (i.e., when not receiving ESM surveys) depending on the respective ESM protocol, we additionally centered the extracted smartphone measures by the respective person's daytime-specific mean value. That is, for each participant, we calculated person-average smartphone usage behavior measures for each daytime interval of the entire study period (i.e., hourly averages for early mornings: 7am–10am; late mornings: 10am–1pm; afternoons: 1pm–4pm; early evenings: 4pm–7pm). Then, we used the smartphone measures of the one-hour interval surrounding the respectively scheduled ESM surveys and the person-average smartphone measures to extract the deviation of participants' ESM-related from their person-average smartphone usage behavior in terms of *total usage time* and *total number of unlocks*.

Finally, as exemplary secondary study outcomes we used different well-being measures assessed during the pre- and post-questionnaires which participants answered before and after the experience sampling period: (1) a subset of the PANAS (Watson et al., 1988) to assess trait positive and negative affect (α = .70/.70; ω = .74/.73 for positive/negative affect); (2) the Patient Health Questionnaire-9 excluding the item on suicidal ideation (PHQ-9, Kroenke et al., 2001) to assess depression (α = .84, ω = .88); (3) the Satisfaction With Life Scale (SWLS, Diener et al., 1985) to assess

satisfaction with life ($\alpha = .88$, $\omega = .90$); and (4) Ryff's Psychological Well-being Scale (PWB, Ryff & Keyes, 1995) to assess general psychological well-being ($\alpha = .83$, $\omega = .86$). Factor scores for all secondary study outcomes were calculated using the lavaan package (Rosseel, 2012).

3.3.5 Data Analysis

All analyses were conducted using the statistical software R (version 4.4.2; R Core Team, 2024). Generalized linear mixed effects regression models were estimated using the packages lme4 (Bates et al., 2015) and lmertest (Kuznetsova et al., 2017). Predicted average response probabilities were obtained using the marginaleffects package (Arel-Bundock et al., 2024). For reproducibility, we used the package renv (Ushey & Wickham, 2024) and uploaded the lockfile to the project's osf repository.

3.3.5.1 Confirmatory Data Analysis. For the confirmatory analysis (RQ1), we used multilevel logistic regression models estimating within-person effects. That is, for each participant we repeatedly observed whether a scheduled ESM survey was answered or not (i.e., binary outcome variable) depending on the respective ESM protocol characteristics timing (dummy-coded: 0 = fixed [f], 1 = varying [v]) and contingency (dummy-coded: 0 = direct [d], 1 = indirect [i]) as well as their interaction. Therefore, we specified random-intercept fixed-slope models. To test our hypotheses, we calculated the average predicted response probabilities for the different types of ESM protocols resulting from the combination of timing and contingency and inspected their contrasts. In more detail, we specified the following hypotheses:

Hypothesis 1:
H1a: $P(Y = 1|f \times i) - P(Y = 1|f \times d) > 0$ AND
H1b: $P(Y = 1|v \times i) - P(Y = 1|v \times d) > 0$
That is, we assume that participants are more likely to answer ESM surveys if they are notified about the ESM survey the next time they use their smartphone (indirect mode) than if they are notified directly at the time the ESM survey was scheduled (direct mode) (for both fixed (*H1a*) and varying timing (*H1b*) protocols).

Hypothesis 2:

$$\Big(P(Y = 1|f \times d) - P(Y = 1|v \times d)\Big) - \Big(P(Y = 1|f \times i) - P(Y = 1|v \times i)\Big) > 0$$

That is, we assume that the increase of ESM survey response probability is stronger in fixed (vs. varying) timing protocols when participants are notified directly at the time the ESM survey was scheduled (direct mode) than when they are notified about the ESM survey the next time they use their smartphone (indirect mode).

3.3.5.2 Exploratory Data Analysis. On an exploratory level, we compared the number of study dropouts across ESM protocols and study weeks. Given that the observed numbers were generally rather small, we present a descriptive account and refrain from conducting any statistical tests.

For the exploratory ESM data quality outcomes (i.e., response duration, contradictory and repetitive response styles, response latency) and primary study outcomes (i.e., absolute levels of and deviation from person-average of state positive affect and state negative affect, deviation of person-average smartphone usage duration and of number of unlocks), we used generalized linear mixed effects models (random-intercept fixed-slope), with timing, contingency, and their interaction as predictors. We standardized the numeric outcome variables.

In addition, we explored whether the use of different ESM protocols was related to differential association patterns between our set of primary study variables (i.e., state affect and smartphone use) and external constructs (i.e., trait affect, depression, satisfaction with life, and psychological well-being). For these analyses, we used the absolute levels of state positive affect, of state negative affect, of smartphone usage duration, and of smartphone unlocks each as outcome variable. We ran individual mixed effects models per ESM protocol with trait affect, depression, satisfaction with life, and psychological well-being as predictors. Thus, in total we fitted 16 models, one for each combination of the four ESM-level primary study outcomes and the four ESM protocols.

3.3.5.3 Robustness Checks. Finally, we conducted different robustness checks. First, we repeated the confirmatory and exploratory analyses separately for the student subsample ($N = 112$, $n = 11{,}661$) and the subsample recruited from the general public

($N = 283$, $n = 29{,}575$). Detailed results along with the full sample results can be found in Appendix B in our OSM. We give brief summaries of the additional analyses in the results section.

Second, in our preregistration, we defined the number of *sent* ESM surveys as baseline for calculating response probabilities (see section *Measures*). Alternatively, we could also have defined the number of *scheduled* ESM surveys as baseline. We therefore conducted a sensitivity analysis with this plausible alternative option to calculate response probabilities. As the results did not differ with respect to our preregistered hypothesis tests, we refrain from extensively presenting the results in the manuscript. However, we report in detail on the rationale and procedure of the sensitivity analysis, as well as the results in Appendix C of the OSM.

3.4 Results

A total of 41,236 ESM surveys were sent in our study with an average of 104.39 ESM surveys per participant over four weeks. On average, participants answered 45.32 across all ESM protocols. As a plausibility check, we found that the average number of ESM surveys sent was slightly lower in the indirect modes, which was expected since the ESM surveys in the indirect mode were only sent if the participants used their smartphones in the time window after the scheduled ESM survey. Detailed figures on ESM surveys sent and their distribution across the various ESM protocols can be found in Table 3.2. A detailed presentation of all results can be found in Appendix B of our OSM.

Table 3.2
Means and Standard Deviations of ESM Response Behavior Depending on the ESM Protocol Across All Users

ESM Protocol	Surveys Sent	Surveys Answered	Response Rate (%)	Response Latency (min)
fixed × direct	26.31 (3.71)	8.36 (4.47)	31.75 (16.88)	4.23 (8.12)
fixed × indirect	25.75 (3.20)	13.95 (5.87)	54.18 (22.10)	36.48 (41.38)
varying × direct	26.58 (3.18)	8.76 (4.52)	33.02 (16.81)	4.12 (20.55)
varying × indirect	25.82 (3.35)	14.27 (6.05)	55.24 (22.17)	34.32 (40.25)
Overall	**104.39 (8.51)**	**45.32 (17.04)**	**43.38 (15.95)**	**23.62 (36.89)**

Note. Values in brackets represent standard deviations.

3.4.1 ESM Data Quantity (RQ1)

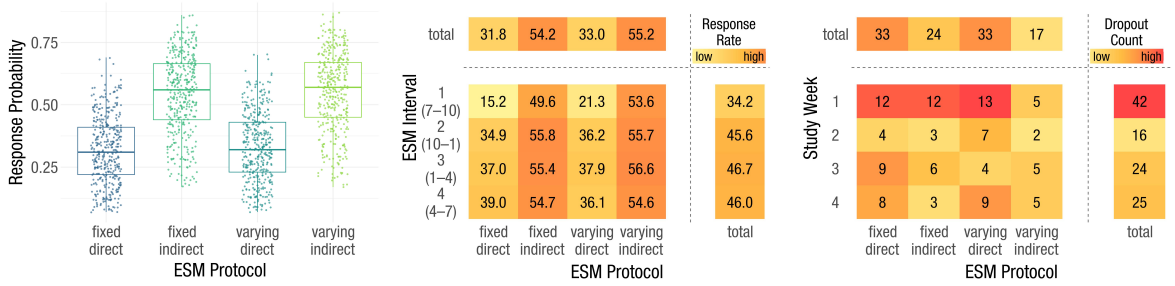
Table 3.2 shows that the number of ESM surveys answered was on average about 1.7 times higher for the indirect (rows 2 and 4) compared to the direct (rows 1 and 3) ESM protocols. That is, participants had on average a significantly higher probability to respond to a sent ESM survey in the indirect (fixed-indirect: 54.2%; varying-indirect: 55.2%) compared to the direct protocols (fixed-direct: 31.8%; varying-direct: 33.0%; see Figure 3.1a). Accordingly, and in line with Hypothesis 1, differences in the response probabilities (RP) between the two contingency variants (i.e., indirect versus direct) were significant across both timing modes (fixed (H1a): ΔRP = 24.6%, $p < .001$; varying (H1b): ΔRP = 24.4%, $p < .001$). We replicated the effect of contingency in both subsamples.

We did not find the expected effect for the combination of timing and contingency (Hypothesis 2). Accordingly, using the fixed compared to the varying ESM protocols did not show a significantly larger increase in response probability in the direct protocols compared to the indirect protocols (H2: ΔRP = -0.2%, $p = .595$). This was also the case in the general public subsample. In the student subsample, we found the proposed effect.

In a first supplemental analysis, we explored the main effect of timing across contingency modes and found that participants' response probability was significantly higher in the varying mode compared to the fixed mode ($OR = 1.06$, $p = .048$). The heatmap of response rates in Figure 3.1b highlights this finding by showing that, overall, varying protocols had higher response rates than fixed protocols in both contingency modes (indirect: 55.2% versus 54.2% and direct: 33.0% versus 31.8%). However, a closer look at the response rates individually per daytime interval reveals that the difference between varying and fixed protocols was greatest in the early morning interval (indirect: 53.6% versus 49.6% and direct: 21.3% versus 15.2%). We therefore excluded all ESM surveys sent in the early morning interval and re-ran the analysis. The main effect for timing then disappeared ($OR = 0.98$, $p = .664$). These findings indicate that the observed main effect for timing may be attributed to a methodological artifact. In the fixed modes, the first survey of the day was by design scheduled for 7am, whereas in the varying modes, surveys were conducted at random times between 7am and 10am.

Figure 3.1
Visualization of ESM Data Quantity Indicators Depending on ESM Protocols

(a) *Participants' Response Probabilities across ESM Protocols* **(b)** *Response Rates across Time Intervals and ESM Protocols* **(c)** *Dropout Counts across Study Weeks and ESM Protocols*



As a result, participants in the fixed mode may have been more likely to miss the ESM survey, as they could still have been asleep or engaged in their morning routines at that earlier time.

In a second supplemental analysis, we aimed to gain a better understanding of participation behavior over the course of the study. As might be expected, participants' response probabilities decreased throughout the course of the study. We found this by including (a) the number of the study days and (b) the number of the ESM surveys as covariates in the model. Although very small, we found significant negative effects for both indicators of study progress on response probability ($OR_a = 0.984$, $p_a < .001$; $OR_b = 0.997$, $p_b < .001$).

As a complementary approach, we also explored participants' drop-out during the study (see Figure 3.1c). On an exclusively descriptive basis, across all ESM protocols, most participants ($n = 42$) dropped out during the first week of the study. The dropout counts fell in week 2 ($n = 16$) and rose again slightly in weeks 3 and 4 ($n = 24$ and $n = 25$). Across all study weeks, the dropout counts were descriptively higher for the direct protocols ($n = 33$, respectively) compared to the indirect protocols, with the varying indirect protocols counting the least ($n = 24$ and $n = 17$).

3.4.2 ESM Data Quality (RQ2)

The time it took participants to complete the ESM surveys hardly differed between the various ESM protocols (between 1.52 minutes and 1.58 minutes). Accordingly, we

found no significant effect of timing (β = -0.03, p = 0.198), contingency (β = -0.04, p = 0.065) or their interaction (β = 0.05, p = 0.088) on participants' response duration. This was also the case when the student and the general public subsample was analyzed separately.

We found n = 58 happy-sad and n = 83 stressed-relaxed semantic antonym cases as well as n = 225 repetitive answer style cases in all completed ESM surveys and among all participants. For none of these indicators we found significant effects for timing (OR between 1.10 and 3.10; p > .05), contingency (OR between 0.75 and 1.98; p > .05), or their interaction (OR between 0.47 and 1.25; p > .05). This was also the case when the student and the general public subsample were analyzed separately.

Finally, Table 3.2 shows that the time between the scheduled time for the ESM survey and the time a participant responded to it was, on average, about 8.5 times longer for the indirect (rows 2 and 4) than for the direct (rows 1 and 3) ESM protocols. Accordingly, we found that the response latency was significantly higher for the indirect ESM protocols compared to the direct ESM protocols (β = 0.86, $p < .001$). The significant interaction effect (β = -0.06, p = .040) shows that the magnitude of the contingency effect varied by timing mode (i.e., higher difference in response latency between direct and indirect in the fixed mode compared to the varying mode, see Table 3.2). We replicated the contingency effect in both subsamples. However, the interaction effect of timing and contingency on response latency was replicated only in the student sample, not in the subsample recruited from the general public.

3.4.3 Bias in Resulting Study Findings (RQ3)

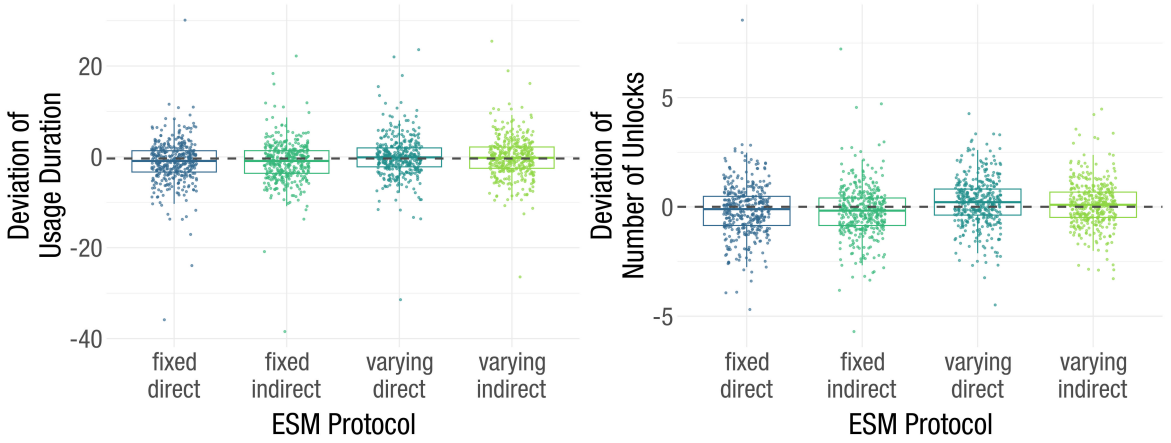
3.4.3.1 Primary Study Outcomes. We found no significant effects of timing, contingency, or their interaction in predicting absolute levels of state positive and negative affect (β between -0.02 and 0.03, $p > .05$). We found the same when predicting individuals' deviation in state positive and negative affect from their person-average (all β between -0.03 and 0.04, $p > .05$).

This was also the case when the student and the general subsample were analyzed separately with the exception that in the student sample there were positive main effects for timing (varying > fixed) and contingency (indirect > direct) on participants'

Figure 3.2
Participants' Smartphone Usage Behavior Deviation Depending on ESM Protocols

(a) *Deviation of Individual Usage Duration from Personal Time-of-Day Average*

(b) *Deviation of Individual Number of Unlocks from Personal Time-of-Day Average*



Note. Dashed line indicates the overall median of all deviations values across all participants.

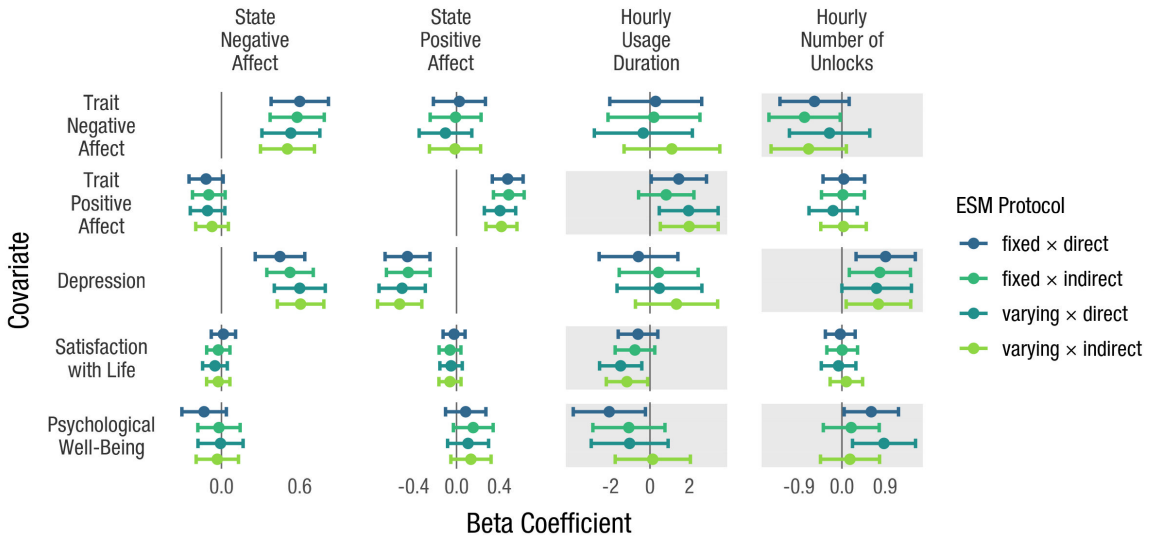
self-reported absolute levels of positive affect and the deviation from their personal average. In addition, there was a negative interaction effect for timing and contingency for both outcomes, meaning that the effect of timing differed by contingency modes.

In comparison to the ESM responses, we found a significant main effect for timing but not for contingency for smartphone usage behavior. That is, participants used their smartphones longer (i.e., 1.1 minutes; $\beta = 0.07$, $p < .001$) and more frequently (i.e., 0.40 unlocks; $\beta = 0.09$, $p < .001$) than was usual for them personally at the relevant time of day when they were in the varying compared to the fixed timing conditions (see also Figures 3.2a and 3.2b). However, the main effect of timing disappeared after excluding the surveys from the early morning intervals (usage duration: $\beta = 0.005$, $p = .747$; usage frequency: $\beta = 0.024$, $p = .141$). This may suggest similar measurement artifacts as for response probability as outcome (see RQ1).

Regarding the deviation in usage duration, we replicated the timing effect for the general public sample but not for the student sample. In addition, we found (opposite) interaction effects for timing and contingency in both subsamples, but not in the full sample. For the deviation of the number of unlocks, we replicated the timing effect in both subsamples and additionally found a contingency effect (direct > indirect) in the student sample.

Figure 3.3

Comparison of Association Patterns Between Primary Study Outcomes and External Constructs Depending on ESM protocols



Note. Grey boxes indicate differences in significance for the covariate-target combination across ESM protocols on the significance level $\alpha = .05$. Factor scores were computed for covariates. All covariates were assessed during the pre- or post-questionnaire and were grand mean centered. No transformation was applied to the four outcome variables.

3.4.3.2 Associations of Primary Study Outcomes With External Constructs.

Figure 3.3 displays the point estimates and 95% confidence intervals for the fixed effects of trait positive and negative affect, depression, satisfaction with life and psychological well-being on our four selected primary study outcomes as a function of the ESM protocols used (shown with different colors). In summary, we found divergent association patterns between ESM protocols for six out of 20 associations (see grey boxes in Figure 3.3). In other words, with a significance level of $\alpha = .05$, in six cases researchers would have come to different conclusions for the significance of associations depending on the applied ESM protocol. These cases were all for associations with the smartphone sensing outcomes. When using a Bonferroni corrected significance level of $\alpha = 0.05/20 = 0.25\%$ to address multiple testing issues, however, we did not find diverging effects for any of the 20 candidate associations.

3.5 Discussion

We examined whether different ESM protocols have side effects on study parameters using a within-subject design, sending ESM surveys with different timing and contingency over four weeks.

As expected with regard to data quantity, response probabilities were significantly higher in indirect protocols (triggered by smartphone unlocking) than in direct protocols (proactive notifications), regardless of timing (H1). The hypothesized effect regarding the combination of contingency and timing (H2) on response probability was not found. In addition, but only on a purely descriptive basis, participants dropped out less in the indirect protocols. These results were robust against different inclusion criteria for scheduled versus sent ESM surveys, as explored in additional sensitivity analyses (see Appendix C in the OSM).

With regard to data quality, we found that participants had longer response latencies in indirect protocols. No other data quality indicators were affected by contingency or timing. With regard to biases in study results, we found that participants used their smartphones more frequently and for longer in the varying protocols compared to the fixed protocols, but neither timing nor contingency affected self-reported primary study outcomes. While the association patterns between the ESM-based primary study outcomes and a diverse set of well-being measures as external criteria demonstrated convergence between the different ESM protocols, this was not the case for the associations with smartphone-based primary study outcomes. These discrepancies disappeared after correcting for multiple testing.

Ultimately, the findings of the study also suggest that side effects of timing and contingency may depend on the sample composition as we will discuss in more detail below. In conclusion, choosing an ESM protocol should involve carefully considering their potential side effects on a diverse set of study parameters. The next sections therefore highlight the key findings to guide ESM protocol design in future studies.

3.5.1 Contingency Side Effects on Efficiency and Ecological Validity of ESM Data Collection

Our results show that ESM data were collected more efficiently when surveys were triggered upon smartphone unlock (indirect mode) compared to when surveys were proactively sent (direct mode). This is evidenced by higher response probabilities in the indirect mode. The design of the indirect protocol makes it unlikely that participants accidentally fail to complete surveys (van Berkel, Goncalves, Lovén, et al., 2019). If they do not respond, it presumably is their deliberate choice (Reiter & Schoedel, 2024). In

addition, the greater efficiency of the indirect mode is also reflected in the lower number of participants who withdrew from the ESM study. It should be noted, however, that we explored the dropout counts only on a descriptive level and conclusions should therefore be drawn only cautiously. Accordingly, future research is needed to substantiate our conclusions.

The higher response probabilities (and descriptively) lower dropout rates may indicate that ESM notifications are less intrusive in the indirect mode (Spathis et al., 2019). Since they are triggered by self-initiated smartphone use, participants may find them less disruptive and more easily integrate them into their routine (Lathia et al., 2013). This is further supported by the fact that indirect protocols did not increase overall smartphone use, making the surveys feel like a by-product of other tasks (van Berkel et al., 2017). Thus, indirect ESM aligns with the call for more interruption-sensitive ESM studies (Mehrotra et al., 2015).

The findings also show that the efficiency of indirect ESM protocols does not negatively impact data quality or bias. The only drawback pointed out by the study is the significantly longer response latency—around 30 minutes more than with direct protocols—due to notifications being sent only when participants next use their smartphones, potentially increasing recall bias (Eisele et al., 2021). While ESM is generally valued for its high ecological validity as self-reports are collected in natural environments (Fuller-Tyszkiewicz et al., 2017), not all ESM protocols meet this standard equally (Ram et al., 2017). Accordingly, indirect protocols that allow ESM surveys to be sent only when the smartphone is actively used can compromise the principle of random sampling. Especially for ESM studies on rapidly changing psychological states, however, random sampling is a key design element. Consequently, the application of indirect protocols may result in biased data limited to certain experiences and contexts (van Berkel & Kostakos, 2021), thus threatening ecological validity.

In addition and from an applied perspective, it has to be noted that designing an ESM study using indirect protocols also comes with increased implementation efforts as ESM has to be combined with passive sensing. This may be the main reason for the prevalent majority of studies relying on direct contingency protocols in the current ESM research landscape.

To summarize, using indirect ESM protocols leads to higher response probabilities but at the same time incurs costs in terms of ecological validity and cost-effectiveness of implementation. Researchers should bear this trade-off in mind when deciding on a specific ESM protocol, depending on their specific research question.

3.5.2 Timing Side Effects and Time of Day Effects

We found significant effects of timing on participants' response probabilities and study outcomes. That is, in the varying mode compared to the fixed mode, participants responded with higher probabilities and used their smartphones more frequently (+0.4 unlocks per hour) and for longer (+1.1 minutes per hour) than their personal average.

This timing effect may be due to the fact that participants in the varying mode may find it more difficult to anticipate the occurrence of ESM surveys compared to those in the fixed mode (Myin-Germeys et al., 2018). As a result, they may check their phones more frequently, simultaneously increasing the likelihood of responding to the ESM surveys when they arrive. This finding suggests that researchers using ESM studies with passive data collection methods like smartphone sensing should be aware of potential interactions between data collection modes. While specific timing protocols may be necessary for certain research topics (Wrzus & Mehl, 2015), they can lead to reactive effects (Eisele et al., 2023), such as participants checking their phones more frequently to avoid missing surveys. This reactive behavior, in turn, can alter natural smartphone use and potentially increase response probability but bias results at the same time.

Alternatively, the timing effects could also just be methodological artifacts caused by the time of day when the ESM surveys were sent. Accordingly, they disappeared when the ESM surveys sent in the early morning interval were excluded from the analysis. As an illustration, smartphone usage was aggregated from 6:30–7:30am in the fixed mode, as the first ESM survey was scheduled at 7am each morning. In the varying mode, in contrast, the vast majority of ESM surveys was scheduled later (between 7am and 10am), shifting the aggregation window of smartphone use. Although we controlled for day-specific variations in smartphone usage, the size of the intervals used for doing so was quite large. The timing effect might therefore be overestimated as the

later it was in the morning interval, the more participants used their smartphone. This could, in turn, also explain the higher usage times and numbers for the varying mode.

Ultimately, we cannot assert with complete certainty that the timing effects were solely a methodological artifact since excluding the morning ESM surveys also led to a reduction in the number of observations, thereby decreasing the statistical power of our analyses. Nonetheless, this finding serves as a valuable reminder of the importance of careful consideration in ESM study design. Depending on the characteristics of the ESM approach, interactions with time of day effects can manifest, as illustrated in our case. This further highlights the need for rigorous testing prior to implementing complex ESM studies (Ebner-Priemer & Santangelo, 2024).

3.5.3 Side Effects of ESM Protocols and Their Dependence on Sample Characteristics

The analyses were conducted separately for the student sample and the general public sample. Given the modest subsample sizes ($N = 112$ and $N = 283$) and the exploratory nature of the analysis, we do not aim to provide a detailed interpretation of all results. However, our comparison suggests that the side effects of ESM protocol design may partially depend on sample characteristics (Eisele et al., 2021; Hasselhorn et al., 2022). For example, in the student sample, even though being the smallest subsample, timing, contingency, and their interaction significantly influenced the reported positive affect levels and deviations from their personal average. This might be due to factors like students being younger, more motivated by compensation, more experienced with studies, or having less structured daily routines. Our findings only offer preliminary evidence that the side effects of ESM protocols may vary by sample. Future studies should systematically investigate these characteristics in greater depth (Wrzus & Neubauer, 2023).

3.5.4 Limitations & Outlook

Our study has some limitations that may serve as starting point for future research. First, our study focused on timing and contingency as two important design characteristics of ESM protocols. At least for contingency, we remained on the surface, focusing only on the distinction between direct and indirect based on smartphone use as a

trigger. Yet, as the combination of ESM and smartphone sensing grows, the nature of the trigger may become more complex. For example, first studies use specific activities such as sitting (based on accelerometer data; Giurgiu et al., 2020) or listening to music (based on music logs; Sust & Schoedel, 2024) to trigger ESM surveys. This interplay of sensing methods and active assessment was already identified as having a large potential for psychological research (Ebner-Priemer & Santangelo, 2024). Thus, our study is only a first step in investigating the side effects of ESM protocols on ESM study parameters.

Second, we examined only a small subset of possible ESM study outcomes. For the primary study outcomes, we limited our focus to affective states and basic indicators of smartphone use. The side effects of ESM protocols may differ for other momentary behaviors, such as social interactions, or cognitive states, such as rumination or fatigue, that are captured by ESM surveys (Eisele et al., 2023). They might also differ for other indicators derived from smartphone sensing such as app usage or physical activity. Again, our study is only a first step, and future studies should investigate generalizability to a larger space of ESM survey parameters.

Third, we compared smartphone usage behavior for the periods around the respective ESM surveys with person- and time-of-day-specific behavior without consideration of ESM surveys. Accordingly, our person-average smartphone measures do not represent a counterfactual assessment of participants' usual smartphone usage behavior. Participants were aware that they are participating in an ESM study in which they were asked to complete as many surveys as possible. This may be a potential source of bias arising from measurement reactivity. To address this, future research should therefore include a baseline assessment week of smartphone use only, without sending any ESM surveys.

Finally, the present study had a rather low overall response rate (i.e., 43.4%) when compared to other (psychological) ESM studies which often have response rates between 70 and 80% (Wen et al., 2017; Wrzus & Neubauer, 2023). Even though we rigorously tested the research app with multiple pilot participants using different Android versions, technical errors cannot be entirely ruled out. Still, we believe that technical issues did not exclusively or primarily cause the low compliance rates due to different reasons.

We used the same app in previous studies with higher response rates (e.g., Reiter & Schoedel, 2024). Moreover, we applied very conservative inclusion criteria from a technical perspective. As preregistered and as a plausibility check, we only included participants for whom we could assure that at least 70 ESM surveys were sent out and that ESM surveys were generally triggered correctly.

Rather, we think that the low overall response rate is related to some of the decisions we made regarding the design of the ESM study. First, it may be due to the high participant burden of the present study. In contrast to other ESM studies, the study included smartphone sensing and lasted four weeks. This duration is more than twice as long as an average psychological ESM study (i.e., 12.4 days, as reported in Wrzus & Neubauer, 2023). Accordingly, our additional analysis showed that response probabilities decreased for later study days.

Second, the low overall response rate may also be related to our decision to start the ESM intervals already at 7am. On a descriptive level, the morning intervals from 7am to 10am showed the lowest response rates within each ESM protocol. In particular for direct protocols, we observed exceptionally low response rates (i.e., 21.3% and 15.2%) in the morning intervals compared to the later time-of-day intervals. This is even exacerbated in the fixed-direct ESM protocol as for all morning intervals ESM surveys were scheduled for 7am and—due to the study procedures—were only available for 15 minutes after they had been sent. This early timing of ESM surveys might not have harmonized well with the daily routine of many of our participants, especially since about a third of our sample were students who might have been asleep at the time.

Third, from a compliance-related perspective, the inclusion criteria for the present study were rather liberal. Participants had to answer 10 ESM surveys only in order to be included. Consequently, our sample also included low-compliance participants which might have further contributed to overall low and rather pessimistic response rates. In addition, the study's compensation strategy may also not have been conducive to participants' compliance. They were required to answer 50% of the ESM surveys to receive full compensation.

3.6 Conclusion

The present study systematically investigated potential side effects of different ESM protocols at large scale taking into account a variety of study outcome parameters. Our findings highlight the importance of carefully balancing data quantity, data quality, and potential bias in primary and secondary study outcomes when selecting an ESM protocol. Overall, the study demonstrates that while indirect ESM protocols can boost compliance and data collection efficiency, they do so at the potential cost of response latencies and ecological validity. In addition, we have shown that new challenges for ESM protocol selection arise when ESM data are combined with passive sensing data and that there may also be interactions between ESM protocols and times of day. Finally, we found side effects of ESM protocols on study parameters to be slightly different when considering samples with different characteristics. Accordingly, we recommend researchers to carefully weigh these trade-offs against their research goals and study populations of interest before adopting a specific ESM design.

3.7 Declarations

Funding and Acknowledgments:

This research was supported by funding provided to M. Back, M. van Zalk, and M. Buehner by the German Research Foundation (grant number: 458597616).

Conflicts of interest/Competing interests:

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval:

Approval was obtained from the ethics committee of LMU Munich. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate:

Informed consent was obtained from all individual participants included in the study.

Consent for publication:

Not applicable as data were anonymized and no identification of participants is possible.

Availability of data, materials, and code:

The pre-processed feature data are available at https://osf.io/a5bg4/ together with the preprocessing and analysis code for the study.

Declaration of Assistance:

When writing the manuscript, the authors used chatgpt (https://chat.openai.com) and DeepL (https://www.deepl.com) to ensure grammatical accuracy and rewording to improve the comprehensibility of the text. After using these tools, the authors reviewed and edited the content and take full responsibility for the content of the publication.

Authors' contributions:

TR: Conceptualization, Investigation, Data Curation, Formal Analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing, Project Administration

SS, JS, JtH: Investigation, Data Curation, Writing - Review & Editing, Project Administration

MvZ, MBa, MBü: Writing - Review & Editing, Project Administration, Funding Acquisition

RS: Conceptualization, Data Curation, Investigation, Methodology, Supervision, Writing – Review & Editing

3.8 References

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software*, *111*(9), 1–32. https://doi.org/10.18637/jss.v111.i09

Bachmann, A., Zetzsche, R., Schankin, A., Riedel, T., Beigl, M., Reichert, M., Santangelo, P., & Ebner-Priemer, U. (2015). Esmac: A web-based configurator for context-aware experience sampling apps in ambulatory assessment. *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, 15–18. https://doi.org/10.4108/eai.14-10-2015.2261679

Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 108–123). Guilford Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Businelle, M. S., Hébert, E. T., Shi, D., Benson, L., Kezbers, K. M., Tonkin, S., Piper, M. E., & Qian, T. (2024). Investigating best practices for ecological momentary assessment: Nationwide factorial experiment. *Journal of Medical Internet Research*, *26*, e50275. https://doi.org/10.2196/50275

Christensen, M. A., Bettencourt, L., Kaye, L., Moturu, S. T., Nguyen, K. T., Olgin, J. E., Pletcher, M. J., & Marcus, G. M. (2016). Direct measurements of smartphone screen-time: Relationships with demographics and sleep. *PLoS ONE*, *11*(11), e0165331. https://doi.org/10.1371/journal.pone.0165331

Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics. *Psychological Assessment*, *24*(3), 713–720. https://doi.org/10.1037/a0026733

Davanzo, A., Seker, S., Moessner, M., Zimmermann, R., Schmeck, K., Behn, A., et al. (2023). Study features and response compliance in ecological momentary assessment research in borderline personality disorder: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *25*(1), e44853. https://doi.org/10.2196/44853

DeSimone, J. A., & Harms, P. (2018). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology*, *33*, 559–577. https://doi.org/10.1007/s10869-017-9514-9

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Ebner-Priemer, U. W., & Santangelo, P. (2024). Viva experience sampling: Combining passive mobile sensing with active momentary assessments. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications* (pp. 311–328). Guilford Press.

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

Eisele, G., Vachon, H., Lafit, G., Tuyaerts, D., Houben, M., Kuppens, P., Myin-Germeys, I., & Viechtbauer, W. (2023). A mixed-method investigation into measurement reactivity to the experience sampling method: The role of sampling protocol and individual characteristics. *Psychological Assessment*. https://doi.org/10.1037/pas0001177

Eisele, G., Vachon, H., Myin-Germeys, I., & Viechtbauer, W. (2021). Reported affect changes as a function of response delay: Findings from a pooled dataset of nine experience sampling studies. *Frontiers in Psychology*, *12*, 580684. https://doi.org/10.3389/fpsyg.2021.580684

Fischer, J. E., Greenhalgh, C., & Benford, S. (2011). Investigating episodes of mobile phone activity as indicators of opportune moments to deliver notifications. *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*, 181–190. https://doi.org/10.1145/2037373.2037402

Fullagar, C. J., & Kelloway, E. K. (2009). Flow at work: An experience sampling approach. *Journal of Occupational and Organizational Psychology*, *82*(3), 595–615. https://doi.org/10.1348/096317908X357903

Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomyn, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment*, *29*(9), 1120. https://doi.org/10.1037/pas0000411

Ghosh, S., Ganguly, N., Mitra, B., & De, P. (2019). Designing an experience sampling method for smartphone based emotion detection. *IEEE Transactions on Affective Computing*, *12*(4), 913–927. https://doi.org/10.1109/TAFFC.2019.2905561

Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, *36*(2), 410–420. https://doi.org/10.1027/1015-5759/a000526

Giurgiu, M., Niermann, C., Ebner-Priemer, U., Kanning, M., et al. (2020). Accuracy of sedentary behavior–triggered ecological momentary assessment for collecting contextual information: Development and feasibility study. *JMIR mHealth and uHealth*, *8*(9), e17852. https://doi.org/10.2196/17852

Gloster, A. T., Miche, M., Wersebe, H., Mikoteit, T., Hoyer, J., Imboden, C., Bader, K., Meyer, A. H., Hatzinger, M., & Lieb, R. (2017). Daily fluctuation of emotions and memories thereof: Design and methods of an experience sampling study of major depression, social phobia, and controls. *International Journal of Methods in Psychiatric Research*, *26*(3), e1578. https://doi.org/10.1002/mpr.1578

große Deters, F., & Schoedel, R. (2024). Keep on scrolling? using intensive longitudinal smartphone sensing data to assess how everyday smartphone usage behaviors are related to well-being. *Computers in Human Behavior*, *150*, 107977. https://doi.org/10.1016/j.chb.2023.107977

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

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Hartmann, J. A., Wichers, M., Menne-Lothmann, C., Kramer, I., Viechtbauer, W., Peeters, F., Schruers, K. R., van Bemmel, A. L., Myin-Germeys, I., Delespaul,

P., et al. (2015). Experience sampling-based personalized feedback and positive affect: A randomized controlled trial in depressed patients. *PLoS ONE*, *10*(6), e0128095. https://doi.org/10.1371/journal.pone.0128095

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2022). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*, *54*, 1541–1558. https://doi.org/10.3758/s13428-021-01683-6

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

Horstmann, K. T. (2021). Experience sampling and daily diary studies: Basic concepts, designs, and challenges. In *The handbook of personality dynamics and processes* (pp. 791–814). Elsevier. https://doi.org/10.1016/B978-0-12-813995-0.00030-3

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114. https://doi.org/10.1007/s10869-011-9231-8

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

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Larsen, R. J., & Kasimatis, M. (1991). Day-to-day physical symptoms: Individual differences in the occurrence, duration, and emotional concomitants of minor daily illnesses. *Journal of Personality*, *59*(3), 387–423. https://doi.org/10.1111/j.1467-6494.1991.tb00254.x

Lathia, N., Rachuri, K. K., Mascolo, C., & Rentfrow, P. J. (2013). Contextual dissonance: Design bias in sensor-based experience sampling methods. *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 183–192. https://doi.org/10.1145/2493432.2493452

Liebherr, M., Schubert, P., Antons, S., Montag, C., & Brand, M. (2020). Smartphones and attention, curse or blessing? - a review on the effects of smartphone usage on attention, inhibition, and working memory. *Computers in Human Behavior Reports*, *1*. https://doi.org/10.1016/j.chbr.2020.100005

Lucas, R. E., Wallsworth, C., Anusic, I., & Donnellan, M. B. (2021). A direct comparison of the day reconstruction method (drm) and the experience sampling method (esm). *Journal of Personality and Social Psychology*, *120*(3), 816–835. https://doi.org/10.1037/pspp0000289

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Mehrotra, A., Vermeulen, J., Pejovic, V., & Musolesi, M. (2015). Ask, but don't interrupt: The case for interruptibility-aware mobile experience sampling. *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*, 723–732. https://doi.org/10.1145/2800835.2804397

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. https://doi.org/10.1177/1745691612441215

Moskowitz, D. S., & Coté, S. (1995). Do interpersonal traits predict affect? a comparison of three models. *Journal of Personality and Social Psychology*, *69*(5), 915–924. https://doi.org/10.1037/0022-3514.69.5.915

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry*, *17*(2), 123–132. https://doi.org/10.1002/wps.20513

Neubauer, A. B., Scott, S. B., Sliwinski, M. J., & Smyth, J. M. (2020). How was your day? convergence of aggregated momentary and retrospective end-of-day affect

ratings across the adult life span. *Journal of Personality and Social Psychology*, *119*(1), 185–203. https://doi.org/10.1037/pspp0000248

Ottenstein, C., & Werner, L. (2021). Compliance in ambulatory assessment studies: Investigating study and sample characteristics as predictors. *Assessment*, *29*(8), 1765–1776. https://doi.org/10.1177/10731911211032718

Pargent, F., Koch, T. K., Kleine, A.-K., Lermer, E., & Gaube, S. (2024). A tutorial on tailored simulation-based sample-size planning for experimental designs with generalized linear mixed models. *Advances in Methods and Practices in Psychological Science*, *7*(4). https://doi.org/10.1177/25152459241287132

Przybylski, A. K., & Weinstein, N. (2017). A large-scale test of the goldilocks hypothesis: Quantifying the relations between digital-screen use and the mental well-being of adolescents. *Psychological Science*, *28*(2), 204–215. https://doi.org/10.1177/0956797616678438

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Research in Human Development*, *14*(3), 253–270. https://doi.org/10.1080/15427609.2017.1340052

Reis, H. T. (2012). Why researchers should think "real-world": A conceptual rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 3–21). Guilford Press.

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

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, *69*(4), 719. https://doi.org/10.1037/0022-3514.69.4.719

Scharbert, J., Humberg, S., Kroencke, L., Reiter, T., Sakel, S., Ter Horst, J., Utesch, K., Gosling, S. D., Harari, G., Matz, S. C., et al. (2024). Psychological well-being in europe after the outbreak of war in ukraine. *Nature Communications*, *15*(1), 1202. https://doi.org/10.1038/s41467-024-44693-6

Scharbert, J., Reiter, T., Sakel, S., Ter Horst, J., Geukes, K., Gosling, S. D., Harari, G., Kroencke, L., Matz, S., Schoedel, R., et al. (2023). A global experience-sampling method study of well-being during times of crisis: The coco project. *Social and Personality Psychology Compass*, *17*(10), e12813. https://doi.org/10.1111/spc3.12813

Scharbert, J., Utesch, K., Reiter, T., ter Horst, J., van Zalk, M., Back, M. D., & Rau, R. (2024). If you were happy and you know it, clap your hands! testing the peak-end rule for retrospective judgments of well-being in everyday life. *European Journal of Personality*. https://doi.org/10.1177/08902070241235969

Schoedel, R., & Mehl, M. (2024). Mobile sensing methods. In H. T. Reis, T. West, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (3rd, pp. 297–321). Cambridge University Press. https://doi.org/10.1017/9781009170123.014

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology*, *125*(6), 1442–1471. https://doi.org/10.1037/pspp0000469

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, *31*(4), 471–481. https://doi.org/10.1177/0894439313479902

Sokolovsky, A. W., Mermelstein, R. J., & Hedeker, D. (2014). Factors predicting compliance to ecological momentary assessment among adolescent smokers. *Nicotine & Tobacco Research*, *16*(3), 351–358. https://doi.org/10.1093/ntr/ntt154

Spathis, D., Servia-Rodriguez, S., Farrahi, K., Mascolo, C., & Rentfrow, J. (2019). Passive mobile sensing and psychological traits for large scale mood prediction.

Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, 272–281. https://doi.org/10.1145/3329189.3329213

Stieger, S., Lewetz, D., & Swami, V. (2021). Emotional well-being under conditions of lockdown: An experience sampling study in austria during the covid-19 pandemic. *Journal of Happiness Studies*, *22*(6), 2703–2720. https://doi.org/10.1007/s10902-020-00337-2

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

Sust, L., & Schoedel, R. (2024). *Investigating everyday music choice on smartphones: The role of personality traits and mood states*. https://doi.org/10.31234/osf.io/4en3j

Tripathi, N., Zhu, J., Jacob, G. H., Frese, M., & Gielnik, M. M. (2020). Intraindividual variability in identity centrality: Examining the dynamics of perceived role progress and state identity centrality. *Journal of Applied Psychology*, *105*(8), 889–906. https://doi.org/10.1037/apl0000465

Ushey, K., & Wickham, H. (2024). *Renv: Project environments* [R package version 1.0.7, https://github.com/rstudio/renv]. https://rstudio.github.io/renv/

van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, *50*(6), 1–40. https://doi.org/10.1145/3123988

van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., & Kostakos, V. (2019). Context-informed scheduling and analysis: Improving accuracy of mobile self-reports. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300281

van Berkel, N., Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. (2019). Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies*, *125*, 118–128. https://doi.org/10.1016/j.ijhcs.2018.12.002

van Berkel, N., & Kostakos, V. (2021). Recommendations for conducting longitudinal experience sampling studies. In E. Karapanos, J. Gerken, J. Kjeldskov, & M. B.

Skov (Eds.), *Advances in longitudinal hci research. human–computer interaction series*. Springer, Cham. https://doi.org/10.1007/978-3-030-67322-2_4

Verhagen, S. J., Daniëls, N. E., Bartels, S. L., Tans, S., Borkelmans, K. W., de Vugt, M. E., & Delespaul, P. A. (2019). Measuring within-day cognitive performance using the experience sampling method: A pilot study in a healthy population. *PLoS ONE*, *14*(12), e0226409. https://doi.org/10.1371/journal.pone.0226409

Ward, M., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Watson, D., & Clark, L. A. (1994). The panas-x: Manual for the positive and negative affect schedule-expanded form. *Unpublished manuscript, University of Iowa*. https://doi.org/10.1037/t04754-000

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, *54*(6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *19*(4), e132. https://doi.org/10.2196/jmir.6641

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, *59*(3), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x

Wong, M. M., & Csikszentmihalyi, M. (1991). Motivation and academic achievement: The effects of personality traits and the duality of experience. *Journal of Personality*, *59*(3), 539–574. https://doi.org/10.1111/j.1467-6494.1991.tb00259.x

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

Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? measuring personality processes and their social consequences. *European Journal of Personality*, *29*(2), 250–271. https://doi.org/10.1002/per.1986

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

4 General Discussion

The present dissertation investigated different methodological aspects of the Experience Sampling Method (ESM). To this end, two studies were conducted to enhance the methodological understanding of ESM and to help inform design-related decisions for applied ESM researchers.

The first study examined the relevance of various factors in predicting survey-level (non-)compliance in ESM studies, using a sample of over 25,000 observations from 592 participants. More than 400 predictor variables carrying information about person, behavior, and context were collected using traditional surveys, experience sampling, and smartphone sensing. Using different machine learning models, participants' (non-)compliance was successfully predicted. This reveals a certain degree of systematic missingness in ESM response data which may introduce compliance bias if not properly accounted for. Moreover, it implies possible approaches for designing ESM protocols when the goal is to increase participants' compliance. Lastly, it can inform applied researchers about which variables to collect and include as control or auxiliary variables in the statistical models to potentially de-bias their causal estimates.

The second study investigated side effects of ESM protocols regarding data quality, data quantity, and bias in subsequent study results from an explanatory modeling perspective. For this, the two key ESM protocol characteristics timing and contingency were introduced and experimentally manipulated during a pre-registered, four-week study with 395 participants. The four different ESM protocols resulting from the 2×2 within-subjects design were evaluated regarding different criteria. As hypothesized for contingency, higher response probabilities but also higher response latencies were observed in indirect protocols. Contrary to our expectations, the combined effect of contingency and timing did not significantly influence response probability. No other effects of timing or contingency on data quality were observed. Smartphone usage behaviors, objectively assessed using smartphone sensing, varied depending on contingency whereas self-reported states were not affected by the choice of ESM protocol. Similar trends were observed in the associations between these primary study outcomes and external criteria such as trait affect and well-being. This implies that researchers should carefully consider potential side effects and trade-offs regarding

different study parameters when selecting or designing their ESM protocol.

A more extensive discussion of the results, implications, and limitations of both studies can be found in the respective chapters. In the remainder of this dissertation, the focus broadens to address overarching themes and areas not directly examined in the individual study manuscripts. First, the multifaceted contributions of the dissertation to methodological ESM research are outlined. This is followed by a discussion of the work's limitations and its broader implications. Finally, the dissertation concludes with a brief outlook and suggested directions for future ESM research.

4.1 Multifaceted Contributions for Methodological ESM Research

This dissertation offers several distinct contributions, which can be viewed from multiple methodological perspectives. The following sections discuss these contributions in a structured manner.

4.1.1 Improving the Understanding and Foundation of ESM Design Decisions

First and foremost, the present dissertation directly contributes to a deeper methodological understanding of ESM. Although the general advantages of experience sampling are well understood and appreciated (Scollon et al., 2003; Verhagen et al., 2016), there often is a lack of empirical evidence when it comes to specific decisions in the design of ESM studies. Consequently, researchers often rely on a priori assumptions or pragmatic considerations when making critical decisions during study design (Fritz et al., 2024; Himmelstein et al., 2019). As a response to this, the present dissertation thoroughly investigated two issues associated with ESM, namely systematic non-response or non-compliance and potential side effects of different ESM protocols regarding data quantity, data quality, or bias in subsequent study results. The findings of this dissertation represent examples of empirical evidence that can be used to guide study design and data analysis in the context of ESM. For example, it can inform applied researchers which variables to collect and include in subsequent analyses to mitigate compliance bias. This may eventually help to increase validity of findings in psychological research. Moreover, this dissertation can provide guidance to applied researchers deciding which ESM protocol to use or which sensor-based triggers to use if the goal is to increase

compliance and thus data quantity. However, this work also raises awareness of possible trade-offs researchers should keep in mind when designing their ESM studies. For example, it was found that triggering ESM surveys based on smartphone sensing reduces the likelihood of participants unintentionally missing surveys. Accordingly, it can represent a promising approach for researchers who want to optimize compliance. For example, this could be beneficial in clinical settings, where researchers or practitioners want to ensure that participants do not overlook surveys regarding high-importance topics such as taking medicine. However, it should be noted that optimizing ESM study designs for compliance might also come with side effects such as increases in response latency or participants' smartphone usage. Accordingly, in the clinical medication example, a practitioner would have to accept certain drawbacks regarding other study outcomes. Moreover, selecting only situations with high predicted compliance may, in turn, reduce the representativeness of the data collected. Although this may not pose a problem when asking participants about medication, it may be problematic for other topics such as participants' mood. One possible solution to handle this could could involve strategically blending surveys scheduled for maximum predicted compliance with surveys scheduled for maximum situational randomness and representativeness. As can be seen from this example, even though ESM is in general valued for its high ecological validity, the present study acts as a reminder to researchers that not all ESM protocols may meet this standard equally (Ram et al., 2017). Weighing the respective strengths and weaknesses of different design choices and determining their appropriateness for answering a given research question remains a substantive decision that must ultimately be made by the applied researcher (Dejonckheere & Erbas, 2021). However, the present dissertation supports this decision-making process at multiple stages by strengthening the empirical and methodological foundation upon which these decisions can be based.

4.1.2 Integrating Smartphone Sensing into Methodological ESM Research

Being appreciated for their objectivity, non-intrusiveness, and high temporal granularity, smartphone sensing methods are increasingly utilized in applied psychological research (Harari, Müller, et al., 2017; Krämer et al., 2024; Müller et al., 2020; Schoedel et al., 2023). This dissertation showcases how not only applied research but also

methodological research can benefit from integrating smartphone sensing. Methodological researchers investigating non-compliance, so far, mostly relied on general participant information collected during one-time surveys (e.g., age, gender, personality traits; Rintala et al., 2019; Silvia et al., 2013), very general contextual information (e.g., weekday or daytime; Csikszentmihalyi & Hunter, 2003; Rintala et al., 2020), or information collected during earlier ESM surveys (e.g., mood or stress reported at previous beeps; Murray et al., 2023; Sokolovsky et al., 2014). Smartphone sensing, as can be seen from the two studies, provides a new way of collecting data —even in cases other data collection approaches fail— which, until now, has not been leveraged in methodological ESM research. Indeed, in the first study information such as whether participants were currently at home, at work, or in a train could be derived from sensed data even if the ESM surveys themselves were not answered. Thus information proved relevant for predicting (non-)compliance and thus to directly benefit methodological ESM research.

The second study introduced another way of how methodological ESM research can benefit from including smartphone sensing by directly integrating smartphone sensing into the design and evaluation process of ESM protocols. Here, particular emphasis was placed on the possibility of using smartphone sensing for designing ESM protocols. This reflects a research approach that partly diverges from the initial use of smartphone sensing in psychological research — namely, to substantiate more applied research questions or predictions with objective data (Harari, Gosling, et al., 2017; Müller et al., 2020; Wang et al., 2018). However, especially in other disciplines such as human-computer-interaction (HCI), researchers already started to creatively integrate smartphone sensing into the design of self-report data collection (e.g., by triggering ESM surveys based on ambient noise, light, CO_2, the number of people in a room or users' typing patterns; Ghosh et al., 2015; Lim et al., 2024). Likewise, first similar advances were made in psychology, for example by triggering surveys based on music listening behavior (Sust & Schoedel, 2024) or incoming calls (Roos et al., 2023). Furthermore, HCI researchers proposed different prototypes for capturing participants' self-reports at well-selected times based on smartphone sensing measures (Bachmann et al., 2015; Ferreira et al., 2014). However, as a discipline, HCI tends to emphasize technical feasibility over thorough evaluation of data collection methods — especially with regard to their effects on psychological study outcomes. As a response to this, the

present dissertation integrates smartphone sensing into the ESM design and evaluation process explicitly focusing both on general aspects of data quantity and quality, and on more traditional psychological study outcomes. Thus, just in 2025 —the very year envisioned by Miller (2012) for the predictions in his *Smartphone Psychology Manifesto*— this dissertation offers empirical examples of how smartphones can drive change in (methodological) psychological research.

4.1.3 Drawing on Statistical Approaches for Explanation and Prediction

Besides combining different data types, the present dissertation also combined different approaches from study design and statistical analysis to investigate the proposed research questions. More precisely, the two studies of this dissertation can be differentiated with respect to whether they relied on observational data vs. experimental data and whether they used an explanatory vs. predictive modeling approach. The second study applied a "traditional" randomized experimentation approach as it comprised randomized allocation of participants to the different possible treatment orders (i.e., orders of ESM Protocols). This was paired with a "classical" (based on what psychologists are usually trained in (Pargent et al., 2023)) explanatory or data modeling approach to uncover the effects of specific ESM design aspects. Randomized experiments are often considered the preferred way of learning about causal effects as they eliminate alternative explanations for a given treatment effect of interest (e.g., Campbell & Stanley, 1963; Grosz et al., 2024). By stating (and preregistering) hypotheses about design effects together with experimental data, the present study followed a "proper" explanatory modeling approach as opposed to the more common use of association-based statistical models applied to observational data (Shmueli, 2010).

Study 1, in turn, followed an algorithmic or predictive modeling approach utilizing observational data and different machine learning algorithms to predict participants' (non-)compliance during an ESM study. Contrary to Study 2, the goal here was less to explain or understand the effects of specific design aspects, but rather to accurately predict whether a participant will answer a given ESM survey using all data available up to this point in time. This aligns more closely with applied settings in which researchers or practitioners aim to trigger surveys at moments that are most convenient or opportune

for participants, patients, or users. In practice, this may be relevant whenever the primary objective is to maximize the likelihood that surveys are completed. Although in Study 1 methods from interpretable machine learning were used to better understand the models' predictions, the overall approach remains fundamentally predictive in nature. This in general prohibits any careless causal or explanatory interpretation of the results (Pargent et al., 2023). Nevertheless, even if sometimes termed unacademic, prediction or predictive modeling serves different scientific functions associated with development and testing of new theories (Shmueli, 2010). Apart from this, following a predictive approach directly responds to a call in psychology about increasing the focus on prediction not least as a result of many psychological theories failing to demonstrate satisfactory predictive performance (Yarkoni & Westfall, 2017). As can be seen, both explanatory and predictive modeling approaches come with their own advantages, challenges, and goals and which one is the most appropriate solution to a problem depends on the problem itself and the data at hand (Breiman, 2001).

Accordingly, the present dissertation serves as an example of how the two approaches can jointly enhance the methodological understanding of experience sampling. In doing so, it may be seen as a step towards the integrative and complementary research strategy advocated by Mahmoodi et al. (2017) in the context of big data in the social and behavioral sciences.

4.2 Limitations and Implications for Methodological ESM Research

While this dissertation offers several contributions, it is also subject to certain limitations. The following sections outline and reflect on these limitations in a structured manner.

4.2.1 Sample Selectivity as Potential Threat to Generalization

As already outlined in the included studies, selective sampling may introduce biased estimates and threaten the generalization of results (Demark-Wahnefried et al., 2011; Elwert & Winship, 2014). The samples of the studies included in this dissertation can be seen from two different levels. On the one hand, the sample can be seen from the level of unique participants, whereas, on the other hand, it can also be seen from the level of ESM surveys nested in participants. Selective sampling may occur at both

levels with the present dissertation explicitly investigating (non-)compliance or selective sampling at the level of ESM surveys nested in participants. This required both studies to account for the nested data structure via appropriate methods proposed for the different modeling approaches. More precisely, Study 1 applied blocked resampling (blocked by participant), which is often proposed to prevent overly optimistic estimates of predictive performance due to both the training and test set containing data points of the same participant (Dragicevic & Casalicchio, 2020). In Study 2, mixed effects models were used to directly reflect the nested data structure by including random effects for the different study participants (Pinheiro & Bates, 2000). Assuming that participants in both studies represent a true random sample of participants would support the generalizability of the findings. However, this is a strong assumption which is rarely met in practice, especially with psychological research often relying on convenience samples, student populations, or participants from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010; Peterson & Merunka, 2014; Sears, 1986). The present dissertation shares this limitation as it only includes participants that were users of an android smartphone and self-selected into the respective smartphone sensing study. This in general threatens generalization of results to populations other than the ones the participants were recruited from. Samples from other populations might differ in characteristics relevant to the question of interest (Henrich et al., 2010). For example, the second study found differences between the student and the general public subsample regarding the effects of timing and contingency on self-reported positive affect. These differences could potentially arise from differences in daily routines or study motivation and experience between the two subpopulations. Such differences substantiate the suspicion that it cannot safely be assumed that the results of this dissertation can be readily generalized to other populations. Still, although this concern can hardly be fully resolved, we hope that the results of this study possess at least some degree of generalizability, for different reasons. For example, when comparing users of the two most common smartphone operating systems, Android and iOS, only negligible to small differences were found when considering key personality traits (Götz et al., 2017). Moreover, a recent study comparing sample characteristics across different stages of self-selection into a study including smartphone sensing (e.g., initial interest, study signup, app installation,

finishing the study) did not suggest strong self-selection biases in studies including mobile sensing (Schoedel et al., 2025). Still, by making the issue of selective sampling explicit, the present dissertation aims to provide full transparency about its limitations and the associated constraints to generality to discourage readers from assuming the broadest possible generalizations (Simons et al., 2017).

4.2.2 Subjectivity in (Pre-)Processing of Sensing Data

Both studies of the present dissertation relied on variables that were gathered via smartphone sensing or more precisely, variables that were extracted from the raw, time-stamped smartphone sensing logs gathered via the PhoneStudy research app (https://www.phonestudy.org/en). Although these smartphone sensing logs themselves —ignoring possible logging errors— can be considered highly objective, the process of feature extraction and data analysis contains a non-negligible amount of subjectivity. As part of the preprocessing and feature extraction pipeline researchers have to make many decisions introducing a large number of researcher degrees of freedom (Langener, Siepe, et al., 2024). One example are the variables about whether participants were currently at home or at work used in the first study which were derived from smartphone sensed GPS data (Langener, Stulp, et al., 2024). GPS data can be considered to contain most of the information necessary for inferring whether a participant is currently at home or at work. Still, the best way of extracting this information from the raw GPS data, as is also the case for many other sensing-based indicators, still lacks guidelines and common standards (Wrzus & Schoedel, 2024). Data transformation or preprocessing heuristics such as labeling the location visited most frequently between 9 a.m. and 4 p.m. as “work” may have intuitive appeal and apparent face validity; however, there will be situations in which this approach fails. This may not necessarily represent an issue in predictive modeling settings as in Study 1. Here the goal usually is to achieve predictive performances as high as possible and "feature engineering" is considered a highly task- and modeler-dependent step during modeling which lacks an indisputable "correct" solution or gold standard (Kuhn, Johnson, et al., 2013; Verdonck et al., 2024). It may, however, represent an issue in cases where the research question is more explanatory in nature as, in this case, the potential for distorted and/or non replicable results is exaggerated (Simmons et al., 2011). To address this issue, different open

science standards and practices were proposed for studies relying on passive sensing or smartphone sensing measures (Langener, Siepe, et al., 2024). This includes the preregistration of study designs, data preprocessing decisions, and hypotheses and the general open-science principle of sharing preprocessing and analysis code or raw and preprocessed data (Harari et al., 2024; Langener, Siepe, et al., 2024; Wrzus & Schoedel, 2024). The present dissertation tried to adhere to these principles as well as possible. Yet, some of them can be considered hard to meet in practice and are unfortunately also not completely satisfied by the present dissertation. For example, the present dissertation did not fully adhere to the principle of open data when it comes to the raw non-preprocessed data. In practice, this principle may to always be feasible due to the sensitivity of the raw sensing data (e.g., raw GPS coordinates) and associated data privacy and data protection legislation. Still, the present dissertation made an effort to implement as many of the proposed measures as possible. This includes the precise preregistration of confirmatory hypotheses, exclusion criteria, study measures considered, and statistical analyses and the general sharing of preprocessed data, analysis code, and codebooks via the Open Science Framework (OSF) repositories for the respective studies. Although the present dissertation may not fully adhere to all standards, it demonstrates that it is possible to balance the multitude of decisions involved in modeling smartphone sensing data with the demands of transparency and reproducibility.

4.2.3 Limited Focus: Empirical Insights for Few Issues out of Many

Lastly, despite providing general guidance and valuable additions to the methodological understanding of ESM protocols, there remain factors and design aspects around ESM that are out of scope of the present work and for which the level of knowledge is still limited. One exemplary factor not addressed at all in the present dissertation and mainly untouched by empirical ESM methodology research in general is the issue of questionnaire time-out or response delay. This describes the duration for which an ESM survey remains available and can be completed by participants. Although questionnaire availability was found to be on average 16.4 minutes (Deakin et al., 2022), there so far is no study which experimentally manipulated and evaluated the effects of different questionnaire availability settings. In line with this, Scollon et al. (2003) note

that, even though providing some uniformity to the data, the often observed 20- or 30-minute time window can be considered an arbitrary cutoff. It could be possible that different settings of questionnaire availability might have effects on the data collected and may consequently affect study findings (Scollon et al., 2003). Accordingly, the methodological understanding of ESM could benefit from studies empirically addressing this design aspect.

Apart from factors exceeding the scope of the present dissertation, even the factors directly addressed may not always result in comprehensive guidelines for study design. For example, predictive performance in Study 1 may still have been improved by inclusion of further additional sensor data (e.g., by including further technical devices such as smartwatches) or extraction or derivation of further additional features. The same holds true for Study 2. Moreover, although Study 2 proposed and empirically investigated the characteristics of timing and contingency for ESM protocols, there remain many other related ESM protocol design characteristics for which empirical knowledge remains limited. For example, smartphone sensing offers the possibility of ESM surveys to be triggered in response to detected activities or contextual factors such as physical activity or being in certain environments like public parks or restaurants. However, these triggering mechanisms —despite some noteworthy implementations in applied studies (e.g., Delobelle et al., 2025; Shevchenko & Reips, 2024; Törnros et al., 2016)— have so far not been comprehensively evaluated regarding aspects such as data missingness or their potentially distorting effects on study results. In summary, there exist pioneering articles providing guidelines for researchers (Fritz et al., 2024; Stone et al., 2023; van Berkel & Kostakos, 2021) planning to conduct an ESM study and the present dissertation contributed to this methodological understanding of ESM at multiple stages. However, there remain decisions for which the present dissertation does not provide further guidance and empirical evidence remains limited. Although each of these design aspects may seem minor in isolation, taken together they may impact study results and thus, at least to a certain extent, could threaten the validity and reproducibility of ESM studies.

4.3 Future Directions and Challenges

Considering the rise of ESM within the last decades in combination with its many proposed advantages, there is reason to expect this trend to continue. This dissertation presents some first use cases demonstrating how methodological ESM research can benefit from integrating smartphone sensing and combining approaches from explanatory and predictive modeling. Given the associated strengths and opportunities presented in this dissertation, it is likely that their integration into methodological ESM research will continue and even grow further. However, there are challenges and obstacles throughout the process of adoption which are briefly addressed in this section. First of all, increased integration of smartphone sensing into research will result in a variety of highly sensitive data which comes with different necessities regarding the establishment of standards and best practices. This includes, for example, the development of open science practices regarding smartphone sensing data but also general guidance regarding aspects such as safe storage or anonymization of data. These considerations are especially important given the sensitivity of such data that may include GPS information and therefore pose a risk of identifying individuals (De Montjoye et al., 2013). In line with this, researchers should always remind themselves to only collect data necessary for addressing their research questions at hand. Moreover, especially with respect to data preprocessing, researchers could benefit from developing general preprocessing and measure validation best-practices or guidelines beyond the mere open sharing of preprocessing pipelines. As a first step, this could especially be put into practice for often-used "standard variables" of smartphone usage. If this challenge is addressed effectively, it may greatly improve transparency, replicability, and eventually validity of psychological research that integrates smartphone sensing.

Second, to properly combine and adopt approaches from both explanatory and predictive modeling, psychologists need to be trained appropriately in both approaches. However, so far psychology or social science students are mostly trained in explanatory rather than predictive modeling — especially when it comes to non-linear models (Pargent et al., 2023). This is further aggravated by the fact that most papers published in psychology are concerned with criteria from the explanatory modeling rather than the predictive modeling world further exacerbating the lack of exposure to

predictive modeling in psychology (Yarkoni & Westfall, 2017). Accordingly, training future psychological researchers multi-methodically for applying and integrating both explanatory and predictive modeling approaches represents one challenge which —if overcome— psychological research could benefit from (Mahmoodi et al., 2017).

Lastly, despite an increased level of awareness, smartphone sensing should still be considered a relatively new approach to data collection and assessment in psychology. Accordingly, psychologists may still require and benefit from training in this specific method, for example by including it in study curricula, in order for it to be successfully added to the psychological toolbox. Positive developments regarding all of the previously named challenges may happen as cascading side-effects of the growing commercial availability of smartphone sensing solutions. Moreover, potentially large synergetic effects could arise from increased efforts in interdisciplinary collaboration and research. The same applies for the studies of the present dissertation, which were planned and designed in close collaboration with researchers from HCI and media informatics. This collaboration allowed different aspects of data collection and app or study design to be evaluated from multiple perspectives, ultimately leading to more suitable and effective outcomes. Moreover, increased collaboration can stimulate innovation and foster synergies in adjacent research areas such as privacy awareness, user behavior quantification, or user-centric design.

4.4 Conclusion

With increased technological tools and opportunities, researchers benefit from novel ways to conduct research. The present dissertation explored smartphone sensing and the role it can play in (methodological) ESM research. Experimentation and both explanatory and predictive modeling were integrated into the research process. Two empirical studies drawing upon the previously named methods and strategies yielded novel insights into (non-)compliance in ESM studies and the side effects of specific protocol characteristics. Thus, this dissertation adds to the body of empirical and methodological ESM research and can act as a reference point for applied researchers guiding certain decisions regarding design and analysis of ESM studies. On a higher level and beyond addressing the specific research questions, the present dissertation showcases possible applications of smartphone sensing and the interplay of explanatory

and predictive modeling with observational and experimental data. This may serve as an example that inspires future psychological research in general and future methodological research in the context of ESM in particular. Despite showcasing and leveraging the advantages of smartphone sensing and the integration of explanatory and predictive modeling, the present dissertation openly discusses the limitations associated with either approach. Challenges such as the currently prevalent lack of best-practices and standards regarding the preprocessing of smartphone sensing data are discussed as potential threats to transparency, comparability, and reproducibility of study findings. In response to this, current suggestions for addressing these issues are showcased and possible future measures for improvement are discussed and motivated. In conclusion, this dissertation contributes to methodological ESM research in two ways: by directly offering empirical insights and guidelines, and by showcasing possible new research approaches.

4.5 References

Bachmann, A., Zetzsche, R., Schankin, A., Riedel, T., Beigl, M., Reichert, M., Santangelo, P., & Ebner-Priemer, U. (2015). Esmac: A web-based configurator for context-aware experience sampling apps in ambulatory assessment. *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, 15–18. https://doi.org/10.4108/eai.14-10-2015.2261679

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally & Company: Chicago.

Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in everyday life: The uses of experience sampling. *Journal of Happiness Studies*, *4*(2). https://doi.org/10.1023/A:1024409732742

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1), 1376. https://doi.org/10.1038/srep01376

Deakin, E., Ng, F., Young, E., Thorpe, N., Newby, C., Coupland, C., Craven, M., & Slade, M. (2022). Design decisions and data completeness for experience sampling methods used in psychosis: Systematic review. *BMC Psychiatry*, *22*(1), 669. https://doi.org/10.1186/s12888-022-04319-x

Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing esm studies* (pp. 33–70). Leuven: Center for Research on Experience Sampling; Ambulatory Methods Leuven.

Delobelle, J., Compernolle, S., Vetrovsky, T., Van Cauwenberg, J., & Van Dyck, D. (2025). Contexts, affective and physical states and their variations during physical activity in older adults: An intensive longitudinal study with sensor-triggered event-based ecological momentary assessments. *International Journal of Behavioral Nutrition and Physical Activity*, *22*(1), 30. https://doi.org/10.1186/s12966-025-01724-9

Demark-Wahnefried, W., Bowen, D. J., Jabson, J. M., & Paskett, E. D. (2011). Scientific bias arising from sampling, selective recruitment, and attrition: The case for improved reporting. *Cancer Epidemiology, Biomarkers & Prevention*, *20*(3), 415–418. https://doi.org/10.1158/1055-9965.EPI-10-1169

Dragicevic, M., & Casalicchio, G. (2020). *Resampling—stratified, blocked and predefined*. Mlr-Org. Retrieved January 4, 2023, from https://mlr-org.com/gallery/basic/2020-03-30-stratification-blocking/

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L., & Dey, A. K. (2014). Contextual experience sampling of mobile application micro-usage. *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services*, 91–100. https://doi.org/10.1145/2628363.2628367

Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., et al. (2024). So you want to do esm? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Science*, *7*(3). https://doi.org/10.1177/25152459241267912

Ghosh, S., Chauhan, V., Ganguly, N., Mitra, B., & De, P. (2015). Impact of experience sampling methods on tap pattern based emotion recognition. *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*, 713–722. https://doi.org/10.1145/2800835.2804396

Götz, F. M., Stieger, S., & Reips, U.-D. (2017). Users of the main smartphone operating systems (ios, android) differ only little in personality. *PLoS ONE*, *12*(5), e0176921. https://doi.org/10.1371/journal.pone.0176921

Grosz, M. P., Ayaita, A., Arslan, R. C., Buecker, S., Ebert, T., Hünermund, P., Müller, S. R., Rieger, S., Zapko-Willmes, A., & Rohrer, J. M. (2024). Natural experiments: Missed opportunities for causal inference in psychology. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231218610

Harari, G. M., Gosling, S. D., Wang, R., Chen, F., Chen, Z., & Campbell, A. T. (2017). Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior*, *67*, 129–138. https://doi.org/10.1016/j.chb.2016.10.027

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

Harari, G. M., Soh, S., & Kroencke, L. (2024). How to conduct mobile sensing research. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications* (pp. 3–27). Guilford Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

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

Krämer, M. D., Roos, Y., Schoedel, R., Wrzus, C., & Richter, D. (2024). Social dynamics and affect: Investigating within-person associations in daily life using experience sampling and mobile sensing. *Emotion*, *24*(3), 878–893. https://doi.org/10.1037/emo0001309

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer. https://doi.org/10.1007/978-1-4614-6849-3

Langener, A. M., Siepe, B. S., Elsherif, M., Niemeijer, K., Andresen, P. K., Akre, S., Bringmann, L. F., Cohen, Z. D., Choukas, N. R., Drexl, K., et al. (2024). A template and tutorial for preregistering studies using passive smartphone measures. *Behavior Research Methods*, *56*, 8289–8307. https://doi.org/10.3758/s13428-024-02474-5

Langener, A. M., Stulp, G., Jacobson, N. C., Costanzo, A., Jagesar, R. R., Kas, M. J., & Bringmann, L. F. (2024). It's all about timing: Exploring different temporal res-

olutions for analyzing digital-phenotyping data. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231202677

Lim, J., Koh, Y., Kim, A., & Lee, U. (2024). Exploring context-aware mental health self-tracking using multimodal smart speakers in home environments. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18. https://doi.org/10.1145/3613904.3642846

Mahmoodi, J., Leckelt, M., van Zalk, M. W., Geukes, K., & Back, M. D. (2017). Big data approaches in social and behavioral science: Four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, *18*, 57–62. https://doi.org/10.1016/j.cobeha.2017.07.001

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*(3), 221–237. https://doi.org/10.1177/1745691612441215

Müller, S. R., Peters, H., Matz, S. C., Wang, W., & Harari, G. M. (2020). Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and gps tracking. *European Journal of Personality*, *34*(5), 714–732. https://doi.org/10.1002/per.2262

Murray, A. L., Brown, R., Zhu, X., Speyer, L. G., Yang, Y., Xiao, Z., Ribeaud, D., & Eisner, M. (2023). Prompt-level predictors of compliance in an ecological momentary assessment study of young adults' mental health. *Journal of Affective Disorders*, *322*, 125–131. https://doi.org/10.1016/j.jad.2022.11.014

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning: A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3). https://doi.org/10.1177/25152459231162559

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*(5), 1035–1041. https://doi.org/10.1016/j.jbusres.2013.08.010

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-plus*. Springer Science Business Media. https://doi.org/10.1007/b98882

Ram, N., Brinberg, M., Pincus, A. L., & Conroy, D. E. (2017). The questionable ecological validity of ecological momentary assessment: Considerations for design and analysis. *Research in Human Development*, *14*(3), 253–270. https://doi.org/10.1080/15427609.2017.1340052

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, *31*(2), 226–235. https://doi.org/10.1037/pas0000662

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2020). Momentary predictors of compliance in studies using the experience sampling method. *Psychiatry Research*, *286*, 112896. https://doi.org/10.1016/j.psychres.2020.112896

Roos, Y., Krämer, M. D., Richter, D., Schoedel, R., & Wrzus, C. (2023). Does your smartphone "know" your social life? a methodological comparison of day reconstruction, experience sampling, and mobile sensing. *Advances in Methods and Practices in Psychological Science*, *6*(3), 1–12. https://doi.org/10.1177/25152459231178738

Schoedel, R., Kunz, F., Bergmann, M., Bemmann, F., Bühner, M., & Sust, L. (2023). Snapshots of daily life: Situations investigated through the lens of smartphone sensing. *Journal of Personality and Social Psychology*, *125*(6), 1442–1471. https://doi.org/10.1037/pspp0000469

Schoedel, R., Reiter, T., Krämer, M. D., Roos, Y., Buehner, M., Richter, D., Mehl, M. R., & Wrzus, C. (2025). *Person-related selection bias in mobile sensing research: Robust findings from two panel studies* [Manuscript submitted for publication. Department of Psychology, LMU Munich].

Scollon, C. N., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*(1), 5–34. https://doi.org/10.1023/A:1023605205115

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515. https://doi.org/10.1037/0022-3514.51.3.515

Shevchenko, Y., & Reips, U.-D. (2024). Geofencing in location-based behavioral research: Methodology, challenges, and implementation. *Behavior Research Methods*, *56*(7), 6411–6439. https://doi.org/10.3758/s13428-023-02213-2

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, *31*(4), 471–481. https://doi.org/10.1177/0894439313479902

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Sokolovsky, A. W., Mermelstein, R. J., & Hedeker, D. (2014). Factors predicting compliance to ecological momentary assessment among adolescent smokers. *Nicotine & Tobacco Research*, *16*(3), 351–358. https://doi.org/10.1093/ntr/ntt154

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, *19*. https://doi.org/10.1146/annurev-clinpsy-080921-083128

Sust, L., & Schoedel, R. (2024). *Investigating everyday music choice on smartphones: The role of personality traits and mood states*. https://doi.org/10.31234/osf.io/4en3j

Törnros, T., Dorn, H., Reichert, M., Ebner-Priemer, U., Salize, H.-J., Tost, H., Meyer-Lindenberg, A., Zipf, A., et al. (2016). A comparison of temporal and location-based sampling strategies for global positioning system-triggered electronic diaries. *Geospatial Health*, *11*(3). https://doi.org/10.4081/gh.2016.473

van Berkel, N., & Kostakos, V. (2021). Recommendations for conducting longitudinal experience sampling studies. In E. Karapanos, J. Gerken, J. Kjeldskov, & M. B. Skov (Eds.), *Advances in longitudinal hci research. human–computer interaction series*. Springer, Cham. https://doi.org/10.1007/978-3-030-67322-2_4

Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2024). Special issue on feature engineering editorial. *Machine Learning*, *113*(7), 3917–3928. https://doi.org/10.1007/s10994-021-06042-2

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. *BMJ Mental Health*, *19*(3), 86–89. https://doi.org/10.1136/ebmental-2016-102418

Wang, W., Harari, G. M., Wang, R., Müller, S. R., Mirjafari, S., Masaba, K., & Campbell, A. T. (2018). Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–21. https://doi.org/10.1145/3264951

Wrzus, C., & Schoedel, R. (2024). Transparency and reproducibility in mobile sensing research. In M. R. Mehl, M. Eid, C. Wrzus, G. M. Harari, & U. W. Ebner-Priemer (Eds.), *Mobile sensing in psychology: Methods and applications* (pp. 53–84). Guilford Press.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393