At the Interface of Personality Psychology and Computational Science

Using Smartphones to Investigate Individual Differences

Ramona Schödel

München 2020

At the Interface of Personality Psychology and Computational Science

Using Smartphones to Investigate Individual Differences

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilians-Universität München

vorgelegt von
Ramona Schödel
aus Hof

München 2020

Erstgutachter:	Prof. Dr. Markus Bühner
Zweitgutachter:	Prof. Dr. Felix Schönbrodt
Tag der Abgabe:	18.03.2020
Tag der mündlichen Prüfung:	09.07.2020

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich in den letzten dreieinhalb Jahren begleitet und unterstützt haben.

Ich möchte mich bei meinem Doktorvater Prof. Dr. Markus Bühner für seine Unterstützung, den Freiraum kreativ arbeiten und mich entfalten zu können, und seine stets ansteckende Begeisterung bedanken. In der Zeit meiner Dissertation habe ich das interdisziplinäre Arbeiten kennen und schätzen lernen dürfen. Vielen Dank geht hierbei an Clemens Stachl, der mir im Rahmen der Projekte NZML und PhoneStudy den Raum dafür gegeben hat. Ein großes Dankeschön geht dabei an Sarah Völkel und Quay Au, meine beiden "Pendants" in der Informatik und der Statistik. Es hat mir immer viel Spaß gemacht mit euch zusammenzuarbeiten und durch euch einen Einblick in eure Disziplinen zu bekommen. Vielen Dank möchte ich in diesem Zusammenhang auch an Prof. Dr. Bernd Bischl und Prof. Dr. Heinrich Hussmann richten, die die interdisziplinäre Zusammenarbeit stets unterstützt haben.

Weiterhin möchte ich mich bei der AUDI AG für die Möglichkeit der Durchführung eines Kooperationsprojekts im Rahmen des Doktorandenprogramms INI.LMU bedanken. Stellvertretend geht hier ein Danke an meine Projektansprechpartner Stephan Hummel und Anna Krahnstöver.

Ein großes Dankeschön geht auch an Julia Graefe, Henrike Haase, Maria Vazquez, Janina Bindschädel, Alena Koch, Regina Oberwesterberger, Noel Lommer, Katja Frey und Michelle Oldemeier, die mich als Praktikanten oder Abschlussarbeitsschreiber in den letzten dreieinhalb Jahren begleitet haben. Ein weiteres großes Dankschön geht an das PhoneStudy Team - ohne euch wäre meine Art von Forschung nicht möglich und ich weiß eure Arbeit sehr wertzuschätzen. In diesem Sinne vielen Dank an Florian Bemmann, Daniel Buschek, Florian Lehmann, Dominik Heinrich, Gayatri Kudchadker, Daniela Becker, Florian Wahl, Peter Mailänder, Peter Ehrich, Marius Herget, Christian Smutek, Theresa Ullmann und Michelle Oldemeier.

Außerdem möchte ich mich bei meinen Lehrstuhl-Kollegen ganz herzlich für die schöne Arbeitsatmosphäre, das angenehme Arbeitsumfeld und die große Bereitschaft, stets mit Rat und Tat weiterzuhelfen, bedanken. Und natürlich war jedes Mittagessen, jeder Kaffee und jeder Karaoke-Abend eine große Inspiration für diese Diss! Mein herzlicher Dank geht dabei an Ricarda Lübke, Florian Pargent, Felix Naumann, Samsad Afrin-Himi, Timo Koch, Felix Schönbrodt, Johannes Albert-von der Gönna, Caroline Zygar-Hoffmann, David Goretzko, Philipp Sckopke, Laura Israel, Lena Schiestel, Larissa Sust, Elisabeth Kraus, Kathryn Eichhorn, Lena-Marei Wiessner und Cora Laugs.

Ebenfalls möchte ich mich bei Fenne große Deters, Ricarda Lübke und Larissa Sust ganz herzlich für das Korrekturlesen des Mantels der Doktorarbeit und das sehr hilfreiche sowie konstruktive Feedback bedanken.

Zuletzt geht ein großes Dankeschön an meine Familie und meine Freunde. Vielen Dank an Anne, C, Gitti, Reinhold, Sandi, Jule, Timo, Martin, Läng, Dodo, und Sonja, dass ihr immer mitgefiebert und meine Work-Life-Balance aufrecht erhalten habt! Außerdem danke ich meinen Eltern ganz herzlich dafür, dass sie mir diesen Weg ermöglicht haben, mir Rückhalt geben und immer für mich da sind. Ein großes Dankeschön geht auch an meine Geschwister Kristina, Markus und Lea dafür, dass sie an mich glauben und ein offenes Ohr für mich haben. Schließlich möchte ich mich ganz herzlich bei meinem engsten Begleiter David bedanken, der die letzten dreieinhalb Jahre immer an meiner Seite war, mich motiviert und aufgebaut, mir zugehört und sehr viel Kraft gegeben hat (das Triple D ist ab jetzt mehr als verdient)!

Contents

List of Figures

List of Tables

List of Abbreviations

API	application programming interface
BFI-2	Big Five Inventory 2
BFSI	Big Five Structure Inventory
CV	cross-validation
EEG	Electroencephalography
GDPR	General Data Protection Regulation
GPS	Global Positioning System
ImpSS	Impulsive Sensation Seeking Scale
JC	Jaccard coefficient
MAE	mean absolute deviation
MCTQ	Munich Chronotype Questionnaire
MEQ	Morningness-Eveningness Questionnaire
MICE	multivariate imputation by chained equations technique
MSE	mean squared error
MSF	midpoint of sleep
$\mathbf{MSF_{corr}}$	corrected midpoint of sleep
NEO-PI-R	Revised NEO Personality Inventory

OSF	Open Science Framework
R^2	coefficient of determination
RBF	radial basis function
RMSE	root mean squared error
SAS	Smartphone Addiction Scale
SSL	Secure Sockets Layer
SSS-V	Sensation Seeking Scale Form V
SVM	support vector machine
wb.ratio	ratio of average within- and between cluster distances
ZKPQ-III-R	Zuckerman-Kuhlman Personality Questionnaire

Abstract

With the increasing digitalization of everyday life, the research landscape has started to change. Disciplines such as psychology are benefiting from the increasing availability of new digital sources and types of data. However, to make full use of these new possibilities, psychological research must face the challenge of incorporating methods of computational science. For example, in personality psychology, research on behavior has been neglected for decades because appropriate data collection methods have been missing so far. The increasing availability of digital records enables to use behavioral data from daily life to investigate personality constructs established in this self-report dominated field of research. In this context, this dissertation presents one example of digital ways for data collection: smartphone sensing. Research apps specially developed for this purpose can be installed on commercially available smartphones, enabling the collection of large amounts of usage data from everyday life. During data preprocessing, meaningful behavioral and situational variables can be extracted from the raw data and subsequently used for statistical analyses. Previous studies have already indicated that smartphone-sensed data provide useful information about personality. Overall, however, work in this area is still in its infancy.

The present dissertation takes up the current state of research and pursues two goals. First, by presenting two empirical studies, the dissertation aims to contribute to the integration of behavioral data into personality psychology. Established personality constructs from previous literature were selected to show how behavioral markers extracted from smartphone sensing data can be used to investigate individual differences. Secondly, a debate is currently underway on the integration of predictive approaches such as machine learning in the statistical modeling culture of psychology, which has been oriented towards explanation traditionally. This discussion is particularly relevant in the context of smartphone sensing data because its complex structure imposes specific requirements for their processing. The dissertation thus contributes to this debate by using methods from both statistical cultures and by outlining their benefits using the example of explorative

research.

To pursue these two goals, various types of data collected in smartphone sensing studies in the field across several weeks were analyzed. Both studies focused on biopsychological personality concepts. Study 1 investigated the trait sensation seeking and its behavioral counterparts in terms of smartphone sensing data. Derived from the literature on the chronotype trait, study 2 focused on behavioral indicators of day-night behavior patterns. The findings of both studies illustrate that the integration of behavioral measures into personality research helps to foster the understanding of individual differences beyond established personality constructs. In this context, the combination of statistical modeling techniques from both the prediction and explanation culture provided new insights for smartphone sensing research in personality psychology. However, the empirical studies also pointed to the current limitations of this research approach ranging from issues in data preprocessing to a lack of validation procedures and limited generalizability of findings due to sample characteristics. Based on this, the dissertation gives an outlook on how smartphone sensing could establish in the future as an alternative method of data collection in personality psychology. The dissertation concludes by discussing the challenges and future directions resulting from working with digital data sources at the interface of psychology and computational science.

Zusammenfassung

Seit Beginn des digitalen Zeitalters hat auch die Forschungslandschaft angefangen, sich zu verändern. An der interdisziplinären Schnittstelle von Psychologie und Informatik entwickelt sich derzeit ein neues Forschungsfeld, die sogenannte *Psychoinformatik*. Disziplinen wie die Psychologie profitieren von der zunehmenden Verfügbarkeit neuer digitaler Datenquellen und Datentypen, weil neue Fragestellungen, die mit der Digitalisierung in Zusammenhang stehen, aber auch altbewährte psychologische Konzepte mit Hilfe neuer Datentypen untersucht werden können. Um diese neuen Möglichkeiten vollumfänglich nutzen zu können, sieht sich die psychologische Forschung mit der Herausforderung konfrontiert, Methoden der rechnergestützten Wissenschaften zu integrieren und interdisziplinäres Arbeiten zu etablieren.

In der Persönlichkeitspsychologie dominiert seit Jahrzehnten der Einsatz von Fragebögen die Forschungsinhalte. Der Selbstbericht eignet sich vor allem zur Erhebung von Merkmalen einer Person, wie z.B. Gedanken, Gefühlen oder Eigenschaften. Dementsprechend liegt der Schwerpunkt der Persönlichkeitsforschung traditionell auf der Untersuchung personenzentrierter Aspekte. Das etablierte Konzept der Persönlichkeitstriade legt jedoch nahe, dass Persönlichkeit das Produkt aus drei Komponenten ist: der Person, der Situation und des Verhaltens. Vor allem die Erforschung der Verhaltenskomponente wurde bisher jedoch vernachlässigt. Die Forschungsliteratur nennt als Grund dafür, dass sich die noch immer dominierende Fragebogenmethode nur bedingt dazu eignet, Verhalten zu erfassen. So werden Selbstauskünfte über Verhalten zum Beispiel durch Erinnerungsfehler oder Effekte sozialer Erwünschtheit verzerrt. Mit der zunehmenden Verbreitung mobiler Technologien eröffnen sich jedoch neue Möglichkeiten zum Sammeln "echter" Verhaltensdaten aus dem alltäglichen Leben. Smartphones werden beispielsweise inzwischen nicht mehr nur zu Kommunikationszwecken genutzt, sondern sie bieten auch zahlreiche andere Funktionalitäten, wie z.B. die Verwendung als Kamera, Kalender oder Navigationsgerät. In ihrer Funktion als Alltagsgegenstand verraten sie eine Menge über den Nutzer. Speziell

für diesen Zweck entwickelte Applikationen zeichnen das natürliche Nutzungsverhalten im Hintergrund auf. Die aufgezeichneten Daten können im Nachgang wiederum analysiert und somit für die Forschung nutzbar gemacht werden.

In der Persönlichkeitspsychologie gibt es bereits erste Studien, die den Zusammenhang zwischen Smartphone-Nutzung und Persönlichkeit untersucht haben. Die Forschung in diesem Bereich steht bisher jedoch noch am Anfang. Die vorliegende Dissertation knüpft an dieser Stelle an und untersucht die Erhebung von Verhaltensdaten mit Hilfe von Smartphones als alternative Methode in der Persönlichkeitspsychologie. Zu diesem Zweck greifen zwei empirische Studien etablierte Persönlichkeitskonzepte aus der Psychologie auf und untersuchen diese auf Basis von Verhaltensvariablen, welche aus Smartphonedaten extrahiert werden. Um an die aktuelle Debatte über die Verwendung von Prädiktionsansätzen in der traditionell erklärungssorientierten Datenmodellierungskultur der Psychologie anzuknüpfen, kommen dabei Methoden beider Modellierungskulturen zum Einsatz und es wird veranschaulicht, wie sich beide Ansätze gegenseitig ergänzen können.

In der ersten Studie wurde das Persönlichkeitsmerkmal *Sensation Seeking* untersucht. Sensation Seeking beschreibt individuelle Unterschiede im Bedürfnis nach externer Stimulation und die damit verbundene Bereitschaft, Risiken einzugehen. Dieses in der Biopsychologie verankerte Konstrukt wurde in den letzten Jahrzehnten umfassend erforscht. Jedoch gibt es bisher nur sehr wenige Arbeiten, die sich mit dem objektiven Verhaltensausdruck von Sensation Seeking im Alltag beschäftigen. Daher stellen per Smartphones gesammelte Nutzungsdaten einen neuen Untersuchungskontext dar. In der ersten Studie wurde somit der Frage nachgegangen, ob die individuelle Ausprägung des Persönlichkeitsmerkmals Sensation Seeking mittels per Smartphone gesammelten, objektiven Nutzungsdaten reliabel vorhergesagt werden kann. Dazu wurde eine 30-tägige Feldstudie durchgeführt. In dieser wurden kontinuierlich pseudonymisierte Smartphone-Nutzungsdaten mittels einer speziell für Android-Geräte programmierten App aufgezeichnet und Persönlichkeitsmerkmale sowie Demografie erfragt. Zunächst wurden in der Literatur Verhaltenskorrelate von Sensation Seeking identifiziert, die bisher größtenteils mittels Selbstauskunft erhoben wurden. Im Anschluss wurden diese Verhaltenskorrelate in Smartphone-Nutzungsparameter übersetzt und aus den Smartphonedaten extrahiert. In einem Benchmark-Experiment wurde die kreuzvalidierte Vorhersagegüte vier verschiedener Machine Learning Algorithmen miteinander verglichen. Nur das nicht-lineare *Random Forest* Modell konnte bessere Vorhersagen als ein Zufallsmodell liefern. Insgesamt erwies sich aber auch bei diesem Modell die Vorher-

sagegüte als sehr gering. Die anschließende Untersuchung der Wichtigkeit der einzelnen, als Prädiktoren in das Modell eingegangenen Variablen, zeigte, dass für die Vorhersagen des Random Forest Algorithmus insbesondere Variablen mit inhaltlichem Bezug zu Telefonieverhalten und nächtlicher Nutzung von Bedeutung waren. In dieser Studie wurden zum ersten Mal Verhaltenskorrelate von Sensation Seeking auf Basis von Verhaltensdaten aus dem Alltag untersucht. Daher können insbesondere die Ergebnisse zur Wichtigkeit der einzelnen Prädiktoren im Vorhersagemodell als Anregung für zukünftige Forschung dienen, die sich mit Erklärungsmodellen des Persönlichkeitsmerkmals Sensation Seeking befasst.

Basierend auf dem Stand der Forschung zum Persönlichkeitsmerkmal *Chronotyp*, wurden in der zweiten Studie Fragestellungen rund um das Thema Tag-Nacht-Verhaltensmuster untersucht. Das Merkmal Chronotyp beschreibt individuelle Unterschiede im circadianen Rhythmus, der durch biologische Faktoren bedingt ist und sich u.a. durch Unterschiede in tageszeitlichem Verhalten ausdrückt. Auch die Untersuchung dieses Persönlichkeitskonzepts war bisher stark von Selbstauskunftsmaßen bestimmt und es wurden nur wenige objektive Verhaltensdaten berücksichtigt. Um die Nützlichkeit von Smartphonedaten für die Erforschung von Tag-Nacht-Verhaltensmustern zu explorieren, wurden exemplarisch drei Forschungsthemen aus der Literatur aufgegriffen und auf Basis des bestehenden Smartphone-Datensatzes eines fortlaufenden Forschungsprojekts explorativ untersucht.

Als erste Fragestellung wurden zwei in der Fragebogenforschung gängige Operationalisierungen des Chronotyps genauer betrachtet. Ein in der Literatur weit verbreiteter Ansatz zur Bestimmung des Chronotyps ist die Abfrage tageszeitlicher Präferenzen. Es wird davon ausgegangen, dass Personen sich in ihren tageszeitlichen Präferenzen unterscheiden und somit unterschiedliche Tag-Nacht-Typen repräsentieren. In Anlehnung an die Fragebogenforschung wurde untersucht, ob sich Personengruppen mit ähnlichen Tag-Nacht-Aktivitätsmustern auf Basis von Smartphone-Nutzungsdaten identifizieren lassen. Zu diesem Zweck wurden die Items eines etablierten Messinstruments zur Bestimmung des Chronotyps in Smartphone-basierte Indikatoren der tageszeitlichen Präferenzen des Nutzers übersetzt. Von diesen Verhaltensvariablen ausgehend, wurden mit Hilfe einer Clusteranalyse unter Anwendung der Bootstrapping-Methode zwei stabile Gruppen identifiziert. Bei der anschließenden Betrachtung deskriptiver Statistiken beider Gruppen stellte sich heraus, dass diese in den für die Clusteranalyse verwendeten Variablen zwar mehrheitlich große Mittelwertsunterschiede aufwiesen, ihre Verteilungen jedoch stark überlappten. Ein weiterer in der Literatur verbreiteter Ansatz zur Operationalisierung des

Chronotyps ist die Abfrage von Tag-Nacht-Gewohnheiten und die anschließende Bestimmung des sog. Schlafmittelpunkts. Zu diesem Zweck wurde ein Algorithmus zur Berechnung eines Näherungsmaßes des Schlafmittelpunkts entwickelt, welcher auf Smartphone-Indikatoren für Tag-Nacht-Zeiten basierte. Deskriptive Statistiken wurden betrachtet, um den Smartphone-basierten Schlafmittelpunkt zu explorieren. Es wurden Zusammenhänge zu Alter, Geschlecht und nächtlicher Inaktivität der Smartphone-Nutzung gefunden.

Als zweite Fragestellung wurde untersucht, ob sich die in der Fragebogenforschung etablierten Befunde zu Zusammenhängen zwischen Tag-Nacht-Aktivität und Persönlichkeit auch zeigen, wenn anstatt der selbstberichteten, Smartphone-basierte Tag-Nacht-Variablen verwendet werden. Anhand der empirischen Korrelationen zeigte sich, dass Gewissenhaftigkeit mit einem nach vorne verschobenem tageszeitlichem Rhythmus zusammenhing.

Als dritte Fragestellung wurden schließlich Auswirkungen der sogenannten Sozialen Jetlag-Hypothese betrachtet. Diese Hypothese besagt, dass der Schlafmittelpunkt unter der Woche und am Wochenende aufgrund sozialer Verpflichtungen voneinander abweichen. Als Folge gleichen Menschen ein Schlafdefizit, das sie unter der Woche anhäufen, durch kompensierenden Schlaf am Wochenende aus. Zu diesem Zweck wurde untersucht, ob die Inaktivitätsdauer der nächtlichen Smartphone-Nutzung am Wochenende durch die in der vorhergehenden Woche und gewohnheitsmäßig unter der Woche gezeigten nächtlichen Inaktivitätsdauer, sowie durch Persönlichkeitsmerkmale, Alter und Geschlecht beeinflusst wird. Da pro Person mehrere Messzeitpunkte über den mehrwöchigen Studienzeitraum vorlagen, wurde ein Mehrebenenmodell gerechnet. Bei der Aufbereitung der Daten stellte sich heraus, dass es für viele der getroffenen Vorverarbeitungsentscheidungen zahlreiche plausible Alternativen gab. Um die Abhängigkeit der Modellergebnisse von den subjektiven Freiheitsgraden in der Datenvorverarbeitung transparent zu machen, wurde deswegen eine sogenannte *Multiverse-Analyse* vorgenommen, d.h. das Mehrebenenmodell wurde für die Kombination aller Alternativentscheidungen einzeln berechnet und berichtet. In der Multiverse-Analyse zeigte sich schließlich, dass nur die gewohnheitsmäßige nächtliche Inaktivitätsdauer der Smartphone-Nutzung unter der Woche einen robusten Einfluss auf die nächtliche Inaktivitätsdauer am Wochenende hatte. Insbesondere für die Variablen Gewissenhaftigkeit, Alter und Geschlecht hingen die Ergebnisse von den in der Vorverarbeitung getroffenen Entscheidungen ab.

Insgesamt zeigen die Ergebnisse dieser Dissertation, dass die Verwendung von Verhaltensdaten einen zusätzlichen Mehrwert für die Untersuchung von Persönlichkeit bietet.

Die beiden explorativen Studien knüpfen dabei an erste, bereits bestehende Pionierarbeiten der Datenerhebung mittels Smartphones im Kontext der Persönlichkeitspsychologie an und liefern neue empirische Erkenntnisse zum alltäglichen Verhaltensausdruck biopsychologischer Persönlichkeitskonstrukte. Die vorliegende Dissertation illustriert damit, dass Smartphones die Erfassung vielfältiger Verhaltensdaten aus dem Alltag für große Stichproben und über einen längeren Studienzeitraum ermöglichen. Somit könnten sie zukünftig die nach wie vor von der Fragebogenforschung dominierte Untersuchung von Persönlichkeit als alternative Datenerhebungsmethode hilfreich ergänzen. Wie die beiden Studien zeigen, ergeben sich beim Einsatz von Smartphones neue Herausforderungen bei der Datenerhebung, Datenvorverarbeitung und Datenanalyse, denen die psychologische Forschung nur durch das Know-How der rechnergestützten Wissenschaften begegnen kann. Die Limitationen dieses neuen Forschungsfeldes und des übergeordneten Kontexts der Psychoinformatik werden aufgezeigt und Ideen für zukünftige Forschung werden diskutiert.

Chapter 1

Introduction

Technological advances in the 21st century have brought about the "datafication" of everyday life (Mayer-Schönberger & Cukier, 2013). We continuously leave behind vast amounts of digital footprints: When we pay for the bus ticket with our credit card, wish friends a happy birthday on social media, and order the new sweater via the shopping app (Lazer et al., 2009). These new digital sources provide data from real-life, at a high frequency, and for large samples, and are both promising and challenging for psychological research at the same time (Montag, Duke, & Markowetz, 2016).

On the one hand, these digital traces are promising because they provide various types of data such as language, activity records, images, or music records (e.g., Eichstaedt et al., 2015; Harari et al., 2016; Kosinski, Stillwell, & Graepel, 2013; Montag & Elhai, 2019; Nave et al., 2018; Thorstad & Wolff, 2019; Y. Wang & Kosinski, 2018; Youyou, Kosinski, & Stillwell, 2015) from naturally occurring data sources such as social media platforms or smartphones (Harari et al., 2016; Kosinski et al., 2013). On the other hand, these digital traces are a challenge because they require a certain amount of technical and statistical know-how (e.g. Chen & Wojcik, 2016; Yarkoni, 2012). Thus, psychology has to include methods of computational science to make these data usable for research (Montag et al., 2016). Montag et al. (2016) argue that this new intersection of psychology and computational science will establish itself as a separate discipline in the upcoming decades. Accordingly, Yarkoni (2012) has introduced the term *psychoinformatics*, which has been described as *"an emerging discipline that uses tools and techniques from the computer and information sciences to improve the acquisition, organization, and synthesis of psychological data"* (p.391). Other researchers used alternative terms such as *computational social sciences* (Lazer et al., 2009) or *digital phenotyping* (Montag et al., 2019). Research in this

field is still in its infancy (Markowetz, Błaszkiewicz, Montag, Switala, & Schlaepfer, 2014; Montag et al., 2016).

This dissertation offers an insight into the interface of psychology and computational science. Its starting point is the observation that personality psychology has neglected the study of actual behavior for decades (Baumeister, Vohs, & Funder, 2007). The integration of new data-intensive approaches providing behavioral measures could remedy this situation (Harari et al., 2020). The dissertation addresses this by using methods from the toolbox of computational science. First, smartphone sensing is introduced as a tool for data collection. Recent literature argued that smartphones are the most important source of knowledge about human behavior, as these portable supercomputers have become constant companions in everyday life and provide much information about daily behavior (Harari et al., 2016; Harari, Müller, Aung, & Rentfrow, 2017; Harari et al., 2020; Montag et al., 2016). The complexity of these smartphone sensing data results in a need for more flexible data processing methods. Therefore, as another tool of computational science, the predictive modeling approach is presented as an alternative to the explanatory approach commonly used in psychology. Through two empirical studies, the dissertation takes up the blind spot of personality psychology, i.e., the investigation of actual behavior. Thus, established personality constructs are investigated using a combination of smartphone sensing and methods from both the prediction and the explanation approach. Based on these two studies, the present dissertation addresses the question of whether integrating the new data-intensive approach of smartphone sensing into personality psychology provides new empirical insights into individual differences.[1]

1.1 Traditional Focus of Personality Psychology

Individuals differ in their characteristic patterns of thinking, acting, and feeling. The theoretical interest of personality psychology is to describe, explain, and understand these differential patterns and their underlying dynamics (Funder, 2001, 2006). Empirical interest lies in the entity of personality which is composed of three components: *person*, *situation*, and *behavior* (Funder, 2001). This conceptualization of personality, which is

[1] A more detailed discussion of using sensing data for personality research mentioned in chapters 1 and 4 can be found in Harari et al. (2020), which the author of this dissertation co-authored. While Harari et al. (2020) present an overall conceptual framework and research agenda for personality sensing, this dissertation focuses on the methodological opportunities and challenges arising from working with smartphone sensing data.

referred to as *personality triad*, argues that all three components are interdependent, resulting from one another (Funder, 2001; Lewin, 1951). According to Funder (2009), it implies that persons are the sum of their behaviors, which they show in different specific situations. Situations build the stages for persons' behaviors. To describe and understand behavior, it is important to consider which persons perform the particular behavior in which situations (Funder, 2006, 2009).

However, at the beginning of the 21st century, Funder (2001) noted that personality research majorly studied one component of the personality triad: the person. This bias was argued to result from the limited availability of data collection methods (Baumeister et al., 2007). Self-report questionnaires, which are the most frequently used data collection tool in personality psychology, are more suitable to study person-centered constructs such as feelings, thoughts, and attitudes than to collect data about situations and behaviors (Baumeister et al., 2007; Funder, 2001). Accordingly, in the previous decades, numerous conceptualizations focusing the person component have been proposed and operationalized via self-report questionnaires (e.g., De Raad, 2000). However, the lack of suitable methods for recording situations and behaviors might have led empirical studies to neglect them (Funder, 2009).

The situation component has only recently gained increasing attention. Rauthmann, Sherman, and Funder (2015) have proposed a conceptual framework for the investigation of situations by emphasizing the subjective experience (*characteristics*) as important information about the situation. A corresponding taxonomy of those characteristics, the so-called situational eight DIAMONDS, has been established and has become fertile ground for the study of person-situation interactions (Rauthmann et al., 2014; Rauthmann et al., 2015).

Baumeister et al. (2007) argue that the imbalance in the empirical investigation of the personality triad was, in particular, at the expense of the behavior component. Traditionally, behavioral data are collected in laboratory settings (e.g., as reaction times) or via questionnaires (Baumeister et al., 2007). However, the discrepancy between self-reported past or hypothetical behavior and actual behavior has often been demonstrated (e.g., Celis-Morales et al., 2012; Junco, 2013). Observation of actual behavior has been too expensive, inconvenient, and time-consuming (Baumeister et al., 2007; Funder, 2009). Consequently, empirical studies incorporating measures of actual behavior have been underrepresented in personality psychology (Baumeister et al., 2007). To fill this research gap, Funder (2009) demanded *"studies in which individuals are each placed into or observed in each of a range of situations, and their behavior in them observed and measured directly. Studies that do*

this are almost unknown in the literature, not really because psychologists do not grasp the need for them, but because they are so difficult and expensive to conduct" (pp. 124-125).

1.2 The Toolbox of Computational Science in Personality Research

1.2.1 Tool 1: Smartphone Sensing

Roughly ten years after Funder's (2009) cited call progressive digitalization has opened a new possibility to study actual behavior (Harari et al., 2020). Think of this familiar scenario: *You wake up from the ringing of the alarm clock app on your phone. During your first coffee, the broadcast app informs you about the news of the day. A glance in your calendar app reminds you of an early work appointment so you use the public transportation app to check for the next possible train connection.*

Nowadays, about 45% of the world's population (O'Dea, 2020) uses smartphones for many different purposes in everyday life: for communication, as a personal organizer, notebook, address book, a navigation device, or music player (Harari et al., 2020; Miller, 2012). Due to their popularity among users, smartphones are also becoming increasingly useful for researchers. Especially for research developed apps can be installed on the own smartphones of the participants to log information about natural usage from these sensors continuously (Harari et al., 2016; Miller, 2012). This in situ-data collection happens unobtrusively as participants are not required to take active actions except using their smartphone in the same way as usual (Harari et al., 2016). The smartphone's tracking in the background reduces participants' awareness of the study setting and thus lessens the likelihood of reactive behavior (Conner & Mehl, 2015; Harari et al., 2017; Miller, 2012). Therefore, Harari et al. (2016) praise smartphones as a valuable tool for conducting ecologically valid research and define *smartphone sensing* as the recording of usage information from sensor technology, which is embedded in regular smartphones. Smartphone sensing enables the efficient aggregation of data over more extended periods and various situations. Additionally, it provides the opportunity to collect data from many participants simultaneously, benefiting the collection of large sample sizes (Miller, 2012).

Data Sources and Types

As commercially available smartphones are equipped with many different types of sensors, they provide a wide range of different data types. Previous literature has proposed various structured overviews and frameworks of data sources, data types, and variables derived for statistical analysis (Beierle et al., 2019; Conner & Mehl, 2015; Cornet & Holden, 2018; Harari et al., 2016; Harari et al., 2017; Harari et al., 2020; Miller, 2012; Mohr, Zhang, & Schueller, 2017; Piwek & Ellis, 2016). Some of them were created for a wider audience and show, for example, how smartphone sensing helps to capture situations or behaviors (Harari, Gosling, Wang, & Campbell, 2015; Harari et al., 2017). Other frameworks are embedded in a narrow research context and are thus very specific. For example, Mohr et al. (2017) report a hierarchical framework with different aggregation levels of smartphone sensing data about clinical traits. Figure 1.1 summarizes these previously reported frameworks while putting smartphone sensing in the general context of *ambulatory assessment*, which is defined as the investigation of persons and their naturally occurring behaviors in a wide range of real-life situations (Conner & Mehl, 2015). Figure 1.1 shows that ambulatory assessment comprises *passive sensing* and *active logging* (Harari et al., 2017; Harari et al., 2020; Seifert, Hofer, & Allemand, 2018).

For passive sensing, stand-alone devices such as electronically activated recorders (logging of ambient sounds; Mehl, Pennebaker, Crow, Dabbs, & Price, 2001), actigraphs (logging of movements; Van De Water, Holmes, & Hurley, 2011), or biosignal recorders (e.g., logging heart rate, respiration patterns, body temperature; Wilhelm & Grossman, 2010) have commonly been used in the past (Conner & Mehl, 2015; Wrzus & Mehl, 2015). Smartphones can combine many of these functionalities in one device and are thus an important tool for ambulatory assessment (Harari et al., 2020; Miller, 2012). The proposed framework in Figure 1.1 comprises available sources and types of data provided by smartphones. Data sources are being assigned to three categories: environment, device status, and usage. The first category provides information about the physical environment of the smartphone and its user (Beierle et al., 2019). The category device status provides information about the smartphone's basic functionality, including its current operating state and connectivity (Beierle et al., 2019). Finally, usage summarizes all data types that are initiated by user interaction (Miller, 2012). Smartphone sensing is only one part of mobile sensing, which subsumes any data acquisition by mobile devices (e.g., wearables, smart shoes, tablets) (Harari et al., 2016). Accordingly, smartphone sensor data have recently been supplemented by other *external sources* such as sensors integrated in portable devices

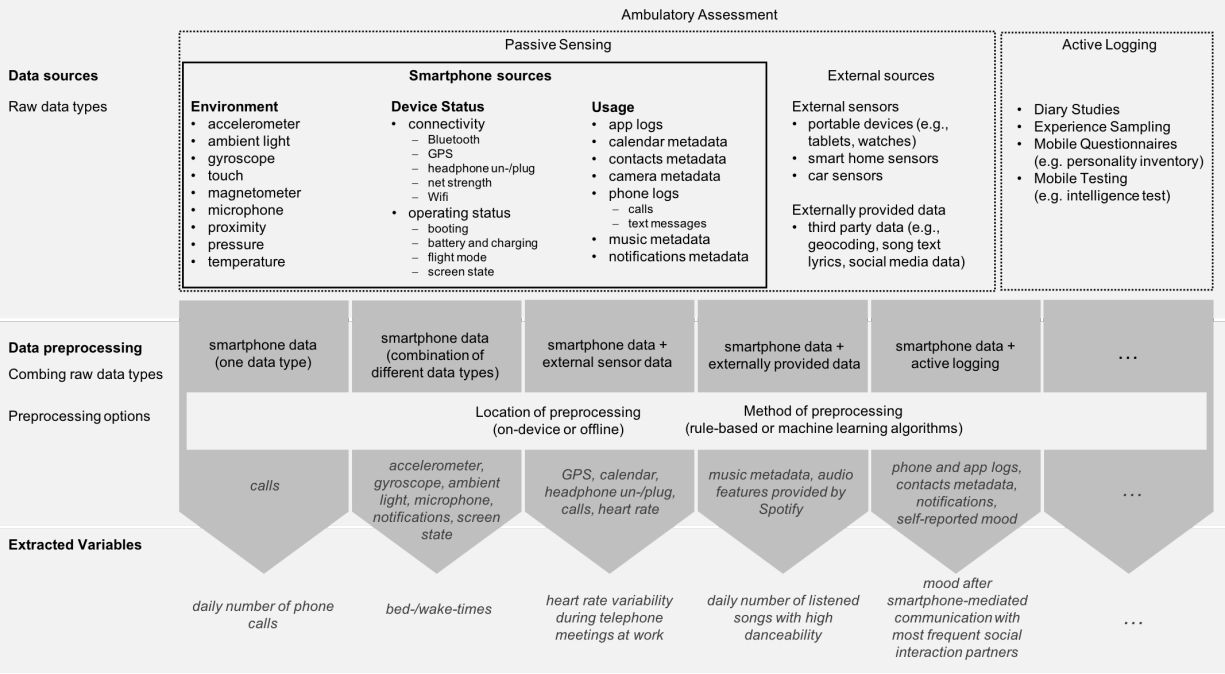

Figure 1.1: Overview of frequently used smartphone sensor sources, raw data types, and preprocessing options in the framework of ambulatory assessment. For illustration, the figure provides a simplified overview and only presents a selection of examples (in italics).

(e.g., wristband sensors; Zenonos et al., 2016), in smart-homes (e.g., motion or ambient temperature sensors; Nelson & Allen, 2018), or in cars (e.g., steering wheel motion sensors; Lee & Chung, 2017) to study psychologically relevant constructs. In addition, smartphone sensing data can be enriched by data from third parties (e.g., labeling of place types based on Global Positioning System (GPS) data by using a Google application programming interface (API); Harari et al., 2020; Mehrotra et al., 2017).

The second component of ambulatory assessment is active logging, which refers to the application of mobile questionnaires and the collection of in-situ self-reported experiences by techniques such as diary studies or experience sampling. The latter describes a method where short questions are repeatedly presented at random times during the day via online survey tools, text messages, or smartphone notifications (Conner & Mehl, 2015). Smartphone sensing and active logging can be used individually, but also in combination (R. Wang et al., 2014). For this purpose, self-report questionnaires or psychological tests can be presented before or after smartphone sensing periods. In addition, the repeated presentation of questionnaires as in diary or experience sampling studies can be integrated into smartphone sensing studies (Wrzus & Mehl, 2015). By covering a broad range of subjective and objective real-life data, passive sensing and active logging have the potential to

inform about all three components of the personality triad and to contribute to the study of personality (Harari et al., 2020; Wrzus & Mehl, 2015).

Data Preprocessing

The framework presented in Figure 1.1 illustrates that researchers have to undertake many decisions and steps before using smartphone sensing data for data analysis. First, meaningful variables have to be extracted from the various types of raw sensor data (Harari et al., 2020). Data from different sensors can either be used individually or in various combinations resulting in differently sophisticated features. For example, Montag et al. (2014) used call logs to extract simple call-related variables, while Min et al. (2014) combined accelerometer, ambient light, battery, microphone, proximity, screen, and app logs to calculate bed- and wake-times. The extraction of variables can take place directly on the smartphone, for example, by extracting conversation behaviors from microphone sensors (Harari et al., 2019). Alternatively, raw data can be stored on servers, and variables can be extracted offline after the data collection phase (e.g., Stachl et al., 2017). The same variables can be extracted via methods of different complexity. For example, Lin et al. (2019) used a rule-based algorithm to determine sleep time, and Min et al. (2014) applied complex machine learning algorithms for the same goal. The selection of examples in the lower part of Figure 1.1 represents a small fraction of the universe of extractable variables. Further examples currently used in empirical studies are presented in chapter 1.3.

1.2.2 Tool 2: Prediction Approaches

In psychology and other social sciences, the *data modeling* approach has been the gold standard for decades (Breiman, 2001). Data modeling aims at describing data with stochastic models. Inferences are based on the fit of a specified model to the respective dataset and are obtained using statistical concepts such as significance and effect sizes (Breiman, 2001; Shmueli, 2010). In research practice, association-based models are often applied to observational data to test hypotheses about underlying constructs (Shmueli, 2010). Thus, the data modeling strategy emphasizes explaining and understanding underlying structures in the data (Breiman, 2001; Mahmoodi, Leckelt, van Zalk, Geukes, & Back, 2017; Shmueli, 2010; Yarkoni & Westfall, 2017). An alternative statistical modeling approach is algorithmic modeling, also called predictive modeling (Breiman, 2001). It focuses on data mining and finding algorithms for highly accurate predictions (Breiman, 2001; Shmueli, 2010).

The overall goal is to create models that adapt as flexible as possible but still not too flexible to the properties of a given dataset, including its noise and, therefore, still provide good predictions on new unseen data (Yarkoni & Westfall, 2017). Thus, the algorithmic modeling strategy emphasizes predicting future observations (Yarkoni & Westfall, 2017).

Prediction algorithms are hard to interpret and, thus, so-called black-box models. Therefore, psychological research, which traditionally sees itself as an explanatory science with methods of the data modeling culture, has neglected prediction for decades (Yarkoni & Westfall, 2017). This has recently led to a debate in psychology (Mahmoodi et al., 2017; Yarkoni & Westfall, 2017). Explanation and prediction are not per se mutually exclusive. However, models with high explanatory power often do not have high predictive power. Conversely, as already mentioned, models that provide accurate predictions are often so complex that they do not have a high explanatory value because they are challenging to interpret (Mahmoodi et al., 2017; Yarkoni & Westfall, 2017). However, some researchers argue that psychology as an explanatory science should not completely neglect the predictive focus, because both foci can complement each other (Yarkoni & Westfall, 2017). Ideally, there should be a cyclical empirical process of prediction and explanation (Mahmoodi et al., 2017). For example, the predictive approach helps to identify aspects that are important to consider in explanatory models. These should then be empirically investigated in the course of explanatory research. Conversely, the explanatory approach provides indications of which variables might be important for accurate predictions (Mahmoodi et al., 2017; Yarkoni & Westfall, 2017).

Accordingly, some researchers have claimed that psychological research should broaden its focus by increasingly including prediction techniques such as machine learning (Yarkoni & Westfall, 2017). This call goes hand in hand with the sophisticated character of new types of data requiring more flexible methods of analysis. In the context of smartphone sensing, some researchers have started to use machine learning for personality prediction (e.g., Chittaranjan, Blom, & Gatica-Perez, 2011, 2013; de Montjoye, Quoidbach, Robic, & Pentland, 2013; Mønsted, Mollgaard, & Mathiesen, 2018; Stachl et al., 2019). The basic principle of these studies is the same: Smartphone sensing delivers behavioral data, which are enriched with standard self-report questionnaires revealing information about participants' personality. Supervised machine learning is used to obtain an algorithm that operates on the behavioral variables to predict self-reported psychological constructs. As can be seen, the first steps have recently been taken to integrate the predictive approach into the traditional explanatory focus of personality research. However, the cyclical em-

pirical process of prediction and explanation has yet to establish itself (Mahmoodi et al., 2017).

1.3 State of the Art: Smartphone Sensing in Personality Psychology

The first generation of studies focused on introducing smartphone sensing as a new data collection tool in psychology (Harari et al., 2015; Harari et al., 2016; Harari et al., 2017). Besides delivering technical descriptions of research apps (Beierle et al., 2019; Montag et al., 2019), the authors have discussed challenges and opportunities of smartphone sensing (Harari et al., 2016; Seifert et al., 2018). However, the following literature overview focuses on the second generation of studies working empirically with smartphone sensing data. Previous smartphone sensing studies investigating personality constructs have mainly focused on the *big five* personality traits (e.g., Harari et al., 2019; Mønsted et al., 2018; Montag et al., 2014; Stachl et al., 2017). The *Five-Factor Model* is one of the person-centered conceptualizations of personality previously mentioned in chapter 1.1 and is the most widely accepted model describing personality in psychological research (De Raad, 2000; McCrae, 2009). Its development is based on the psycholexical hypothesis, which assumes that individual human differences are reflected in everyday language use (McCrae, 2009). The factor-analytical reduction of dictionary-derived lists containing characterizing human adjectives reveals factors describing fundamental individual differences (McCrae, 2009). After decades of work by many researchers, this approach has led to the establishment of the following big five factors to describe and assess the underlying latent personality constructs: openness, conscientiousness, extraversion, agreeableness, and emotional stability (De Raad, 2000). The trait openness relates to curiosity, interest in new experiences, creativity, and aesthetic sensation. Conscientiousness relates to love of order, diligence, high motivation, and reliability. Extraversion relates to being sociable and outgoing, assertive, and active. Agreeableness relates to compassion, politeness, compliance, and trust. Finally, emotional stability relates to balanced emotional reactions, tolerance of frustration and stress, and calmness (Danner et al., 2016; De Raad, 2000).

1.3.1 Studies in the Tradition of Explanation

In the tradition of explanatory research, various types of behavior extracted from smartphone sensing data have been investigated as behavioral counterparts of the big five personality traits. For example, extraversion was found to be related to social behavior, including conversations extracted from microphone sensors, and communication via text messages or phone calls (Harari et al., 2019; Montag et al., 2014; Stachl et al., 2017). Furthermore, Stachl et al. (2017) have investigated individual differences in app usage behavior. Among other associations, they found that extraversion is positively associated with the frequency of using photography and communication apps. Associations were also found for other personality traits: More conscientious people used gaming apps less often, and more agreeable people used transportation apps more frequently. Another illustration is the study of daily spatial behavior (Ai, Liu, & Zhao, 2019; Alessandretti, Lehmann, & Baronchelli, 2018). Ai et al. (2019) found that the number of different places visited at weekends is positively related to extraversion but negatively related to conscientiousness, and the range of movement at weekends is positively related to agreeableness. As can be seen from these empirical examples, smartphone sensing provides a variety of behaviors that can be studied in terms of individual differences (Harari et al., 2020).

1.3.2 Studies Using Prediction Approaches

In the context of smartphone sensing, the investigation of personality traits using methods of the prediction approach has recently also gained increasing interest (Harari et al., 2020). So far, a variety of smartphone sensing variables have been proposed to predict the big five personality traits. Logging data ranging from app usage, phone usage (texting, calling), Bluetooth and WiFi scans, GPS positions, screen state, microphone, accelerometer, or battery sensors have been aggregated using various quantification measures of central tendency, between- and within-person variation, regularity, or diversity (Chittaranjan et al., 2011, 2013; de Montjoye et al., 2013; Kambham, Stanley, & Bell, 2018; Mønsted et al., 2018; W. Wang et al., 2018). However, each study used only subsets of these smartphone usage variables for prediction (Chittaranjan et al., 2011, 2013; de Montjoye et al., 2013; Kambham et al., 2018; Mønsted et al., 2018). Stachl et al. (2019) presented the most comprehensive set of variables to date, covering app usage, music consumption, communication behavior, spatial behavior, and general smartphone usage. They made a further distinction between day and night activities (Stachl et al., 2019). In summary, there are

first indications that the big five personality traits can be predicted from smartphone usage behavior better than chance. According to Mønsted et al. (2018), however, the results of some previous studies on smartphone sensing should be interpreted with caution, as the simultaneous use of small samples and a large number of variables resulted in overestimating the algorithmic model performance (Chittaranjan et al., 2013; de Montjoye et al., 2013). In addition, the results indicate that the predictive power for each personality trait depends on the variety of smartphone sensor variables used. Accordingly, Mønsted et al. (2018) focused exclusively on communication variables extracted from call and text logs and concluded that smartphone sensor data can only predict the trait extraversion better than chance. Nevertheless, the recent large-scale analysis by Stachl et al. (2019) shows that a wider variety of smartphone usage patterns can predict not only extraversion but also openness and conscientiousness above chance, albeit with low prediction performance.

One reason for this limited prediction accuracy could be the size of the used samples. Machine learning algorithms usually unfold their predictive performance with a large number of observations. However, recent smartphone sensing studies have reached a maximum number of about 630 participants (Mønsted et al., 2018; Stachl et al., 2019). Another reason could be the types of smartphone sensing data available so far (Harari et al., 2020). Due to technical and data privacy restrictions mainly quantitative aspects of smartphone usage such as the frequency and duration of app and phone usage but not qualitative aspects of smartphone usage such as contents and emotional valence of text messages or usage histories have been tracked and used for personality prediction (e.g., Chittaranjan et al., 2013; de Montjoye et al., 2013; Mønsted et al., 2018; Stachl et al., 2019; W. Wang et al., 2018). However, these quantitative aspects might not be suitable for predicting all personality traits equally well. For example, compared to the other big five traits, the predictive accuracy for extraversion is relatively high. One of the main functionalities of smartphones is communication, and merely sensing its quantity (e.g., frequency and duration of calls, text messages, or the usage of communication and social media apps) already reveals much about extraversion (Mønsted et al., 2018; Stachl et al., 2019). In contrast, for example, agreeableness has not been predicted above chance (Stachl et al., 2019). Accordingly, previous research indicates that agreeableness is characterized by qualitative aspects of behaviors such as the quality of social interactions (Park et al., 2015), which have not been included in smartphone sensing studies so far.

1.4 The Present Dissertation

1.4.1 Rationale

In recent literature, smartphones have been presented as a useful tool for data collection, ultimately leading to new insights into human personality (Harari et al., 2020). In contrast to the established questionnaire-based personality research, however, the investigation of individual differences using smartphone sensing has only just begun. Individual differences in the context of smartphone sensing had so far mainly been investigated using the Five-Factor Model of personality. However, Funder (2001) argued that the big five traits do not paint an exhaustive picture of human personality. Although many other personality constructs are somehow related to the big five, they are not a direct derivative of the Five-Factor Model but cover distinct aspects of personality. One alternative view comes from the biopsychological perspective suggesting that individual differences are associated with biological functioning and structures (Funder, 2001). To demonstrate that smartphone sensing data are not only interesting concerning the big five traits but also provide a wide range of possibilities for investigating other concepts established in personality psychology, in this dissertation, exemplary personality constructs with a biopsychological basis were selected. In addition, from a legal and technical perspective, smartphone sensing is currently limited in its range of available data types. For this reason, only those personality constructs were selected, which, to a large extent, are likely to be covered by quantitative aspects of behavior. Using these exemplary selected personality concepts, the present dissertation aims to illustrate the potential of integrating computational science in personality psychological research. Taking the perspective of psychoinformatics, the goal of the present dissertation is twofold.

First, the selected established person-centered constructs are studied by using real-world behavior. Previous research had strongly neglected this approach (Baumeister et al., 2007). For this purpose, smartphone sensing was applied as an assessment tool. Previous literature has built on a range of self-reported behavioral and contextual variables for the study of personality traits. The present dissertation turns this approach around by translating established behavioral variables from survey research into smartphone-sensed equivalents.

Second, the research presented in this dissertation is explorative, as the work with smartphone sensing data in the field of personality psychology is still relatively new. With its exploratory character, the present dissertation aims to integrate methods from both

the data and the algorithmic modeling tradition. Regarding the ongoing debate about explanation and prediction approaches in psychological research (Mahmoodi et al., 2017; Yarkoni & Westfall, 2017), it aims to illustrate that, especially in this explorative stadium of research, both modeling cultures can be useful to gain first insights into new research topics.

In detail, the two empirical studies of this dissertation contribute to these goals as follows:

Study 1 Study 1 uses smartphone sensing data to investigate the biopsychological personality trait sensation seeking. This construct has previously been described as the need for external stimulation (Zuckerman, 1994). It was selected to be investigated in the present dissertation because smartphones could potentially contribute to this external stimulation. Quantitative aspects of smartphone usage could, in turn, provide useful behavioral counterparts of the trait sensation seeking. Previous literature provided a selection of assumed analog behavioral correlates of the trait seeking sensation, that translate into meaningful quantitative aspects of smartphone usage. A supervised machine learning approach was used to investigate whether self-reported personality traits can be predicted based on smartphone-sensed behavioral markers. Finally, novel methods of interpretable machine learning were applied to inspire future explanatory research concerning the investigation of behavioral counterparts of sensation seeking.

Study 2 The research questions of study 2 were derived from literature about the biopsychological trait chronotype. This trait was chosen because it is well suited to study quantitative aspects of smartphone usage by using day-night activity patterns. More specifically, it was investigated whether "morning larks" and "night owls" manifest in day-night patterns of smartphone usage behavior, how day-night patterns relate to big five traits, and whether traits and day-night activity patterns during the week are associated with day-night activity on weekends. For this purpose, smartphone sensing variables indicating day-night activity patterns were derived from previous literature based on self-reports. Unsupervised machine learning was used to identify groups of persons with similar day-night activity patterns. Methods from the data modeling tradition, such as exploratory factor analysis, correlations, and multilevel modeling, were used to gain first insights into inter- and intraindividual differences in day-night activity patterns based on smartphone usage data.

1.4.2 Parts of the Dissertation and Author Contributions

Table 1.1 lists the empirical studies that contribute to this dissertation. All authors have contributed significantly to the research presented in the articles. The right column of Table 1.1 shows the respective individual contributions.

Table 1.1: Publications in the dissertation and author contributions

Study	Publication	Author Contributions
1	**Schoedel, R.**, Au, Q., Völkel, S.T., Lehmann, F., Becker, D., Bühner, M., Bischl, B., Hussmann, H., & Stachl, C. (2018). Digital Footprints of Sensation Seeking: A Traditional Concept in the Big Data Era. In *Zeitschrift für Psychologie, 226(4)*, 232-245. https://doi.org/10.1027/2151-2604/a000342	**R.S.** designed research; F.L., D.B., and S.T.V. programmed the app; S.T.V. managed app development; **R.S.** and C.S. conducted research; **R.S.** and Q.A. preprocessed data; **R.S.** conducted data analysis; Q.A. and B.B. supervised data analysis; **R.S.** wrote the manuscript; all authors gave feedback to the manuscript; C.S. helped to improve the manuscript; M.B., B.B., and H.H. provided resources
2	**Schoedel, R.**, Pargent, F., Au, Q., Völkel, S. T., Schuwerk, T., Bühner, M., & Stachl, C. (2020). To Challenge the Morning Lark and the Night Owl: Using Smartphone Sensing Data to Investigate Day-Night Behavior Patterns. In *European Journal of Personality*, https://doi.org/10.1002/per.2258.	**R.S.** designed research; S.T.V. managed app development; **R.S.**, T.S., and C.S. conducted research; Q.A. provided basic code for data preprocessing from previous projects; **R.S.** preprocessed data; **R.S.** conducted data analysis; F.P. gave feedback on data analysis; **R.S.** wrote the manuscript; all authors gave feedback to the manuscript; **R.S.** improved the manuscript; M.B. provided resources

Note. Contributions of the author of this dissertation are in bold.

1.4.3 Open Science Statement

The research presented in this dissertation is based on the idea of open science. Study 1 was exploratory. The research question and methodological procedures were preregistered before data analysis. The open science center honored this article with the preregistration challenge prize (Center for Open Science, 2019). Study 2 was also exploratory, but not preregistered as parts of the data have already been inspected during study 1. For both articles, raw data cannot be published due to the re-identification risk of personal data.

However, aggregated datasets and supplemental materials are available on the respective Open Science Framework (OSF) project pages. The chapters on the respective studies contain the corresponding links.

1.5 References

Ai, P., Liu, Y., & Zhao, X. (2019). Big five personality traits predict daily spatial behavior: Evidence from smartphone data. *Personality and Individual Differences*, *147*, 285–291. doi:10.1016/j.paid.2019.04.027

Alessandretti, L., Lehmann, S., & Baronchelli, A. (2018). Understanding the interplay between social and spatial behaviour. *EPJ Data Science*, *7*(1). doi:10.1140/epjds/s13688-018-0164-6

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. doi:10.1111/j.1745-6916.2007.00051.x

Beierle, F., Tran, V. T., Allemand, M., Neff, P., Schlee, W., Probst, T., ... Pryss, R. (2019). What data are smartphone users willing to share with researchers? *Journal of Ambient Intelligence and Humanized Computing*, 1–13. doi:10.1007/s12652-019-01355-6

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. doi:10.1214/ss/1009213726

Celis-Morales, C. A., Perez-Bravo, F., Ibañez, L., Salas, C., Bailey, M. E. S., & Gill, J. M. R. (2012). Objective vs. self-reported physical activity and sedentary time: Effects of measurement method on relationships with risk biomarkers. *PLoS ONE*, *7*(5), e36345. doi:10.1371/journal.pone.0036345

Center for Open Science. (2019, February 4). *December 2018 winners of the Prereg Prize*. Retrieved from https://cos.io/about/news/December-2018-Prereg-Winners/

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. doi:10.1037/met0000111

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011). Who's who with big-five: Analyzing and classifying personality traits with smartphones. *15th Annual International Symposium on Wearable Computers*, 29–36. doi:10.1109/ISWC.2011.29

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433–450. doi:10.1007/s00779-011-0490-1

Conner, T. S., & Mehl, M. R. (2015). Ambulatory assessment: Methods for studying everyday life. In R. A. Scott, M. C. Buchmann, & S. M. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–15). doi:10.1002/9781118900772.etrds0010

Cornet, V. P., & Holden, R. J. (2018). Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics*, *77*, 120–132. doi:10.1016/j.jbi.2017.12.008

Danner, D., Rammstedt, B., Bluemke, M., Treiber, L., Berres, S., Soto, C., & John, O. (2016). Die deutsche version des big five inventory 2 (bfi-2) [german version of the big five inventory (bfi-2)]. *Zusammenstellung Sozialwissenschaftlicher Items und Skalen*, Advance online publication. doi:10.6102/zis247

De Raad, B. (2000). *The big five personality factors: The psycholexical approach to personality.* Göttingen, Germany: Hogrefe & Huber Publishers.

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 48–55. doi:10.1007/978-3-642-37210-0_6

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... Seligman, M. E. P. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. doi:10.1177/0956797614557867

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*(1), 197–221. doi:10.1146/annurev.psych.52.1.197

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, *40*(1), 21–34. doi:10.1016/j.jrp.2005.08.003

Funder, D. C. (2009). Persons, behaviors and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality*, *43*(2), 120–126. doi:10.1016/j.jrp.2008.12.041

Harari, G. M., Gosling, S. D., Wang, R., & Campbell, A. T. (2015). Capturing situational information with smartphones and mobile sensing methods. *European Journal of Personality*, *29*(5), 509–511. doi:10.1002/per.2032

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. *Perspectives on Psychological Science*, *11*(6), 838–854. doi:10.1177/1745691616650285

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., ... Gosling, S. D. (2019). Sensing sociability: Individual differences in young adults' conversation,

calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, Advance online publication. doi:10.1037/pspp0000245

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. doi:10.1016/j.cobeha.2017.07.018

Harari, G. M., Vaid, S. S., Müller, S. R., Stachl, C., Marrero, Z., Schoedel, R., ... Gosling, S. D. (2020). Personality sensing for theory development and assessment in the digital age. *European Journal of Personality*. doi:10.1002/per.2273. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/per.2273

Junco, R. (2013). Comparing actual and self-reported measures of facebook use. *Computers in Human Behavior*, *29*(3), 626–631. doi:10.1016/j.chb.2012.11.007

Kambham, N. K., Stanley, K. G., & Bell, S. (2018). Predicting personality traits using smartphone sensor data and app usage data, 125–132. doi:10.1109/iemcon.2018.8614854

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. doi:10.1073/pnas.1218772110

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., ... Alstyne, M. V. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Lee, B.-G., & Chung, W.-Y. (2017). Wearable glove-type driver stress detection using a motion sensor. *Transactions on Intelligent Transportation Systems*, *18*(7), 1835–1844. doi:10.1109/tits.2016.2617881

Lewin, K. (1951). *Field theory in social science*. New York, USA: Harper.

Lin, Y.-H., Wong, B.-Y., Lin, S.-H., Chiu, Y.-C., Pan, Y.-C., & Lee, Y.-H. (2019). Development of a mobile application (app) to delineate "digital chronotype" and the effects of delayed chronotype by bedtime smartphone use. *Journal of Psychiatric Research*, *110*, 9–15. doi:10.1016/j.jpsychires.2018.12.012

Mahmoodi, J., Leckelt, M., van Zalk, M., Geukes, K., & Back, M. (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. doi:10.1016/j.cobeha.2017.07.001

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-informatics: Big data shaping modern psychometrics. *Medical Hypotheses*, *82*(4), 405–411. doi:10.1016/j.mehy.2013.11.030

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Boston, USA: Houghton Mifflin Harcourt.

McCrae, R. R. (2009). The five-factor model of personality traits: Consensus and controversy. In P. J. Corr & G. Matthews (Eds.), *The Cambridge Handbook of Personality Psychology* (pp. 148–161). doi:10.1017/cbo9780511596544.012

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517–523. doi:10.3758/bf03195410

Mehrotra, A., Müller, S. R., Harari, G. M., Gosling, S. D., Mascolo, C., Musolesi, M., & Rentfrow, P. J. (2017). Understanding the role of places and activities on mobile phone interaction and usage patterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–22. doi:10.1145/3131901

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

Min, J.-K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., & Hong, J. I. (2014). Toss'n'turn: Smartphone as sleep and sleep quality detector. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 477–486. doi:10.1145/2556288.2557220

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, *13*(1), 23–47. doi:10.1146/annurev-clinpsy-032816-044949

Mønsted, B., Mollgaard, A., & Mathiesen, J. (2018). Phone-based metric as a predictor for basic personality traits. *Journal of Research in Personality*, *74*, 16–22. doi:10.1016/j.jrp.2017.12.004

Montag, C., Baumeister, H., Kannen, C., Sariyska, R., Meßner, E.-M., & Brand, M. (2019). Concept, possibilities and pilot-testing of a new smartphone application for the social and life sciences to study human behavior including validation data from personality psychology. *J—Multidisciplinary Scientific Journal*, *2*(2), 102–115. doi:10.3390/j2020008

Montag, C., Błaszkiewicz, K., Lachmann, B., Andone, I., Sariyska, R., Trendafilov, B., ... Markowetz, A. (2014). Correlating personality and actual phone usage. *Journal of Individual Differences*, *35*(3), 158–165. doi:10.1027/1614-0001/a000139

Montag, C., Duke, É., & Markowetz, A. (2016). Toward psychoinformatics: Computer science meets psychology. *Computational and Mathematical Methods in Medicine*, *2016*, 1–10. doi:10.1155/2016/2983685

Montag, C., & Elhai, J. D. (2019). A new agenda for personality psychology in the digital age? *Personality and Individual Differences*, *147*, 128–134. doi:10.1016/j.paid.2019.03.045

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., & Rentfrow, J. (2018). Musical preferences predict personality: Evidence from active listening and facebook likes. *Psychological Science*, *29*(7), 1145–1158. doi:10.1177/0956797618761659

Nelson, B. W., & Allen, N. B. (2018). Extending the passive-sensing toolbox: Using smart-home technology in psychological science. *Perspectives on Psychological Science*, *13*(6), 718–733. doi:10.1177/1745691618776008

O'Dea, S. (2020, February 28). Smartphone users worldwide 2016-2021. Retrieved March 5, 2020, from https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. doi:10.1037/pspp0000020

Piwek, L., & Ellis, D. A. (2016). Can programming frameworks bring smartphones into the mainstream of psychological science? *Frontiers in Psychology*, *7*(1252). doi:10.3389/fpsyg.2016.01252

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., ... Funder, D. C. (2014). The situational eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. doi:10.1037/a0037250

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*(3), 363–381. doi:10.1002/per.1994

Seifert, A., Hofer, M., & Allemand, M. (2018). Mobile data collection: Smart, but not (yet) smart enough. *Frontiers in Neuroscience*, *12*. doi:10.3389/fnins.2018.00971

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi:10.2139/ssrn.1351252

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., ... Bühner, M. (2019). Behavioral patterns in smartphone usage predict big five personality traits. *Psyarxiv*. doi:10.31234/osf.io/ks4vd

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., ... Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722. doi:10.1002/per.2113

Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big data approach. *Behavior Research Methods*, *51*(4), 1586–1600. doi:10.3758/s13428-019-01235-z

Van De Water, A. T., Holmes, A., & Hurley, D. A. (2011). Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review. *Journal of Sleep Research*, *20*(1pt2), 183–200. doi:10.1111/j.1365-2869.2009.00814.x

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., ... 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. doi:10.1145/2632048.2632054

Wang, W., Harari, G. M., Wang, R., Müller, S. R., Mirjafari, S., Masaba, K., & Campbell, A. T. (2018). Sensing behavioral change over time. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–21. doi:10.1145/3264951

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, *114*(2), 246–257. doi:10.1037/pspa0000098

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, *84*(3), 552–569. doi:10.1016/j.biopsycho.2010.01.017

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. doi:10.1002/per.1986

Yarkoni, T. (2012). Psychoinformatics. *Current Directions in Psychological Science*, *21*(6), 391–397. doi:10.1177/0963721412457362

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

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. doi:10.1073/pnas.1418680112

Zenonos, A., Khan, A., Kalogridis, G., Vatsikas, S., Lewis, T., & Sooriyabandara, M. (2016). HealthyOffice: Mood recognition at work using smartphones and wearable sensors. *International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. doi:10.1109/percomw.2016.7457166

Zuckerman, M. (1994). *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. New York, USA, Melbourne, Australia: Cambridge University Press.

Chapter 2

Digital Footprints of Sensation Seeking: A Traditional Concept in the Big Data Era

Schoedel, R., Au, Q., Völkel, S.T., Lehmann, F., Becker, D., Bühner, M., Bischl, B., Hussmann, H., & Stachl., C. (2018). Digital Footprints of Sensation Seeking: A Traditional Concept in the Big Data Era. In *Zeitschrift für Psychologie, 226(4)*, 232-245. https://doi.org/10.1027/2151-2604/a000342.

2.1 Abstract

The increasing usage of new technologies implies changes for personality research. First, human behavior becomes measurable by digital data, and second, digital manifestations to some extent replace conventional behavior in the analog world. This offers the opportunity to investigate personality traits by means of digital footprints. In this context, the investigation of the personality trait sensation seeking attracted our attention as objective behavioral correlates have been missing so far. By collecting behavioral markers (e.g., communication or app usage) via Android smartphones, we examined whether self-reported sensation seeking scores can be reliably predicted. Overall, 260 subjects participated in our 30-day real-life data logging study. Using a machine learning approach, we evaluated cross-validated model fit based on how accurate sensation seeking scores can be predicted in unseen samples. Our findings highlight the potential of mobile sensing techniques in personality research and show exemplarily how prediction approaches can help to foster an increased understanding of human behavior.

Keywords: Sensation Seeking, Machine Learning, Big Data, Behavior, Smartphone Sensing

2.2 Introduction

Only recently researchers have started to discover the potential of big data for research in psychology. E. E. Chen and Wojcik (2016) pointed out that the rather theory-driven field of psychology could benefit from an additional focus on big data methods such as prediction modeling (Yarkoni & Westfall, 2017). Inversely, Cheung and Jak (2016) highlighted that the discipline of psychology traditionally aims to explain complex issues and consequently could help to develop an understanding of big data. Although only a small number of studies has so far combined both approaches, an increasing potential for such studies exists. These days, people produce vast amounts of user data throughout their daily lives by means of increased technology usage (E. E. Chen & Wojcik, 2016). Thereby, human behavior becomes more and more quantifiable in terms of data (e.g., mobility can be measured via GPS data; Harari et al., 2016). Furthermore, according to Mayer-Schönberger and Cukier (2013), digital behavior even replaces formerly "analog" behavior (e.g., using gaming apps on a smartphone instead of playing a card game). Such digital footprints can be used for personality research as they offer the opportunity for traits to manifest in a new context and to investigate those manifestations in terms of daily usage behavior.

2.2.1 The Personality Trait of Sensation Seeking

Why do some people go skydiving, while others read detective stories to feel aroused? Systematic, individual differences in the need for external stimulation have been described as the personality trait sensation seeking (Zuckerman, 1994). Initially proposed by Zuckerman, it refers to "seeking of varied, novel, complex, and intense sensations and experiences, and the willingness to take physical, social, legal, and financial risks for the sake of such experience" (Zuckerman, 1994, p.27). The construct of sensation seeking has been defined from a biopsychological personality perspective and is explained by genetic, biological, psychophysiological, but also social factors (Roberti, 2004; Zuckerman, 1994). Accordingly, age and sex were found to be related to sensation seeking, namely younger and male individuals showed higher trait scores (Roberti, 2004). After reviewing the vast amount of existing studies on sensation seeking, we have identified three key issues that provide room for new research.

First, the majority of studies has dealt with an unsocialized form of sensation seeking. This term refers to actions like criminal behaviors, alcohol and substance usage, excessive gambling, risky sexual activities, or reckless driving (Roberti, 2004). However, Zuckerman

(1994) also postulated the existence of a non-impulsive, socialized type of sensation seeking. This type was described by individual characteristics such as being against conventionalism, lacking planning skills (Glicksohn & Abulafia, 1998), and by an affinity for unfamiliar international travel destinations (Lepp & Gibson, 2008).

Second, most previous studies have been focused on high-risk activities including taking financial risks (Zabel, Christopher, Marek, Wieth, & Carlson, 2009) or doing extreme sports (Jack & Ronan, 1998). Thus, Guszkowska and Bołdak (2010) found that individual levels of sensation seeking are positively related to practicing sports like parachuting, snowboarding, or alpinism. However, according to Roberti (2004), sensation seeking is not limited to the seeking of risks per se. Rather, a certain amount of risk is accepted to obtain an ideal level of arousal. In contrast to research focusing on high-risk activities, studies about everyday expressions of sensation seeking have been rare and have investigated, for example, the association between sensation seeking and the need for social stimulation (Weisskirch & Murphy, 2004).

Third, traditionally the collection of data about actual behavior has been very difficult and costly to achieve. Behavioral correlates of sensation seeking like reckless driving (Dahlen, Martin, Ragan, & Kuhlman, 2005) or smartphone usage (Leung, 2008) have almost exclusively been measured via retrospective self-reports. However, it is commonly known that self-report questionnaires are subject to a series of biases, such as memory and social desirability (Ziegler & Buehner, 2009). Accordingly, Baumeister, Vohs, and Funder (2007) argued that self-reported behavior can greatly differ from actual behavior and highlighted the necessity to investigate behavior directly. To summarize our three key points, previous studies have mainly focused on self-reports of unsocialized, and high-risk-related types of sensation seeking. This motivates our research effort to re-investigate the socialized and everyday expression of sensation seeking by using objective behavioral data collected via smartphone sensing.

2.2.2 Smartphone Sensing and Automated Trait Recognition

Within the last few years, smartphone sensing has established itself as an active area of research within the field of psychology (Harari et al., 2016). An increasing number of consumer electronics are equipped with sensors capable of logging data about its user's natural mobility and everyday activities, and habits. These developments enable researchers to develop applications (apps) to collect extensive records of individual behavior in an efficient and unobtrusive manner (Harari et al., 2016). Smartphone sensing seems especially promis-

ing for personality psychology, as more and more behaviors (e.g., shopping, listening to music, playing games) can be exerted via smartphones, reflecting potential dimensions of individual difference. Accordingly, a growing body of research has investigated associations of smartphone usage and individual traits. So far, there has been consensus that individual traits are related to smartphone usage behavior in some way. Andone et al. (2016) reported that age and gender were systematically related to individual smartphone usage. Montag et al. (2015) reported associations of extraversion and conscientiousness with daily WhatsApp usage. Smartphone usage in a broader sense was examined by Stachl et al. (2017). They evaluated the predictive performance of personality traits, fluid intelligence, and demographic variables for the frequency and duration of categorical app usage.

Beyond mere association, patterns in sensing data could also be used to directly predict individual trait levels. The idea of inferring states and traits from the everyday digital technology usage has recently gained importance in the field of psychology. So far, studies have focused on the investigation of social network data (e.g., Kosinski, Stillwell, & Graepel, 2013; Youyou, Kosinski, & Stillwell, 2015).

Researchers from other fields have started to investigate the automatic inference of traits based on data collected via smartphones. Chittaranjan, Blom, and Gatica-Perez (2013) and de Montjoye, Quoidbach, Robic, and Pentland (2013) used machine learning algorithms to predict Big Five traits based on smartphone logging data. Whereas Chittaranjan et al. (2013) focused on features derived from app, text message, and call logs, de Montjoye et al. (2013) additionally included features based on location data. Despite their slightly different approaches, both Chittaranjan et al. (2013) and de Montjoye et al. (2013) reported that their machine learning algorithms could predict personality traits above chance.

If successful, the automated recognition of trait variables from usage data could have impact on both the academic and industrial sector. First, predicted traits could be used in recommender systems to develop personalized services or interfaces (Brinkman & Fine, 2005; Tkalcic & Chen, 2015). Second, the recognition of pathological traits like depression could help to develop smartphone-based prevention programs (Saeb et al., 2015). Third, Yarkoni and Westfall (2017) argued that prediction approaches could also help to understand and consequently explain systematic variations in human behavior. It might be promising to revisit theory-based findings with objective data within a machine learning framework to detect possible underlying mechanisms of individual differences in human behavior.

2.2.3 Rationale

The aim of this study was to investigate the traditional concept of sensation seeking as reflected in natural smartphone usage. We think that for observing objective behavioral manifestations of sensation seeking in everyday contexts, appropriate investigation methods have been missing so far. We therefore combined smartphone sensing data with traditional self-report measures, to gain new insights into the behavioral manifestations of sensation seeking. Using a large number of literature-derived predictor variables, we evaluated whether individual sensation seeking scores can be reliably predicted from the data. Additionally, we compared the prediction performance of different machine learning algorithms and investigated the importance of single variables for the models. Moreover, we want to replicate the often reported finding that sensation seeking is related to age as well as gender.

2.3 Method

This study was preregistered prior to analyzing the data. The preregistration form and all supplemental materials are available in our open science framework project (OSF; Schoedel, Au, Völkel, Bühner, & Stachl, 2018) [1]. Our data was collected within the framework of the larger, ongoing "PhoneStudy" project – an interdisciplinary research project between the chair of psychological assessment and the working groups computational statistics as well as media informatics at Ludwig-Maximilians-Universität München (LMU), Germany (see Stachl et al., 2018). The present study obtained approval from the responsible institutional review board and data protection office.

2.3.1 Participants

All participants were recruited by student researchers during a seminar. Participation requirements included speaking German fluently as well as a minimum age of 18 years. For technical reasons, only participants with smartphones running Android 4.4 or higher could participate in the study. Initially, our dataset contained data entries from 361 participants. However, as defined in our preregistration, we only included participants with completed questionnaire data and at least 15 days of logging data in our analyses. This resulted in a final sample size of $N = 260$ participants (68% women). Participants' age ranged from 18 to 72 with an average age of 24 years ($SD = 8.84$). The sample was skewed toward younger and highly educated participants as recruitment took mainly place in the university context. Accordingly, 73% of all participants had a high school degree; 16% had a university degree.

2.3.2 Data Collection Procedure

After being informed about the study, the participants provided informed consent via an online form. In the consequent 30-day data collection period, rich behaviorally focused log data was collected on the participants' smartphones. Participants were instructed to

[1] All supplemental files are now accessible via an open science framework project link: https://osf.io/v4xrf/. *data.csv*: contains the dataset with aggregated features used for prediction modeling; *benchmark.R*: contains the R code for reproducing the reported results; *features.pdf*: lists our features derived from a literature review; *app_categories.csv*: contains our categorization of apps and their definition; *summary_descriptives.pdf*: contains descriptive statistics for all variables; *packages.pdf*: lists all used R packages including version information; *cforest_analyses.pdf*: contains additional analyses regarding the conditional forest learner.

answer a series of self-report questionnaires integrated in the app at a time convenient for them during the study period. The PhoneStudy research app enables unobtrusive data logging utilizing background services to monitor smartphone usage and location tracking. For this study, we focused on the logging of app usage, phone calls, and GPS data. For privacy reasons, we did not collect content-related data (e.g., text or notification contents). App usage and phone calls were recorded event based, location data time based every 15 min. Data was synchronized hourly, if users were connected to WiFi. In the case of missing WiFi connectivity, synchronization was forced using any available network connection after one week. The data was synchronized with a backend server using Secure Sockets Layer (SSL) encryption. Data was stored in encrypted form on the backend server and secured via two-factor authentication. The entire data collection for this study took place between October 2017 and January 2018.

2.3.3 Measures

Self-Report Measures

In previous studies, a series of sensation seeking questionnaires had been used. Although the 40-item Sensation Seeking Scale Form V (SSS-V; Zuckerman, Eysenck, & Eysenck, 1978) was used in most studies, this scale shows weakness in terms of psychometric properties and its factorial structure (Beauducel, Strobel, & Brocke, 2003). Thus, we employed the impulsive sensation seeking (ImpSS) subscale of the Zuckerman–Kuhlman personality questionnaire (ZKPQ-III-R Zuckerman, 2002) which represents a more reliable and valid alternative (Roberti, 2004). Zuckerman (2002) reports good internal consistency (Cronbach's $\alpha = .83$ for a German subsample). The ImpSS consists of 19 items (e.g., "I am an impulsive person"), and participants are instructed to indicate if statements are either *true* or *false*. The ImpSS is defined by two facets: impulsivity (8 items) and sensation seeking (11 items). According to Zuckerman and Aluja (2015), facets can be cumulated to one score due to their joint biological basis. Therefore, we summed up the 19 individual item scores to one ImpSS score (ranging between 0 and 19). For our sample, we found Cronbach's $\alpha = .80$, $CI_{95\%}$ [0.77, 0.84]. Moreover, participants were asked to indicate their demographics. In addition, participants completed the German version of the Big Five Structure Inventory (BFSI; Arendasy, 2009), the newer German version of the Big Five Inventory 2 (BFI-2; Danner et al., 2016), and the Smartphone Addiction Scale (SAS; Kwon et al., 2013). As those questionnaires were used for additional research, not covered in this

study, we will not continue to elaborate on it in this article.

Behavioral Measures and Extracted Features

Originally, the data existed as time-stamped event data. Each row represented a registered event (e.g., call, app usage), each column an event characteristic (e.g., outgoing, time stamp, duration, contact-hash). Thus, before modeling, we preprocessed our dataset in order to create meaningful predictors (also called features in machine learning) for our models. The feature extraction was carried out with specifically created aggregation functions from an R-package, currently under development by the working group of computational statistics at LMU.

Identification and Quantification of Behavioral Categories Initially, we performed an extensive literature review to identify behaviors characteristic for sensation seeking. As we could not find research about sensation seeking and smartphone usage, we identified behavioral manifestations of sensation seeking from "traditional" literature and matched those to measures of possibly equivalent smartphone usage. For example, sensation seeking was commonly associated with gambling in previous studies (McDaniel, 2002). Consequently, we "translated" gambling into gaming app usage behavior. Afterward, we quantified the literature-derived categories (e.g., gaming app usage) by following previous research investigating the relationship of smartphone usage and user characteristics (e.g., Chittaranjan, Blom, & Gatica-Perez, 2011; de Montjoye et al., 2013; Stachl et al., 2017). Used quantification measures were for example mean/variation of frequency and duration, entropy, irregularity, ratio, or radius of gyration. For their detailed explanation, see Table A1 in the Appendix. The complete feature list was preregistered prior to data analyses and is available in our OSF project.

Categorization of Apps In order to effectively analyze app usage data, we chose to categorize all used apps into a finite number of categories. The Google Play store offers a categorization of apps (Google, 2018). However, this categorization is based on the subjective labeling by app developers and might be influenced by reasons like popularity of certain app categories. We therefore predefined our own app categories relevant for our research question: gaming, dating, communication, social media, listening to music/audio clips, watching video clips, planning and organizing, traveling, trading, browsing, shopping, reading news, personalizing the own smartphone, informing about risky driving behavior,

and apps related to running as well as to outdoor sporting activities. In order to increase transparency of our categorization approach, we provide the full list of apps, assigned labels and the definition of all categories in our OSF project.

In the course of data pre-processing, all apps were categorized manually by one coder who read the descriptions provided in the Google Play store. A second coder checked the reliability of these codings and ambiguous cases were discussed with a third coder. Only apps available in the Google Play store at the time of recategorization (18.01.2018) were included. Background and launcher apps were excluded, as they do not reflect intentional app usage behavior.

2.3.4 Data Preprocessing

In order to prepare the dataset for prediction modeling, we applied a series of preprocessing steps according to Kuhn and Johnson (2013) and Schiffner et al. (2016). We removed predictors with more than 90% missing values and predictors with zero or near-zero variance (10% cut-off). To avoid overfitting and to get a reliable estimate of the predictive performance on new data, the preprocessing steps transformation (scaling and centering) and imputation of missing values were performed within the respective inner resampling iterations. In our preregistration, we planned to use a k-nearest neighbor's algorithm for imputation. Due to software-related bugs, we had to use the median for imputation.

2.3.5 Data Analysis

First, we aimed to replicate the often reported finding that impulsive sensation seeking is related to both age and gender (Roberti, 2004). To do so, we calculated Bonferroni corrected pairwise Spearman correlations. In addition, we calculated simple pairwise correlations between impulsive sensation seeking scores and the self-reported Big-Five personality scores. As suggested in previous literature (Yarkoni, 2010), we consistently used Spearman's correlation coefficients due to non-normally distributed data. Second, we computed descriptive statistics related to smartphone usage and app usage in particular. Third, we used a machine learning approach to predict self-reported sensation seeking scores from the features described in the method section.

Machine Learning Algorithms

Within algorithmic modeling culture, it is assumed that there is no single best model (Wolpert & Macready, 1997). Rather, various models perform differently well, dependent on the unknown true relationship between predictors and outcome. Therefore, we carried out a benchmark experiment in which we compared the generalized predictive performance of different algorithms (also called learners) against a common guessing baseline. This baseline is also called "featureless learner" and constantly predicts the mean value of the training data's outcome value. The learners we chose for the benchmark experiment represent various trade-off levels between interpretability and expected prediction performance. First, we used an elastic net model (J. Friedman, Hastie, & Tibshirani, 2010; Zou & Hastie, 2005). It is a linear regression method applying a mixed L1-L2 regularization which allows to model linear relationships on high-dimensional spaces. Furthermore, the L1 penalty drives irrelevant predictor variables out of the model for model sparsity and therefore better interpretability. We chose the elastic net model because it has often been proven to be competitive in contrast to nonlinear methods and provides well interpretable coefficients (Zou & Hastie, 2005). Second, we included a random forest (Breiman, 2001; Wright & Ziegler, 2017) which is an ensemble technique of multiple bootstrapped, decorrelated decision trees. The random forest as non-linear model is an all-rounder, which can handle high-dimensional feature spaces and small sample sizes usually very well. Third, a support vector machine (SVM) with radial basis function (RBF) kernel was used (Karatzoglou, Smola, Hornik, & Zeileis, 2004; Vapnik, 1999). Through its kernel function, the SVM implicitly maps the training observations into a high-dimensional feature space, where a linear decision boundary is learnt. This results in a non-linear decision boundary in the original feature space. We included the SVM because it is the most prevalent one used for personality prediction in psychological research (Chittaranjan et al., 2013; de Montjoye et al., 2013). Forth, we used extreme gradient boosting (xgboost; T. Chen, He, & Benesty, 2015; J. H. Friedman, 2001). This method is again an ensemble technique based on trees, which are combined via sequential gradient boosting. Currently, xgboost is considered one of the most powerful prediction algorithms in the machine learning community.

Evaluation Metrics

We consider metrics that are typically used to measure the predictive performance of regression models: mean squared error (MSE), root mean squared error ($RMSE$), mean absolute deviation (MAE), and the coefficient of determination (R^2) (e.g., James, Witten,

Hastie, & Tibshirani, 2013; Kuhn & Johnson, 2013). For the three metrics *MSE*, *RMSE*, and *MAE*, it is valid that lower values (approaching zero) indicate better model performance. The measure of R^2 is also referred to as the coefficient of determination. According to the conventional, in psychological research prevalent definition of R^2, its values range between 0 and 1, whereby the closer R^2 is to 1, the better the model explains the data. However, if model training and model evaluation happens on different datasets (e.g., in cross-validation), the mean of the response values between the training and the validation dataset can vary greatly, and therefore, R^2 can become negative (Alexander, Tropsha, & Winkler, 2015). According to Alexander et al. (2015), negative values indicate that model fit is poor and that the number of observations is too small. As there is no consensus in literature which metric is superior to others, we follow Chai and Draxler (2014) and consider a combination of all metrics for model evaluation. In addition, we will present correlation coefficients between actual and predicted sensation seeking scores.

Resampling Procedure

For each learner, the optimal choice of hyperparameters is data-dependent (Schiffner et al., 2016). To avoid overfitting, we applied a nested resampling strategy selecting optimal hyperparameters within inner resampling loops. The predictive performance of the tuned learners is then evaluated within separate outer resampling loops. This ensures a strict separation of training and test data while allowing for the tuning of hyperparameters as well as preprocessing. More information about the detailed tuning procedure is included in the R code of the benchmark experiment which can be found as a supplemental file in our OSF project. For the inner resampling loops, we used simple holdout validation for all learners. In the outer resampling loop, 10 times repeated 10-fold cross-validation was performed. Tuning for all learners was optimized on the MSE performance metric.

Variable Importance

We selected the best prediction model with regard to the presented performance measures. To achieve a better understanding of which variables were important for prediction success, methods-inherent variable importance measures and partial dependence plots are presented (Schiffner et al., 2016). The plots help to explore the partial dependence of the trained function by selecting a subset of the predictor space. That means, the curves show how a trained function takes the values of features into consideration in order to predict sensation seeking scores.

Statistical Software

Data processing and statistical analyses were performed with the statistical software R 3.4.3 (R Core Team, 2017). For preprocessing, we used the car and dplyr packages (Fox et al., 2012; Wickham, Francois, Henry, & Müller, 2017). For modeling, we used functions from the mlr, caret, and psych packages (Bischl et al., 2016; Kuhn, 2017; Revelle, 2017). See the OSF project link in footnote 1 for a complete overview of all R packages used including version information.

2.4 Results

2.4.1 Descriptive Statistics

Our dataset contained 222 predictors before and 178 variables after preprocessing. On average, 1,263 daily events were recorded for each participant. The participants used a total number of 2,205 different apps during the course of the study. Table 2.1 shows information about the overall usage frequencies of app categories. The most frequently used app categories were related to communication, social media usage, and browser usage. The number of different apps within one category was highest for gaming apps. Due to the scope of this article, summary statistics for all included variables are provided as supplemental files in our OSF project.

2.4.2 Impulsive Sensation Seeking and Demographics

On average, participants reported an ImpSS score of M = 7.91 (SD = 4.22) which is in line with previous literature (e.g., Aluja, Garcia, & Garcia, 2003). Contrary to our assumptions, neither age (r_s = -0.04, $CI_{95\%}$ [-0.15, 0.07]), nor gender (r_s = -0.02, $CI_{95\%}$ [-0.15, 0.14]) were significantly related to sensation seeking.

2.4.3 Prediction of Individual Sensation Seeking Scores

Benchmark Results

Table 2.2 presents the results of our benchmark experiment. The mean performance measures *MSE*, *RMSE*, and *MAE* were lowest, and R^2 was highest for the random forest compared to extreme gradient boosting, the support vector machine, and the elastic net. The mean *MSE* of the random forest was 10% lower; the mean *RMSE* and *MAE* were 5%

Table 2.1: Descriptive statistics of app category usage

App category	$M_{Freq.total}$	$SD_{Freq.total}$	Num_{Users}	Num_{Apps}	M_{ImpSS}	SD_{ImpSS}
Communication apps	1,522.88	1,576.08	234	59	7.97	4.24
Social media apps	485.34	715.83	203	70	8.18	4.20
Browser apps	210.35	316.01	231	22	8.02	4.24
Music and audio apps	183.97	522.02	190	85	8.13	4.20
Planning tool apps	92.85	185.26	209	70	7.97	4.22
Gaming apps	87.18	199.74	140	415	7.71	4.28
Video watching apps	80.43	155.22	210	48	8.00	4.25
Trip planning apps	38.60	61.20	208	52	8.04	4.17
News apps	31.63	138.31	125	48	7.80	4.22
Shopping apps	22.73	60.87	107	60	7.89	4.03
Dating apps	22.38	132.15	24	13	9.21	4.86
Trading apps	14.27	177.11	8	27	11.38	6.61
Personalization apps	12.88	147.34	47	34	8.06	4.72
Running sports apps	9.03	78.59	37	8	7.08	3.74
Risky driving apps	0.00	0.06	1	2	9.00	NA
Outdoorsports apps	0.05	0.74	1	8	12.00	NA

Note. $M_{Freq.total}$ = average total usage count within 30 days; $SD_{Freq.total}$ = standard deviation of average total usage count within 30 days; Num_{Users} = number of all users that have ever used an app of the respective agg category; Num_{Apps} = number of different apps within one category. App categories are sorted in descending order of $M_{Freq.total}$.

lower than the guessing baseline. However, the dispersion of the MSE and R^2 across all 100 iterations (see Figure 2.1) shows that in some iterations, R^2 was negative for the random forest, indicating poor fit (Alexander et al., 2015). Despite the relatively low mean R^2 (0.06) of the random forest model, we assume that the model grasped systematic variance in sensation seeking-related behaviors. Due to both the constantly better performance measures and a Pearson correlation of $r = 0.31$ between true and predicted test data, we consider the random forest provided predictions even if only slightly better than predicting by chance.

Table 2.2: Summary of mean performance measures of the 10x10 CV benchmark experiment

Measures	FL	RF	XG	SVM	EN
MSE	17.83	16.03	16.71	17.35	17.43
MAE	3.52	3.34	3.37	3.45	3.44
$RMSE$	4.22	4.00	4.09	4.17	4.18
R^2	-0.04	0.06	0.02	-0.02	-0.01
r	NA	0.31	0.27	0.18	0.17

Note. FL = featureless learner; RF = random forest; XG = extreme gradient boosting; SVM = support vector machine; EN = elastic net. MSE = mean squared error; $RMSE$ = root mean squared error; MAE = mean absolute deviation; R^2 = coefficient of determination; r = Pearson correlation.

Variable Importance

Permutation-Based Feature Importance To gain a better understanding of how the random forest model predicted new cases, we investigated the permutation-based feature importance measures for the top ten predictors. According to Breiman (2001), the idea behind is that first, the initial relation of one feature with the criterion variable is dissolved by randomly permuting the respective feature. Second, the permutated feature and all other remaining (unchanged) features are used to predict the criterion. The variable importance measure is the result of taking into account the difference in the prediction performance before and after permuting the respective feature. The larger the reduction in the prediction performance is, the stronger is the initial relation between the particular feature and the criterion variable, and consequently, the more important is this respective feature in the model (Breiman, 2001). Table 2.3 displays the top ten predictors with the highest permutation-based feature importance (Wright & Ziegler, 2017). To illustrate the prediction direction of features, we added pairwise Spearman correlations between predictors and sensation seeking to the table.

The list suggests that the top ten features for predicting sensation seeking belonged to two primary categories: calling and day/night time activity. Calling activity included outgoing and missed calls, represented via different quantification metrics. For example, the random forest judged participants as higher sensation seekers if they initialized or missed calls more often. In addition, the entropy of calling turned out to be important. Spearman coefficients suggest positive relationships between entropy of contactsrelated

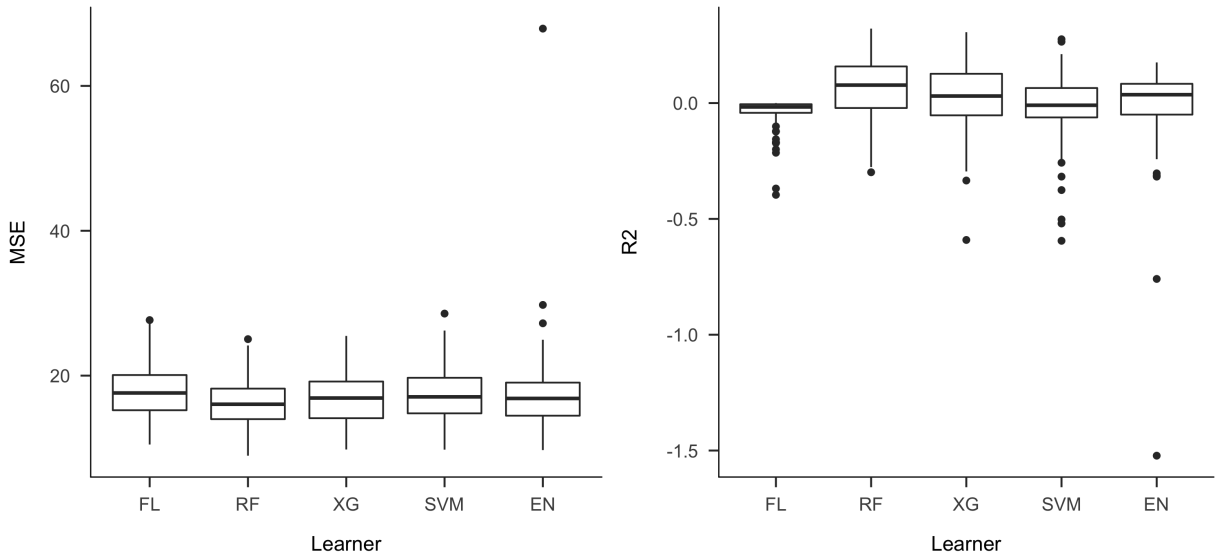

Figure 2.1: Distribution of the mean squared error (*MSE*) and the coefficient of determination (R^2) across all resampling iterations. FL = featureless learner; RF = random forest; XG = extreme gradient boosting; SVM = support vector machine; EN = elastic net.

variables and sensation seeking scores. Another set of important features for the random forest was related to individual day and night time activity. Spearman correlations suggest positive associations between night time activity indicators and sensation seeking levels. As an illustration, predicted sensation scores were higher, if the average point of time of the last smartphone usage on Friday/Saturday or on Sunday was late and if the mean range of motion was high during night at weekends.

Some of our features were highly inter-correlated. Multicollinearity has been proven not to be an issue for the predictive performance of the random forest (and other machine learning algorithms), but to be likely to bias variable importance measures (James et al., 2013; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). We therefore conducted an additional analysis using the conditional forest which is a learner taking into account the correlated structure of features (Strobl et al., 2008). As neither prediction performance nor variable importance measures differed considerably between the conditional versus the random forest, and not to go beyond the scope of this article, we only report the results for the random forest here. However, corresponding additional analysis including detailed background information can be found as a supplemental file in our OSF project.

Table 2.3: Variable importance and Spearman correlations for the top 10 predictors

Feature	I	r_s
mean frequency of missed calls per day	0.62	0.32
entropy of contacts for outgoing calls	0.51	0.33
entropy of contacts for missed calls	0.41	0.29
variation of frequency of outgoing calls per day	0.32	0.26
mean time of the last event on Friday/Saturday	0.21	0.18
variation of the time of the first event from Monday to Friday	0.17	0.12
mean number of intended events during night on Friday/Saturday	0.14	0.16
mean radius of gyration during night on Friday/Saturday	0.14	0.31
mean time of the last event on Sunday	0.14	0.21
mean frequency of outgoing calls per day	0.13	0.24

Note. I = permutation-based variable importance. Variables are in descending order of importance scores.

Partial Dependence Plots In addition to feature importance values, partial dependence plots can help to better understand how values of individual features on average influenced the prediction model (see Figure 2.2). The curves show how predicted sensation seeking scores (y-axis) changed with regard to values of the respective predictor variable (x-axis).

In the top left of Figure 2.2, the mean frequency of missed calls per day is plotted against sensation seeking scores. The plot shows that the average frequency of missed calls per day led to an increase in predicted sensation seeking scores for very low-frequency values, but did not change noticeably if a mean value of about 0.4 missed calls per day was exceeded.

At the top right of Figure 2.2, a partial dependence plot for entropy of contacts for outgoing calls is visible. Increasing values in contact-entropy on average resulted in higher predicted sensation seeking scores. This increase got sharper with rising entropy values.

As shown in the bottom left of Figure 2.2, with a rising mean number of intended events on Fridays/Saturdays nights predicted sensation seeking scores first slowly and from a value of about 15 intended events sharply increased. Events were counted as “intended” when they were carried out intentionally by the participant.

The curve in the bottom right of Figure 2.2 displays, that the mean time of the last

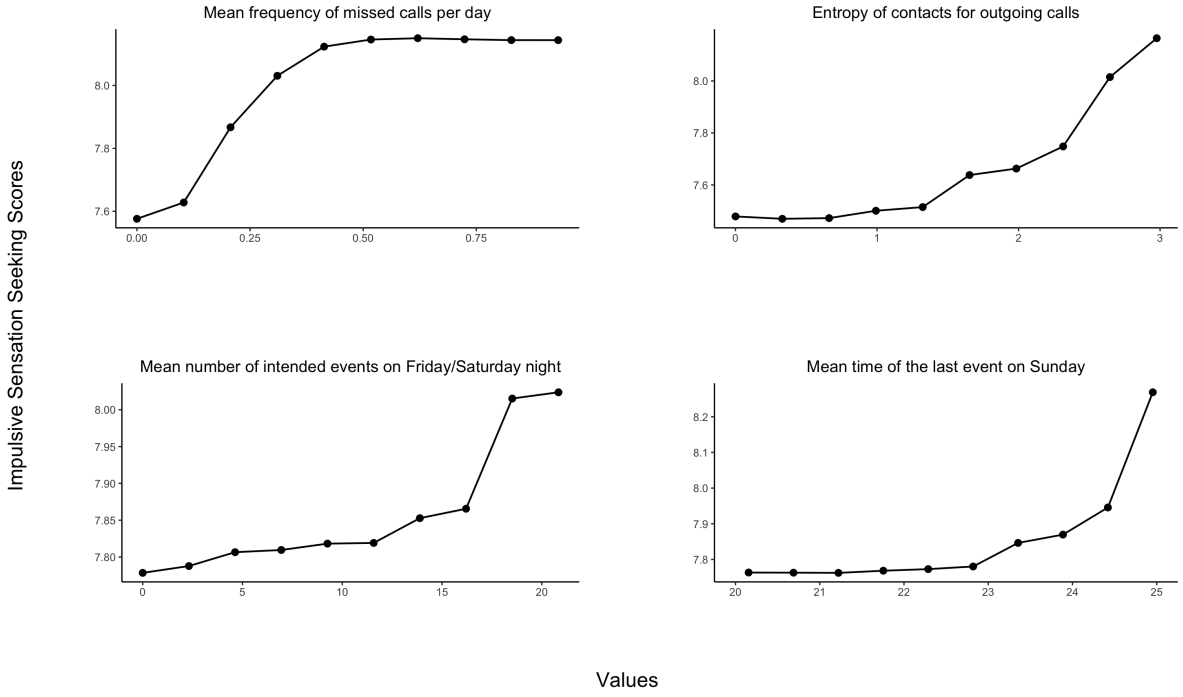

Figure 2.2: Plots displaying the partial dependence of the random forest learner function for four exemplary selected features. The curves show how predicted sensation seeking scores (on the vertical axis) change with increasing values (on the horizontal axis) of the respective displayed features. A last event value of 25 in the bottom right means, that the smartphone was used at 1am in the night between Sunday and Monday.

event on Sunday on average led to higher predicted sensation seeking scores, when they occurred after around 11 pm.

2.5 Discussion

The results of the present study indicate that individual scores of self-reported sensation seeking can be predicted from digital records of smartphone behavior above chance. Precisely, our results suggest that variables related to calling behavior as well as day and night time activity were particularly predictive for individual sensation seeking scores. In the following subsections, we will critically discuss the results within the context of the used machine learning approach and will try to give some post hoc explanations for important variables in the model. Please note that those interpretations are partially drawn post hoc and should therefore not be easily generalized.

2.5.1 A Timely Approach to a Traditional Concept

In contrast to all previous studies on sensation seeking, we used data about actual behavior to predict sensation seeking with a machine learning approach. Thus, we compared different statistical models based on their ability to accurately predict sensation seeking scores from unseen data.

Despite a relatively low overall accuracy, our results suggest that the flexible, non-linear random forest model outperformed all the other models. This suggests that research in the field of individual differences might benefit from the additional usage of flexible models for the investigation of behavior-trait relationships (Benson & Campbell, 2007).

Previous studies investigating the automatic inference of personality traits based on features extracted from smartphone logs also reported prediction performances, ranging in size from 6% to 25% points depending on respective traits and used preprocessing procedures (Chittaranjan et al., 2011, 2013; de Montjoye et al., 2013). Contrarily, those studies used classification approaches and binned participants in low, average, and high on the Big-Five personality traits. Although this hinders a direct comparison with those studies, we argue that the prediction of continuous trait scores can be more adequately modeled with a regression approach. Related to this, the findings of Kosinski et al. (2013) offer a possibility to put the present results into perspective. Kosinski et al. (2013) predicted personality traits from digital footprints (Facebook likes) and reported correlations between predicted and actually observed trait values in the range of 0.29 for conscientiousness and 0.43 for openness. Despite very different sample sizes, our analyses produced coefficients in a similar range, suggesting comparable prediction performance. Please note that the prediction accuracies of Kosinski et al. (2013) were exceeded in a later study (Youyou et al., 2015).

Although the obtained prediction performance is comparable to previous studies, one question still remains: What is the meaning of being 10% better than guessing? First, we want to point out that in psychological research, it is often investigated how well a model fitted a given dataset (e.g., by considering in-sample R^2), and therefore, how trustworthy it is. In contrast, in the context of prediction modeling, "good" and "trustworthy" are independent criteria. Good model fit refers to how well new, unseen cases can be predicted with a trained model and "trustworthy" indicates the correct application of methodological procedures.

Consequently, with regard to the question how "good" our model fit is, it has to be considered that the prediction performance in the large majority of our folds was above

the guessing baseline, indicating that something more than randomness was going on. However, the overall prediction performance was low. Reasons for this could be the relatively small sample size or that self-reported sensation seeking scores cannot be perfectly predicted from the behavioral indicators used in our study. We carefully selected these indicators by identifying manifestations of sensation seeking assessed via self-reports in previous literature. Though they were not reflected in objective behavioral data to the extent, we would have expected from previous research. This could in turn suggest that the theoretical conceptualization of sensation seeking might benefit from additional research efforts in future studies. To sum up, we argue that our results are very well trustworthy in the sense that they indicate that sensation seeking cannot be predicted very accurately from the used behavioral predictors.

In addition, we think that howmuch better the prediction performance of a learner compared to a guessing baseline has to be is a context-related question. With regard to practical applications (e.g., mobile computing), our model is certainly far away from being good and therefore applicable. However, in the context of psychological research effect sizes are usually very small, and therefore, we would argue that our obtained mean (out of sample) R^2 is not unusually small.

2.5.2 The Trait Sensation Seeking and its Correlates

Beyond the evaluation of prediction performance, our analyses provided more detailed insights into the behavioral correlates of sensation seeking. Following previous studies, we hypothesized that both age and gender are related to sensation seeking (Roberti, 2004). However, associations of demographics with sensation seeking were not present in our data. We suspect that the absence of those effects could be related to our sample characteristics (predominately young females). Although previous studies reported similar gender ratios (Roberti, 2004), age ranges were larger. Possibly, but it can only by suspected, gender and age differences in socialized forms of sensation seeking might also not be as pronounced as in unsocialized forms (e.g., risky driving). However, those post hoc explanations should be tested in future studies.

Although machine learning algorithms are often labeled as "black-box models," they can provide additional information beyond prediction performance. In our study, the inspection of variable importance measures suggested that variables related to calling as well as day and night time activity were particularly important for predicting sensation seeking scores.

The variables of our prediction model were in advance derived from a literature review. For example, we "translated" the finding that sensation seeking is positively associated with a self-reported preference for social contexts (Roberti, 2004) into predictors related to calling activity. Variables regarding day and night time activity were based on findings about the relation between self-reported preferences for later bedtimes and sensation seeking (Tonetti, Fabbri, & Natale, 2009). The reviewed literature was exclusively based on self-reported behavioral correlates. But our model showed that these variables also turned out to be important for the prediction of sensation seeking when they are operationalized by means of behavioral data collected via smartphones. We think that our results can partially help to underpin questionnaire-based research with objective behavioral data.

Additionally, our prediction modeling approach provides new insights into behavioral manifestations of sensation seeking. Although our study cannot raise any claims of causality or explanation, it can foster the postulation of new hypotheses for future studies. For example, two of the three most important features for the random forest's sensation seeking prediction were related to missed calls. As a mental game, one could deduce the hypothesis that people scoring high in sensation seeking are very active and busy in their everyday life and therefore miss incoming calls. Such hypotheses can be tested in futures studies and aid the understanding of behavioral expressions of sensation seeking. To take up the current debate whether novel prediction-focused approaches are contradictory to the explanatory goal of psychology (Yarkoni & Westfall, 2017), we argue that our prediction framework suggests otherwise. Following Yarkoni and Westfall (2017), the present study highlights that prediction approaches can help to better understand the structure of objective behavioral data and can help to generate new hypotheses for confirmatory research.

As discussed in the previous two subsections, our study illustrates how psychology and big data can work together. Psychological theories and findings can help to understand and interpret what machine learning algorithms do (Cheung & Jak, 2016). But conversely, prediction models could help to understand basic structures in complex behaviors (E. E. Chen & Wojcik, 2016). At this point, we want to emphasize that our analyses cannot be considered big data due to the relatively small sample size. However, our data collection tool, the "PhoneStudy app," with its vast variety of collected variables as well as the methods used in this study hopefully highlight some potential of the big data approach in psychological research and inspire future work.

2.5.3 Limitations and Outlook

The present study has some limitations which we will discuss below. The categorization of apps holds the problem that they can be used ambiguously. Hence, an app can be used to fulfill different purposes and needs. For example, browser apps can be used to do online shopping, to read news, to visit social media channels, and so on. As the PhoneStudy app only provides meta-data, we only know that participants used certain apps belonging to predefined categories. However, it remains unclear what participants used the app for. Accordingly, we think that for improving the prediction performance of machine learning algorithms, the inclusion of content-related logging data such as user preferences (e.g., genres of listened music), browsing histories, or notification texts might be a promising strategy. Although more fine-grained data will likely improve trait prediction accuracy, the protection of individual privacy rights must be prioritized.

Our sample was primarily collected in the university context. Thus, younger and higher educated participants were overrepresented in our sample. Accordingly, some of our literature-derived features (usage of counteracting risky driving apps) were automatically excluded in the preprocessing as only single participants used respective apps. As our sample mainly consisted of younger people, car ownership might be systematically underrepresented. A more representative sample (including elders) could therefore provide more variance in behaviors related to sensation seeking.

Furthermore, machine learning algorithms only perform really well with large samples. Relatedly, the negative range of R^2 values of the featureless learner could indicate that our sample size was too small. This study should be replicated with a larger sample size (maybe 10 times), to fully benefit from the predictive capabilities of those methods.

Finally, as already stated by Chittaranjan et al. (2013) and Kosinski et al. (2013), personality trait prediction is a challenging task. Traits are defined as latent constructs and can only be measured roughly, via self-report questionnaires. Therefore, prediction efforts using self-reported trait scores as ground truth, can only achieve accuracies that mimic those of self-report questionnaires. As we know that self-report questionnaires are also affected by a series of biases, this problem needs to be addressed eventually. Nevertheless, trait prediction can be improved in many ways. As the biopsychological trait sensation seeking was found to be related to individually as optimally considered levels of arousal (Roberti, 2004), physiological thresholds might be meaningful indicators. It might be helpful to include measures reflecting physiological processes in prediction models of sensation seeking. Measures of heart rate and electro-thermal activity could be provided

by wearables. Even though we are aware that the performance of our prediction model has to be higher to reach practical importance (e.g., for mobile computing), we think that our study can be a starting point for future research. Accordingly, it is one of the first studies working with such a broad variety of data collected via smartphone sensing in the field of trait prediction, and especially in the context of sensation seeking.

2.6 Conclusion

The present study combined smartphone sensing data with traditional self-report measures, to gain new insights into behavioral manifestations of sensation seeking. The present study shows that self-reported sensation seeking scores can be predicted by smartphone logging data above the level of chance. Despite limited prediction accuracies our results highlight novel behavioral indicators of sensation seeking and the potential of big data for psychological research.

2.7 Acknowledgments

We thank all the student researchers for supporting us with recruiting participants, Henrike Haase for her persistence in categorizing apps, Florian Pargent for his insightful modeling advice, and the PhoneStudy team (Daniel Buschek, Marius Herget, Ferdinand Hof, Peter Ehrich, Miriam Metz, Theresa Ullmann) for making this kind of research possible.

2.8 Appendix

Table 2.4: Description of quantifications of behavioral data collected via smartphones

Quantification	Description
App usage	
Frequency/duration	Usage frequencies and durations of app usage (Stachl et al., 2017) were aggregated as daily mean and variation per day. As the logging of app usage is generally prone to logging errors, robust estimators were used: the huber mean as measure of central tendency (Kafadar, 2003) and the robust location-free scale estimate Q_n as a measure of dispersion (Rousseeuw & Croux, 1993). Robust estimators are less sensitive to outliers which are possibly caused by faulty logs.
Phone usage	
Frequency/duration	Frequencies and durations of incoming/outgoing/missed calls/text messages were aggregated as daily mean and variation per day. We pre-registered to use robust measures for phone logging data, too. However, as their inspection did not reveal the same potential logging errors as for app usage data, we used the arithmetic mean and variance for feature calculation, because they are more precise estimators if outliers are not an issue.
Response rate	The response rate was defined as percentage of missed calls and text messages people responded to by calling back within 24 hours (de Montjoye, Quoidbach, Robic, & Pentland, 2013).
App and phone usage	
Entropy	The entropy describes how many categories one variable has (e.g. total number of contacts), while regarding how equally events (e.g. calls) are distributed across these categories. Therefore, entropy of contacts is high if a person called a broad range of contacts equally often within the study period. We also considered entropy of used apps, measuring how equally often participants used their individual spectrum of installed apps (de Montjoye, Quoidbach, Robic, & Pentland, 2013).
Ratio	The ratio indicates the extent of certain behavioral categories in relation to the overall smartphone usage. For example, we considered the ratio of duration of dating app usage and overall smpartphone usage (de Montjoye, Quoidbach, Robic, & Pentland, 2013).
Irregularity	We computed irregularity of the point of time first and last events happen per day. As defined by Williams, Whitaker, and Allen (2012), this measure represents the dissimilarity of events in a time course. If events happen to very similiar points across time (e.g. every day at 10am), dissimilarity and consequently irregularity are very low.
Mobility	
Radius of gyration	The radius of gyration was used for the quantification of mobility behavior (Canzian & Musolesi, 2015). It quantifies a person's range of mobility by considering the deviation from the centre of all GPS positions, visited within per day.
Total distance covered	The total distance covered was defined as summed distance between sequent GPS points per day (Canzian & Musolesi, 2015).
Maximum distance covered	The maximum distance covered was defined as maximum strech of way per day (Canzian & Musolesi, 2015).

Note. Unlike stated in our pre-registration, we had to exclude the predictors "number of contacts at the beginning of the study" and "number of contacts added within 30 days" as our contact logging data turned out to be corrupted for the majority of our participants due to logging errors. Only after completion of our pre-registration, we conceived the "total usage frequency of app categories within 30 days" as additional important predictors and therefore decided to add them to our feature list.

2.9 References

Alexander, D. L. J., Tropsha, A., & Winkler, D. A. (2015). Beware of r2: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling*, *55*(7), 1316–1322. doi:10.1021/acs.jcim.5b00206

Aluja, A., Garcia, O., & Garcia, L. (2003). Psychometric properties of the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ-III-R): A study of a shortened form. *Personality and Individual Differences*, *34*(7), 1083–1097. doi:10.1016/s0191-8869(02)00097-1

Andone, I., Błaszkiewicz, K., Eibes, M., Trendafilov, B., Montag, C., & Markowetz, A. (2016). How age and gender affect smartphone usage. *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*, 9–12. doi:10.1145/2968219.2971451

Arendasy, M. (2009). BFSI: Big-Five Struktur-Inventar (Test & Manual). *Mödling, Austria: SCHUHFRIED GmbH.*

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. doi:10.1111/j.1745-6916.2007.00051.x

Beauducel, A., Strobel, A., & Brocke, B. (2003). Psychometrische Eigenschaften und Normen einer deutschsprachigen Fassung der Sensation Seeking-Skalen, Form V. *Diagnostica*, *49*, 61–72. doi:10.1026//0012-1924.49.2.61

Benson, M. J., & Campbell, J. P. (2007). To be, or not to be, linear: An expanded representation of personality and its relationship to leadership performance. *International Journal of Selection and Assessment*, *15*(2), 232–249. doi:10.1111/j.1468-2389.2007.00384.x

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... Jones, Z. M. (2016). mlr: Machine learning in r. *Journal of Machine Learning Research*, *17*(170), 1–5. Retrieved from http://jmlr.org/papers/v17/15-066.html

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

Brinkman, W.-P., & Fine, N. (2005). Towards customized emotional design: An explorative study of user personality and user interface skin preferences. *Proceedings of the 2005 Annual Conference on European Association of Cognitive Ergonomics*, 107–114.

Canzian, L., & Musolesi, M. (2015). Trajectories of depression. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, 1293–1304. doi:10.1145/2750858.2805845

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. doi:10.5194/gmd-7-1247-2014

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. doi:10.1037/met0000111

Chen, T., He, T., & Benesty, M. (2015). Xgboost: Extreme gradient boosting. *R package version 0.4-2*. Retrieved from https://cran.r-project.org/web/packages/xgboost/index.html

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, *7*. doi:10.3389/fpsyg.2016.00738

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2011). Who's who with big-five: Analyzing and classifying personality traits with smartphones. *15th Annual International Symposium on Wearable Computers*, 29–36. doi:10.1109/ISWC.2011.29

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433–450. doi:10.1007/s00779-011-0490-1

Dahlen, E. R., Martin, R. C., Ragan, K., & Kuhlman, M. M. (2005). Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving. *Accident Analysis & Prevention*, *37*(2), 341–348. doi:10.1016/j.aap.2004.10.006

Danner, D., Rammstedt, B., Bluemke, M., Treiber, L., Berres, S., Soto, C., & John, O. (2016). Die deutsche version des big five inventory 2 (bfi-2) [german version of the big five inventory (bfi-2)]. *Zusammenstellung Sozialwissenschaftlicher Items und Skalen*, Advance online publication. doi:10.6102/zis247

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 48–55. doi:10.1007/978-3-642-37210-0_6

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ... Graves, S., et al. (2012). Package 'car'. *Vienna: R Foundation for Statistical Computing*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232. doi:10.1214/aos/1013203451

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

Glicksohn, J., & Abulafia, J. (1998). Embedding sensation seeking within the big three. *Personality and Individual Differences*, *25*(6), 1085–1099. doi:10.1016/s0191-8869(98)00096-8

Google. (2018, January 1). Google play store. Retrieved January 18, 2018, from https://play.google.com/store?hl=de

Guszkowska, M., & Bołdak, A. (2010). Sensation seeking in males involved in recreational high risk sports. *Biology of Sport*, *27*(3), 157–162. doi:10.5604/20831862.919331

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. *Perspectives on Psychological Science*, *11*(6), 838–854. doi:10.1177/1745691616650285

Jack, S., & Ronan, K. R. (1998). Sensation seeking among high-and low-risk sports participants. *Personality and Individual Differences*, *25*(6), 1063–1083. doi:10.1016/S0191-8869(98)00081-6

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, USA: Springer.

Kafadar, K. (2003). John tukey and robustness. *Statistical Science*, *18*(3), 319–331. doi:10.1214/ss/1076102419

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20. doi:10.18637/jss.v011.i09

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. doi:10.1073/pnas.1218772110

Kuhn, M. (2017). Caret: Classification and regression training. *R package version 6.0-78*. Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, USA: Springer.

Kwon, M., Lee, J.-Y., Won, W.-Y., Park, J.-W., Min, J.-A., Hahn, C., ... Kim, D.-J. (2013). Development and validation of a smartphone addiction scale (SAS). *PLoS ONE*, *8*(2), e56936. doi:10.1371/journal.pone.0056936

Lepp, A., & Gibson, H. (2008). Sensation seeking and tourism: Tourist role, perception of risk and destination choice. *Tourism Management*, *29*(4), 740–750. doi:10.1016/j.tourman.2007.08.002

Leung, L. (2008). Leisure boredom, sensation seeking, self-esteem, and addiction. In E. A. Konijn, S. Utz, M. Tanis, & S. B. Barnes (Eds.), *Mediated interpersonal communication* (pp. 359–381). New York, USA, London, UK: Routledge.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Boston, USA: Houghton Mifflin Harcourt.

McDaniel, S. R. (2002). Investigating the roles of gambling interest and impulsive sensation seeking on consumer enjoyment of promotional games. *Social Behavior and Personality*, *30*(1), 53–64. doi:10.2224/sbp.2002.30.1.53

Montag, C., Błaszkiewicz, K., Sariyska, R., Lachmann, B., Andone, I., Trendafilov, B., ... Markowetz, A. (2015). Smartphone usage in the 21st century: Who is active on whatsapp? *BMC Research Notes*, *8*(1), 331. doi:10.1186/s13104-015-1280-z

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Revelle, W. (2017). Psych: Procedures for psychological, psychometric, and personality research. *R package version 1.7.8*. Retrieved from https://CRAN.R-project.org/package=psych

Roberti, J. W. (2004). A review of behavioral and biological correlates of sensation seeking. *Journal of Research in Personality*, *38*(3), 256–279. doi:10.1016/S0092-6566(03)00067-9

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. doi:10.2307/2291267

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, *17*(7), e175. doi:10.2196/jmir.4273

Schiffner, J., Bischl, B., Lang, M., Richter, J., M Jones, Z., Probst, P., ... Kotthoff, L. (2016). Mlr tutorial. *Arxiv*. doi:1609.06146

Schoedel, R., Au, Q., Völkel, S., Bühner, M., & Stachl, C. (2018). Digital footprints of sensation seeking: A traditional concept in the big data era. *OSF*. doi:10.17605/osf.io/v4xrf

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., ... Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722. doi:10.1002/per.2113

Stachl, C., Schoedel, R., Au, Q., Völkel, S., Buschek, D., Hussmann, H., ... Bühner, M. (2018). The phonestudy project. *OSF*. doi:10.17605/osf.io/ut42y

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307. doi:10.1186/1471-2105-9-307

Tkalcic, M., & Chen, L. (2015). Personality and recommender systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 715–739). doi:10.1007/978-1-4899-7637-6_21

Tonetti, L., Fabbri, M., & Natale, V. (2009). Relationship between circadian typology and big five personality domains. *Chronobiology International*, *26*(2), 337–347. doi:10.1080/07420520902750995

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, *10*(5), 988–999. doi:10.1109/72.788640

Weisskirch, R. S., & Murphy, L. C. (2004). Friends, porn, and punk: Sensation seeking in personal relationships, internet activities, and music preference among college students. *Adolescence*, *39*(154), 189–201.

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). Dplyr: A grammar of data manipulation. *R package version 0.7.4*. Retrieved from https://CRAN.R-project.org/package=dplyr

Williams, M. J., Whitaker, R. M., & Allen, S. M. (2012). Measuring individual regularity in human visiting patterns. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 117–122. doi:10.1109/SocialCom-PASSAT.2012.93

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. doi:1089-778X(97)03422-X.

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. doi:10.18637/jss.v077.i01

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. doi:10.1016/j.jrp.2010.04.001

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

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4), 1036–1040. doi:10.1073/pnas.1418680112

Zabel, K. L., Christopher, A. N., Marek, P., Wieth, M. B., & Carlson, J. J. (2009). Mediational effects of sensation seeking on the age and financial risk-taking relationship. *Personality and Individual Differences*, *47*(8), 917–921. doi:10.1016/j.paid.2009.07.016

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69*(4), 548–565. doi:10.1177/0013164408324469

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. doi:1369âĂŞ7412/05/67301

Zuckerman, M. (1994). *Behavioral Expressions and Biosocial Bases of Sensation Seeking*. New York, USA, Melbourne, Australia: Cambridge University Press.

Zuckerman, M. (2002). Zuckerman-Kuhlman Personality Questionnaire (ZKPQ): an alternative five-factorial model. In B. de Raad & M. Perugini (Eds.), *Big five assessment* (pp. 377–396). Seattle, USA: Hogrefe & Huber Seattle.

Zuckerman, M., & Aluja, A. (2015). Measures of sensation seeking. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 352–380). doi:10.1016/B978-0-12-386915-9.00013-9

Zuckerman, M., Eysenck, S. B., & Eysenck, H. J. (1978). Sensation seeking in england and america: Cross-cultural, age, and sex comparisons. *Journal of Consulting and Clinical Psychology*, *46*(1), 139–149. doi:10.1037//0022-006X.46.1.139

Chapter 3

To Challenge the Morning Lark and the Night Owl: Using Smartphone Sensing Data to Investigate Day-Night Behavior Patterns

Schoedel, R., Pargent, F., Au, Q., Völkel, S. T., Schuwerk, T., Bühner, M., and Stachl, C. (2020) To Challenge the Morning Lark and the Night Owl: Using Smartphone Sensing Data to Investigate Day–Night Behaviour Patterns. In *European Journal of Personality*, https://doi.org/10.1002/per.2258.

3.1 Abstract

For decades, day-night patterns in behavior have been investigated by asking people about their sleep-wake timing, their diurnal activity patterns, and their sleep duration. We demonstrate that the increasing digitalization of lifestyle offers new possibilities for research to investigate day-night patterns and related traits with the help of behavioral data. Using smartphone sensing, we collected in-vivo data from 597 participants across several weeks and extracted behavioral day-night pattern indicators. Using this data, we explored three popular research topics. First, we focused on individual differences in day-night patterns by investigating whether "morning larks" and "night owls" manifest in smartphone-sensed behavioral indicators. Second, we examined whether personality traits are related to day-night patterns. Finally, exploring social jetlag, we investigated whether traits and work weekly day-night behaviors influence day-night patterns on weekends. Our findings highlight that behavioral data play an essential role in understanding daily routines and their relations to personality traits. We discuss how psychological research can integrate new behavioral approaches to study personality.

Keywords Day-Night Behavior Patterns, Smartphone Sensing Data, Personality, Chronotype, Diurnal Activity

3.2 Introduction

Are there times of day when you do not use your smartphone at all? Most likely at night. As our everyday companions, smartphones can provide much information about people's day-night patterns (Harari et al., 2016). So far, behavioral manifestations of the underlying circadian system like sleep-wake timing, diurnal activity, or sleep duration have mainly been assessed via self-reports (Adan et al., 2012). However, self-reports about behavior are known to differ from actual records of behavior (Baumeister, Vohs, & Funder, 2007; Gosling, John, Craik, & Robins, 1998). Emphasizing this dilemma, Lauderdale, Knutson, Yan, Liu, and Rathouz (2008) correlated behaviorally assessed sleep duration with self-reports and concluded that people systematically misjudge it. An alternative approach is to collect actigraphy-based data to study sleep behavior: Movements and environmental factors like ambient brightness are recorded with wristbands and are jointly converted to indicators for sleep-wake timing by special algorithms (e.g., Križan & Hisler, 2019; Tonetti et al., 2016; Vitale et al., 2015). Regarding the trade-off between measurement accuracy and ecological validity, another interesting complement for studying sleep behavior could be the use of smartphone sensing data. These data cannot provide a direct measurement of sleep-wake phases, but only periods of nightly inactivity of smartphone use in which physiological sleep occurs. In contrast to actigraphy, these measurements do not take body signals such as movements or pulse into account. However, first studies have indicated that smartphone data provide useful information about sleep-wake timing as smartphones are meanwhile considered to be part of new sleeping habits (Borger, Huber, & Ghosh, 2019; Chen et al., 2013; Lin et al., 2019; Min et al., 2014). Borger et al. (2019) have shown that indicators for sleep on- and offset derived via actigraphy and smartphone touch interactions are highly correlated. In addition, independence from sensors worn on the body also offers advantages in terms of ecological validity. With the help of commercially available smartphones, behavioral indicators for sleep-wake timing can be collected efficiently and unobtrusively in everyday life over a more extended period, even for large samples. To illustrate this, we use smartphone-sensed indicators for sleep-wake timing to investigate traits related to day-night patterns. For this purpose, we chose to study three frequently researched questions, which we will introduce in the following sections.

3.2.1 Individual Differences in Behavioral Day-Night Patterns

The human circadian system has been studied for decades by interdisciplinary research teams. The most prominent finding across all research disciplines is that individuals show stable differences in day-night patterns, a stable trait that is often referred to as the *chronotype* (e.g., Adan et al., 2012; Cavallera & Giudici, 2008; Roenneberg, Wirz-Justice, & Merrow, 2003). Literature frequently describes two extremes: The *morning type* ("morning lark") wakes up and goes to bed early, feels fit after getting up, and performs best early in the day. The *evening type* ("night owl") wakes up and goes to bed later, feels tired after waking up, and performs best towards the end of the day (for extensive reviews, see Adan et al., 2012; Cavallera & Giudici, 2008; Takano, Sakamoto, & Tanno, 2014). The chronotype has been argued to be a genetically predisposed trait with various biological manifestations like body temperature or hormone levels (Bailey & Heitkemper, 2001; Horne & Östberg, 1976; Katzenberg et al., 1998; Roenneberg et al., 2003). In addition, chronotype should be distinguished from sleep duration, which has been argued to be an independent trait (Roenneberg et al., 2007).

Based on the distinction between variable- and person-centered personality assessment (Asendorpf, 2003), one might assume that chrono"types" refer to distinct groups of individuals with similar manifestations in chronotype-related behaviors. However, Putilov (2017) points out in his review that researchers have not yet reached an agreement on the number and content of underlying dimensions, the resultant number of types, and whether the conceptualization as types makes sense at all (Roenneberg et al., 2003). Two different operationalizations of chronotype are most prominent in the literature (see Table 3.1).

Dating back to (Horne & Östberg, 1976), chronotype is described as *circadian* or *morningness-eveningness preferences*. The term "circadian typology" is often used synonymously and shows the emphasis on the categorization of chronotypes in this research tradition (e.g., Adan et al., 2012; Lipnevich et al., 2017). In comparison, Roenneberg et al. (2015) accentuate the chronotype as a continuous variable and describe it as a trait reflecting the *phase of entrainment*, which represents individual differences in the synchronization of the internal circadian rhythm to environmental factors (e.g., light/dark cycle, diurnal temperature curve, social interaction). Despite their different understanding of the underlying construct of chronotype (Roenneberg, 2015), both operationalizations have been found to be strongly correlated (Zavada, Gordijn, Beersma, Daan, & Roenneberg, 2005). In the present study, we take the structural ambiguity of chronotype as our starting point to investigate how smartphone sensing data reflecting day-night activity patterns

could help to inform chronotype research, as operationalized both in the Horne-Östberg- and in the Roenneberg-tradition.

In the Horne-Östberg-tradition, the Morningness-Eveningness Questionnaire (MEQ; Horne & Östberg, 1976) still represents the gold standard for chronotype assessment (Putilov, 2017). The MEQ asks for circadian preferences and categorizes people according to ad-hoc specified cut-off values (Horne & Östberg, 1976). In the development of the MEQ, neither the grouping nor the factorial structure was investigated. Cut-off values were determined using a small but not representative sample (Caci, Deschaux, Adan, & Natale, 2009). Meanwhile, various derivates and short scales of the MEQ have been published (Adan et al., 2012; Putilov, 2017). Assumptions on the underlying structure of circadian preferences range from a continuum with two extremes (Natale & Cicogna, 2002; Tonetti et al., 2016), over two dimensions (morningness and eveningness as separate dimensions; Lipnevich et al., 2017) to a multidimensional construct with up to four factors (Adan et al., 2012; Caci et al., 2009; Randler, Díaz-Morales, Rahafar, & Vollmer, 2016). Recently, Preckel et al. (2019) have published pioneering work on a typology of circadian preferences providing empirical evidence on the possible number of types. In an adolescent sample, they found evidence for four types resulting from the combination of the two independent dimensions of morningness and eveningness preference. Joining this search for structure, we translate the questionnaire items typically used to determine the Horne-Östberg chronotype into behavioral smartphone sensing equivalents. Smartphone usage variables can approximate many of them. Following Putilov's (2017) recommendation to consider behavioral markers for circadian preferences, we investigate whether we can find types of individuals with similar smartphone usage patterns indicating circadian preferences. Finally, we explore the factorial structure of the behavioral indicators.

In the Roenneberg-tradition, freely chosen sleep-wake timing is considered the best approximation of the internal circadian rhythm. Therefore, sleep-wake habits for both work and free days are assessed while controlling for alarm clock usage (Roenneberg et al., 2015; Roenneberg et al., 2003). In this taxonomy, the midpoint point between sleep on- and offset determines the chronotype. This reference point for sleep has proven to coincide with nocturnal melatonin production, which in turn controls sleep-wake timing (Roenneberg et al., 2015; Roenneberg et al., 2007; Roenneberg et al., 2003; Terman, Terman, Lo, & Cooper, 2001). In this context, the Munich Chronotype Questionnaire (MCTQ), which has been repeatedly validated by behavioral (actigraphy) and biological (melatonin, cortisol) circadian system markers, is primarily used (Roenneberg et al., 2007; Roenneberg et al.,

2003). Only recently, Lin et al. (2019) took up the idea to determine the Roenneberg chronotype by using smartphone sensing data and provided first indications that there is a considerable overlap between sleeping times assessed via smartphones and self-reports. However, their algorithm for characterizing a *digital chronotype* does not explicitly correspond to Roenneberg's chronotype criteria, as they did not differentiate between work and free days and were restricted to the use of a very limited range of data (screen and notification events; Lin et al., 2019). We propose a more fine-grained algorithm for determining a smartphone sensing-based proxy by using only free days without alarm clock usage. To explore our smartphone-chronotype, we look at descriptives and correlational analyses that were presented by Roenneberg's group to describe the MCTQ based chronotype. For example, Roenneberg et al. (2007) found that sleep duration depends on chronotype if analyzed separately for work and free days and that chronotype is related to age and gender.

Table 3.1: Description of the two most popular approaches to chronotype

Feature	Horne-Östberg chronotype	Roenneberg chronotype
Assessment	Morningness-Eveningness Questionnaire (MEQ) by Horne & Östberg (1976)	Munich Chronotype Questionnaire (MCTQ) by Roenneberg et al. (2003)
Chronotype as	time of day preferences	phase of entrainment
Items	ask for imagined free days: preferred sleeping times, preferred times for mental/physical activity, subjective feeling in the morning/evening, self-reported chronotype	ask for both free and work days: habitual sleeping times
Determination of chronotype	cut-off values classify participants according to their 19-items sum score	midpoint of sleep for free days without alarm clock usage
Emphasized structure	4 dimensions (peak time, morning affect, retiring, rising) according to Caci et al. (2009)	continuous variable

Note. The structure for the Horne-Östberg chronotype refers to the original chronotype assessment with the MEQ. However, several derivates of the MEQ have been developed and there is no consensus in research about the factorial structure of the chronotype approximated by the assessment of circadian preferences. Solutions range from 1 to 4 dimensions.

3.2.2 Behavioral Day-Night Patterns and Personality Traits

Important research questions are associations between day-night patterns, personality, and demographics. Different aspects of day-night behavior have been addressed in this context. For example, the *morningness preference* has been linked to personality. Higher values in this dimension indicate a preference for getting up and going to bed early, feeling fit in the morning, and achieving peak performance earlier in the day (Lipnevich et al., 2017). The most established findings in meta-analyses are that conscientiousness and agreeableness are positively related to morningness (Lipnevich et al., 2017; Tsaousis, 2010). No or only small relationships in a specified direction can be found for neuroticism and openness (Adan et al., 2012; Lipnevich et al., 2017; Tsaousis, 2010). Negative relationships between morningness and extraversion were found, but only if the trait extraversion was described with Eysenck's Three-Factor model (Adan et al., 2012; Tsaousis, 2010). Using the Five-Factor model, this association is almost zero (Tsaousis, 2010). For the sake of completeness, please note that morningness has also been found to be related to personality styles, or more precisely with thinking and behaving styles (Díaz-Morales, 2007). Furthermore, age has been robustly related to morningness. Shifts towards eveningness in adolescence and towards morningness with increasing age (at around 50) have been reported (e.g., Adan et al., 2012; Cavallera & Giudici, 2008). Regarding gender, a meta-analysis has found that the preference for morningness is slightly higher for females compared to males (Randler, 2007). However, complex interactions between age and gender have been reported in previous literature. For example, girls at the age of 13 and 14 have a lower tendency towards morningness than their male counterparts (Mateo, Diaz-Morales, Barreno, Prieto, & Randler, 2012), and their peak towards eveningness is earlier (e.g., Adan et al., 2012). In addition, Randler and Engelke (2019) have shown a complex interaction between age and gender with regard to morningness preferences: Young females were more and older females less morning-oriented than young or older males.

In addition, associations between *sleep duration* and personality traits have been investigated, but findings have been ambiguous so far. For example, there is some evidence that individuals with higher values in neuroticism report to sleep longer (Duggan, Friedman, McDevitt, & Mednick, 2014). According to Križan and Hisler (2019), neuroticism is not related to the mean sleep duration but positively related to the intra-individual variation in sleep duration. Some studies reported correlations between sleep duration and conscientiousness, agreeableness, or openness but not extraversion (Križan & Hisler, 2019; Randler, 2008). In contrast, other researchers did not find any evidence that sleep duration and big

five personality traits are associated (Gray & Watson, 2002; Randler, Schredl, & Göritz, 2017; Sutin, Gamaldo, Stephan, Strickhouser, & Terracciano, 2019). Sleep duration decreases with age (Randler, 2008) but was not found to be related to gender (Randler et al., 2017).

In summary, past research provides some evidence for associations between personality traits and day-night behavior, but past findings are inconsistent. One possible reason for this could be that the majority of studies (except Križan & Hisler, 2019; Sutin et al., 2019) asked participants about their habits but did not include any behavioral measures of sleep. Not only might people differ in their ability to estimate their sleep duration, personality traits themselves might play a role in the evaluation of their day-night behaviors. To circumvent this issue here, we use data from smartphone sensing to derive indicators for sleep-wake behavior and to consequently investigate their relationship with big five personality traits on factor and facet level. Additionally, we explore *sleep continuity*, which has been defined as a measure of how well people fall asleep and sleep through (Ohayon et al., 2017). Recent actigraphy-based research has found, for example, that conscientiousness and extraversion were negatively related to behavioral indicators of sleep continuity, such as wake after sleep onset. In contrast, higher scores in neuroticism were associated with more wakening (Sutin et al., 2019). As a rough smartphone-based approximation measure, we look at two aspects of sleep continuity: how often and for how long people check their smartphones during the night. Additionally, we analyze smartphone activity-logs to explore how *alarm clock usage* - particularly "snoozing" - is related to personality.

3.2.3 Intra- and Interindividual Differences in Day-Night Patterns: The Social Jetlag

Finally, we explore the so-called *social jetlag* hypothesis (e.g., Adan et al., 2012; Wittmann, Dinich, Merrow, & Roenneberg, 2006). Roenneberg et al. (2007) surveyed the sleep habits of more than 55,000 people using the MCTQ and found that sleep behavior differs for work-free versus workdays. Specifically, their findings suggest that people, on average, go to bed and awake earlier on work than on free days. Furthermore, the proportion of sleep on- and offset is smaller for workdays than for free days. It has been suggested that this effect is induced by social obligations (Wittmann et al., 2006). Thus, the pairing of late bedtimes with consistent wake-up times leads to a sleep deficit for a week. As a consequence, sleep is compensated on weekends (Roenneberg et al., 2015). This misalignment of the internal biological and the external social clock is associated with health risk be-

haviors (e.g., increased BMI and smoking; Roenneberg, Allebrandt, Merrow, & Vetter, 2012; Wittmann et al., 2006). According to Wittmann et al. (2006) and Roepke and Duffy (2010), late chronotypes are particularly affected by the social jetlag as they stay up until late at night but have to get up early to go to work or to pursue other social obligations on the following day. The assessment of individuals' daily routines through the analysis of smartphone activity-logs for several weeks allows us to investigate compensatory nightly rest by considering intra- and interindividual factors. Using these indicators, we want to explore whether the smartphone-sensed proxies for sleep duration on weekends and respective weeks are related and whether interindividual factors like the Roenneberg chronotype, demographics, and personality traits have an impact.

3.2.4 Rationale

Our study aims to re-investigate selected topics regarding day-night pattern related traits by using smartphone sensing data. Since we use a new type of data in this field of research, this is exploratory work. A handful of studies have started to use smartphone data in this context (e.g., Chen et al., 2013; Lin et al., 2019; Min et al., 2014). However, these studies have mostly been limited in terms of sample size and types of sensing data.

Here, we show how behavioral records from smartphones can be used to investigate individual differences in day-night patterns, how they relate to personality traits, and how they are influenced by intra- and interindividual factors. Besides the examination of whether "morning larks" and "night owls" manifest in indicators of sleep-wake timing and diurnal activity patterns, we explore the smartphone-based operationalization of the Roenneberg chronotype. We investigate the associations of day-night behavior patterns and personality traits. Finally, we illustrate how continuously logged behavioral data can be used to investigate the contribution of both intra- and interindividual factors to predict indicators for sleep behavior on weekends, using the social jetlag hypothesis as an example.

3.3 Method

Our analyses are based on data collected within the long-time project *PhoneStudy* (Stachl et al., 2018). This ongoing interdisciplinary research project at LMU Munich uses the continuously developed smartphone sensing application *PhoneStudy* for Android smartphones for collecting natural smartphone usage behaviors in the field. Data about app usage, calling activity, general phone usage (e.g., calendar, music, power supply), and connectivity (e.g., Bluetooth, WiFi) are logged whenever the respective events occur. GPS data is usually recorded once every 15 minutes. Data is synchronized hourly to the backend server via SSL-encryption, whenever a WiFi connection is available. The responsible institutional review board and data protection office approved the project and all associated studies. All materials and aggregated data can be found in our open science framework project (OSF; Schoedel et al., 2020)[1]. To protect the data privacy rights of our participants, the raw sensing data cannot be made available due to their granularity.

3.3.1 Description of Dataset

We combined data resulting from three studies conducted between 2014 and 2018. In Table 3.2 we show some basic information about the included studies. Despite some marginal differences, data collection procedures of all studies followed the same principle: After giving informed consent, participants were asked to install the *PhoneStudy* app for at least 30 days on their private smartphones and to complete several questionnaires before, during, or after the smartphone logging period. Participants were mostly recruited in the university context via flyers, mailings lists, social media, and personal contact in Munich, Germany. For more detailed information about study procedures, see also Harari et al. (2019), Schoedel et al. (2018), Schuwerk, Kaltefleiter, Au, Hoesl, and Stachl (2019), Stachl et al. (2019), Stachl et al. (2017).

We applied several exclusion criteria to our initial dataset of 743 participants. We excluded participants with fewer than 21 days of sensing data, more than 50% missing values across all variables, and if questionnaire data was not available. We included data from a maximum of 32 days of continuous logging. This resulted in a final sample size of 597 (61% females). As recruitment took place in the university context, participants were, on average well-educated (71% with a high school and 20% with a university degree). With a mean age of 23.56 years ($SD = 6.55$; $Min = 18$, $Max = 72$) the sample was skewed

[1] https://osf.io/a4h3b/

Table 3.2: Description of datasets used in the study

Dataset	References	*N*	Study Period	Compensation
1	Stachl et al. (2017), Harari et al. (2019)	132 (137)	09/2014 - 08/2015	individualized personality profile and 30 € or course credits
2	Schuwerk et al. (2019)	240 (245)	08/2016 - 08/2017	up to 35 € and lottery (smartphone or tablet worth 400 €)
3	Schoedel et al. (2018)	225 (361)	10/2017 - 01/2018	individualized personality profile and user activity feedback, course credits and lottery (10 x 50 €)

Note. N indicates the size of the sample of the respective study after application of our inclusion criteria. The total number of subjects per study is given in brackets.

towards younger participants (18-21: 39%; 22-25: 34%; 26-30: 12%; 31-40: 5%; 41 and older: 3%). For a more detailed description of the sample, according to studies, see Table 3.3.

3.3.2 Measures

Self-Report Measures

We administered various self-report questionnaires. However, we limit our report to the ones used in our statistical analyses. Besides demographics, personality traits were assessed with the Big Five Structure Inventory (BFSI Arendasy, 2009). Each of the big five factors openness, conscientiousness, extraversion, agreeableness, and emotional stability was measured on respectively six subscales (Table 3.8). Participants were asked to rate 300 personality describing adjectives and short phrases on a four-point Likert scale with the labels *untypical for me*, *rather untypical for me*, *rather typical for me*, and *typical for me*. Compared to the widely used structure inventory NEO-PI-R (Costa & McCrae, 2008), the BFSI is supposed to have better psychometric properties: Cronbach α values (ranging between 0.72 and 0.92) are partly higher, and subscales are unidimensional in the original paper (Arendasy, 2009). In addition, the BFSI should be less dependent on the participant's reading comprehension ability as it uses short and simple items (Arendasy, 2009). The construction of the BFSI does not follow the classical test theory, but the item response theory framework. Accordingly, the BFSI has been developed in conformity with the par-

Table 3.3: Description of the sample according to studies

Dataset	*N*	Age	Education	Students	Employment Status
1	132	23.61 (4.73)	no qualification: 0.00% secondary school: 3.79% high school: 65.15% university: 31.06%	no data available	no data available
2	240	22.94 (4.57)	no qualification: 0.00% secondary school: 9.58% high school: 72.50% university: 17.92%	73.50%	no data available
3	225	24.20 (8.86)	no qualification: 0.44% secondary school: 8.88% high school: 72.44% university: 16.44%	77.33%	unemployed: 4.89% in training: 24.89% minor employm.: 41.33% part-time: 10.67% full-time: 15.56% other: 0.88%

Note. *N* indicates the size of the samples according to studies. The column Age presents the mean value, and standard deviations are given in brackets. As procedures slightly varied across studies, not all demographic variables are available for all datasets. The category *other* in the column Employment Status comprises retraining and pension.

tial credit model (Masters, 1982) which is a probabilistic model describing an individual's observable score on a single item as the result of the functional relationship between the individual's latent trait value (person parameter) and latent item thresholds which indirectly determine item difficulty (item parameter; Arendasy, 2009). Correspondingly, we used the person parameter estimates as personality scores in all our analyses.

Day-Night Behavioral Measures

Raw smartphone sensing data are sequences of timestamped event data. Whenever a usage event happens, a data entry specified by several event characteristics (e.g., date, study day, details about the event like app package name or type of call) is created. To get an idea of the raw data structure, see also the supplemental codebook (Schoedel et al., 2020). To investigate the research questions specified above, we created variables by reviewing the literature and translating behavioral sleep indicators into smartphone sensing behaviors. Based on our smartphone sensing data, we computed proxy-variables to estimate sleep-related behaviors. Please note that our variables are likely to overestimate actual sleep as

the last smartphone usage event in the evening has to be before the physiological onset of sleep, and the first smartphone usage event in the morning occurs with delay after waking up. As smartphone sensing data are prone to logging errors, we extracted robust behavioral estimators when appropriate for the respective variable (Kafadar, 2003; Rousseeuw & Croux, 1993). To stay within the scope of this article, we only summarize our procedure and the engineered variables in the following sections. However, note that variable extraction is usually the most complex and time-consuming task in analyses of smartphone sensing data, and the process includes many researchers' degrees of freedom. For transparency, we provide all code in our OSF project, and the variable extraction procedure is described in detail in the supplemental codebook (Schoedel et al., 2020).

General Indicators for Sleep-related Behaviors We computed the following variables daily while distinguishing between days during the week versus the weekend (Roenneberg et al., 2007). Based on the algorithm specified in Table 3.7, we determined the *first and last events* according to individual study days and calculated mean and intra-individual variation variables. We defined the smartphone proxy for sleep duration, *nightly inactivity*, as the period between the last event of the day and the first event of the following day. To explore social jetlag, we calculated the average daily inactivity during the night for weekdays and weekends for all study weeks individually.

In addition, we translated two aspects of sleep continuity, sleep fragmentation and waking up after bed, into smartphone usage behavior by calculating the average number and duration of *checking events* at night. At this point, we would like to point out that our measures do not fully meet the definition of sleep fragmentation and wake after sleep onset by Ohayon et al. (2017). Hence, our measurements only give a rough estimate, taking into account the occurrence of very short smartphone checking events during the nightly inactivity period of smartphone use, which was not part of a more extended usage period in the evening and the next morning. Accordingly, we defined nightly checking events as short periods of less than two minutes of smartphone usage during otherwise nightly inactivity. Due to the lack of empirical data in the literature, we have set this threshold value considering that smartphone usage of fewer than two minutes might be caused by less significant actions such as checking the clock during the night.

Finally, we calculated some variables related to using the smartphone as an alarm clock: the *mean point of time of alarm app ringing*, and the mean daily number, and duration of *snoozing events* (snoozing was defined as the repetition of alarm app events in the morning).

Horne-Östberg Chronotype Variables To operationalize circadian preferences in terms of smartphone usage behavior, we computed variables following the items of the MEQ (see Table 3.1). We translated preferred sleeping times as *mean points of time of the first and the last smartphone usage event on weekends*, as weekends are likely to be organized freely. Following this assumption, we also specified preferred times for activity as diurnal smartphone activity patterns. In this context, we distinguished between different behavioral categories: social communication (social media/communication app usage, calls, and texting), entertainment (browser, gaming, music/video, and news app usage), and general smartphone usage (all active smartphone usage events). To take into account the distribution of usage events throughout the day, we computed the first quartile, the median, and the third quartile of usage events according to the behavioral categories for each day. In other words, we extracted timestamps that indicate when *25%, 50%, and 75% of the daily events of the respective usage category* took place. Then we computed the mean across all study days for each of the three quantiles. Finally, to depict the subjective feeling of sleepiness in the morning, we considered the *mean number and duration of snoozing events during the week* to indicate how readily people get up in the morning.

Roenneberg Chronotype Variables Similar to the assessment of the chronotype using the MCTQ, we calculated the *midpoint of sleep* (*MSF*) which is the mean half-way point in time between the last event of a day and the first event of the next day for free (weekend) days without alarm app usage. In addition, we determined the *corrected midpoint of sleep* (MSF_{corr}) which has been proposed by Roenneberg et al. (2007) to correct for the sleep-debt collected during the week. According to them the MSF_{corr} is better suited for estimating the true underlying chronotype.

3.3.3 Data Analysis

Clustering

In the following, we give a short overview of the applied methods. More detailed information can be found in the Appendix 3.8.1. To investigate whether participants can be assigned to groups of similar smartphone usage behaviors indicating circadian preferences, we used clustering as an unsupervised machine learning method. We applied the commonly used *k-means* clustering algorithm with the euclidean distance as proximity measure. Clustering aims to reduce complexity by finding meaningful structures within

the data. According to their similarity in a pre-defined set of variables, participants are clustered in within-homogeneous groups that are well-separated from participants of other clusters (Tan, Steinbach, & Kumar, 2006). However, one disadvantage of clustering algorithms is that they sometimes identify random and, therefore, non-replicable structures (Tan et al., 2006). In line with the literature, we address this problem by using a data-driven approach to determine the number of clusters (Tibshirani & Walther, 2005) and by evaluating the stability and validity of the identified clusters based on bootstrapped metrics (Hennig, 2007, 2008; Tan et al., 2006). We followed the recommendations of Hennig (2018) and used 100 bootstrap iterations. For evaluating cluster stability we considered the Jaccard coefficient (*JC*, indicates stability if values exceed 0.85) and the criteria of *recovery* and *dissolution* which count how often each cluster has been successfully recovered and dissolved across all bootstrap iterations (Hennig, 2007, 2008). For evaluating the internal validity of clusters we looked at metrics indicating how similiar participants within each cluster are (within-compact) and how different participants from different clusters are (between-separated): the ratio of average within- and between-cluster distances (*wb.ratio* Tan et al., 2006), the *silhouette* coefficient (Rousseeuw, 1987), and the *dunn* index (Dunn, 1974; Halkidi, Batistakis, & Vazirgiannis, 2001). Clusters are within-compact and between-separated if the ratio of distances is small, the silhouette index is close to 1, and the dunn index is high (Hennig, 2018; Tan et al., 2006). As the *k-means* algorithm cannot handle missing values, we used the multivariate imputation by chained equations technique and specified a random forest imputation model (MICE; van Buuren & Groothuis-Oudshoorn, 2011).

Exploratory Factor Analysis

To explore the factorial structure of our smartphone-based proxy for the Horne-Östberg chronotype, we conducted an exploratory factor analysis based on the averaged correlation matrix of the imputed datasets. We determined the number of factors using the *empirical* Kaiser criterion, which has been shown to perform well for short scales (Braeken & Van Assen, 2017).

Multilevel Modeling

Measures for nightly inactivity of smartphone usage were repeatedly measured across several study weeks. Considering the intra-individual data dependency, we used multilevel regression modeling with behavioral measures on a weekly basis reflecting level 1 variables

that were nested within individuals (level 2). Therefore, we specified a *random-intercept-random-slope model* predicting the mean nightly inactivity duration on weekends based on the mean nightly inactivity duration of the respective preceding workweek (level 1). The averaged nightly inactivity duration, the Roenneberg chronotype, the big five traits, age, and gender, were included as predictors on level 2.

Regarding data preprocessing, we were faced with the challenge of selecting one path from a series of plausible steps. To do justice to these many researcher degrees of freedom and to increase research transparency, we follow the suggestion of Steegen, Tuerlinckx, Gelman, and Vanpaemel (2016) and present a *multiverse analysis*: for each possible combination of plausible preprocessing steps, a "new" dataset is constructed, and the same multilevel model is estimated for each of those datasets. The multiverse analysis illustrates how much the results depend on the choice of specific preprocessing steps or vice versa, which results are robust across all preprocessing options (Simonsohn, Simmons, & Nelson, 2015; Steegen et al., 2016). Our preprocessing choices include the *coding of the weekend* (Friday to Sunday versus Friday to Monday), the selection of the *number of repeated measurements* (3 versus 4 weeks), the handling of *outliers* (median versus winsorization), and the handling of *missing values* (listwise deletion versus multiple imputation). A detailed description of the alternatives for each decision can be found in the supplemental method section in Appendix 3.8.1. Combining all described decisions resulted in 2 x 2 x 2 x 2 = 16 choice combinations (see left side in Figure 3.4).

We used the uncorrected version of the Roenneberg chronotype as a predictor, as we explicitly control for a nightly inactivity deficit in the multilevel model. Gender was dummy-coded (0 = male, 1 = female) and all continuous predictor variables were z-standardized based on the grand-mean. The level 1 predictor duration of nightly inactivity during the week was centered around the individual mean, which in turn was entered as level 2 predictor (Curran & Bauer, 2011). For a more detailed description of the equation of the multilevel model, we refer the interested reader to the supplemental method section in Appendix 3.8.1.

Statistical Software

Data preprocessing and analyses were conducted using R 3.5.0 (R Core Team, 2018). We used *packrat* (Ushey, McPherson, Cheng, Atkins, & Allaire, 2018) for package management. For extracting behavioral variables we mainly used the R packages *dplyr* (Wickham, François, Henry, & Müller, 2019) and *fxtract* (Au, 2019). Multiple imputation was done by

using the package *mice* (van Buuren & Groothuis-Oudshoorn, 2011). In addition, we used the following packages to conduct our main analyses: *fpc* for clustering (Hennig, 2018), *psych* for exploratory factor analysis (Revelle, 2018), *lme4* and *lmerTest* for multilevel modeling (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017). For data visualization, we applied *ggplot2* (Wickham, 2016) and *corrplot* (Wei & Simko, 2017) and created *raincloud plots* (Allen, Poggiali, Whitaker, Marshall, & Kievit, 2019). The complete list of used R packages can be found in our OSF project.

3.4 Results

3.4.1 Descriptives

We recorded a mean of 22,547 events (SD = 24,368) for each participant across the whole study period. Participants had on average smartphone records for 21 (SD = 1.57) weekdays and 8 (SD = 0.92) weekend days. The mean number of logs per study day was 765 (SD = 804.70). As can be seen in Table 3.4, the average time of first and last smartphone usage was later for weekends than weekdays, and the duration of nightly inactivity was about 20 minutes longer on weekends than on weekdays. However, the mean number and duration of checking events during the night were similar for weekends and weekdays. 91% of our participants used alarm clock apps in the morning, at 7.19 a.m. on average during the week and about thirty minutes later on weekends. Note that 38% of participants did not use alarm clock apps on any weekend during the entire study period. The number and duration of snoozing events were similar for weekdays and weekends. Descriptive statistics for big five personality traits can be found in Table 3.8 in the Appendix.

3.4.2 Individual Differences in Behavioral Day-Night Patterns

Person- and Variable-Centered Structure of the Horne-Östberg Chronotype

In a first step, we determined the number of clusters. Following the suggestions of Tibshirani and Walther (2005), we looked for solutions resulting in a prediction strength above 0.80. Doing so, in 49 out of 50 imputed datasets, the data-driven proposed number for clustering based on smartphone-proxies for circadian preferences was 1. However, decreasing the prediction strength criterion to a value of 0.75 yielded a 2-cluster solution for all imputed datasets. Although the recommended predictive power was slightly missed, we

Table 3.4: Descriptive statistics for day-night behavior patterns

Variable	Week		Weekend		
	Mean	SD	Mean	SD	Cohens's d [$CI_{95\%}$]
Mean First Event Week	7.89	1.31	8.96	1.30	0.82 [0.70, 0.93]
Mean Last Event Week	23.15	1.23	23.79	1.42	0.49 [0.37, 0.60]
Mean Duration Nightly Inactivity Week [h]	8.68	1.20	9.02	1.45	0.26 [0.14, 0.37]
Mean Number Checking Events Week	5.59	3.97	5.61	5.37	0.00 [-0.11, 0.12]
Mean Duration Checking Events Week [sec]	26.07	26.12	25.29	40.72	-0.02 [-0.14, 0.09]
Mean First Alarm Event Week	7.19	1.29	7.47	1.68	0.19 [0.06, 0.33]
Mean Number Snoozing Week	1.33	1.76	1.33	2.04	0.00 [-0.13, 0.13]
Mean Duration Snoozing Week [min]	23.26	23.61	23.89	34.26	0.02 [-0.11, 0.16]

Note. The coefficients for first and last events represent times of the day. The decimal places indicate the percentage of a full hour. For example, 7.89 means 7:53 a.m. or 23.15 means 11:09 p.m.

further investigated k-means clustering with $k = 2$. The averaged bootstrapped performance measures for the cluster-wise stability assessment show that each component of the 2-cluster solution turned out to be highly stable (cluster 1 $_{n = 296}$: $JC = 0.94$, *dissolved* = 0, *recovered* = 100; cluster 2 $_{n = 301}$: $JC = 0.93$, *dissolved* = 0, *recovered* = 100). However, the internal cluster validation coefficients indicated that the two clusters were poorly separable from each other and were not compact in themselves (*wb.ratio* = 0.73, *silhouette* = 0.25, *dunn* = 0.06). To get a better understanding of the identified structure in the daily smartphone usage timing, descriptive statistics of the variables that were considered for clustering are displayed in Table 3.5. On average, participants assigned to cluster 2 had later first and last smartphone usage events on weekends and the daily 25%, 50%, and 75% timestamps for general, social interaction, and entertainment usage events on weekends were on average about 2 hours later. The mean number of snoozing events was similar in both groups, but participants of cluster 1 on average snoozed approximately 3.5 minutes longer. As an external criterion, we considered the smartphone-based Roenneberg chronotype. The mean midpoint of sleep was $M = 3.90$ ($SD = 1.15$) for cluster 1 and $M = 5.19$ ($SD = 1.38$) for cluster 2.

Table 3.5: Descriptive statistics for smartphone usage indicating circadian preferences by clusters

Variable	Cluster 1		Cluster 2		
	Mean	SD	Mean	SD	Cohens's d [$CI_{95\%}$]
First/last events on weekends					
Mean time of the first event	8.35	1.16	9.58	1.13	1.07 [0.90, 1.25]
Mean time of the last event	23.09	1.18	24.51	1.29	1.15 [0.97, 1.32]
Mean on weekends daily timestamp of					
25% general usage	12.28	1.29	14.38	1.37	1.57 [1.39, 1.76]
50% general usage	15.34	1.30	17.62	1.20	1.82 [1.63, 2.01]
75% general usage	18.37	1.34	20.62	1.17	1.79 [1.60, 1.98]
25% social interaction usage	12.51	1.28	14.47	1.22	1.57 [1.38, 1.76]
50% social interaction usage	15.38	1.34	17.53	1.16	1.72 [1.53, 1.91]
75% social interaction usage	18.18	1.53	20.29	1.16	1.56 [1.37, 1.74]
25% entertainment usage	12.91	1.76	15.22	2.02	1.21 [1.03, 1.39]
50% entertainment usage	15.07	1.86	17.70	1.75	1.46 [1.27, 1.64]
75% entertainment usage	17.25	2.05	20.03	1.74	1.47 [1.28, 1.65]
Snoozing events on weekdays					
Mean number of snoozing events	1.31	1.88	1.35	1.65	0.02 [-0.15, 0.20]
Mean duration of snoozing events	21.53	22.01	24.91	24.97	0.14 [-0.03, 0.32]

Note. Except the snoozing variables, the coefficients represent times of the day and the corresponding standard deviations are given in hours. The decimal places indicate the percentage of a full hour. The mean daily timestamp of 25% general usage indicates that 25% of all activities on a given day had happend at this point in time. The mean number of snoozing events means the daily mean absolute frequency and the snoozing duration is in minutes.

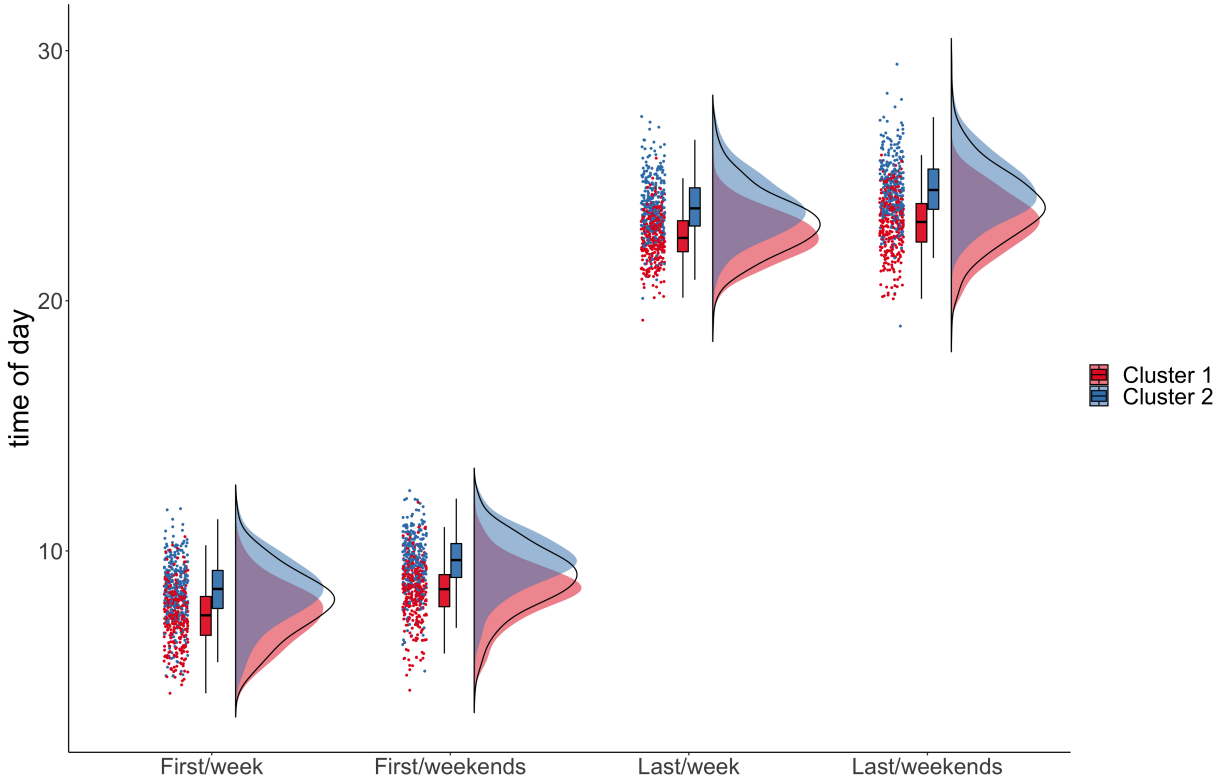

Figure 3.1: Plots displaying the distribution of mean daily first and last events on weekdays versus weekends by cluster. The black line shows the distribution based on the total sample. The ordinate axis goes beyond midnight, because last events after midnight were added to 24. An event at 26 therefore means it happened at 2.00 a.m.

To return to the question of whether we found different groups of individuals with similar smartphone usage patterns indicating circadian preferences, we refer to Table 3.5. Effect sizes for variables indicating sleep-wake timing are large, suggesting that participants assigned to cluster 2 have noticeable back-shifted diurnal smartphone usage patterns in comparison to participants assigned to cluster 1. Figure 3.1 shows, however, that the distributions of the two cluster groups overlap. A considerable proportion of participants could not be clearly assigned to one of the two clusters. Accordingly, the distribution based on the entire sample was not bi- but only unimodal.

In a second step, we also explored the factorial structure of the smartphone-based proxies for the Horne-Östberg chronotype. The empirical Kaiser criterion suggested a 3-factorial solution accounting for 62% of the variance. The obliquely (oblimin) rotated factor matrix is displayed in Table 3.6. Factor 1 explained 23% of the variance and was

Table 3.6: Exploratory factor analysis of the smartphone-sensed circadian preferences

Variable	F1	F2	F3	U
Mean time of the first event on weekends	0.09	0.06	**0.49**	0.67
Mean time of the last event on weekends	**0.54**	0.03	0.04	0.66
Mean daily timestamp of 25% general usage on weekends	0.05	0.10	**0.84**	0.16
Mean daily timestamp of 50% general usage on weekends	**0.44**	0.17	**0.47**	0.20
Mean daily timestamp of 75% general usage on weekends	**0.74**	0.14	0.11	0.23
Mean daily timestamp of 25% social interaction usage on weekends	0.26	-0.01	**0.68**	0.30
Mean daily timestamp of 50% social interaction usage on weekends	**0.63**	-0.01	**0.38**	0.22
Mean daily timestamp of 75% social interaction usage on weekends	**0.88**	0.03	0.03	0.19
Mean daily timestamp of 25% entertainment usage on weekends	-0.20	**0.77**	**0.33**	0.24
Mean daily timestamp of 50% entertainment usage on weekends	0.02	**0.99**	0.00	0.01
Mean daily timestamp of 75% entertainment usage on weekends	**0.34**	**0.77**	-0.17	0.21
Mean daily number of snoozing events on weekdays	0.26	0.01	-0.24	0.94
Mean daily duration of snoozing events on weekdays	0.27	0.00	-0.19	0.94
F2	0.46	1.00		
F3	0.52	0.47	1.00	

Note. Maximum likelihood factor analysis, obliquely rotated (oblimin) with 3 factors. Loadings greater than the amount of 0.30 are in bold. The correlations between the factors are displayed at the bottom of the table. F1 = Factor 1; F2 = Factor 2, F3 = Factor 3; U = Uniqueness.

comprised of behavioral indicators describing markers for later diurnal smartphone usage. In contrast, the behavioral variables loading high on factor 3 (19% variance explanation) described markers characteristic for early diurnal smartphone usage. The 50%-timestamps for daily (general and social interaction) smartphone usage considerably loaded on both, factor 1 and 3. Finally, factor 2 explained 20% of the variance and reflected behavioral indicators of smartphone usage for entertainment purposes independent of the time of the day. The two snoozing items did not load considerably on any factor. All factors were correlated (see Table 3.6).

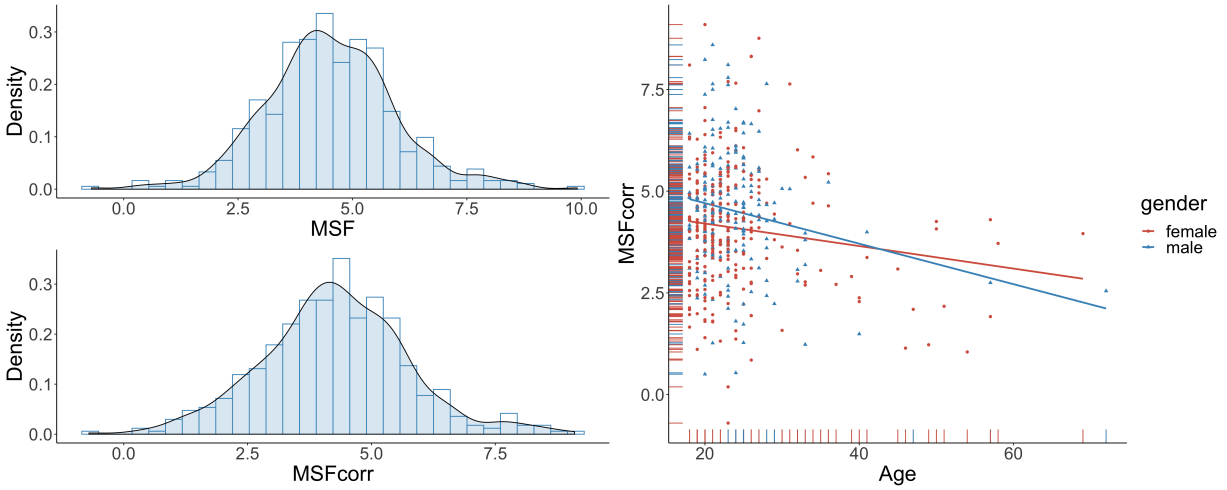

Figure 3.2: Plots displaying the distributions of the local time of the midpoint of sleep (MSF) and its sleep-debt corrected version (MSF_{corr}) and its relationship with age divided by gender.

The Roenneberg Chronotype and its Correlates

The smartphone-based midpoint of sleep (*MSF*) and the sleep-debt corrected version MSF_{corr} which both indicate the Roenneberg chronotype were approximately unimodally symmetrically distributed (see Figure 3.2). As no weekends without alarm clock usage were available for some participants, their *MSF* could not be computed. Therefore, the following results are based on a subsample of $n = 497$ participants. On average, the mean *MSF* was at 4.52 a.m. ($SD = 1.42$) and the MSF_{corr} slightly earlier at 4.26 a.m. ($SD = 1.47$). The *MSF* and MSF_{corr} ranged between 0.75 p.m. and 9.91 a.m. The *MSF* was weakly negatively related to nightly inactivity duration during the weeks (r = -0.13, $CI_{95\%}$ [-0.21, -0.04]) as well as the weekends (r = -0.11, $CI_{95\%}$ [-0.20, -0.03]). As suggested by Roenneberg et al. (2007), we used the MSF_{corr} for investigating the relationship of chronotype and demographics. Age (r = -0.16, $CI_{95\%}$ [-0.24, -0.07]) and gender (r = -0.15, $CI_{95\%}$ [-0.23, -0.06]) were both negatively related to the corrected midpoint of sleep, indicating that older and female participants had on average earlier chronotype values. However, the age-correlation should be interpreted with caution, as the plot on the right side of Figure 3.2 indicates that it was probably caused by data points of older participants of whom we only had few in the sample ($Q_3 = 25$). The correlation disappears (r_s = -0.03, $CI_{95\%}$ [-0.12, 0.06]), when computing the spearman correlation which is only based on ranks.

3.4.3 Day-Night Behaviors and Personality Traits

Since our analysis of relationships between behavioral day-night patterns and personality is exploratory, we do not perform any hypothesis tests, nor do we speculate about correlations on a variable-by-variable basis. Instead, based on the correlation plot displayed in Figure 3.3, we want to show the general result pattern and address some conspicuities. Overall, spearman correlations ranged between r_s = -0.24 (mean time of last events during the week and sense of duty) and r_s = 0.15 (mean time of the first event on weekends and carefreeness). As can be seen in Figure 3.3, the most striking aspect is that conscientiousness and its facets (except competence) were related to various day-night behaviors. First, more conscientious people on average had earlier mean and less varying daily points of time of first and last smartphone usage events both during weeks and on weekends. Furthermore, their duration of nightly inactivity varied less on weekdays and they had lower values on the Roenneberg chronotype. Finally, individuals with higher values on the facet sense of duty snoozed on average less often and shorter on weekdays.

Further but less coherent patterns in Figure 3.3 can be seen for openness, extraversion, and emotional stability. For example, openness to imagination showed some positive relations to day-night behavioral indicators. Openness to value and norm system was associated positively with the mean number and duration of snoozing events, especially on weekdays. Higher extraversion was related to longer smartphone checking events during nights on weekdays. Furthermore, carefreeness as a facet of emotional stability was associated positively with later day-night activity patterns. Regarding demographics, females' first use on weekends and general last use was on average earlier. Accordingly, they also had lower Roenneberg chronotype values. However, no correlations of considerable size were found for age.

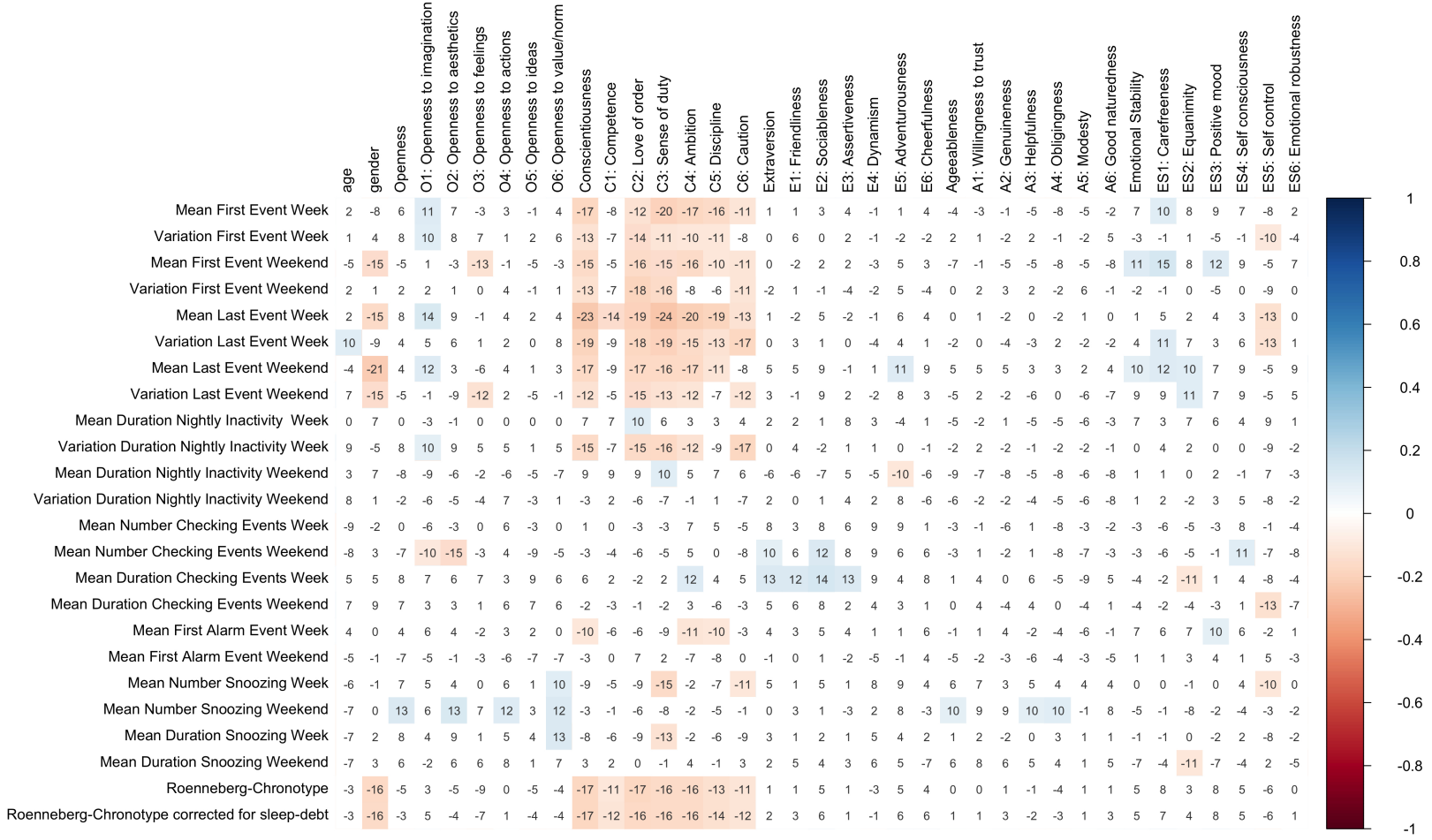

Figure 3.3: Pairwise complete spearman correlations between smartphone-sensed day-night activities for weekdays versus weekends and personality traits. Males were coded as 0. As not all participants used alarm clock apps, the sample size for respective correlations was reduced ($n_{week} = 506$, $n_{weekend} = 371$). The color of the squares indicates the direction and the strength of the respective correlations. For better readability correlations are presented as percentage (e.g., a value of 3 means $r_s = 0.03$). Additionally, only correlations with greater absolute values than 0.10 are highlighted in color.

3.4.4 Using Multilevel Modeling to Explore Social Jetlag

To investigate social jetlag, we explored compensatory sleep on weekends approximated as nightly inactivity duration by multilevel modeling. The duration of nightly inactivity on weekends was predicted by the duration of nightly inactivity during the week and the inter-individual variables Roenneberg chronotype, big five personality traits, age, gender, and the averaged individual mean duration of nightly inactivity. The results are presented in the twelve panels in Figure 3.4, which show the estimates and their 95% confidence intervals across all multiverse datasets for each predictor in the model. Some aspects were evident across all datasets. There were no relationships between the nightly inactivity duration on weekends and the variables Roenneberg chronotype, openness, extraversion, agreeableness, emotional stability, and the interaction between the Roenneberg chronotype and the nightly weekday inactivity. Second, the averaged nightly inactivity duration across the study weeks (level 2) was positively associated with the nightly inactivity period on weekends. Nevertheless, estimates for the individual nightly inactivity duration on weekdays (level 1) and conscientiousness, age, and gender (all level 2) varied across the multiverse datasets. Depending on the preprocessing steps, individuals with longer nightly inactivity duration on weekdays in the corresponding week, higher conscientiousness, higher age, and male gender had, on average, longer nightly inactivity periods on weekends.

As can be seen in Figure 3.4, some patterns can be identified in the multiverse results across different variables: The coding of the weekend seemed to have an influence. In conditions in which the weekend was coded as nights between Friday and Monday, the mean duration of nightly inactivity on weekends was, on average lower compared to the conditions in which weekends were coded as nights between Friday and Sunday. Also, for gender, a pattern can be determined depending on the coding of the weekend. For conscientiousness, estimates in conditions including 3 weeks were, on average higher than conditions comprising 4 weeks. Regarding the average duration of nightly inactivity during the week (level 2), estimates were higher when winsorized and imputed.

To get a better understanding of the results concerning social jetlag, we calculated an additional multiverse analysis. For this purpose, we considered a variant of the multilevel model without personality traits and demographics as covariates. As results did not considerably differ and not to go beyond the scope of this paper, they can be found as a supplementary analysis in our OSF project.

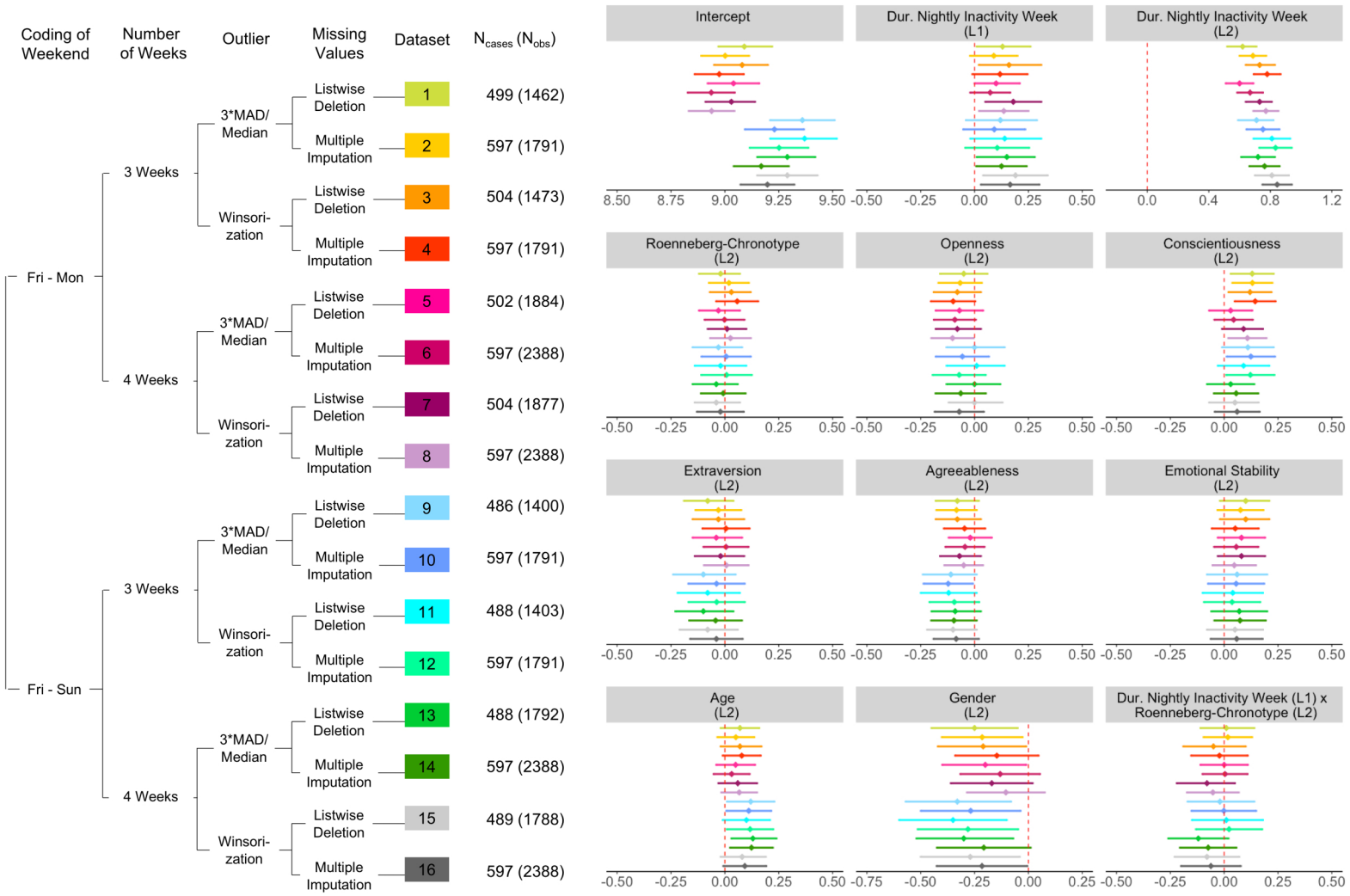

Figure 3.4: The decision tree on the left side shows how the multiverse of 16 datasets was created. The twelve panels on the right display the estimates and their 95% confidence intervals for the intercept and each predictor, resulting from multilevel modeling across the multiverse of 16 datasets. L1 = level 1 predictors (z-standardized and person-mean-centered); L2 = level 2 predictors (z-standardized, except gender). Males were coded as 0. The individual mean of the level 1 predictor was additionally entered as level 2 predictor. Each dataset and the corresponding model are coded the same color.

3.5 Discussion

We investigated three prominent research questions related to common behavioral day-night patterns by using smartphone sensing data. First, we focused on individual differences in day-night activity patterns. Based on behavioral indicators of circadian preferences, we explored the structure underlying our smartphone-proxy for the Horne-Östberg chronotype. Regarding the search for a smartphone-chronotype, we found non-discrete groups of individuals with similar diurnal smartphone usage patterns. In addition, our smartphone-based proxy for the Horne-Östberg chronotype turned out to be a multidimensional construct. In addition, we presented an algorithm for computing the chronotype as defined by Roenneberg et al. (2003). We used smartphone-based indicators for the midpoint of sleep and found associations with age, gender, and duration of nightly inactivity. Regarding personality traits, we found associations of conscientiousness with smartphone-sensed indicators for day-night behavior. Finally, we explored social jetlag by examining whether people were inactive longer during weekend nights if they accumulated a deficit of nightly inactivity during the preceding workweek while controlling for individual differences. Our findings suggest that nightly inactivity duration on weekends was mainly related to individuals' general level of nightly inactivity across all study weeks. We will critically discuss our results in the following sections. Since our research was explorative, explanations drawn post-hoc should not be easily generalized but be confirmed by preregistered hypotheses testing in future studies.

3.5.1 Smartphone Sensing in the Context of Behavioral Day-Night Patterns

Individual Differences in Day-Night Activity Patterns

In contrast to previous research based on self-reports, we used smartphone-sensed behavioral data to investigate the structure of chronotype and to inform both the variable- and the person-centered approach to chronotype. Emphasizing chronotype as a continuous dimension reflecting circadian habits, Roenneberg et al. (2003) have suggested computing the midpoint of sleep. Instead of assessing these habits by questionnaires (e.g., Roenneberg et al., 2007; Roenneberg et al., 2003), we followed Lin et al. (2019) and used smartphone sensing data to determine a smartphone-equivalent for the Roenneberg chronotype. We compared our resulting measure to the findings reported by Roenneberg et al. (2007) and

found similar descriptive parameters (distribution, mean) and associations with external criteria like gender and sleep duration during the week. In accordance with our assumption that smartphone-based sleep-wake timing indicators overestimate sleep times, the range of values was slightly larger for our measure. Regarding age and chronotype, we found a negative correlation, which was caused by a few older participants with lower chronotype values. However, since the age composition of our sample was highly skewed towards younger participants, we do not want to over-interpret this finding. A non-matching result was that whereas Roenneberg et al. (2007) found a positive correlation between chronotype and sleep duration on weekends, we found a negative association. Roenneberg et al. (2007) argued that later chronotypes sleep longer on weekends because they collect a sleep-debt during the workweek. In contrast to the large representative sample of their epidemiological study, our sample consisted mainly of students who are more likely to have fewer social obligations during the week than people who have a nine to five job. Accordingly, compared to nights during the week, our participants' nightly inactivity (indicating sleep) did not differ considerably on weekends. Therefore, one interpretation of our results could be that students have the opportunity to be more flexible in their daily routines during the week following their chronotype. Therefore late chronotypes do not need disproportionately more sleep on weekends. Accordingly, previous studies have shown that many students report napping after lunch during the week (Vela-Bueno et al., 2008). These naps could serve to use both the weekend and the week for sleep compensation (Gradisar, Wright, Robinson, Pain, & Gamble, 2008). In line with our interpretation, students with late chronotypes have been found to nap more extensively than students with early chronotypes (Zimmermann, 2011). Please note that this is only our post-hoc interpretation and further confirmatory research using behavioral data to study the interplay of sleep duration, chronotype, and work schedules.

Keeping the focus on variable-centered trait assessment (Asendorpf, 2003), but following the Horne-Östberg tradition, we operationalized circadian preferences as diurnal smartphone usage behaviors and explored the underlying factorial structure. We found three correlated dimensions reflecting early use of the smartphone during the day, late use of the smartphone during the day, and entertainment usage. In comparison, findings of previous studies investigating the structure of self-reported chronotype have resulted in one to four factors (e.g., Caci et al., 2009; Lipnevich et al., 2017; Natale & Cicogna, 2002). In their recent meta-analysis, Lipnevich et al. (2017) concluded that the preferences for morningness versus eveningness are not the extreme poles of one dimension but two

interdependent dimensions. Accordingly, our two correlated dimensions reflecting early and late diurnal smartphone usage activity align with their findings. Regarding our factor entertainment usage, we think that this could be regarded as a methodological artifact, as the content entertainment might have overlaid the diurnal character of the respective behavioral circadian indicators.

Dimensional approaches to personality, such as the two described above, offer the advantage to focus on individual differences. However, in contrast to person-centered approaches, they are not able to describe the structure of traits within persons (e.g., Asendorpf, 2003; Asendorpf & van Aken, 1999). In addition, types might have an advantage for applied purposes as the classification as "morning larks" or "night owls" is widely anchored in the popular science literature and scientific research. Therefore, besides examining dimensionality, we also explored the existence of types of individuals with similar diurnal smartphone usage patterns by using unsupervised machine learning. We found two groups that showed earlier versus later smartphone usage over the day. As the effect sizes show, these two groups considerably differed in indicators of diurnal smartphone usage patterns. However, our results also indicate that despite the high average group differences, a large number of participants could not easily be assigned to one of these two groups which overlapped considerably in the behavioral indicators used. Therefore, we asked ourselves whether we should call the structure we found types. In previous chronotype literature, types had often been considered as empirically validated, if the resulting groups subsequently proved to be different concerning external criteria (e.g., body temperature, EEG recordings; Horne & Östberg, 1976; Putilov, Donskaya, & Verevkin, 2015). In contrast, we did not determine any cut-off values but searched for non-random structures in the data. Only recently, Preckel et al. (2019) followed a similar approach identifying four chronotypes in an adolescent sample. However, since circadian preferences change with age (Roenneberg et al., 2007), and our sample was older, and we focused on smartphone-sensed rather than self-reported circadian habits, we argue that the results are not fully comparable.

From a statistical point of view, the existence of types is only justified if underlying variables are multimodally distributed (Fleiss, Lawlor, Platman, & Fieve, 1971; Hicks, 1984), which was not the case for our behavioral day-night indicators. However, previous research in the social sciences has revealed that non-overlapping types hardly exist for human behaviors (Costa Jr, Herbst, McCrae, Samuels, & Ozer, 2002; Meehl, 2004). Accordingly, Asendorpf and van Aken (1999) distinguish between discrete and non-discrete types in the context of personality research. Thus, the criteria for defining types are not uniformly

defined and applied in the literature. Our results are in line with this argument. Even if there were discrete underlying chronotype groups, it is unlikely that they would appear so clearly in everyday behavioral indicators due to social obligations and societal demands. Nevertheless, the identified non-discrete groups in our study can be a good starting point towards a smartphone-based behavioral proxy of chronotype operationalized as circadian preferences. Future research should replicate the structure in diurnal smartphone-usage indicators across different samples and use external validity criteria.

Conscientiousness and Differences in Behavioral Day-Night Patterns

In contrast to the majority of previous studies, we used behavioral markers for day-night activity patterns to investigate associations with personality traits and demographics. In line with past studies showing females' preference for morningness (Randler, 2007), women in our study were earlier in the day, and their day-night activity timing varied less. Besides, our results were consistent with previous research showing a majorly coherent pattern of day-night activity and conscientiousness (Adan et al., 2012; Lipnevich et al., 2017) but less clear relations for other big five personality traits (e.g., Gray & Watson, 2002; Randler et al., 2017). Precisely, highly conscientious participants on average showed lower and less varying sleep-wake timing indicators and lower Roenneberg chronotype values. Following questionnaire-based research (Adan et al., 2012; Križan & Hisler, 2019; Lipnevich et al., 2017; Tsaousis, 2010), our results indicate that more conscientious people on average are active earlier during the day and have longer nightly rest periods on weekends. Compared to findings from a meta-analysis ($r = .33$ according to Tsaousis, 2010), our correlations were smaller. However, our findings show that more conscientious people, who describe themselves as dutiful, ambitious, and disciplined (Arendasy, 2009), also act accordingly in everyday life (e.g., getting up early in the morning, longer nightly rest on weekends). Accordingly, Spears, Montgomery-Downs, Steinman, Duggan, and Turiano (2019) found in a recent longitudinal study, that conscientiousness was associated with mortality risk after ten years and that this association was mediated by sleep duration as an everyday expression of behavior.

In contrast to previous findings, conscientiousness, and emotional stability were not related to indicators for sleep continuity, but extraversion was (Križan & Hisler, 2019; Sella, Carbone, Toffalini, & Borella, 2020; Sutin et al., 2019). These recent studies measured sleep continuity using actigraphy and therefore used completely different operationalizations of the related indicators sleep fragmentation and wake up after bed (Križan & Hisler,

2019; Sella et al., 2020; Sutin et al., 2019). For example, Sella et al. (2020) defined sleep fragmentation as the number of awakenings exceeding a certain duration. In contrast to actigraphy, smartphone sensing does not provide continuous measurement of wakefulness but approximates this measure via active smartphone usage. This requires the determination of a specific threshold value to classify smartphone usage either as part of a continuous usage phase belonging to the last or first event of the day or as a short usage event during the period of otherwise nightly inactivity. Determining a threshold value according to this principle, our approach has two significant drawbacks. First, using two minutes as a threshold was a subjective decision due to the lack of empirical data from previous literature. Second, the derived variable checking duration is restricted in its variance by a maximum value of two minutes. Consequently, individual differences in the actual wake after sleep onset might be masked by our smartphone-based operationalization, which in turn could explain the differences in findings compared to actigraphy.

In addition, we did not find some of the relationships which have previously been reported. For example, in our data, we did not find associations between a preference for morningness and agreeableness (Adan et al., 2012; Tsaousis, 2010), or age (Adan et al., 2012). As already discussed in the previous section, our results regarding age should be interpreted with caution due to the restricted variability of age in our sample. Overall, the differing findings could result from the usage of actual behavioral variables in contrast to self-reported preferences in most previous studies. Additionally, differences with past studies might not be surprising considering that previous questionnaire-based research is not clear either (e.g., Duggan et al., 2014; Gray & Watson, 2002). Besides, to the best of our knowledge, we have been the first to explore differences in alarm clock app usage. Our results provide first indications about the relation of snoozing behavior and personality facets (sense of duty and openness to value and norm system). They should be further investigated in future research.

Individual Differences in Compensatory Nightly Inactivity on Weekends

To explore social jetlag, we investigated which intra- and inter-individual factors predict the duration of nightly inactivity of smartphone usage (assumed to indicate sleep duration) on weekends. To explore this research question and to get an impression of the robustness of our estimates, we created a multiverse of 16 datasets resulting from combining different choices of plausible preprocessing steps. In the following, we focus only on those aspects that have been demonstrated across all datasets. Individuals who had higher overall levels

of smartphone inactivity during nights on weekdays were also inactive longer on weekend nights. Even though our inactivity measure is not identical to sleep, our results indicate that individuals differ in their nightly rest duration. These findings support the notion that sleep duration is an independent trait (Ferrara & De Gennaro, 2001; Roenneberg et al., 2007). In contrast to the assumptions of social jetlag (Roenneberg et al., 2015; Wittmann et al., 2006), we neither found compensatory nightly inactivity on weekends nor any impact of the Roenneberg chronotype. As already discussed in the section above, our sample was highly skewed towards students. Thus, maybe their social obligations during the week are less pronounced, and therefore, we could not find their need for compensatory sleep on weekends. In addition, previous studies often used self-reports to investigate social jetlag (e.g., Roenneberg et al., 2012; Wittmann et al., 2006). Even though participants are instructed to indicate their habits for the last four weeks (Roenneberg et al., 2003), their answers might be biased towards a more general judgment of sleep-wake timing or influenced by short-term experiences like the sleep behavior of the previous night. In contrast, we looked at behavioral snippets of three or four concrete weeks.

Finally, our multiverse analysis showed that the results depend on the selected preprocessing steps. Especially for the predictors age, gender, and conscientiousness, the size of the estimates differed depending on the constructed datasets. Our study, therefore, points to two problems. First, for behavioral indicators extracted from smartphone sensing data, the definition of the weekend and the number of weeks included made a difference to the results. Future research in the field of smartphone sensing should, therefore, carefully explore and report whether decisions made in the preprocessing have an impact on the results. Second, our study highlights the issue of selective reporting in research articles (Simonsohn et al., 2015; Steegen et al., 2016). We could just as well have reported only one of the paths and the results of the corresponding model, and the choice of each path would have been equally plausible. However, depending on the preprocessing decisions, we might or might not have emphasized the effect of conscientiousness or gender or age at this point. In line with Simonsohn et al. (2015) and Steegen et al. (2016), we argue that decisions that might affect the results should be made transparent.

3.5.2 Limitations and Outlook

Our study exemplifies the usage of smartphone sensing data in the research field of behavioral day-night patterns. Strictly speaking, the assessment of day-night structures in everyday life and, therefore, sleep-wake phases would require the collection of EEG data

(Shambroom, Fábregas, & Johnstone, 2012). For reasons of efficiency, self-report questionnaires have so far been used to approximate sleep-related behaviors. We propose smartphone sensing as an alternative to collect proxies for these behaviors. However, our approach has some limitations.

First, similar to questionnaires (Lauderdale et al., 2008), our behavioral markers are only proxies for actual sleep-wake timing. In our dataset, only app-, phone-, screen-, and notification-events were available to determine the nightly inactivity period. Thus, actual sleep times were estimated based on active smartphone usage behaviors. However, for improving the accuracy of smartphone-based sleep-wake indicators, it would be helpful to include sensor data that do not require active usage, for example, brightness and ambient noise (Min et al., 2014). An even better estimate of sleep could be obtained by integrating the idea of actigraphy into the smartphone sensing approach. Meanwhile, many commercial wearables, which can also be used conveniently during bedtime, offer an open interface to integrate motion and physiological data like heart rate variability or galvanic skin response into research apps used for smartphone sensing.

Second, we defined new behavioral variables, which we extracted from smartphone sensing data. Although we derived our variables from previous literature, we had many degrees of freedom. Which period is defined as a weekend? What does active smartphone usage mean? How can daily values be aggregated? - These questions are only a few examples for the vast amount of decisions we had to make during data preprocessing. To make this process as transparent as possible, we provide an extensive codebook and analyze a multiverse of datasets where appropriate. However, the researcher community should develop a common standard for sensing data so that the results obtained do not depend on the respective data preprocessing decisions in individual studies.

One further limitation of our study was the skewed sample. In comparison to previous epidemiological studies, it was skewed in terms of age and occupation. As age and work schedules are related to sleep-wake timings (Adan et al., 2012), future studies using smartphone sensing data should use more representative samples.

Finally, in our study, we only focused on smartphone sensing data. Although resulting indicators cannot be equated one-to-one with physiological sleep, smartphone sensing can nevertheless unobtrusively collect data in the field over a long period. This is very beneficial as far as day-night habits are investigated. However, in research focusing on constructs like sleep quality (Križan & Hisler, 2019), it is essential to measure a possible mismatch between behavioral sleep indicators in contrast to individual perceptions and

feelings about sleep-wake timings. Consequently, the integration of the experience sampling method (e.g., Takano et al., 2014) could help to gain further interesting insights in individual differences into behavioral day-night patterns. Future studies could additionally benefit from combining actigraphy and smartphone sensing. Both methods assess actual behavior but highlight different aspects of day-night activity patterns (Borger et al., 2019). In summary, we do not want to discuss whether self-reports, smartphone sensing, or actigraphy are better suitable for depicting actual behavioral day-night patterns. We think that all data collection approaches have their place and could be very fruitfully combined to gain better insights into human day-night behavior patterns.

3.6 Conclusion

We used smartphone sensing data to extract behavioral variables usually assessed by self-reports in the context of day-night behaviors. Our study contributes to gain new insights into traits related to day-night behavior patterns. First, we investigated two prominent operationalizations of chronotype: Based on indicators for sleep-wake timing and diurnal activity, we found two overlapping groups of smartphone-based "morning larks" and "night owls" and two correlated dimensions that were similar to previously reported questionnaire-based factors. By computing a smartphone-based proxy, we presented a smartphone-sensed measure for the Roenneberg chronotype. Second, conscientiousness was related to earlier day schedules. In addition, we found individuals to differ in their overall level of nightly rest. We argue that it is important to understand individual differences in behavioral day-night patterns, as they previously have been found to be related to individuals' well-being and health. This work demonstrates that smartphone sensing provides an efficient and ecologically valid tool that can help to foster this understanding.

3.7 Acknowledgments

We want to thank all PhoneStudy team members for making this research possible. Special thanks go to Theresa Ullmann, Michelle Oldemeier, Florian Bemmann, Daniel Buschek, Florian Lehmann, Daniela Becker, and Peter Ehrich. In addition, we thank all involved students for supporting us by recruiting participants. Finally, we want to thank David Goretzko and Caroline Zygar-Hoffmann for their helpful advice.

3.8 Appendix

3.8.1 Supplemental Method

Measures

Table 3.7: Description of the algorithm for detecting nightly inactivity

Step	Description
1	Exclude passive smartphone events (GPS logs, notifications, and related screen events)
2	Exclude active usage events lasting shorter than two minutes and label them as checking behavior
3	Search for the maximum distance between consecutive events
4	Label the starting point of the maximum distance as last event of the day and the end point as first event of the next day

Note. To avoid longer periods of inactivity being detected during the day, the time frame for maximum distance detection was limited to 6.00 pm to 2.00 pm of the following day. We defined and filtered checking behavior, because we wanted to exclude less significant actions like checking the clock or notification texts.

Clustering

K-Means Algorithm For clustering, we used the *k-means* algorithm, which is one of the most frequently used algorithms for clustering (Tan et al., 2006). In the following section, we only describe the basic principles behind *k-means* clustering and refer the interested reader to Tan et al. (2006) for a detailed explanation. After the user has defined the expected number of clusters k, k points in the sample data are randomly determined and represent initial centroids. In a second step, all remaining data points are assigned to the centroid for which the euclidean distance is lowest. Afterward, the centroids in each of the k clusters are updated by calculating the arithmetic mean of all points in the respective clusters. Step-by-step the procedures are repeated as long as the centroids do not change anymore, which indicates that the grouping structure in the data has been identified. As the centroid represents the data points within the clusters, *k-means* clustering is also often referred to as prototype-based or partitional clustering (Tan et al., 2006).

Evaluation Metrics To ensure cluster validity, we took several steps to find non-random structures in our data. The first step is to determine the appropriate number of clusters. Tibshirani and Walther (2005) proposed to re-frame clustering as a supervised prediction problem by splitting the data into a training and a test set and estimating the number of pairwise cases that are assigned to the same cluster in the test set based on centroids of the training set. The associated prediction strength measure defined by Tibshirani and Walther (2005) can be used to determine an optimal number of clusters. Another important aspect is cluster stability (Hennig, 2007). If clusters disappear when data is slightly modified, they are not regarded as stable and consequently might reflect only random structure. Hennig (2007), therefore, suggests bootstrapping the data and considering the Jaccard coefficient (JC) for each cluster separately. The JC gives the proportion of data points (participants) that are assigned to the same cluster across the bootstrapped iterations, thus expressing the similarity of cluster solutions across bootstrapped datasets on a cluster-wise basis (Hennig, 2007). Further descriptive measures of cluster stability are the criteria of *recovery* and *dissolution* which count how often each cluster has been successfully recovered and dissolved across all bootstrap iterations (Hennig, 2007, 2008). As recommended by Hennig (2018), we used 100 bootstrap replications and interpreted clusters as stable if the JC exceeded values above 0.85.

Imputation of Missing Values Based on a variable-by-variable procedure, missings are replaced by values of a conditional distribution, which results from estimating imputation models using the remaining variables of the dataset (van Buuren & Groothuis-Oudshoorn, 2011). We chose the random forest as an imputation algorithm as it has been proven useful for complex, incomplete data problems (Shah, Bartlett, Carpenter, Nicholas, & Hemingway, 2014). To reduce the imputation bias caused by stochastic variation, we specified 50 imputation models. For each of the resulting 50 datasets, we performed a separate cluster analysis and report the mean/modus of the performance coefficients and cluster membership across datasets (Basagaña, Barrera-Gómez, Benet, Antó, & Garcia-Aymerich, 2013).

Multilevel Modeling

Decisions in the Multiverse For constructing the data multiverse (Steegen et al., 2016) we considered the following decisions concerning preprocessing steps:

Decision 1: Coding of Weekend In an earlier draft of the manuscript, we defined the weekend not as a period from *Friday to Sunday*, but from *Friday to Monday*. We found it challenging to decide whether Sunday evening and the following night still belong to the weekend or whether it is more of a weekday in terms of sleep-wake behavior. In sleep research, the nights from Friday to Saturday and from Saturday to Sunday are considered as weekends traditionally. Since on Monday, one usually has to attend to social obligations again, sleep behavior during the night from Sunday to Monday is assumed not to be chosen as freely and used to balance the weekly sleep deficit as the other two weekend nights (Roenneberg et al., 2007). Despite the standard in sleep research, we want to include both variants in our multilevel modeling and thus make our research process transparent.

Decision 2: Number of Weeks We considered the number of repeated measurements to be plausible as both *3* and *4 weeks* because we noticed during the aggregation of the raw timestamped event data that some participants had only partially participated in the last weekend (e.g., only on Saturday, no longer on Sunday).

Decision 3: Outliers For the handling of outliers, we found two points of view plausible. First, smartphone sensing derived variables are usually susceptible to distortion due to data errors, which do not matter if enough data is aggregated using robust measures over a longer period. However, as for week-based variables, only a few single data points can be summarized, outliers due to data errors are more problematic. Therefore, we identified outliers as cases *deviating more than three times the mean absolute deviation from the median* and replaced them by the *person-specific median* of the corresponding variable. Second, the identification of outliers arising from the underlying smartphone usage behavior can be emphasized. In this case, it would be plausible to use a method for outlier handling that limits the variability of the smartphone indicators less than using the median. To cover this aspect, we used *winsorization* as the second alternative.

Decision 4: Missing Values Dealing with missing values in multilevel models is a challenging task. Traditionally, *listwise deletion* has been used which uses only complete observations for estimating the model (e.g., Newman, 2014). Besides the disadvantage of the reduced sample and power, results are likely to be biased if the incomplete observations differ systematically from complete observations (Grund, Lüdtke, & Robitzsch, 2018; Newman, 2014). An alternative approach to deal with missing data is to apply *multiple imputation*. However, in the context of multilevel models, this is not a trivial task as

the imputation model itself should consider the multilevel structure. Current methods and software implementations are reaching their limits if more complicated use cases like random slopes or cross-level interactions are included in the model (Grund et al., 2018). For our analyses, we used the multivariate imputation by chained equations technique and implemented a random slope imputation model with group-level variables as proposed by (Grund et al., 2018). Please note the imputation bias since we were unable to integrate cross-level interactions with existing software implementations. In addition, Grund et al. (2018) point out that this area of research is still ongoing and that there are no clear recommendations for dealing with missing data in use cases such as ours.

Model Description To comprehensibly illustrate the multilevel model used for the multiverse analysis, we present the pseudo-model equation using the lmer syntax of the *lme4 package* in R (Bates et al., 2015). We specified a *random-intercept-random-slope model* predicting the mean duration of nightly inactivity on weekends based on the mean nightly inactivity duration during the previous week (level 1). Chronotype, the big five traits, age, and gender were included as level 2 predictors. The level 1 predictor duration of nightly inactivity during the week was person-centered and the individual mean was entered as level 2 predictor (Curran & Bauer, 2011). Besides, the cross-level interaction of the mean nightly inactivity duration during the previous week and chronotype was added:

$$\begin{aligned}\text{NightlyInactivity}_{\text{weekend}} \sim 1 + \text{NightlyInactivity}_{\text{week}}(\text{L1}, \text{z}, \text{pc}) + \\ \text{Chronotype}(\text{L2}, \text{z}, \text{gc}) + \text{NightlyInactivity}_{\text{week}}(\text{L2}, \text{z}, \text{gc}) + \\ \text{Openness}(\text{L2}, \text{z}, \text{gc}) + \text{Conscientiousness}(\text{L2}, \text{z}, \text{gc}) + \\ \text{Extraversion}(\text{L2}, \text{z}, \text{gc}) + \text{Agreeableness}(\text{L2}, \text{z}, \text{gc}) + \\ \text{EmotionalStability}(\text{L2}, \text{z}, \text{gc}) + \text{Age}(\text{L2}, \text{z}, \text{gc}) + \text{Gender}(\text{L2}, \text{dc}) + \\ \text{NightlyInactivity}_{\text{week}}(\text{L1}, \text{z}, \text{pc}) * \text{Chronotype}(\text{L2}, \text{z}, \text{gc}) + \\ (1 + \text{NightlyInactivity}_{\text{week}}(\text{L1}, \text{z}, \text{pc}) | \text{userid})\end{aligned} \tag{3.1}$$

where L1 denotes predictors on level 1, L2 denotes predictors on level 2, z denotes that predictors were z-standardized, pc denotes that predictors were person-mean-centered, gc denotes that predictors were grand-mean-centered, and dc denotes that gender was dummy-coded (0 = male, 1 = female).

3.8.2 Supplemental Results

Big Five Personality Traits

Table 3.8: Descriptive statistics of personality factors and facets

Variable	*M*	*SD*	alpha CI95%
Openness	-0.05	0.71	[0.93, 0.94]
O1: Openness to imagination	1.28	1.41	[0.84, 0.87]
O2: Openness to aesthetics	0.37	1.29	[0.85, 0.88]
O3: Openness to feelings	2.05	2.09	[0.91, 0.93]
O4: Openness to actions	1.35	1.4	[0.84, 0.87]
O5: Openness to ideas	1.65	1.42	[0.82, 0.86]
O6: Openness to value/norm system	0.9	1.02	[0.73, 0.79]
Conscientiousness	-0.09	0.74	[0.95, 0.96]
C1: Competence	0.84	1.22	[0.76, 0.82]
C2: Love of order	1.1	1.58	[0.87, 0.90]
C3: Sense of duty	1.93	1.41	[0.80, 0.85]
C4: Ambition	1.83	1.68	[0.86, 0.89]
C5: Discipline	1.45	1.46	[0.81, 0.86]
C6: Caution	1.51	1.34	[0.80, 0.84]
Extraversion	-0.01	0.74	[0.95, 0.96]
E1: Friendliness	1.45	1.29	[0.80, 0.84]
E2: Sociableness	1.3	1.74	[0.89, 0.92]
E3: Assertiveness	0.45	1.38	[0.84, 0.87]
E4: Dynamism	1.2	1.59	[0.85, 0.88]
E5: Adventurousness	0.45	1.49	[0.88, 0.91]
E6: Cheerfulness	1.97	1.64	[0.86, 0.89]
Ageeableness	-0.06	0.75	[0.92, 0.94]
A1: Willingness to trust	0.4	1.43	[0.86, 0.89]
A2: Genuineness	1.01	0.94	[0.61, 0.70]
A3: Helpfulness	1.65	1.38	[0.77, 0.82]
A4: Obligingness	1.17	1.31	[0.81, 0.85]
A5: Modesty	0.77	1.13	[0.79, 0.84]
A6: Good naturedness	2.1	1.77	[0.84, 0.88]
Emotional Stability	-0.03	0.71	[0.93, 0.94]
ES1: Carefreeness	0.12	1.3	[0.82, 0.86]
ES2: Equanimity	0.57	1.07	[0.78, 0.83]
ES3: Positive mood	0.95	1.43	[0.84, 0.88]
ES4: Self consciousness	0.66	1.18	[0.83, 0.86]
ES5: Self control	0.64	1	[0.74, 0.81]
ES6: Emotional robustness	0.65	1.19	[0.80, 0.85]

Note. $N = 597$; Alpha CI95% = 95% bootstrapped confidence intervals for Cronbach alpha coefficients.

3.9 References

Adan, A., Archer, S. N., Hidalgo, M. P., Di Milia, L., Natale, V., & Randler, C. (2012). Circadian typology: A comprehensive review. *Chronobiology International*, *29*(9), 1153–1175. doi:10.3109/07420528.2012.719971

Allen, M., Poggiali, D., Whitaker, K., Marshall, T., & Kievit, R. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Res*, *4*, 63. doi:10.12688/wellcomeopenres.15191.1

Arendasy, M. (2009). BFSI: Big-Five Struktur-Inventar (Test & Manual). *Mödling, Austria: SCHUHFRIED GmbH.*

Asendorpf, J. B. (2003). Head-to-head comparison of the predictive validity of personality types and dimensions. *European Journal of Personality*, *17*(5), 327–346. doi:10.1002/per.492

Asendorpf, J. B., & van Aken, M. A. G. (1999). Resilient, overcontrolled, and undercontroleed personality prototypes in childhood: Replicability, predictive power, and the trait-type issue. *Journal of Personality and Social Psychology*, *77*(4), 815–832. doi:10.1037/0022-3514.77.4.815

Au, Q. (2019). Fxtract: Feature extraction from grouped data. *R package version 0.9.1.* Retrieved from https://github.com/QuayAu/fxtract

Bailey, S. L., & Heitkemper, M. M. (2001). Circadian rhythmicity of cortisol and body temperature: Morningness-eveningness effects. *Chronobiology International*, *18*(2), 249–261. doi:10.1081/CBI-100103189

Basagaña, X., Barrera-Gómez, J., Benet, M., Antó, J. M., & Garcia-Aymerich, J. (2013). A framework for multiple imputation in cluster analysis. *American Journal of Epidemiology*, *177*(7), 718–725. doi:10.1093/aje/kws289

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

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. doi:10.1111/j.1745-6916.2007.00051.x

Borger, J. N., Huber, R., & Ghosh, A. (2019). Capturing sleep–wake cycles by using day-to-day smartphone touchscreen interactions. *NPJ Digital Medicine*, *2*(1), 1–8.

Braeken, J., & Van Assen, M. A. (2017). An empirical kaiser criterion. *Psychological Methods*, *22*(3), 450–466. doi:10.1037/met0000074

Caci, H., Deschaux, O., Adan, A., & Natale, V. (2009). Comparing three morningness scales: Age and gender effects, structure and cut-off criteria. *Sleep Medicine*, *10*(2), 240–245. doi:10.1016/j.sleep.2008.01.007

Cavallera, G., & Giudici, S. (2008). Morningness and eveningness personality: A survey in literature from 1995 up till 2006. *Personality and Individual Differences*, *44*(1), 3–21. doi:10.1016/j.paid.2007.07.009

Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., ... Campbell, A. T. (2013). Unobtrusive sleep monitoring using smartphones. *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, 145–152. doi:10.4108/icst.pervasivehealth.2013.252148

Costa Jr, P. T., Herbst, J. H., McCrae, R. R., Samuels, J., & Ozer, D. J. (2002). The replicability and utility of three personality types. *European Journal of Personality*, *16*(S1), S73–S87. doi:10.1002/per.448

Costa, P. T., & McCrae, R. R. (2008). The Revised Neo Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment* (pp. 179–198). doi:10.4135/9781849200479.n9

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, *62*, 583–619. doi:10.1146/annurev.psych.093008.100356

Díaz-Morales, J. F. (2007). Morning and evening-types: Exploring their personality styles. *Personality and Individual Differences*, *43*(4), 769–778. doi:10.1016/j.paid.2007.02.002

Duggan, K. A., Friedman, H. S., McDevitt, E. A., & Mednick, S. C. (2014). Personality and healthy sleep: The importance of conscientiousness and neuroticism. *PloS ONE*, *9*(3), e90628. doi:10.1371/journal.pone.0090628

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104. doi:10.1080/01969727408546059

Ferrara, M., & De Gennaro, L. (2001). How much sleep do we need? *Sleep Medicine Reviews*, *5*(2), 155–179. doi:10.1053/smrv.2000.0138

Fleiss, J. L., Lawlor, W., Platman, S. R., & Fieve, R. R. (1971). On the use of inverted factor analysis for generating typologies. *Journal of Abnormal Psychology*, *77*(2), 127.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers.

Journal of Personality and Social Psychology, *74*(5), 1337–1349. doi:10.1037/0022-3514.74.5.1337

Gradisar, M., Wright, H., Robinson, J., Pain, S., & Gamble, A. (2008). Adolescent napping behavior: Comparisons of school week versus weekend sleep patterns. *Sleep and Biological Rhythms*, *6*(3), 183–186. doi:10.1111/j.1479-8425.2008.00351.x

Gray, E. K., & Watson, D. (2002). General and specific traits of personality and their relation to sleep and academic performance. *Journal of Personality*, *70*(2), 177–206. doi:10.1111/1467-6494.05002

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149. doi:10.1177/1094428117703686

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2-3), 107–145. doi:10.1023/A:1012801612483

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. *Perspectives on Psychological Science*, *11*(6), 838–854. doi:10.1177/1745691616650285

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., ... Gosling, S. D. (2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, Advance online publication. doi:10.1037/pspp0000245

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, *52*(1), 258–271. doi:10.1016/j.csda.2006.11.025

Hennig, C. (2008). Dissolution point and isolation robustness: Robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis*, *99*(6), 1154–1176. doi:10.1016/j.jmva.2007.07.002

Hennig, C. (2018). Fpc: Flexible procedures for clustering. *R package version 2.1-11.1*. Retrieved from https://CRAN.R-project.org/package=fpc

Hicks, L. E. (1984). Conceptual and empirical analysis of some assumptions of an explicitly typological theory. *Journal of Personality and Social Psychology*, *46*(5), 1118–1131. doi:10.1037/0022-3514.46.5.1118

Horne, J. A., & Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology*, *4*(2), 97–110.

Kafadar, K. (2003). John tukey and robustness. *Statistical Science*, *18*(3), 319–331. doi:10.1214/ss/1076102419

Katzenberg, D., Young, T., Finn, L., Lin, L., King, D. P., Takahashi, J. S., & Mignot, E. (1998). A clock polymorphism associated with human diurnal preference. *Sleep*, *21*(6), 569–576. doi:10.1093/sleep/21.6.569

Križan, Z., & Hisler, G. (2019). Personality and sleep: Neuroticism and conscientiousness predict behaviourally recorded sleep years later. *European Journal of Personality*, *33*, 133–153. doi:10.1002/per.2191

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., & Rathouz, P. J. (2008). Self-reported and measured sleep duration. *Epidemiology*, *19*(6), 838–845. doi:10.1097/ede.0b013e318187a7b0

Lin, Y.-H., Wong, B.-Y., Lin, S.-H., Chiu, Y.-C., Pan, Y.-C., & Lee, Y.-H. (2019). Development of a mobile application (app) to delineate "digital chronotype" and the effects of delayed chronotype by bedtime smartphone use. *Journal of Psychiatric Research*, *110*, 9–15. doi:10.1016/j.jpsychires.2018.12.012

Lipnevich, A. A., Credé, M., Hahn, E., Spinath, F. M., Roberts, R. D., & Preckel, F. (2017). How distinctive are morningness and eveningness from the big five factors of personality? a meta-analytic investigation. *Journal of Personality and Social Psychology*, *112*(3), 491. doi:10.1037/pspp0000099

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/bf02296272

Mateo, M. J. C., Diaz-Morales, J. F., Barreno, C. E., Prieto, P. D., & Randler, C. (2012). Morningness-eveningness and sleep habits among adolescents: Age and gender differences. *Psicothema*, *24*(3), 410–415.

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology*, *113*(1), 39. doi:10.1037/0021-843X.113.1.39

Min, J.-K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., & Hong, J. I. (2014). Toss'n'turn: Smartphone as sleep and sleep quality detector. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 477–486. doi:10.1145/2556288.2557220

Natale, V., & Cicogna, P. (2002). Morningness-eveningness dimension: Is it really a continuum? *Personality and Individual Differences*, *32*(5), 809–816. doi:10.1016/S0191-8869(01)00085-X

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

Ohayon, M., Wickwire, E. M., Hirshkowitz, M., Albert, S. M., Avidan, A., Daly, F. J., ... Gozal, D., et al. (2017). National sleep foundation's sleep quality recommendations: First report. *Sleep Health*, *3*(1), 6–19. doi:10.1016/j.sleh.2016.11.006

Preckel, F., Fischbach, A., Scherrer, V., Brunner, M., Ugen, S., Lipnevich, A. A., & Roberts, R. D. (2019). Circadian preference as a typology: Latent-class analysis of adolescents' morningness/eveningness, relation with sleep behavior, and with academic outcomes. *Learning and Individual Differences*, in press, available online. doi:10.1016/j.lindif.2019.03.007

Putilov, A. A. (2017). Owls, larks, swifts, woodcocks and they are not alone: A historical review of methodology for multidimensional self-assessment of individual differences in sleep-wake pattern. *Chronobiology international*, *34*(3), 426–437. doi:10.1080/07420528.2017.1278704

Putilov, A. A., Donskaya, O. G., & Verevkin, E. G. (2015). How many diurnal types are there? a search for two further "bird species". *Personality and Individual Differences*, *72*, 12–17. doi:10.1016/j.paid.2014.08.003

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Randler, C. (2007). Gender differences in morningness–eveningness assessed by self-report questionnaires: A meta-analysis. *Personality and Individual Differences*, *43*(7), 1667–1675. doi:10.1016/j.paid.2007.05.004

Randler, C. (2008). Morningness–eveningness, sleep–wake variables and big five personality factors. *Personality and Individual Differences*, *45*(2), 191–196. doi:10.1016/j.paid.2008.03.007

Randler, C., Díaz-Morales, J. F., Rahafar, A., & Vollmer, C. (2016). Morningness–eveningness and amplitude–development and validation of an improved composite scale to measure circadian preference and stability (messi). *Chronobiology international*, *33*(7), 832–848. doi:10.3109/07420528.2016.1171233

Randler, C., & Engelke, J. (2019). Gender differences in chronotype diminish with age: A meta-analysis based on morningness/chronotype questionnaires. *Chronobiology International*, *36*(7), 888–905. doi:10.1080/07420528.2019.1585867

Randler, C., Schredl, M., & Göritz, A. S. (2017). Chronotype, sleep behavior, and the big five personality factors. *Sage Open*, *7*(3), 1–9. doi:10.1177/2158244017728321

Revelle, W. (2018). Psych: Procedures for psychological, psychometric, and personality research. *R package version 1.8.12*. Retrieved from https://CRAN.R-project.org/package=psych

Roenneberg, T. (2015). Having trouble typing? what on earth is chronotype? *Journal of Biological Rhythms*, *30*(6), 487–491. doi:10.1177/0748730415603835

Roenneberg, T., Allebrandt, K. V., Merrow, M., & Vetter, C. (2012). Social jetlag and obesity. *Current Biology*, *22*(10), 939–943. doi:10.1016/j.cub.2012.03.038

Roenneberg, T., Keller, L. K., Fischer, D., Matera, J. L., Vetter, C., & Winnebeck, E. C. (2015). Human activity and rest in situ. *Methods in Enzymology*, *552*, 257–283. doi:10.1016/bs.mie.2014.11.028

Roenneberg, T., Kuehnle, T., Juda, M., Kantermann, T., Allebrandt, K., Gordijn, M., & Merrow, M. (2007). Epidemiology of the human circadian clock. *Sleep Medicine Reviews*, *11*(6), 429–438. doi:10.1016/j.smrv.2007.07.005

Roenneberg, T., Wirz-Justice, A., & Merrow, M. (2003). Life between clocks: Daily temporal patterns of human chronotypes. *Journal of Biological Rhythms*, *18*(1), 80–90. doi:10.1177/0748730402239679

Roepke, S. E., & Duffy, J. F. (2010). Differential impact of chronotype on weekday and weekend sleep timing and duration. *Nature and Science of Sleep*, *2*, 213–220. doi:10.2147/NSS.S12572

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi:10.1016/0377-0427(87)90125-7

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. doi:10.2307/2291267

Schoedel, R., Au, Q., Völkel, S. T., Lehmann, F., Becker, D., Bühner, M., ... Stachl, C. (2018). Digital Footprints of Sensation Seeking. *Zeitschrift für Psychologie*, *226*(4), 232–245. doi:10.1027/2151-2604/a000342

Schoedel, R., Pargent, F., Au, Q., Völkel, S., Schuwerk, T., Bühner, M., & Stachl, C. (2020). To challenge the morning lark and the night owl: Using smartphone sensing data to investigate day-night behavior patterns. *OSF*. Retrieved from https://osf.io/a4h3b/

Schuwerk, T., Kaltefleiter, L. J., Au, J.-Q., Hoesl, A., & Stachl, C. (2019). Enter the wild: Autistic traits and their relationship to mentalizing and social interaction in everyday life. *Journal of Autism and Developmental Disorders*, *49*(10), 4193–4208. doi:10.1007/s10803-019-04134-6

Sella, E., Carbone, E., Toffalini, E., & Borella, E. (2020). Personality traits and sleep quality: The role of sleep-related beliefs. *Personality and Individual Differences*, *156*, 109770. doi:10.1016/j.paid.2019.109770

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using mice: A caliber study. *American Journal of Epidemiology*, *179*(6), 764–774. doi:10.1093/aje/kwt312

Shambroom, J. R., Fábregas, S. E., & Johnstone, J. (2012). Validation of an automated wireless system to monitor sleep in healthy adults. *Journal of Sleep Research*, *21*(2), 221–230. doi:10.1111/j.1365-2869.2011.00944.x

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN*. doi:10.2139/ssrn.2694998

Spears, S. K., Montgomery-Downs, H. E., Steinman, S. A., Duggan, K. A., & Turiano, N. A. (2019). Sleep: A pathway linking personality to mortality risk. *Journal of Research in Personality*, *81*, 11–24. doi:10.1016/j.jrp.2019.04.007

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., ... Bühner, M. (2019). Behavioral patterns in smartphone usage predict big five personality traits. *Psyarxiv*. doi:10.31234/osf.io/ks4vd

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., ... Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722. doi:10.1002/per.2113

Stachl, C., Schoedel, R., Au, Q., Völkel, S., Buschek, D., Hussmann, H., ... Bühner, M. (2018). The phonestudy project. *OSF*. doi:10.17605/osf.io/ut42y

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. doi:10.1177/1745691616658637

Sutin, A. R., Gamaldo, A. A., Stephan, Y., Strickhouser, J. E., & Terracciano, A. (2019). Personality traits and the subjective and objective experience of sleep. *International Journal of Behavioral Medicine*. doi:10.1007/s12529-019-09828-w

Takano, K., Sakamoto, S., & Tanno, Y. (2014). Repetitive thought impairs sleep quality: An experience sampling study. *Behavior Therapy*, *45*(1), 67–82. doi:10.1016/j.beth.2013.09.004

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Cluster analysis: Basic concepts and algorithms. In P.-N. Tan, M. Steinbach, & V. Kumar (Eds.), *Introduction to data mining* (pp. 487–568). Boston, USA: Addison-Wesley.

Terman, J. S., Terman, M., Lo, E.-S., & Cooper, T. B. (2001). Circadian time of morning light administration and therapeutic response in winter depression. *Archives of General Psychiatry*, *58*(1), 69–75. doi:10.1001/archpsyc.58.1.69

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528. doi:10.1198/106186005X59243

Tonetti, L., Pascalis, V. D., Fabbri, M., Martoni, M., Russo, P. M., & Natale, V. (2016). Circadian typology and the alternative five-factor model of personality. *International Journal of Psychology*, *51*(5), 332–339. doi:10.1002/ijop.12170

Tsaousis, I. (2010). Circadian preferences and personality traits: A meta-analysis. *European Journal of Personality*, *24*(4), 356–373. doi:10.1002/per.754

Ushey, K., McPherson, J., Cheng, J., Atkins, A., & Allaire, J. (2018). Packrat: A dependency management system for projects and their r package dependencies. *R package version 0.4.9-3*. Retrieved from https://CRAN.R-project.org/package=packrat

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. doi:10.18637/jss.v045.i03

Vela-Bueno, A., Fernandez-Mendoza, J., Olavarrieta-Bernardino, S., Vgontzas, A. N., Bixler, E. O., de la Cruz-Troca, J. J., ... Oliván-Palacios, J. (2008). Sleep and behavioral correlates of napping among young adults: A survey of first-year university students in madrid, spain. *Journal of American College Health*, *57*(2), 150–158. doi:10.3200/jach.57.2.150-158

Vitale, J. A., Roveda, E., Montaruli, A., Galasso, L., Weydahl, A., Caumo, A., & Carandente, F. (2015). Chronotype influences activity circadian rhythm and sleep: Differ-

ences in sleep quality between weekdays and weekend. *Chronobiology International*, *32*(3), 405–415. doi:10.3109/07420528.2014.986273

Wei, T., & Simko, V. (2017). R package corrplot: Visualization of a correlation matrix. *R package version 0.84*. Retrieved from https://github.com/taiyun/corrplot

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. doi:10.1007/978-3-319-24277-4

Wickham, H., François, R., Henry, L., & Müller, K. (2019). Dplyr: A grammar of data manipulation. *R package version 0.8.0.1*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wittmann, M., Dinich, J., Merrow, M., & Roenneberg, T. (2006). Social jetlag: Misalignment of biological and social time. *Chronobiology International*, *23*(1-2), 497–509. doi:10.1080/07420520500545979

Zavada, A., Gordijn, M. C., Beersma, D. G., Daan, S., & Roenneberg, T. (2005). Comparison of the Munich Chronotype Questionnaire with the Horne-Östberg's Morningness-Eveningness Score. *Chronobiology International*, *22*(2), 267–278. doi:10.1081/CBI-200053536

Zimmermann, L. K. (2011). Chronotype and the transition to college life. *Chronobiology International*, *28*(10), 904–910. doi:10.3109/07420528.2011.618959

Chapter 4

General Discussion

The present dissertation employed smartphone sensing to investigate how established biopsychological personality concepts are reflected in actual behavior. Methods of computational science were used, and statistical models of both the data and algorithmic modeling tradition were applied to gain first insights in the course of this exploratory research. Study 1 investigated whether the trait sensation seeking can be predicted based on behavioral indicators of smartphone usage. By comparing different machine learning algorithms, sensation seeking could be predicted above chance, but the overall prediction accuracy was rather low. Methods of interpretable machine learning revealed that smartphone usage indicating calling activity and day-night patterns was particularly important for the algorithm's predictions of self-reported sensation seeking scores. In study 2, it was explored whether differences in similar day-night patterns manifest in smartphone usage activity. Two stable, but non-discrete groups of smartphone-sensed "morning larks" and "night owls" were identified. Besides, behavioral markers of day-night activity patterns were related to conscientiousness, and individuals' habitual nightly usage inactivity during the week influenced their nightly usage inactivity on weekends.

The main findings of both empirical studies have already been discussed in detail in the respective chapters. Therefore, the general discussion focuses on the overall contribution of the present dissertation. In the sense of scientific modesty, it is discussed what the present work can and what it cannot contribute to the existing smartphone sensing literature in personality psychology. Finally, challenges and future directions in the newly emerging field of psychoinformatics are discussed.

4.1 Overall Contribution of the Present Dissertation

4.1.1 Investigation of Actual Behavior

One aim of the present dissertation was to illustrate how smartphone sensing as an assessment tool could foster new insights into personality. Previous research in personality psychology has traditionally been centered on the person component, while especially the behavior component has been neglected so far (Funder, 2001). This imbalance has been encouraged by decades of questionnaire assessment dominating personality research (Baumeister, Vohs, & Funder, 2007). As the assessment of actual behavior has been cumbersome, self-reports were often used despite known biases (Baumeister et al., 2007; Funder, 2009). The studies presented in this dissertation are among the first to empirically investigate actual behavior in real situations collected over a more extended period, resulting in ecologically valid findings (Harari et al., 2016). More generally, the present dissertation illustrates that new technologies, such as smartphone sensing, enable the collection of a wide range of data reflecting behavior in everyday life. Notably, smartphone sensing has the potential to inform two different behavioral approaches (Harari et al., 2020). First, sensing enables the investigation of behavioral data directly reflecting (mainly quantitative) aspects of smartphone usage. Variables representing this approach in the present dissertation are the frequency and duration of using specific app categories or of calling behavior in study 1. Second, smartphone-sensed data also enable the extraction of behavioral markers indicating real-life behaviors beyond smartphone usage. Examples for this approach are the GPS-based variables indicating daily mobility behaviors in study 1 or the nightly inactivity duration of smartphone usage, indicating sleep duration in study 2. The two approaches are not mutually exclusive. Rather, smartphone sensor data have a dual function, i.e., they represent both the behavior on the smartphone and proxies of the real behavior at a more abstract level. For example, in the first study of this dissertation, the calling behavior was considered as smartphone usage behavior. However, in a study by Harari et al. (2019), it was considered as a proxy for daily social behavior.

4.1.2 Combining Statistical Modeling Approaches: Explanation and Prediction

The present dissertation contributes to the ongoing debate about the explanatory versus the predictive focus of psychological research (Mahmoodi, Leckelt, van Zalk, Geukes, &

Back, 2017; Yarkoni & Westfall, 2017). Psychologists have previously neglected the predictive focus leading to many models with high explanatory power, which nevertheless poorly predict future behavior (Yarkoni & Westfall, 2017). Some researchers have proposed explanations and predictions to be mutually complementary approaches that interact in a cyclical empirical research process, but have also found that this process has hardly been implemented in research practice yet (Mahmoodi et al., 2017). The two studies presented in this dissertation provide a first glance at this cyclical research process. For example, the prediction in study 1 is not without theoretical foundation. Instead, we used findings of previous explanatory research to extract meaningful variables informing the prediction model. In turn, the applied methods of interpretable machine learning indicate that behaviors related to calling and day-night activity might be attractive candidates for future explanatory research.

Study 2 gives further indication that statistical methods from both the prediction and explanatory approach fruitfully complement each other in psychological research. The multiverse analysis showed that when working with smartphone sensing data, different preprocessing choices have different effects on the results (and conclusions drawn on them). In algorithmic modeling, preprocessing such as feature engineering or variable transformation is part of the statistical modeling procedure (Kuhn & Johnson, 2020). Therefore, trying out different preprocessing steps is an explicit and deliberate step to achieve the best prediction accuracy possible (Kuhn & Johnson, 2013, 2020). In contrast, the data modeling approach commonly used in psychology has been unaware of the effects of different preprocessing decisions on the robustness of results and conclusions based on them. Recently, the open science movement encouraged some researchers to present modeling techniques such as the multiverse to raise awareness of the dependence of research results on preprocessing steps in the data modeling culture (Simonsohn, Simmons, & Nelson, 2015; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Of course, the application of methods from the algorithmic modeling culture per se does not guarantee a transparent representation of data preprocessing. However, exemplary for the data-intensive context of smartphone sensing, this dissertation demonstrates how the data modeling tradition can benefit from adopting the explicitness in data preprocessing from the algorithmic modeling culture to prevent questionable research practices such as selective reporting (Steegen et al., 2016).

4.1.3 Using Machine Learning for Personality Research

Finally, the present dissertation illustrates two different approaches to employ machine learning for personality research based on digital data. The principle idea behind this emerging field of machine learning-based personality assessment is that personality traits influence the way persons interact with technology. Digital records, in turn, reflect users' personalities and can, therefore, be used for personality assessment (Bleidorn & Hopwood, 2018; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Stachl, Pargent, et al., 2019). Following this idea, many previous studies have used supervised machine learning techniques to predict the big five personality traits based on smartphone sensing data (Chittaranjan, Blom, & Gatica-Perez, 2013; de Montjoye, Quoidbach, Robic, & Pentland, 2013; Mønsted, Mollgaard, & Mathiesen, 2018; Stachl, Au, et al., 2019). The dissertation adds to this previous work by using both supervised and unsupervised machine learning approaches to assess biopsychological personality traits. In line with methodological procedures of previous work, study 1 applied supervised machine learning (e.g., Chittaranjan et al., 2013; de Montjoye et al., 2013; Mønsted et al., 2018): Behavioral variables extracted from smartphone sensing data served as input to predict the personality criterion sensation seeking assessed via self-report questionnaires. In contrast, study 2 applied unsupervised machine learning to uncover structures in the day-night behavior patterns. This person-centered approach enables the assessment of personality types, which means groups of persons with similar behavioral patterns. Both studies thus show that smartphone sensing data, in combination with supervised and unsupervised machine learning approaches, provide a broad range of opportunities for personality research.

4.2 Limitations of the Present Dissertation and Implications for Smartphone Sensing Research

This dissertation has some limitations, some of which reflect the fact that the integration of tools of the computational science into personality research is still in its infancy.

4.2.1 Extraction of Meaningful Variables

Recent literature has argued that ambulatory assessment tools could help to restore balance in the study of the personality components person, situation, and behavior (Harari et al., 2020; Wrzus & Mehl, 2015). However, the empirical studies of this dissertation only

exploited a fraction of the opportunities offered by smartphone sensing. Accordingly, all three personality triad components were only considered in a simplified manner: The person component was operationalized as a trait and therefore assessed once via self-report questionnaires. The situation component was not considered except the time of the week and the day. The behavior component was aggregated over all situations during the period of investigation. However, to obtain a comprehensive understanding of the personality triad and its underlying dynamics, it would be necessary to investigate which specific personal characteristics are related to the execution of certain behaviors in certain situations (Funder, 2001). Therefore, future studies should design assessment methods and the subsequent variable extraction in more sophisticated ways to cover each of the three components comprehensively.

Regarding the person component, besides stable personality traits, internal and variable person characteristics such as thoughts or feelings could additionally be captured by integrating active logging via experience sampling in smartphone sensing studies (Harari, Gosling, Wang, & Campbell, 2015; Harari et al., 2020). Only a few empirical studies have combined passive sensing and active logging for assessing person characteristics so far, but it is a promising approach for future research (e.g., R. Wang et al., 2014).

In addition, smartphone sensing has the potential to draw a broad picture of the situation component (Harari et al., 2015; Harari et al., 2020). Beyond time, as considered in the present work, objective characteristics of situations such as locations, objects, or activities could be detected (Rauthmann, Sherman, & Funder, 2015). For this purpose, future research should develop even better sensor technology and invest more effort in data preprocessing (Seifert, Hofer, & Allemand, 2018). For example, ambient sounds detected by microphone sensors (Harari et al., 2019), locations resulting from GPS labeling (Mehrotra et al., 2017), Bluetooth scans of surrounding devices (Chittaranjan et al., 2013), and physical activity data (W. Wang et al., 2018) could be combined to detect social situations. The detection of these objective situational cues could, in turn, be combined with event-triggered experience sampling. For example, if the passive sensing detected a social situation, participants could be notified and asked about their subjective experience in this particular class of situation. Thus, the integration of active logging into passive sensing would be a promising approach to study the psychological meaning of situations (Rauthmann et al., 2015).

The detection of situations, in turn, comes along with new opportunities for the behavior component. For example, behaviors can be extracted depending on situational classes.

To stay with the example mentioned above: If a social situation is detected, behavioral variables such as the participants' speech rate or smartphone usage frequency could be extracted and aggregated to behaviors expressed in social situations.

In addition, future research could address the longitudinal character offered by smartphone sensing data. To some extent, study 2 of the present dissertation considered this aspect by extracting the nightly inactivity of smartphone usage on a week-wise basis. Nevertheless, the gold standard in personality research using smartphone sensing data is to aggregate behavioral variables for the entire study period (e.g., Montag et al., 2014; Stachl et al., 2017). In doing so, the smartphone sensing method loses its beneficial longitudinal character. Fostering the investigation of *intra*individual besides *inter*individual aspects of personality, future studies should consider the extraction of variables in the form of repeated measures (Harari et al., 2020). For example, behaviors displayed in social situations could be extracted for every single day of the week, and behavior tendencies throughout the workweek could be investigated (Harari et al., 2019).

4.2.2 Validation of Measures Reflecting Real-World Behavior

A further limitation concerning data preprocessing procedures is that smartphone-sensed variables that serve as markers of behavior beyond smartphones, such as day-night patterns, have not been validated by other measurements. Therefore, the conclusions based on the present results are restricted to smartphone usage patterns (e.g., duration of nightly smartphone usage inactivity) and cannot readily be generalized to other real-world behavior (e.g., sleep duration). Although the plausibility of the results tempts to draw more profound conclusions, future research should first address the question of how well behavioral markers extracted from smartphone sensing data reflect the respective real-world behavior. Personality psychological research has not yet addressed corresponding validation studies. However, only recently, the German Council for Social and Economic Data published first suggestions addressing the validity and reliability of data collected via smartphones and wearable devices (RatSWD, 2020). It is proposed, for example, to check construct validity by comparing the smartphone-sensed data to the previous gold standard of measurement. External validity should be examined by comparing smartphone sensing with other measurement methods (RatSWD, 2020). Considering an example from study 2, nightly inactivity could be validated by an Electroencephalography (EEG) as the gold standard for determining sleep or by actigraphy for sleep measurement in the field. In addition, the external validity could be investigated by combining smartphone sensing with daily

experience sampling questionnaires and a questionnaire after the study period to assess self-reported sleeping times. In sum, personality psychological research has used smartphone sensing data without addressing the quality of sensor-based sensing methods. To make a sustainable contribution, future research should establish psychometric standards, as known from survey research (Harari et al., 2020).

4.2.3 Characteristics of Smartphone Sensing Studies

The empirical studies 1 and 2 have already discussed the limitation of biased student samples. However, at this point, it should be highlighted again that this sample composition reflects a general problem of smartphone sensing research in personality psychology (e.g., Harari et al., 2019; Mønsted et al., 2018; Montag et al., 2014; Stachl et al., 2017; R. Wang et al., 2014). The generalizability of previous research results is, therefore, only limited to young and well-educated individuals. Student samples are more comfortable to recruit and also provide the first essential insights. However, to establish smartphone sensing research in personality psychology, one next step in the field should be to replicate findings with more representative samples.

Less of a limitation and more as an outlook for future studies, it should be noted that research presented in this dissertation is exclusively observational. The previous goal of smartphone sensing research in personality psychology was to record everyday smartphone usage as unobtrusive as possible (e.g., Harari et al., 2019; Stachl et al., 2017). However, smartphones have been ascribed with the potential to become mobile laboratories to study humans (Miller, 2012). For example, future research could use the numerous functionalities of smartphones to run interactive experiments such as presenting audio or image stimuli and measuring participants' reaction times in non-standardized settings outside the laboratory (Miller, 2012). In addition, Mohr, Zhang, and Schueller (2017) propose to incorporate behavioral interventions into smartphone sensing technology. A three-step study procedure could help to do so: In the first stage, researchers can use smartphones to sense naturally occurring behaviors (e.g., physical activity). The second stage can contain interventions, for example, by providing participants with psycho-educational content via the smartphone. In the third and final stage, the smartphone senses if changes in behavior occur (Mohr et al., 2017).

4.3 Challenges and Future Directions in Psychoinformatics

Smartphone sensing research is one example of newly emerging sources of large-scale data resulting from the daily use of digital technologies. It enables psychologists not only to investigate new research topics related to technology interaction but also to investigate established psychological constructs with new types of data, as shown in the present dissertation. Psychological research at the interface of computer science and statistics still has to overcome some hurdles to establish itself as psychoinformatics gradually and to fully profit from this increasing availability of digital data sources (Lazer et al., 2009).

4.3.1 Ethical Handling of Digital Data

In contrast to other disciplines conducting data-intensive research such as engineering or physics, social science in general and psychology in particular use sensitive human data about individuals' lives such as daily behavior or mental well-being. This entails a special responsibility in many respects. First, social sciences must establish mechanisms and rules for the ethical handling of these new sensitive data types (Harari et al., 2020; Seifert et al., 2018). Technical advancements make it possible to obtain detailed records of digital life (Reeves, Robinson, & Ram, 2020). This raises concerns to what extend research should make use of these opportunities. As an example, the Human Screenome project takes screenshots from participant's smartphones every five seconds (Reeves et al., 2020). The corresponding research article discusses how this method of data recording could be used to investigate different questions such as the influence of media use on well-being or changes in people's lives through media use (Reeves et al., 2020). Thereby, the article gives the Human Screenome project the impression of being a data mining approach, i.e., the actual research objectives are not defined a priori. Thus, it remains unclear whether this high resolution of data recording is necessary to achieve the research goal. In contrast, data protection laws such as the General Data Protection Regulation (GDPR) in Europe suggest that personal data should be collected according to the principle of minimum (European Commission, 2016). This means that only data necessary for answering the research questions at hand should be collected. Accordingly, Mahmoodi et al. (2017) argue that with this new availability of large amounts of data, researchers should not forget to integrate the theoretical work that has become established in the behavioral and social sciences. The Human Screenome project, therefore, illustrates the need for

psychoinformatics to strike a balance between the almost infinite possibility of gaining knowledge through the use of innovative technologies and collecting data following the principle of minimum to protect participants' privacy, but also to save technical resources.

In this context, the responsible handling of the knowledge resulting from the work with new data types poses another challenge (Harari et al., 2020). Publishing articles in this line of research can have far-reaching consequences: For example, a paper describing how digital data predict the risk of depression can be misused by insurance companies to avoid insuring people with specific digital usage patterns. On the other hand, publishing is an essential part of scientific work. It draws the public's attention to research that has probably been going on for years in private companies collecting large datasets of their users (Lazer et al., 2009; Miller, 2012). However, researchers should be careful not to misuse their results for commercial or political purposes. One prominent case highlighting researchers' responsibilities to handle research results ethically is Cambridge Analytica (Isaak & Hanna, 2018). Researchers from the University of Cambridge presented psychographic profiling, i.e., predicting the personality of Facebook users from their usage data. Based on these findings, the company Cambridge Analytica, with the participation of one of the researchers, retrieved the data of millions of Facebook users without their knowledge and created customized advertising in the US election campaign 2014 intending to influence voting behavior (Isaak & Hanna, 2018). In the long term, these media-effective cases likely contribute to a decline in the public's trust in science. Accordingly, psychoinformatics is challenged by the reactance of people to participate in data-intensive studies for research purposes (Lazer et al., 2009).

4.3.2 Data Privacy and Data Security

First steps to establish responsible handling of personal data in psychoinformatics are related to data privacy and data security (Seifert et al., 2018). Data privacy describes measures at an organizational level, including, for example, to make data processing transparent to participants (European Commission, 2016). In data-intensive studies, vast amounts of personal data are collected, which often makes anonymization impossible (Lazer et al., 2009; Miller, 2012; Seifert et al., 2018). For example, in smartphone sensing studies knowing participants' age and gender and where they stay during the night according to GPS data makes it usually possible to identify participants (Mohr et al., 2017). Therefore, researchers must carefully coordinate data collection procedures with data protection officers and ethical committees of the respective institutions. In addition, participants should

be informed extensively about the purpose of the study and the collected types of data (Beierle et al., 2019; Harari et al., 2016; Wrzus & Mehl, 2015). Doing so should enable participants to give *informed* consent (Harari, 2020; Harari et al., 2020). This requirement introduces new research questions at a meta-level into the field of psychoinformatics. For example, there are first indications that participants do not read provided data privacy information carefully, impairing the concept of informed consent (Keusch, Struminskaya, Antoun, Couper, & Kreuter, 2019; Kreuter, Haas, Keusch, Bähr, & Trappmann, 2018; Piwek & Ellis, 2016). Hence, future research should address how to inform participants about the data collection procedure properly and how to give them as much control as possible over their data (e.g., Beierle et al., 2019; Harari, 2020; Mohr et al., 2017). A further question for future meta-research is to explore which incentives attract people to take part in data-intensive studies and how research can create added value for participation (Harari et al., 2017). Another meta-question could focus on investigating the generalizability of data-intensive studies. It would be interesting to investigate whether individual differences cause a pre-selection of the samples, i.e., whether persons with certain trait levels (e.g., high openness) are more likely to participate in the studies than others.

Besides data privacy, another critical step for responsible research in psychoinformatics is data security. Data security describes technical measures to handle personal data including, for example, the secure storage of the data without access for third parties and also the deletion of sensitive data as soon as they fulfilled their research purposes (European Commission, 2016; Harari et al., 2016; Miller, 2012; Wrzus & Mehl, 2015). In this context, an active field of research in computer science is how data can be processed immediately on the device without storing raw data (Harari et al., 2020; Markowetz, Błaszkiewicz, Montag, Switala, & Schlaepfer, 2014; Wrzus & Mehl, 2015). Once again, referring to the Human Screenome project, one could refrain from storing the screenshots, i.e., the raw image data, and classify them on-device in real-time via image processing algorithms. At this point, psychology depends on advances in computer science, but can also contribute in the form of interdisciplinary work to identify which data to extract for other disciplines.

4.3.3 Interdisciplinarity

Another challenge is the development of an interdisciplinary infrastructure to establish the field of psychoinformatics (Lazer et al., 2009; Montag, Duke, & Markowetz, 2016). On the one hand, this includes providing training opportunities for psychologists to deepen their existing statistical knowledge enabling themselves to work with big datasets (Miller,

2012). Training programs should adapt to new requirements such as, for example, data management, data preprocessing, and algorithmic modeling techniques such as machine learning (Seifert et al., 2018; Yarkoni, 2012). On the other hand, the technical infrastructure poses another big challenge for psychoinformatics. Most psychologists have primary education in statistics but not in computer science. However, tracking digital data requires programming skills (Miller, 2012; Seifert et al., 2018), which is why interdisciplinary research teams should be formed (Lazer et al., 2009). However, the example of smartphone sensing demonstrates that interdisciplinarity is a challenging task as scientific goals often differ between disciplines. The smartphone sensing approach comes from computer science, which is concerned with showing that novel logging procedures work (Piwek & Ellis, 2016). Therefore, computer scientists often focus on a proof of concept of the logging technique with smaller samples and not about building stable software that other disciplines can use for their content-related research in the long term. Accordingly, the psychological research landscape has introduced several sensing apps, but they only resulted in single studies before disappearing again (e.g., Beierle et al., 2019; R. Wang et al., 2014). This leads Lazer et al. (2009) to conclude that publication incentives in science have to change to promote sustainable interdisciplinarity and, consequently, psychoinformatics.

4.4 Conclusion

The present dissertation builds a bridge between personality psychology and computational science. Two presented empirical studies used smartphone sensing data to investigate established biopsychological personality concepts. In contrast to earlier research, which mainly used self-report questionnaires, both studies focused on the extraction of behavioral counterparts of the respective personality construct under investigation. The findings of both empirical studies point out that integrating measures of actual behaviors in personality research could be a promising way to foster the understanding of individual differences beyond established personality constructs. From a methodological perspective, the present dissertation used statistical modeling approaches from both the prediction and the explanation culture. In doing so, it demonstrated that exploratory research could fruitfully combine prediction and classical data modeling approaches and that machine learning techniques offer various options for personality research.

Through the two empirical studies, the present dissertation also demonstrates some of the current limitations and challenges of smartphone sensing in particular and psychoin-

formatics in general. For example, logging and variable extraction procedures have the potential for improvement in many respects, validation procedures for behavioral variables have not yet been established, and study characteristics of previous research limit the generalizability of findings. Further challenges not restricted to smartphone sensing but referring to psychoinformatics, in general, are the ethical handling of personal data, data privacy, and the establishment of interdisciplinary infrastructure. In summary, both smartphone sensing research in personality psychology and psychoinformatics are still in its infancy. However, by providing empirical illustrations, this dissertation contributes to a better understanding of some of the limitations and current challenges. If future research overcomes these hurdles, smartphone sensing, in combination with active logging methods, could establish as an ecological valid ambulatory assessment tool in psychology, provide essential insights in the study of human personality, and contribute to the establishment of psychoinformatics.

4.5 References

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. doi:10.1111/j.1745-6916.2007.00051.x

Beierle, F., Tran, V. T., Allemand, M., Neff, P., Schlee, W., Probst, T., ... Pryss, R. (2019). What data are smartphone users willing to share with researchers? *Journal of Ambient Intelligence and Humanized Computing*, 1–13. doi:10.1007/s12652-019-01355-6

Bleidorn, W., & Hopwood, C. J. (2018). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. doi:10.1177/1088868318772990

Chittaranjan, G., Blom, J., & Gatica-Perez, D. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433–450. doi:10.1007/s00779-011-0490-1

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 48–55. doi:10.1007/978-3-642-37210-0_6

European Commission. (2016, April 5). 2018 reform of eu data protection rules. Retrieved March 6, 2020, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*(1), 197–221. doi:10.1146/annurev.psych.52.1.197

Funder, D. C. (2009). Persons, behaviors and situations: An agenda for personality psychology in the postwar era. *Journal of Research in Personality*, *43*(2), 120–126. doi:10.1016/j.jrp.2008.12.041

Harari, G. M. (2020). A process-oriented approach to respecting privacy in the context of mobile phone tracking. *Current Opinion in Psychology*, *31*, 141–147. doi:10.1016/j.copsyc.2019.09.007

Harari, G. M., Gosling, S. D., Wang, R., & Campbell, A. T. (2015). Capturing situational information with smartphones and mobile sensing methods. *European Journal of Personality*, *29*(5), 509–511. doi:10.1002/per.2032

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. *Perspectives on Psychological Science*, *11*(6), 838–854. doi: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. *Social Psychological and Personality Science*, *8*(5), 479–492. doi:10.1177/1948550617712033

Harari, G. M., Müller, S. R., Stachl, C., Wang, R., Wang, W., Bühner, M., ... Gosling, S. D. (2019). Sensing sociability: Individual differences in young adults' conversation, calling, texting, and app use behaviors in daily life. *Journal of Personality and Social Psychology*, Advance online publication. doi:10.1037/pspp0000245

Harari, G. M., Vaid, S. S., Müller, S. R., Stachl, C., Marrero, Z., Schoedel, R., ... Gosling, S. D. (2020). Personality sensing for theory development and assessment in the digital age. *European Journal of Personality*. doi:10.1002/per.2273. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/per.2273

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, *51*(8), 56–59. doi:10.1109/mc.2018.3191268

Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P., & Kreuter, F. (2019). Willingness to participate in passive mobile data collection. *Public Opinion Quarterly*, *83*(S1), 210–235. doi:10.1093/poq/nfz007

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. doi:10.1037/a0039210

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2018). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*. doi:10.1177/0894439318816389

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, USA: Springer.

Kuhn, M., & Johnson, K. (2020). *Feature Engineering and Selection*. New York, USA: Chapman and Hall/CRC. Retrieved from https://doi.org/10.1201%2F9781315108230

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., ... Alstyne, M. V. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Mahmoodi, J., Leckelt, M., van Zalk, M., Geukes, K., & Back, M. (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. doi:10.1016/j.cobeha.2017.07.001

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-informatics: Big data shaping modern psychometrics. *Medical Hypotheses*, *82*(4), 405–411. doi:10.1016/j.mehy.2013.11.030

Mehrotra, A., Müller, S. R., Harari, G. M., Gosling, S. D., Mascolo, C., Musolesi, M., & Rentfrow, P. J. (2017). Understanding the role of places and activities on mobile phone interaction and usage patterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–22. doi:10.1145/3131901

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

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, *13*(1), 23–47. doi:10.1146/annurev-clinpsy-032816-044949

Mønsted, B., Mollgaard, A., & Mathiesen, J. (2018). Phone-based metric as a predictor for basic personality traits. *Journal of Research in Personality*, *74*, 16–22. doi:10.1016/j.jrp.2017.12.004

Montag, C., Błaszkiewicz, K., Lachmann, B., Andone, I., Sariyska, R., Trendafilov, B., ... Markowetz, A. (2014). Correlating personality and actual phone usage. *Journal of Individual Differences*, *35*(3), 158–165. doi:10.1027/1614-0001/a000139

Montag, C., Duke, É., & Markowetz, A. (2016). Toward psychoinformatics: Computer science meets psychology. *Computational and Mathematical Methods in Medicine*, *2016*, 1–10. doi:10.1155/2016/2983685

Piwek, L., & Ellis, D. A. (2016). Can programming frameworks bring smartphones into the mainstream of psychological science? *Frontiers in Psychology*, *7*(1252). doi:10.3389/fpsyg.2016.01252

RatSWD. (2020). Datenerhebung mit neuer Informationstechnologie. Empfehlungen zu Datenqualität und -management, Forschungsethik und Datenschutz. *RatSWD Output 6 (6). Berlin, Rat für Sozial- und Wirtschaftsdaten (RatSWD)*. doi:10.17620/02671.47

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*(3), 363–381. doi:10.1002/per.1994

Reeves, B., Robinson, T., & Ram, N. (2020). Time for the human screenome project. *Nature*, *577*(7790), 314–317. doi:10.1038/d41586-020-00032-5

Seifert, A., Hofer, M., & Allemand, M. (2018). Mobile data collection: Smart, but not (yet) smart enough. *Frontiers in Neuroscience*, *12*. doi:10.3389/fnins.2018.00971

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN*. doi:10.2139/ssrn.2694998

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., ... Bühner, M. (2019). Behavioral patterns in smartphone usage predict big five personality traits. *Psyarxiv*. doi:10.31234/osf.io/ks4vd

Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., ... Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*(6), 701–722. doi:10.1002/per.2113

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., ... Bühner, M. (2019). Personality research and assessment in the era of machine learning. *Psyarxiv*. doi:10.31234/osf.io/efnj8

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. doi:10.1177/1745691616658637

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., ... 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. doi:10.1145/2632048.2632054

Wang, W., Harari, G. M., Wang, R., Müller, S. R., Mirjafari, S., Masaba, K., & Campbell, A. T. (2018). Sensing behavioral change over time. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–21. doi:10.1145/3264951

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. doi:10.1002/per.1986

Yarkoni, T. (2012). Psychoinformatics. *Current Directions in Psychological Science*, *21*(6), 391–397. doi:10.1177/0963721412457362

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