

Anna Schnauber-Stockmann; Teresa K. Naab; Daniela Schlütz; Veronika Karnowski

Wiederholte In-situ-Befragungen: Experience Sampling und verwandte Forschungsdesigns in der Kommunikationswissenschaft. Ein Open-Access-Lehrbuch zu Grundlagen, Durchführung und Auswertung kommunikationswissenschaftlicher In-situ-Studien

2026

<https://doi.org/10.25969/mediarep/25266>

Veröffentlichungsversion / published version

Empfohlene Zitierung / Suggested Citation:

Nutzungsbedingungen:

Dieser Text wird unter einer Creative Commons - Namensnennung - Weitergabe unter gleichen Bedingungen 4.0/ Lizenz zur Verfügung gestellt. Nähere Auskünfte zu dieser Lizenz finden Sie hier:

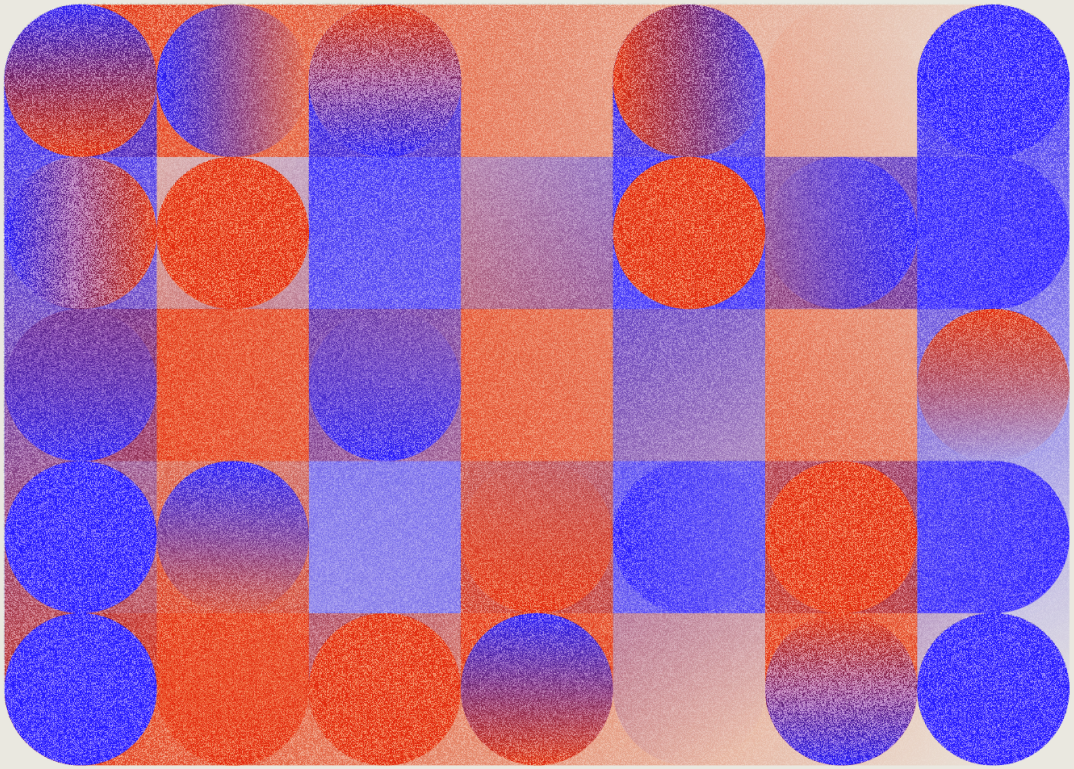
<https://creativecommons.org/licenses/by-sa/4.0/>

Terms of use:

This document is made available under a creative commons - Attribution - Share Alike 4.0/ License. For more information see:

<https://creativecommons.org/licenses/by-sa/4.0/>

Anna Schnauber-Stockmann
Teresa K. Naab
Daniela Schlütz
Veronika Karnowski



WIEDERHOLTE IN-SITU BEFRAGUNGEN

Ein Open Access Lehrbuch zu Grundlagen, Durchführung und
Auswertung kommunikationswissenschaftlicher In-situ-Studien

**Wiederholte
In-situ-Befragungen:
Experience Sampling und
verwandte
Forschungsdesigns in der
Kommunikations-
wissenschaft**

**Ein Open-Access-Lehrbuch zu Grundlagen,
Durchführung und Auswertung
kommunikationswissenschaftlicher In-situ-Studien**

Anna Schnauber-Stockmann Teresa K. Naab
Daniela Schlütz Veronika Karnowski

©2026 Anna Schnauber-Stockmann, Teresa K. Naab, Daniela Schlütz & Veronika Karnowski

Dieses Werk ist unter der Lizenz „Attribution-ShareAlike 4.0 International“ (CC BY-SA 4.0) lizenziert. Diese Lizenz erlaubt es Ihnen, das Material für jeden Zweck, zu vervielfältigen, weiterzuverbreiten, zu verändern und darauf aufzubauen, sofern die Urheberinnen genannt werden, jedoch nicht in einer Weise, die den Eindruck erweckt, dass diese Sie oder Ihre Nutzung des Werks befürworten. Die Namensnennung sollte folgende Angaben enthalten:

Schnauber-Stockmann, Anna, Naab, Teresa K., Schlütz, Daniela & Karnowski, Veronika (2026). *Wiederholte In-situ-Befragungen: Experience Sampling und verwandte Forschungsdesigns in der Kommunikationswissenschaft: Ein Open Access Lehrbuch zu Grundlagen, Durchführung und Auswertung kommunikationswissenschaftlicher In-situ-Studien*. <https://doi.org/10.25969/mediarep/25266>

Die Online-Version des Buches sowie der Anhang sind unter <https://lehrbuch-insitu.de> abrufbar.

Verlag: Anna Schnauber-Stockmann, c/o Johannes Gutenberg-Universität Mainz, Institut für Publizistik, Jakob-Welder-Weg 12, 55099 Mainz, anna.schnauber-stockmann@uni-mainz.de

Herstellung: epubli – ein Service der neopubli GmbH, Köpenicker Straße 154a, 10997 Berlin

Kontaktadresse nach EU-Produktsicherheitsverordnung: produktsicherheit@epubli.com

Umschlaggestaltung: Tinett Kähler

Satz und technische Umsetzung: Erstellt mit Quarto und R

Ziel des Buches

Ziel dieses Lehrbuchs ist es, eine anschauliche und praxisorientierte Einführung in In-situ-Designs für die Kommunikations- und Medienwissenschaft (KMW) zu bieten. Damit meinen wir Befragungen, die sich auf die aktuelle Situation der Teilnehmer:innen beziehen und nicht nur einmalig, sondern wiederholt im Alltag ausgefüllt werden. Leser:innen werden aufbauend auf den theoretischen Grundlagen an die Prinzipien der Methodik herangeführt und lernen, die Aussagekraft und Belastbarkeit von In-situ-Daten einzuschätzen. Darüber hinaus versetzt sie das Buch unter Rückgriff auf zahlreiche Praxisbeispiele aus verschiedenen kommunikations- und medienwissenschaftlichen Anwendungskontexten in die Lage, eigene In-situ-Studien durchzuführen, und zwar von der Entwicklung einer geeigneten Forschungsfrage und deren Dimensionalisierung über die Konstruktion des Erhebungsinstrumentes bis hin zur Datenaufbereitung und -analyse und Ergebnisinterpretation. Das Lehrbuch ist als kompakter, didaktisch aufbereiteter Einstieg für Studierende, Promovierende und etablierte Forschende konzipiert. Gleichzeitig kann das Buch Forschenden, die bereits Erfahrung mit In-situ-Designs haben, als hilfreiche Ressource dienen, indem es den aktuellen Wissensstand zum Einsatz von In-situ-Verfahren für kommunikations- und medienwissenschaftliche Fragestellungen sammelt und sichert.

Aufbau des Buches

Das Buch gliedert sich in elf Kapitel, die in ihrer Gesamtheit den Forschungsprozess abdecken, und wird durch einen Anhang mit Steckbriefen exemplarischer Studien, dem Auswertungscodex in R sowie hilfreichen Ressourcen ergänzt. Dieser ist unter <https://lehrbuch-insitu.de> abrufbar. Wir empfehlen insbesondere Leser:innen, die das erste Mal eine eigene In-situ-Studie durchführen, die Lektüre des gesamten Buches, um den Forschungsablauf insgesamt zu erfassen. Durch die umfassende Auseinandersetzung mit jedem Thema werden die Leser:innen in die Lage versetzt, eine eigene In-situ-Studie

durchzuführen. Hinweise auf weiterführende Literatur ergänzen die Ausführungen an geeigneten Stellen und bieten die Möglichkeit für eine vertiefte Auseinandersetzung.

Wir beginnen mit einer kontextualisierenden Einführung, in deren Rahmen wir auf die zunehmende Komplexität aktueller Medienökosysteme und die damit einhergehenden methodischen Herausforderungen eingehen (Abschnitt 1.1), um daran anschließend die Entwicklung von In-situ-Designs (Abschnitt 1.2) sowie ihre zentralen Charakteristika (Abschnitt 1.3) und die Abgrenzung zu anderen Verfahren (Abschnitt 1.4) darzustellen. Kapitel 2 stellt die Forschungslogik von In-situ-Designs vor und grenzt die Varianten von In-situ-Designs – Tagebuchstudien und Experience Sampling-Studien – voneinander ab. In Kapitel 3 widmen wir uns typischen Forschungsfragen auf Personen- (Abschnitt 3.1) und Situationsebene (Abschnitt 3.2), die mit In-situ-Studien beantwortet werden können. Kapitel 4 geht auf rechtliche und ethische Aspekte im Allgemeinen (Abschnitt 4.1) und von In-situ-Studien im Besonderen (Abschnitt 4.2) ein. Kapitel 5 erörtert, wie In-situ-Designs in verschiedene Untersuchungsanlagen integriert werden können – dies umfasst kombinierte Designs (Abschnitt 5.1), Measurement-Burst-Designs (Abschnitt 5.2) sowie Interventions- und Experimentaldesigns (Abschnitt 5.3). Kapitel 6 befasst sich mit Stichprobenziehung: Neben einer näheren Betrachtung von Stichprobengröße, Erhebungsdichte und -zeitraum (Abschnitt 6.1) gehen wir auf die Besonderheiten der Stichprobenziehung auf zwei Ebenen, nämlich Personen- (Abschnitt 6.2) und Situationsstichprobe (Abschnitt 6.3) ein. Kapitel 7 behandelt Konstruktionsprinzipien der Erhebungsinstrumente (Abschnitt 7.1) sowie Herausforderungen der praktischen Umsetzung (Abschnitt 7.2). Kapitel 8 befasst sich mit der Durchführung einer In-situ-Studie und geht dabei auf den Pretest (Abschnitt 8.1), die Incentivierung (Abschnitt 8.2) sowie die Schulung, die Feldphase und das Debriefing (Abschnitt 8.3) ein. Die Auswertung einer In-situ-Studie wird ausführlich in Kapitel 9 behandelt. Dabei erklären wir die hierarchische Datenstruktur (Abschnitt 9.1), das Datenmanagement (Abschnitt 9.2) und grundlegende Auswertungsmöglichkeiten (Abschnitt 9.3) so-

wie spezifische Auswertungen auf der Personen- (Abschnitt 9.4) und Situationsebene (Abschnitt 9.5). Kapitel 10 gibt nach deiner Einführung zu Open-Science-Praktiken (Abschnitt 10.1) Hinweise zu Präregistrierung (Abschnitt 10.2) und Reporting (Abschnitt 10.3). Das finale Kapitel 11 enthält eine Checkliste für die eigene Studienplanung und -durchführung.

Didaktische Elemente

Folgende didaktische Elemente erschließen den Text:

- Wichtige Begriffe und Definitionen („MERKE“) werden gerahmt erläutert.
- Durch im Online-Anhang zum Buch enthaltene Studiensteckbriefe illustrierte Fallbeispiele veranschaulichen komplexe Sachverhalte und die praktische Umsetzung von In-situ-Studien in der KMW.
- Zudem illustriert ein durchlaufendes Praxisbeispiel zur Messung der Nutzung von True-Crime-Podcasts in einer In-situ-Studie die theoretischen Ausführungen (Schlütz et al., 2027). Der zugehörige Studiensteckbrief sowie alle Materialien, Daten etc. finden sich im Online-Anhang zum Buch.
- Exkurse geben weiterführende Hinweise.

Danksagungen

Unser Dank für die Idee, Konzeption und Erstellung des Buches gilt allen Mitgliedern des In-situ-Netzwerks sowie insbesondere Michael Scharkow und einer weiteren Reihe von Testleser:innen für ihr wertvolles Feedback. Weiterhin danken wir Michael Scharkow und Felix Meissner für ihre Unterstützung bei der Umsetzung in Quarto.

Inhaltsverzeichnis

1	Einführung	9
1.1	Komplexe Medienumgebungen und die Untersuchung von Situationen	9
1.2	Historie von In-situ-Verfahren	13
1.3	Zentrale Charakteristika von In-situ-Designs	15
1.4	Abgrenzung zu anderen Erhebungsverfahren	18
2	Forschungslogik und Typen von In-situ-Designs	22
3	Typen von Forschungsfragen	29
3.1	Unterschiede zwischen Personen: Fragen nach inter-individueller Varianz	29
3.2	Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz	35
3.2.1	Beschreibung der situativen Variation eines Konstrukts und zeitliche Dynamiken	36
3.2.2	Zusammenhänge zwischen mehreren situativ variierenden Konstrukten	38
3.2.3	Zeitversetzte Zusammenhänge und Autokorrelationen	40
3.2.4	Fragen zum Zusammenspiel von Personen- und Situationsmerkmalen	41
3.2.5	Klassen-/Typenbildung von Personen und Situationen	43
4	Ethische und rechtliche Aspekte von In-situ-Studien	44
4.1	Leitprinzipien forschungsethischer Reflexion	45
4.2	Forschungsethische Besonderheiten von In-situ-Studien	49

5	In-situ-Designs in verschiedenen Untersuchungsan-	56
	lagen	
5.1	Kombination von In-situ-Designs mit anderen Erhe-	56
	bungsmethoden	
5.2	Measurement-Burst-Designs	61
5.3	Interventions- und Experimentaldesigns	62
6	Ziehung von Personen- und Situationsstichproben	65
6.1	Stichprobengröße, Erhebungsdichte und -zeitraum .	66
6.2	Ziehung der Personenstichprobe	69
	6.2.1 Grundlagen	69
	6.2.2 Praktische Umsetzung und Herausforderungen	74
6.3	Ziehung der Situationsstichprobe	80
	6.3.1 Grundlagen	80
	6.3.2 Praktische Umsetzung und Herausforderungen	86
7	Konstruktion der Erhebungsinstrumente	96
7.1	Grundlagen der Fragebogenkonstruktion	96
	7.1.1 Allgemeine Prinzipien der Operationalisierung	97
	7.1.2 Spezifika des In-situ-Protokolls	101
7.2	Herausforderungen für die praktische Umsetzung . .	107
	7.2.1 Reaktivität	107
	7.2.2 Mangelnde Qualität einzelner Antworten . . .	110
8	Durchführung einer In-situ-Studie	113
8.1	Pretest	113
8.2	Incentivierung	114
8.3	Schulung, Feldphase und Debriefing	117
	8.3.1 Informationen im Vorfeld der Studie	117
	8.3.2 Betreuung während der Feldphase und Debrief-	
	ing	120
9	Auswertung einer In-situ-Studie	123
9.1	Struktur von In-situ-Daten	124

Inhaltsverzeichnis

9.2	Datenmanagement	128
9.2.1	Dateninspektion, -prüfung und -bereinigung .	128
9.2.2	Datenaufbereitung	137
9.3	Grundlegende Auswertungen	149
9.4	Unterschiede zwischen Personen: Fragen nach inter-individueller Varianz	149
9.5	Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz	152
9.5.1	Deskriptive Auswertungen	152
9.5.2	Mehrebenenanalysen zur Auswertung von In-situ-Daten	153
9.5.3	Klassen-/Typenbildung von Situationen und Personen	170
10	Open Science und Reporting	174
10.1	Einführung	174
10.2	Präregistrierung	176
10.3	Reporting-Standards und Veröffentlichung von Studienmaterialien und -daten	179
11	Checkliste	185
	Literaturverzeichnis	194

1 Einführung

1.1 Komplexe Medienumgebungen und die Untersuchung von Situationen

Auf dem Weg zur Arbeit in der S-Bahn einen unterhaltsamen Podcast hören, dabei auf der Suche nach einem Geburtstagsgeschenk durch Instagram scrollen und parallel eine private WhatsApp-Konversation im Gruppen-Chat führen, anschließend am Büroschreibtisch berufliche E-Mails lesen, während diverse Kurznachrichtendienste Meldungen zum aktuellen politischen Geschehen pushen, in der Mittagspause die letzte Folge der Lieblingsserie auf dem iPad verfolgen – das sind heutzutage gewöhnliche Mediennutzungssituationen. Sie sind geprägt von flächendeckender Verbreitung unterschiedlicher Medienangebote, ihrer ständigen Verfügbarkeit und mobilen Nutzungsmöglichkeiten, konvergenter Technik, flüchtigen Inhalten und zum Teil sehr kurzen Nutzungsepisoden. Mit anderen Worten: Sie zeichnen sich durch hohe situative Unterschiedlichkeit aus. Die Nutzung variiert zwischen aktiven und passiven, rezipierenden und produzierenden, synchronen und asynchronen Formen der Kommunikation sowie zwischen auditiven, audiovisuellen und visuellen Modi. Sie erstreckt sich über eine Vielzahl von Plattformen und umfasst professionelle ebenso wie nutzergenerierte Inhalte. Die Nutzenden selbst befinden sich in wechselnden räumlichen, zeitlichen und sozialen Kontexten. Damit entsteht eine noch nicht dagewesene Vielfalt von Nutzungsmustern, die situativ erheblich variieren.

Auswahl, Rezeption und Wirkung von Medien beruhen damit nicht nur auf den individuellen Merkmalen und Präferenzen der nutzenden Person, sondern maßgeblich auch auf den situativen Bedingungen (Karnowski et al., 2019): Ob uns ein True-Crime-Podcast

1 Einführung

in die Geschichte transportiert, hängt nicht nur von unseren Genvervorlieben oder von unserer grundsätzlichen Disposition ab, uns in Geschichten zu versenken (der sog. Transportabilität, Dal Cin et al., 2004) , sondern z. B. auch von der Länge der Nutzungsepisode, unserer momentanen Aufmerksamkeit und dem Ablenkungspotenzial durch Umgebungsgeräusche oder Parallelbeschäftigungen. Für eine empirisch adäquate Abbildung dieser komplexen Mediennutzungssituationen ist es nötig, auf Messverfahren zurückzugreifen, die die Situation erfassen können. Darum geht es in diesem Buch: Im Fokus steht die Untersuchung von Mediennutzungssituationen, die sowohl von externen Kontextmerkmalen wie dem Aufenthaltsort, den medienseitigen Kontextmerkmalen, etwa dem verwendeten Abspielgerät, als auch von internen Aspekten der Nutzenden wie der psychischen Verfassung, also z. B. von ihrer aktuellen Stimmung, geprägt sind. Auf diese Gemengelage muss die wissenschaftliche Forschung angemessen reagieren, um das Mediennutzungsverhalten valide und reliabel abzubilden. Eine Möglichkeit besteht darin, wiederholt einzelne Situationen möglichst unmittelbar während ihres Geschehens zu erfassen – daher der Begriff „in-situ“ –, statt sich auf rekonstruierende Aussagen der Nutzer:innen zu verlassen (vgl. Naab et al., 2019; Shalev et al., 2025).

Ein solcher methodischer Zugang ermöglicht die Bearbeitung von Forschungsfragen nach situativen Unterschieden in der Mediennutzung – nicht nur, aber auch im Verhältnis zu interindividuellen Differenzen. Dieser Aspekt ist sowohl theoretisch als auch methodisch von entscheidender Bedeutung (Schnauber-Stockmann, Scharnow, et al., 2025). Abbildung 1.1 verdeutlicht dies: Die linke Seite der Abbildung zeigt, dass sich Mediennutzung zum einen zwischen Personen unterscheiden kann (Personenebene). So haben manche Menschen z. B. eine grundsätzliche Vorliebe für Podcasts (symbolisiert durch die Kopfhörer), andere aber nicht. Die rechte Seite illustriert die Situationsebene: Die Person in der obersten Zeile ist z. B. kein Podcastfan (links) und nutzt dieses Medium auch nie (rechts), während die Person in der zweiten Zeile eine ausgeprägte Vorliebe hat, die sie im Tagesverlauf konstant auslebt. Bei diesen beiden Personen zeigt

1.1 Komplexe Medienumgebungen und die Untersuchung von Situationen

sich keine Variation auf der Situationsebene – die eine Person hört nie, die andere hört immer Podcasts. Anders sieht es bei den Personen in der dritten und vierten Zeile aus: In beiden Fällen variiert die Podcastnutzung je nach Situation.

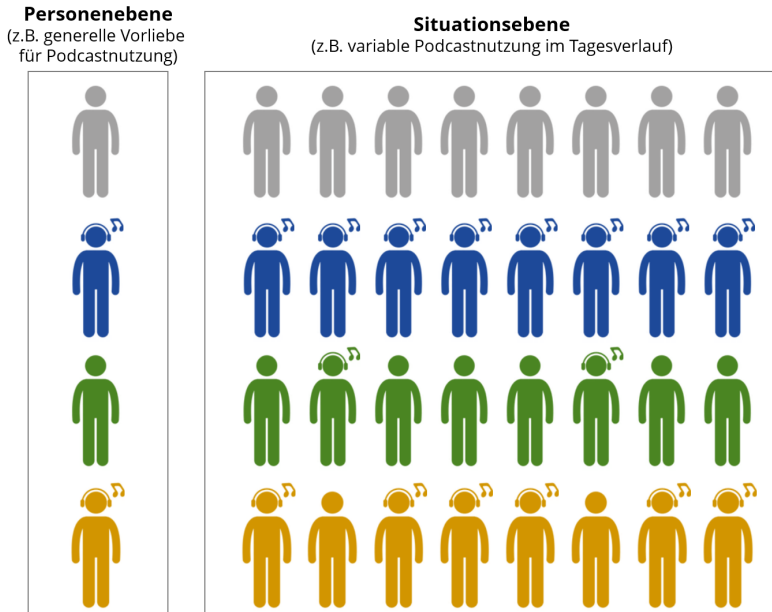


Abbildung 1.1: Personen- und situationsabhängige Varianz (inter-/intraindividuell)

Studien können so angelegt werden, dass die Personenebene, also die sog. **interindividuelle Varianz** im Vordergrund steht. Dieser Fall liegt bei herkömmlichen querschnittlichen Befragungen vor, in denen Personen hinsichtlich ihrer Kognitionen, Emotionen oder Verhaltensweisen miteinander verglichen werden. Situative Besonderheiten spielen i. d. R. keine Rolle, die Abfrage erfolgt generalisiert („im Durchschnitt“, „üblicherweise“, „im Allgemeinen“). Fragt man sich

aber, ob die Auswahl eines bestimmten Podcasts neben generellen persönlichen Präferenzen (also z. B. der Vorliebe für das Genre True Crime) auch mit externen oder internen situativen Gegebenheiten (also z. B. dem Pendeln in einer voll besetzten S-Bahn oder dem akuten Bedürfnis nach ein paar Minuten Ruhe) zusammenhängt, geht es um die **intraindividuelle Varianz**. Hier spielt neben der Personen- auch die Situationsebene eine Rolle für das Verständnis von Medien-nutzung und situativ variierende Faktoren müssen miterfasst werden, und zwar wiederholt. Teilnehmer:innen an solchen Studien werden über einen längeren Zeitraum mehrmals gebeten, über ihre aktuelle Mediennutzung, Stimmung, ihren Aufenthaltsort etc. zu berichten. So erhalten die Forschenden einen Einblick in den (vielfältigen) Alltag der Teilnehmer:innen. Forschungsfragen nach der intraindividuellen Varianz stehen im Mittelpunkt dieses Buches.

Forschungsfragen, die die intraindividuelle Varianz berücksichtigen oder beide Ebenen adressieren, müssen mit spezifischen Forschungsdesigns bearbeitet werden (Bolger & Laurenceau, 2013; Hektner et al., 2007; Kubey & Csikszentmihalyi, 1990; Mehl & Conner, 2012; Myin-Germeyns & Kuppens, 2022b). Wir verwenden dafür den Oberbegriff **In-situ-Studiendesigns** und fokussieren dabei auf längsschnittliche Designs, in denen die Teilnehmer:innen wiederholt aufgefordert werden, entweder über ihre momentane situative Verfasstheit oder über einen spezifischen, aktuellen Zeitraum Auskunft zu geben.

In der KMW gibt es unterschiedliche Forschungstraditionen, die diese Verfahren anwenden und die wiederholt Selbstberichte über das alltägliche Leben der Teilnehmer:innen erheben (vgl. Schnauber-Stockmann & Karnowski, 2020). Darunter fallen zum einen Verfahren, die Bezeichnungen wie Experience Sampling Method (ESM, z. B. Larson & Csikszentmihalyi, 1983) bzw. aktueller Mobile Experience Sampling Method (MESM, vgl. Berkel et al., 2017), Ecological Momentary Assessment (EMA, z. B. Stone et al., 2023), Intensive Longitudinal Methods (z. B. Bolger & Laurenceau, 2013) oder Ambulatory Assessment (AA, z. B. Roekel et al., 2019) tragen. Diese Begriffe werden in der Methodenliteratur zwar weitgehend synonym verwen-

det, stammen jedoch aus unterschiedlichen Forschungstraditionen und setzen teilweise unterschiedliche Schwerpunkte – etwa stärker psychologisch (ESM), klinisch (EMA) oder mit breiterem Einbezug physiologischer Messungen (AA). Wir haben uns in Anlehnung an Schnauber-Stockmann und Karnowski (2020) entschieden, den Begriff **Experience Sampling Method (ESM)** zu verwenden, weil dieser in der KMW am weitesten verbreitet ist. Außerdem gehören zu In-situ-Studiendesigns auch Verfahren, die als Diary Method (z. B. Hoplamazian et al., 2018) bzw. Tagebuch bezeichnet werden. Sie unterscheiden sich von ESM-Designs u. a. durch ihren etwas schwächer ausgeprägten In-situ-Charakter (siehe Kapitel 2). Obwohl viele Aspekte in diesem Buch auch auf Tagebuchstudien angewendet werden können, liegt unser Fokus auf ESM-Designs. Darüber hinaus gibt es eine erhebliche Heterogenität der Vorgehensweisen, die wir im Verlauf des Buches ausführlich erklären.

1.2 Historie von In-situ-Verfahren

Die Geschichte situativer Feldforschungsmethoden lässt sich bis zu Brunswik (1955) und Lewin (1939) zurückverfolgen. Schon diesen Pionieren ging es im Kern darum, die Forschung sowohl räumlich als auch zeitlich enger an die Untersuchungsgegenstände zu binden. Sie verließen daher die Labore, um ihre Daten in der alltäglichen Lebenswelt zu erheben und so die externe Validität der Studien zu erhöhen. Auch in aktuellen In-situ-Studien geht es um die situative Kontextualisierung der untersuchten Konstrukte und darum, ein möglichst repräsentatives Abbild menschlicher Erfahrungen innerhalb eines bestimmten Zeitraums zu ermitteln. Empirisch umgesetzt wird dies durch wiederholte Selbstberichte, entweder zu bestimmten, (häufig) zufällig gewählten Zeitpunkten oder anlässlich spezifischer Ereignisse (etwa der Nutzung einer Onlineplattform wie Spotify). Die Auslösung der Selbstberichte fand früher mithilfe eines Funkempfängers statt (z. B. ein Pager, vgl. Schlütz, 2002), heute wird dazu i. d. R. das Smartphone eingesetzt. Die Teilnehmer:innen werden

1 Einführung

gebeten, möglichst unmittelbar im Anschluss an ein entsprechendes Signal oder ein Ereignis einen kurzen Fragebogen auszufüllen. Auf diese Weise wird über die Zeit eine große Menge an situativen Daten von einer Reihe von Personen gesammelt.

Aufbauend auf dieser Grundidee waren In-situ-Designs Ende der 1970er Jahre zunächst vor allem in der Psychologie verbreitet (im Überblick vgl. Hektner et al., 2007). Vorreiter waren Csikszentmihalyi und Kolleg:innen, die ESM systematisch in der psychologischen Forschung einführten (vgl. z. B. Larson & Csikszentmihalyi, 1983). Im Fokus dieser Arbeiten standen sog. Flow-Erlebnisse im Alltag, also die Erfahrung von Selbstvergessenheit, die durch eine ideale Balance von Anforderung und Kompetenz ausgelöst wird (vgl. Csikszentmihalyi & Larson, 2014). Der Einsatz situativer Messungen diente dazu, besser zu verstehen, wie Menschen im Alltag funktionieren, was sie denken, fühlen und wie sie in bestimmten Situationen handeln. Csikszentmihalyi stellt hierzu fest: „One way to think about ESM is that it tries to provide psychology with a method and a theoretical perspective that makes it possible to study whole individuals functioning in their everyday environments, both as behaving organisms and as acting, conscious beings” (Hektner et al., 2007, S. 292). Obschon sich diese methodische Herangehensweise in der psychologischen Forschung früh etablierte, waren Medien lange Zeit kein Forschungsthema. Zu den Ausnahmen zählen eine Studie von Csikszentmihalyi und Kubey (1981) zur Nutzung des Fernsehens im Alltag und eine Arbeit von Kubey und Larson (1996) über die Nutzung neuer Medien durch Kinder und Jugendliche. In die KMW wurde das Forschungsdesign erst deutlich später eingeführt (Schlütz & Scherer, 2001). In diesen frühen Studien ging es vorwiegend um die Qualität von Erlebnissen und den Zusammenhang mit persönlichen Eigenschaften und bestimmten Situationskontexten. Erforscht wurden beispielsweise die Erlebnisqualität von Gaming (Schlütz, 2002) oder mobile Fernsehnutzung im Alltag (Karnowski & Von Pape, 2009). Heute werden ESM-Studien in der KMW für ein deutlich breiteres Themenfeld eingesetzt. Zahlreiche KMW-Studien befassen sich z. B. mit den situativen Bedingungen der Mediennutzung, etwa mit dem

1.3 Zentrale Charakteristika von In-situ-Designs

Ablenkungspotenzial von Smartphones (Siebers et al., 2024), den Effekten der Social-Media-Nutzung auf das situative Selbstwertgefühl (Desjarlais, 2024) oder Prokrastination (Meier, 2022). Weit verbreitet sind In-situ-Designs auch in der positiven Psychologie, wenn es um den Einfluss mobiler Mediennutzung auf das Wohlbefinden geht (z. B. Beyens et al., 2024; Johannes et al., 2021; Karsay et al., 2023; Meier, 2022). Eine Systematisierung von Forschungsfragen bietet Kapitel 3.

Rasant wächst in jüngerer Zeit die Anzahl der Studien, die In-situ-Designs einsetzen (Myin-Germeyns & Kuppens, 2022a). Stone et al. (2023) beziffern die Gesamtzahl bis zum Jahr 2023 auf über 1.700 Publikationen, allein in den Verhaltenswissenschaften und im Gesundheitswesen (vgl. auch Gabriel et al., 2018 für die Organisationswissenschaft). Mit der wachsenden Anzahl der Studien differenziert sich das Forschungsfeld sowohl begrifflich als auch hinsichtlich methodischer Spezifika (Schnauber-Stockmann & Karnowski, 2020). Gemein ist aber allen In-situ-Designs, dass sie in der aktuellen Situation – d. h. in der natürlichen Umgebung und mit geringem zeitlichem Abstand – Kognitionen, Emotionen und Verhalten sowie den externen Kontext erfassen. Das Ziel lautet: “to capture daily life as it is directly perceived from one moment to the next” (Hektner et al., 2007, S. 6). Es geht also darum, die für die jeweilige Forschungsfrage zentralen Variablen auf der Situationsebene zu erfassen.

1.3 Zentrale Charakteristika von In-situ-Designs

Die Grundidee der situativen Messung wird in unterschiedlichen Disziplinen und Forschungstraditionen auf vielfältige Weise umgesetzt. Die Ansätze zeichnen sich durch eine Gemeinsamkeit aus: „people’s current experiences and behaviors are assessed repeatedly, in their natural environments, without or only with minimal latency” (Wilhelm et al., 2012, S. 62). Daraus ergeben sich spezifische Vor- und Nachteile, die hier nur kurz skizziert und in den Folgekapiteln ausführlicher behandelt werden. Die zentralen Vorteile sind:

1 Einführung

- In-situ-Studien erfassen relevante situative Merkmale im Alltag selbst. Dadurch weisen sie eine hohe ökologische Validität auf – sie zeichnen sich durch ihre realitätsnahe Messung aus – gerade im Unterschied zu Laborexperimenten (siehe Abschnitt 1.4).
- Die Validität wird zudem durch Messwiederholung verbessert: Einmalige In-situ-Messungen haben mit der Einschränkung zu kämpfen, dass die erhobene Situation unter Umständen nicht typisch ist. Wiederholte Messungen decken verschiedene Situationen ab – im Idealfall sogar ein repräsentatives Abbild aller Situationen (siehe Kapitel 2 und Abschnitt 6.3).
- Die situative Messung bringt Mediennutzung und Erhebung zeitlich enger zusammen, was die Reliabilität von Selbstberichtbefragungen verbessert, weil Erinnerungsverzerrungen und heuristische Antworteffekte minimiert werden (Naab et al., 2019; Shalev et al., 2025). Häufig sind alltägliche, „banale“ Nutzungsepisoden von Interesse, die sich retrospektiv schwer erfassen lassen, weil sie oft nicht so bewusst geschehen, dass sie verlässlich zu erinnern sind und berichtet werden können. Effekte selektiver Erinnerung begünstigen wichtige oder besondere Nutzungsepisoden, die aber tendenziell eher selten vorkommen. In-situ-Methoden, die die Mediennutzung wiederholt und situativ erfragen, können solche kognitiven Verzerrungen minimieren (siehe Kapitel 2).
- Schließlich sind situative Längsschnittdesigns mit vielen Messzeitpunkten geeignet, die Dynamik interessierender Phänomene zu erfassen. Durch die hohe zeitliche Auflösung der Messungen können Entwicklungen beobachtet und zeitliche Zusammenhänge abgebildet werden. Die Befunde sind insbesondere für die Medienwirkungsforschung von Interesse (siehe Kapitel 3).

Diese Vorteile gehen mit spezifischen Nachteilen von In-situ-Designs einher (vgl. Masur, 2019b; Schlütz & Scherer, 2001; Scollon

et al., 2009):

- Die Sicherstellung der Güte der Stichprobe hinsichtlich Größe und Repräsentativität ist eine große Herausforderung bei In-situ-Studien. Diese Einschränkung gilt sowohl für die Auswahl auf der Personen- als auch auf der Situationsebene. Denn der Aufwand für die Teilnehmer:innen ist hoch und die Teilnahmebereitschaft entsprechend gering (Smyth et al., 2021) (siehe Kapitel 6 und Kapitel 8).
- Wie alle Befragungen, die auf Selbstauskunft beruhen, ist die Methode reaktiv, d. h. das Messverfahren beobachtet soziale Realität nicht nur, es verändert sie. Dadurch besteht die Gefahr, Forschungsartefakte zu verursachen. Durch die Dauer und Intensität der Selbstbeobachtung bei In-situ-Studien im Längsschnitt steigt die Wahrscheinlichkeit, dass Teilnehmer:innen durch die Messung für ihr Verhalten sensibilisiert werden, was zu systematischen Fehlern führen kann (siehe Abschnitt 6.3.2.2).
- Wie alle Längsschnittdesigns sind In-situ-Studien auch für die Forschenden aufwändig. Neben dem Aufwand, der mit der Rekrutierung der Teilnehmer:innen und der Betreuung während der Feldzeit einhergeht (siehe Abschnitt 6.2.2.2 und Kapitel 8), müssen je nach Fragestellung auch die Besonderheiten der Datenaufbereitung und -auswertung bedacht werden (siehe Kapitel 9).

Diese Herausforderungen gilt es zu bedenken, wenn Forschende eine In-situ-Studie planen. Wir werden auf sie und den Umgang damit im Verlauf des Buches genauer eingehen. Abhängig vom jeweiligen Forschungsinteresse können die geschilderten Vorteile oftmals die Nachteile aufwiegen. Insbesondere für Fragestellungen, die sich auf das Medienhandeln im Alltag, auf situative Unterschiede und Zeitverläufe beziehen, sind In-situ-Designs den herkömmlichen Vorgehensweisen in vielerlei Hinsicht überlegen.

1.4 Abgrenzung zu anderen Erhebungsverfahren

Um ihre Einsatzmöglichkeiten besser zu verstehen, vergleichen wir im Folgenden wiederholte In-situ-Designs mit anderen Studiendesigns und arbeiten ihre besonderen Merkmale heraus.

Eines der am häufigsten gewählten Designs in kommunikations- und medienwissenschaftlichen Studien ist die **Querschnittsbefragung**. Einen ausführlichen Überblick über die Methode geben z. B. Möhring & Schlütz (2025). In solchen Befragungen berichten die Teilnehmer:innen einmalig über ihre Kognitionen, Emotionen und ihr Verhalten. Im Vergleich zu In-situ-Studien ist diese Vorgehensweise mit relativ geringem Aufwand verbunden. Wenn kein Forschungsinteresse an intraindividuellen Veränderungen über die Zeit hinweg besteht, ist das Design für viele Fragestellungen vollkommen ausreichend. Im Rahmen von Querschnittsbefragungen werden die Teilnehmer:innen in der Regel um retrospektive Einschätzungen gebeten (Ha et al., 2015). Solche Messungen sind vor allem für Konstrukte geeignet, bei denen geringe Erinnerungs- und Aggregationsfehler zu erwarten sind (siehe hierzu auch Abschnitt 3.1).

Zudem ist es möglich, im Rahmen einer Querschnittsbefragung nach dem aktuellen Moment (*in situ*) oder einer kurz zurückliegenden Mediennutzungsperiode zu fragen. In einer Studie zur Podcastnutzung könnten wir die Teilnehmer:innen beispielsweise bitten, eine konkrete, kurz zurückliegende Nutzungsperiode zu beschreiben. Dieses Vorgehen geschieht dann allerdings nur einmalig, die intraindividuelle Varianz kann also in Querschnittsbefragungen nicht erfasst werden.

In **Panelbefragungen** beantworten dieselben Personen über einen längeren Zeitraum hinweg mehrfach dieselben Fragen (im Überblick z. B. Eveland & Morey, 2010; Lynn, 2009; Scherer & Naab, 2013). Paneldesigns werden unter anderem eingesetzt, um intraindividuelle Veränderungen im Zeitverlauf zu erfassen. Zum Beispiel werden politische Präferenzen vor und während eines Wahlkampfes sowie nach der Wahl erhoben (z. B. in der Grundlagenstudie *The People's Choice*, Lazarsfeld et al., 1944). So lässt sich die veränder-

liche Zustimmung zu Parteien und Themen nicht nur im Aggregat über alle Wähler:innen erkennen – dafür müsste man lediglich mehrere repräsentative Stichproben der Wähler:innen befragen, aber nicht notwendigerweise wiederholt dieselben Personen. Vielmehr können in Paneldesigens auch intraindividuelle Veränderungen der Präferenzen wie Wählerbewegungen von einer zur anderen Partei untersucht werden (z. B. Zimmermann & Kohring, 2020). Wiederholte In-situ-Befragungen decken sich insofern mit Panelerhebungen, als in beiden Designs dieselben Personen mehrfach Selbstauskünfte geben. Die meisten Panelstudien sind allerdings auf wenige Wiederholungen (sog. Panelwellen) beschränkt, während viele In-situ-Studien eine deutlich größere Anzahl von Messwiederholungen haben. Außerdem liegen die Panelwellen oft einige Wochen, wenn nicht sogar Jahre auseinander, während die Protokolle einer In-situ-Studie häufig mehrfach pro Tag ausgefüllt werden. Der Gesamterhebungszeitraum einer wiederholten In-situ-Studie ist damit auch deutlich kürzer (siehe aber sog. Measurement-Burst-Designs, Abschnitt 5.2). Fragen in Panelerhebungen zielen häufig nicht auf einen aktuellen Moment ab, sondern erfassen weniger fluktuierende Konstrukte, die vermutlich nicht von Moment zu Moment schwanken (z. B. Parteipräferenz).

In-situ-Studien, wie wir sie im vorliegenden Buch beschreiben, messen üblicherweise standardisiert. Das bedeutet, die Forschenden legen vorab einheitliche Verfahrensweisen für den Ablauf der Erhebung fest, die für alle Teilnehmer:innen gelten, insbesondere in Bezug auf die Formulierung und Reihenfolge der Fragen und Antwortmöglichkeiten. Das Ziel ist, die Vergleichbarkeit der Antworten sicherzustellen, so dass die Daten einer Vielzahl von Teilnehmer:innen und mehrerer Erhebungszeitpunkte statistisch verglichen werden können. Auch die Häufigkeit und der Zeitpunkt des Antwortens werden i. d. R. von den Forschenden vorgegeben (siehe Abschnitte 6.1 und 6.3). In täglichen Tagebuchstudien werden die Teilnehmer:innen beispielsweise gebeten, einmal täglich (häufig zu einem vorgegebenen Zeitpunkt, z. B. abends) einen vordefinierten Fragebogen auszufüllen. **Offene bzw. qualitative Tagebücher** (im Überblick z. B. Hyers, 2018; Wagner et al., 2022) haben dagegen ein

1 Einführung

geringeres Maß an Standardisierung. Die Teilnehmer:innen wählen selbst, wann sie etwas eintragen. Auch hinsichtlich der berichteten Erfahrungen sind sie freier. Entsprechend werden die Daten üblicherweise mithilfe qualitativer Methoden analysiert. Der Standardisierungsgrad ist abhängig vom Forschungsinteresse, vom Vorwissen über den Untersuchungsgegenstand, vom Interesse an den subjektiven Bedeutungszuschreibungen durch die Teilnehmer:innen und von der Absicht, die Ergebnisse zu verallgemeinern. Sowohl für offene als auch für hochgradig standardisierte Tagebücher und andere In-situ-Designs besteht die Gemeinsamkeit, dass sie sich auf Selbstberichte der Teilnehmer:innen stützen. Sie fertigen diese Selbstberichte in ihrer natürlichen Umgebung mit geringem Abstand zum berichteten Phänomen an. Der Erhebungszeitraum umfasst außerdem mehrere Tage, so dass auch intraindividuelle Unterschiede sichtbar werden. Ebenso wie quantitative In-situ-Studien verstärkt auf Smartphones als Erhebungstool setzen, stehen auch für offene Tagebücher Smartphoneanwendungen zur Verfügung (z. B. C. Koch et al., 2021; *MeTag* / *MeSoftware*, o. J.).

Bei Verfahren der **digitalen Verhaltensbeobachtung** (im Überblick z. B. Harari et al., 2016) geben die Teilnehmer:innen keine Selbstberichte ab, sondern ein beobachtbares Verhalten wird anhand digitaler Verhaltensspuren gemessen und aufgezeichnet. Diese Verhaltensdaten können unterschiedlicher Natur sein, wie Trace-Daten zur Online- oder Smartphonennutzung (z. B. Log-Files), Mobilitätsdaten, die per GPS von Smartphones oder Wearables aufgezeichnet und übermittelt werden, oder Social Sensing, bei dem per Bluetooth ermittelt wird, in welchen sozialen Konstellationen eine Person sich befindet (siehe z. B. Harari et al., 2016; Harari et al., 2017; Keusch & Kreuter, 2021; Müller et al., 2021; Ohme et al., 2024). Wie bei In-situ-Studien findet die Datenerhebung in natürlicher, alltäglicher Umgebung statt. Den Daten wird aus diesem Grund eine hohe ökologische Validität zugeschrieben. Außerdem werden Limitationen von Selbstauskünften wie Erinnerungsfehler, sozial erwünschtes Antwortverhalten und Aufwand für die Teilnehmer:innen verringert. Allerdings muss sich die Erhebung auf

1.4 Abgrenzung zu anderen Erhebungsverfahren

Konstrukte beschränken, für die es valide apparativ beobachtbare Indikatoren gibt.

2 Forschungslogik und Typen von In-situ-Designs

In-situ-Studien sammeln Erfahrungen über die Realität anhand konkreter Situationen im Alltag von Menschen. Damit legen sie den Fokus auf die direkten und möglichst unmittelbaren Selbstauskünfte über Kognitionen, Emotionen und Verhalten in der natürlichen Umgebung (vgl. Hektner et al., 2007; Larson & Csikszentmihalyi, 1983). In-situ-Studien erfassen nicht nur eine, sondern mehrere Situationen im Alltag, um ein möglichst realistisches Abbild des Alltags zu schaffen (zur Frage der Auswahl dieser Situationen siehe Abschnitt 6.3). Die Erhebung erfolgt möglichst zeitnah, also in situ, um eine zuverlässige Messung des menschlichen Erlebens zu ermöglichen. Beide Aspekte – die Messung im Alltag und die zeitnahe Erhebung – sind wichtige Merkmale von In-situ-Studien, die dazu beitragen, dass sie eine hohe ökologische Validität – eine Facette der externen Validität – aufweisen, also realitätsnahe Ergebnisse liefern.

Exkurs: Externe und interne Validität

Externe Validität bezeichnet die Allgemeingültigkeit der Ergebnisse, ihre Generalisierbarkeit und Repräsentativität. Sie ist abhängig vom gesamten Studiendesign, vor allem aber von der Stichprobenziehung und -realisierung. Davon abzugrenzen ist die interne Validität, also die Möglichkeit, einen strengen Kausalschluss zwischen Ursache und Wirkung zu ziehen. Streng genommen gelingt dieser Anspruch ausschließlich unter künstlichen Bedingungen, also in Laborexperimenten. Interne und externe Validität stehen in einem „hydraulischen Verhältnis“ (T. Koch et al., 2025, S. 57) zueinander: Je stren-

ger die interne Validität kontrolliert wird und Störvariablen ausgeschlossen werden, umso weniger realitätsnah ist eine Studie und umgekehrt.

Beabsichtigen wir also beispielsweise, mehr über die Podcastnutzung junger Erwachsener zu erfahren, so würde eine In-situ-Studie nicht allgemein danach fragen, welche Podcasts die Befragten normalerweise nutzen und wie häufig sie dies tun, sondern die Teilnehmer:innen würden mehrmals zu der aktuellen oder einer kurz zurückliegenden Situation befragt, in der sie einen Podcast gehört haben. Durch diese Art der Erhebung ermöglichen In-situ-Studien die Betrachtung der Vielfältigkeit menschlichen Alltags nicht nur mit Blick auf Unterschiede zwischen verschiedenen Menschen (interindividuelle Unterschiede), sondern auch hinsichtlich der Vielfältigkeit des Verhaltens, der Kognitionen und Emotionen einer einzelnen Person (intraindividuelle Unterschiede).

MERKE

In-situ-Studien erfassen wiederholt und zeitnah per Selbstbericht Alltagssituationen. Sie weisen eine hohe ökologische Validität auf und erlauben Untersuchungen von Unterschieden zwischen Personen (interindividuell) sowie von (zeitlichen) Veränderungen innerhalb von Personen (intraindividuell).

Innerhalb dieser Grundlogik des Studiendesigns finden sich zahlreiche Varianten. Diese Heterogenität wird durch unterschiedliche Begriffsverwendungen noch unübersichtlicher. Oftmals werden die gleichen Vorgehensweisen unterschiedlich bezeichnet oder umgekehrt unterschiedliche Vorgehensweisen mit dem gleichen Begriff belegt. Wenn wir Unterschiede von In-situ-Designs differenzieren, lehnen wir uns in diesem Buch an die Begrifflichkeiten von Schnauber-Stockmann und Karnowski (2020) an, verweisen aber an den jeweiligen Stellen auf alternative Begrifflichkeiten und

2 Forschungslogik und Typen von In-situ-Designs

Einteilungen.

Zunächst lassen sich zwei Haupttypen von In-situ-Studien unterscheiden (vgl. Abbildung 2.1): Tagebuchstudien und ESM-Studien.

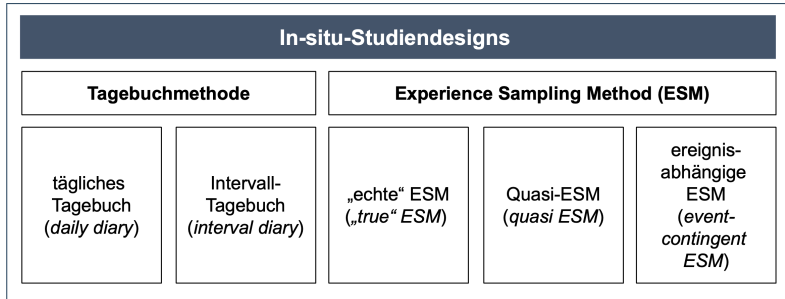


Abbildung 2.1: Typen von In-situ-Studien

Tagebuchstudien lassen sich in tägliche Tagebuch- und Intervalltagebuchstudien unterteilen. Wie es der Name nahelegt, geben die Teilnehmer:innen in täglichen Tagebuchstudien (*daily diaries*) einmal am Tag, meist am Tagesende, Auskunft über ihr Verhalten und ihr kognitives bzw. emotionales Erleben im Laufe dieses Tages oder eines bestimmten Abschnitts des Tages. Üblicherweise bestimmen die Teilnehmer:innen selbst, wann genau sie das Tagebuch ausfüllen. Tägliche Tagebuchstudien erheben also wiederholt (in der Regel an mehreren aufeinanderfolgenden Tagen) bei denselben Teilnehmer:innen Informationen, die sich auf einen bestimmten Tag beziehen, nicht aber auf eine spezifische Situation. In einer Daily-Diary-Studie zur Podcastnutzung würden die Teilnehmer:innen also einmal am Tag zu einem selbst gewählten Zeitpunkt, der in den Abendstunden liegt, berichten, ob sie an diesem Tag Podcasts gehört haben, wenn ja welche, wie lange die Nutzung insgesamt gedauert hat etc. Ein Beispiel für eine tägliche Tagebuchstudie ist die Studie von Schnauber-Stockmann & Naab (2019), die als Studiensteckbrief im Online-Anhang zum Buch vorliegt.

Im Unterschied zu täglichen Tagebuchstudien geben die Teilnehmer:innen in **Intervalltagebuchstudien** (*interval diaries*) – meist über mehrere Tage hinweg – mehrmals täglich Auskunft über ihr Verhalten, ihre Kognitionen und Emotionen. Sie beziehen sich dabei immer auf vordefinierte Zeitabschnitte (Intervalle) innerhalb dieses Tages. Auch in diesen Studien handelt es sich um Selbstauskünfte der Teilnehmer:innen, allerdings nicht in Bezug auf eine spezifische Situation, sondern auf längere Zeitabschnitte im Laufe des Tages. Die Teilnehmer:innen erhalten durch ein Signal jeweils zum Ende des Intervalls die Aufforderung, den Tagebucheintrag vorzunehmen. Im Vergleich zu täglichen Tagebüchern ist der In-situ-Charakter von Intervalltagebuchstudien ausgeprägter, weil die Teilnehmer:innen zeitlich näher zum untersuchten Geschehen Auskunft über ihr entsprechendes Verhalten bzw. ihr kognitives und emotionales Erleben geben. Eine Intervalltagebuchstudie zur Podcastnutzung könnte also beispielsweise vier Zeitabschnitte im Laufe des Tages definieren: 6.00 bis 10.00 Uhr, 10.00 bis 14.00 Uhr, 14.00 bis 18.00 Uhr und 18.00 bis 22.00 Uhr. Über mehrere Tage hinweg würden die Teilnehmer:innen dann jeweils zum Ende eines Intervalls, also um 10.00, 14.00, 18.00 und 22.00 Uhr, aufgefordert, darüber zu berichten, ob sie in den vorhergehenden vier Stunden Podcasts gehört haben und wenn ja, welche sie ausgewählt haben, wie lange die Nutzung insgesamt gedauert hat etc.

Im Unterschied zu Tagebuchstudien, die längere Zeitintervalle bis hin zu einem ganzen Tag abdecken, konzentrieren sich **ESM-Studien** auf eine bestimmte Situation oder einen bestimmten Moment, über den die Teilnehmer:innen Auskunft geben. Sie werden über mehrere Tage hinweg mehrmals täglich aufgefordert, Auskunft über ihr Verhalten, ihre Kognitionen und Emotionen zu geben. Im Unterschied zu Tagebuchstudien interessieren in ESM-Studien jedoch nicht Verhalten, Kognitionen oder Emotionen über einen längeren Zeitabschnitt hinweg, sondern nur das Erleben in einem konkreten Moment oder in einem kurz zurückliegenden Zeitabschnitt. Der In-situ-Charakter dieser Studien ist also nochmals deutlicher ausgeprägt, weil die Selbstauskunft unmittelbarer an das kognitive

bzw. emotionale Erleben der Teilnehmer:innen anschließt.

Auch bei ESM-Studien lassen sich mehrere Unterformen unterscheiden. Wir differenzieren wiederum in Anlehnung an Schnauber-Stockmann und Karnowski (2020; siehe auch Masur, 2019b) drei Unterformen: echte ESM, Quasi-ESM und ereignisabhängiges ESM.

Echte ESM-Studien fordern die Teilnehmer:innen über mehrere Tage hinweg mehrmals täglich zu zufällig ausgewählten Zeitpunkten auf, Auskunft über ihr Verhalten, ihre Kognitionen und Emotionen in exakt diesem Moment zu geben (Kubey et al., 1996). Die Grundidee dieser Form der Datenerhebung besteht darin, durch die zufällige Auswahl von Situationen aus dem Alltag der Teilnehmer:innen ein repräsentatives Abbild dieses alltäglichen Verhaltens, der Kognitionen und Emotionen zu generieren (zur Stichprobenziehung auf Situationsebene siehe Abschnitt 6.3). Das tatsächlich zu messende Phänomen und die Selbstauskunft der Teilnehmer:innen fallen in echten ESM-Studien also zeitlich zusammen. Eine echte ESM-Studie zur Podcastnutzung könnte die Teilnehmer:innen über mehrere Tage hinweg beispielsweise dreimal am Tag alarmieren (zur technischen Umsetzung siehe Abschnitt 6.3.2.1) und fragen, ob sie exakt in diesem Moment einen Podcast nutzen, in welcher Umgebung sie sich gerade befinden, wie sie sich gerade fühlen usw.

Wenn aber die Teilnehmer:innen ausschließlich Auskunft über den aktuellen Moment geben, stellt dies manche Forschungsarbeiten vor Herausforderungen. Denn nicht selten untersucht die KMW nur gelegentlich auftretende Phänomene. Beispielsweise hören die meisten Menschen Podcasts nur in wenigen Momenten des Tages. Wenn wir uns also nur für das kognitive und emotionale Erleben interessieren, während die Befragten Podcasts nutzen (und nicht für Nichtnutzungssituationen), sind nicht alle der zufällig aus dem Alltag gezogenen Situationen für unser Forschungsinteresse relevant. Im Gegenteil, vermutlich werden die Teilnehmer:innen überwiegend berichten, dass sie jetzt gerade in diesem Moment keinen Podcast hören und daher auch keine weiteren Auskünfte geben können, die für unser Forschungsinteresse relevant wären. Aus diesem Grund sind viele erhobene Messungen wertlos.

Quasi-ESM-Studien sollen dieses Problem abmildern, indem die Teilnehmer:innen ebenfalls über mehrere Tage hinweg mehrmals täglich alarmiert werden. Allerdings werden sie nicht aufgefordert, über den aktuellen Moment Auskunft zu geben, sondern über die zuletzt zurückliegenden Situation, die für das Forschungsinteresse relevant ist. Dabei schränken die Studien üblicherweise ein, wie weit diese Situation maximal zurückliegen darf, beispielsweise innerhalb der letzten 30 Minuten, innerhalb der letzten zwei Stunden oder auch seit dem vorhergehenden Signal. Durch diese Einschränkung versuchen Quasi-ESM-Studien den In-situ-Charakter der Datenerhebung, also die möglichst große Nähe zwischen dem untersuchten Phänomen und der Selbstauskunft darüber, zu bewahren, gleichzeitig aber mehr relevante Situationen einzubeziehen. In der Podcastingstudie, die uns in diesem Buch als durchlaufendes Beispiel dient (siehe Studiensteckbrief im Online-Anhang zum Buch), wurden die Teilnehmer:innen über zwölf Tage hinweg viermal täglich zu zufälligen Zeitpunkten alarmiert und um Selbstauskunft gebeten. Sie wurden aufgefordert, an den Zeitraum seit dem letzten Signal zurückzudenken und anzugeben, ob sie in dieser Zeit einen True-Crime-Podcast gehört haben. Falls die Teilnehmer:innen die Frage bejahten, wurden sie gebeten, sich an diese Situation zu erinnern und weitere Fragen dazu zu beantworten.

Im Unterschied zu echten ESM-Studien und Quasi-ESM-Studien stützt sich die Auswahl der zu berichtenden Situationen bei **ereignisabhängigen ESM-Studien** allein auf den Alltag der Teilnehmer:innen. Die Selbstauskünfte finden wiederum über mehrere Tage hinweg zu verschiedenen Zeitpunkten statt. Die Teilnehmer:innen werden aber nicht anhand eines vordefinierten Stichprobenplans (vgl. Abschnitt 6.3) alarmiert, sondern die Selbstauskünfte orientieren sich an Ereignissen im Alltag der Teilnehmer:innen, indem sie z. B. gebeten werden, immer dann ein Protokoll auszufüllen, wenn sie eine Podcastepisode gehört haben. Die Teilnehmer:innen geben also immer dann Auskunft, wenn ein für das Forschungsinteresse relevantes Ereignis auftritt. Ähnlich wie bei echten ESM-Studien fallen bei ereignisabhängigen ESM-Studien das untersuchte Phänomen

2 Forschungslogik und Typen von In-situ-Designs

und die Selbstauskunft nahezu zusammen. Der Unterschied besteht darin, dass der Zeitpunkt, zu dem die Selbstauskunft stattfindet, nicht zufällig ausgewählt wird, sondern sich aus dem Alltag der Teilnehmer:innen ergibt.

MERKE

In-situ-Studien lassen sich in Tagebuch- und ESM-Studien einteilen. Hinsichtlich ihres In-situ-Charakters, also konkret hinsichtlich der zeitlichen Nähe von Situation und Messung, unterscheiden sich die verschiedenen Formen: Während Tagebuchstudien sich rückblickend auf ganze Tage oder Zeitintervalle während eines Tages beziehen, legen ESM-Studien den Fokus auf spezifische Momente im Alltag der Teilnehmer:innen. Der In-situ-Charakter von ESM-Studien ist daher deutlicher ausgeprägt als derjenige von Tagebuchstudien.

3 Typen von Forschungsfragen

Dieses Kapitel befasst sich mit unterschiedlichen Typen von Forschungsfragen, die sich mit In-Situ-Verfahren beantworten lassen (siehe auch Bolger et al., 2003, Tabelle 1; Kuppens & Myin-Germeys, 2022). Wir nennen konkrete Beispiele aus dem Bereich der KMW und beginnen mit Fragen nach der interindividuellen Varianz von Phänomenen. Solche Fragen interessieren sich für die Ausprägungen von Variablen bei einzelnen Personen und Unterschieden zwischen Personen hinsichtlich dieser Merkmale. In-situ-Verfahren sind für ihre Beantwortung nicht zwingend erforderlich, wir werden aber Argumente präsentieren, die für ihren Einsatz sprechen, auch wenn die Daten anschließend nicht situativ ausgewertet werden. Anschließend betrachten wir Fragen nach der intraindividuellen Varianz von Phänomenen. Für solche Untersuchungen sind wiederholte situative Messungen bei denselben Personen zwingend notwendig. Schließlich kommen wir zu Fragen nach Interaktionen zwischen Phänomenen auf der Personen- und Situationsebene.

3.1 Unterschiede zwischen Personen: Fragen nach interindividueller Varianz

Viele Forschungsfragen in der KMW adressieren Unterschiede zwischen Personen, sie untersuchen also interindividuelle Varianz. Sie untersuchen, wie Menschen denken, fühlen und handeln, ohne sich im Detail für die Veränderlichkeit dieser Konstrukte zu interessieren. Vielmehr stehen allgemeine Aussagen über diese Kognitionen, Emotionen und Verhaltensweisen im Zentrum; situative Unterschiede werden nicht berücksichtigt. Die Untersuchungseinheiten dieser Ebene sind entsprechend verschiedene Personen.

Beispielsweise können wir uns dafür interessieren, wie oft und wie

3 Typen von Forschungsfragen

lange Personen am Tag Podcasts nutzen und welche Personen sie länger oder häufiger nutzen als andere. In ähnlicher Weise können wir daran interessiert sein, mit welchen Motiven Menschen Podcasts nutzen und welche Gratifikationen sie ihnen zuschreiben (Chan-Olmsted & Wang, 2022). Zur Beantwortung solcher Fragen benötigen wir nicht zwingend wiederholte Messungen bei denselben Personen; sie können auch mit einer Querschnittsbefragung angegangen werden (z. B. Chan-Olmsted & Wang, 2022; Craig et al., 2023; Perks & Turner, 2019; Samuel-Azran et al., 2019). Allerdings weisen diese einmaligen, oft retrospektiven Messungen eine geringe Validität auf. Teilnehmer:innen müssen zunächst alle interessierenden Ereignisse in der Vergangenheit erinnern und über die Zeit aggregieren, wenn sie in Befragungen rückblickend Auskunft geben sollen. Vergangenes Verhalten, kognitives und emotionales Erleben sind – verglichen z. B. mit aktuellen Ereignissen – jedoch schwierig zu erinnern (Bradburn et al., 1987; Krosnick, 1991; Nisbett & Wilson, 1977; Prior, 2009). Die kognitive Zugänglichkeit ist vor allem bei hoch frequenten und alltäglichen Ereignissen gering, die einzelnen Ereignisse „blend into one global knowledge-like representation that lacks specific time or location markers“ (Schwarz & Oyserman, 2001, S. 137). Deshalb wenden Befragte häufig Inferenz- und Schätzstrategien an, um ihre Antwort herzuleiten (Schwarz & Oyserman, 2001). Diese Strategien sind v. a. schwierig und fehleranfällig, wenn Ereignisse stark über die Zeit variieren (Lee et al., 2008) oder häufig vorkommen (Blair & Burton, 1987; Burton & Blair, 1991). Damit sind gerade verallgemeinernde Angaben über das Mediennutzungsverhalten anfällig für Messfehler (Boase & Ling, 2013; Naab et al., 2019; Shalev et al., 2025). Lediglich herausragende, intensive, emotionale, ungewöhnliche oder persönlich bedeutsame Ereignisse werden besser erinnert (Reis & Gable, 2000). Eine methodische Alternative besteht darin, Teilnehmer:innen nach einzelnen aktuellen Ereignissen zu befragen; sie müssen dann weder Vergangenes erinnern noch mehrere Ereignisse überschlägig zusammenfassen. Solche Abfragen gelten als weniger verzerrt (Kahlor et al., 2003; Lee et al., 2008). Allerdings sind die Antworten besonders von der einen konkreten Messsituation abhän-

gig, die atypisch im Vergleich zu sonstigen Verhaltensweisen, Emotionen und Kognitionen sein kann (Chang & Krosnick, 2003).

In-situ-Studien verbinden den Selbstbericht über konkrete, einzelne, aktuelle Situationen mit der wiederholten Messung über verschiedene Situationen hinweg. Auf diese Weise können wir z. B. Konstrukte wie die tägliche Nutzungsdauer oder -häufigkeit oder die tatsächlich aus der Nutzung gezogenen Gratifikationen nicht nur einmalig, sondern wiederholt in verschiedenen Situationen erheben. Wir können die daraus ermittelten Messwerte einer Person über mehrere Situationen zu einem Wert für die Person zusammenfassen bzw. aggregieren. So können wir z. B. über einen Zeitraum von mehreren Tagen Teilnehmer:innen bitten, immer nachdem sie Podcast gehört haben, einen Fragebogen auszufüllen (ereignisabhängiges ESM, siehe Kapitel 2). Die Antworten geben z. B. Auskunft über die Frage, wie lange und wie aufmerksam sie den Podcast gehört haben und wie intensiv sie sich in die Geschichte hineingezogen fühlten. Alle Messwerte über die Hördauer können zu einem Durchschnittswert für jede Person aggregiert und alle Messwerte der Aufmerksamkeit bzw. Transportation zu einem durchschnittlichen Aufmerksamkeits- bzw. Transportationswert für jede Person im Erhebungszeitraum verrechnet werden. Diese Werte haben eine größere Validität als die o. g. Erhebungen, weil sie weniger anfällig für Erinnerungs- und Schätzfehler sowie atypische Erhebungssituationen sind.

Neben methodischen können auch inhaltliche Gründe für den Einsatz wiederholter In-situ-Messungen sprechen. Schlütz (2002; siehe auch Scherer & Schlütz, 2002) unterscheidet z. B. zwischen generalisierten und konkreten erhaltenen Gratifikationen. Erstere vermitteln eine grundsätzliche Einschätzung darüber, welche Bedürfnisse ein Medienangebot erfüllt (z. B. „Nachrichten-Podcasts informieren mich“). Letztere spiegeln konkret erhaltene Gratifikationen, die in einer Nutzungssituation durch ein bestimmtes Angebot erfüllt wurden (z. B. „Genau dieser Podcast hat mich jetzt informiert“). Befragungsstudien sind häufig an den konkreten Gratifikationsleistungen von Medienangeboten interessiert. In retrospektiven Befragungen zeigen sich jedoch v. a. post hoc rationalisierte Erwartungen, die eher ge-

3 Typen von Forschungsfragen

neralisierte Gratifikationen reflektieren. Auch Naab et al. (2019) zeigen, dass retrospektive Bewertungen der Bedürfnisbefriedigung von Social-Media-Plattformen deutlich von In-situ-Bewertungen abweichen (siehe Studiensteckbrief im Online-Anhang zum Buch).

Bei einer geringen situativen Varianz einer Variable ist dagegen auch der Bedarf für wiederholte In-situ-Messungen gering. Dieser Fall liegt beispielsweise vor, wenn eine Person immer ähnlich lange Podcasts hört, dabei ähnlich aufmerksam ist und eine ähnlich intensive Transportation erlebt. Alle einzelnen Messungen dürften zu einem ähnlichen Ergebnis führen, das sich wiederum im Aggregat widerspiegelt, sodass der retrospektiv gewonnene Messwert wenig verzerrt sein dürfte. Einfach lässt sich dieser Befund an der Messung der Geräteausstattung verdeutlichen. Die Medienausstattung der meisten Menschen ändert sich zwar über die Zeit, weil sie neue Geräte anschaffen, allerdings eher in längeren Zyklen von Monaten oder Jahren. In einer kurz angelegten In-situ-Studie ist es daher kaum sinnvoll, die Ausstattung mehrfach – und schon gar nicht in jeder Situation – abzufragen. Anders verhält es sich mit dem jeweils genutzten Gerät, das durchaus situativ variieren kann. Persönlichkeitsmerkmale gelten zwar als langfristig wandelbar (z. B. B. W. Roberts et al., 2006), in kürzeren Lebensabschnitten bleiben sie jedoch stabil und werden nur einmalig abgefragt (z. B. zur Messung der Big Five, Rammstedt & John, 2007). Diese Annahme gilt auch für viele weitere Dispositionen, die Menschen mitbringen, z. B. Transportabilität, also die Eigenschaft, sich grundsätzlich (mehr oder weniger) gut in Geschichten hineinzusetzen (Dal Cin et al., 2004) oder die grundsätzliche kognitive Leistungsfähigkeit unabhängig von der Aufmerksamkeit in einer spezifischen Situation.

Bei hoher situativen Varianz einer Variable ist hingegen – wie bereits erläutert – die mehrmalige In-situ-Messung sinnvoll. Um mit den Ergebnissen aus In-situ-Studien Forschungsfragen auf Personenebene beantworten zu können, werden die Werte häufig zu Summen oder Mittelwerten aggregiert. Allerdings sind auch andere mathematische Operationen möglich. Wir könnten zum Beispiel per Aggregation die Varianz aller In-situ-Werte einer Person berechnen. Dann

3.1 Unterschiede zwischen Personen: Fragen nach interindividueller Varianz

erhalten wir ein Maß für die intraindividuelle Schwankungsbreite der einzelnen Werte. Es könnte z. B. quantifiziert werden, ob eine Person beim Hören eines Podcasts immer ähnlich aufmerksam ist (geringe Varianz aller In-situ-Werte) oder die Aufmerksamkeit im Zeitablauf sehr schwankt (große Varianz aller In-situ-Werte).

MERKE

Um interindividuelle Varianz valide zu erfassen, werden zahlreiche Messpunkte einer Person aggregiert. Unter Aggregation sind mathematische Operationen zu verstehen, bei denen mehrere Werte zu einer Kennzahl verrechnet werden. Beispiele sind die Bildung der Summe, des Mittelwerts oder der Varianz aus mehreren Einzelwerten. Nach einer Aggregation haben die Daten eine geringere Granularität, d. h. eine geringere Detailgenauigkeit. Werden In-situ-Daten auf Personenebene aggregiert, gehen intraindividuelle Schwankungen zwischen den Messwiederholungen einer Person verloren.

Natürlich befragen wir in In-situ-Studien nicht nur eine einzige Person, sondern eine Gruppe von Teilnehmer:innen. Im Mittelpunkt steht die Frage, wie sich das betrachtete Phänomen über alle Personen verhält, z. B. wie lange Menschen durchschnittlich Podcasts nutzen (Mittelwert der aggregierten Personenwerte), wie aufmerksam sie im Durchschnitt sind oder wie transportiert sie sich im Durchschnitt fühlen. Wir können außerdem fragen, ob es bei diesen Phänomenen große Unterschiede oder Gemeinsamkeiten zwischen den betrachteten Nutzer:innen gibt. Oder anders formuliert: Es wird die Frage untersucht, wie groß die interindividuelle Varianz (*between-person variance*) des Konstrukts ist. Aus statistischer Sicht geht es um die Streuung der aggregierten Personenwerte.

3 Typen von Forschungsfragen

MERKE

Forschungsfragen auf Personenebene interessieren sich für interindividuelle Differenzen, also für Unterschiede zwischen Personen (*between-person variance*). Wiederholte In-situ-Messungen sind dafür nicht zwingend notwendig, können ggf. aber die Validität der Messung verbessern. Die wiederholten Messungen einer Person werden anschließend aggregiert.

Wenn wir feststellen, dass die Teilnehmer:innen sich tatsächlich in dem untersuchten Phänomen unterscheiden, können wir uns auf die Suche nach Zusammenhängen dieser Variable mit anderen Variablen machen. Wir stellen also Fragen nach dem multivariaten Zusammenspiel mehrerer Phänomene, von denen alle oder einige wiederholt *in situ* gemessen wurden. Beispielsweise erhoben Valkenburg et al. (2021) in einer dreiwöchigen ESM-Studie sechsmal täglich die Nutzung sozialer Medien und das situative Selbstwertgefühl der Teilnehmer:innen (siehe Studiensteckbrief im Online-Anhang zum Buch). Sie ermittelten unter anderem eine negative Korrelation zwischen der mittleren Social-Media-Nutzung und dem mittleren Selbstwertgefühl der Teilnehmer:innen. Personen mit (im Durchschnitt über alle ihre situativen Antworten) geringer Social-Media-Nutzung hatten ein (im Durchschnitt über alle ihre situativen Antworten) höheres Selbstwertgefühl, Personen mit hoher Nutzung dagegen ein geringes Selbstwertgefühl (vgl. allerdings Abschnitt 3.2.2 für die abweichenden Resultate auf situativer Ebene). Im Rahmen derselben Erhebung untersuchten Siebers et al. (2022) neben der Nutzung sozialer Medien die von den Teilnehmer:innen selbst wahrgenommene Ablenkung (siehe Studiensteckbrief im Online-Anhang zum Buch). Sie fanden eine mittlere Korrelation der aggregierten In-situ-Messwerte einer Person in dem Sinne, dass Jugendliche, die mehr Zeit mit sozialen Medien verbrachten als ihre Peers auch über eine höhere Ablenkung berichteten (vgl. Abschnitt 3.2.2 zur Korrelation auf situativer Ebene).

3.2 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

Auch die Studie von Naab et al. (2019) kann diesbezüglich als Beispiel dienen (siehe Studiensteckbrief im Online-Anhang zum Buch): Die Autorinnen berechnen Korrelationen der retrospektiv erhobenen Angaben zur Nutzung von Social-Media-Plattformen mit *in situ* erhobenen und anschließend aggregierten Angaben derselben Teilnehmer:innen in einer ESM-Studie über 14 Tage mit drei zufälligen Signalen pro Tag. Sie zeigen für verschiedene Konstrukte mäßig starke Zusammenhänge.

Übertragen wir diese Typen von Forschungsfragen auf die Podcastnutzung, könnten wir die Frage untersuchen, ob Menschen, die grundsätzlich die Eigenschaft haben, sich in Geschichten hineinzuversetzen (einmalig gemessen als Transportabilität, Dal Cin et al., 2004), häufiger oder länger Podcasts hören (wiederholt *in situ* gemessen und aggregiert). Eine weitere zu untersuchende Frage könnte sein, ob Menschen, die häufiger Podcasts nutzen, insgesamt besserer Stimmung sind als Menschen, die seltener Podcasts nutzen.

Auch wenn wir Podcastnutzer:innen typologisieren, arbeiten wir auf der Personenebene (z. B. Karnowski et al. (2024), siehe Studiensteckbrief im Online-Anhang zum Buch). Beispielsweise könnten wir in einer ESM-Studie erfragen, in welcher Umgebung Podcasts gehört wurden, ob nebenbei Paralleltätigkeiten ausgeführt und ob Kopfhörer eingesetzt wurden. Mit den Messwerten können wir Nutzungstypen unterscheiden, deren Podcastnutzung typischerweise in verschiedenen Kontexten stattfindet.

3.2 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

Die Besonderheit von In-situ-Studien besteht darin, dass sie Phänomene über mehrere Situationen hinweg wiederholt bei denselben Teilnehmer:innen erfassen. Es ist daher möglich, Forschungsfragen nach intraindividuellen Veränderungen, also Veränderungen innerhalb derselben Person (*within-person variance*), zu beantworten. Im Folgenden betrachten wir sowohl Fragen nach der intraindividuellen

3 Typen von Forschungsfragen

Varianz einer einzelnen Variablen als auch nach dem Zusammenspiel mehrerer Variablen.

3.2.1 Beschreibung der situativen Variation eines Konstrukts und zeitliche Dynamiken

Nehmen wir an, wir wissen z. B. aus einer einmaligen Befragung von einer Person, dass sie durchschnittlich pro Tag eine Stunde Podcasts hört. Eventuell interessiert uns nun zusätzlich die Frage, ob die Person tatsächlich jeden Tag eine Stunde lang Podcasts hört. Oder nutzt sie Podcasts nur unter der Woche, hat am Wochenende jedoch keine Zeit dafür? Wie verteilt sich die Nutzungszeit über den Tag? Hört die Person nur morgens auf ihrer langen Fahrt zur Arbeit eine Stunde am Stück oder auch abends noch? Diese Fragen beziehen sich auf die Situationsebene und stellen darauf ab, wie veränderlich ein Konstrukt bei derselben Person über verschiedene Situationen hinweg ist. Mit anderen Worten: Wir untersuchen, wie groß die situative bzw. intraindividuelle Varianz, d. h. die *within-person variance* des Konstrukts ist. Aus statistischer Sicht geht es darum, wie stark die einzelnen In-situ-Messwerte einer Person um den Mittelwert aller In-situ-Messwerte dieser Person streuen (siehe Abschnitte 9.5.1 und 9.5.2.7). Diesen Zusammenhang haben wir bereits an anderer Stelle angesprochen, als es um die Aggregation von Messwerten einer Person ging (siehe Abschnitt 3.1).

MERKE

Forschungsfragen auf der Situationsebene sind Fragen nach der Veränderlichkeit von Konstrukten bei derselben Person über verschiedene Situationen hinweg. Sie adressieren intraindividuelle Differenzen, also Unterschiede innerhalb einer Person zwischen verschiedenen Situationen (*within-person variance*). Dafür sind wiederholte In-situ-Messungen bei derselben Person notwendig.

3.2 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

Wiederholte Messungen sind vor allem sinnvoll, wenn intraindividuelle Varianz zu erwarten ist, wie z. B. bei über die Woche oder über den Tag variierender Mediennutzung. Basierend auf fundierten theoretischen Überlegungen und einer ausführlichen Recherche empirischer Befunde müssen wir zunächst klären, ob intraindividuelle Varianz bei den in einem spezifischen Forschungsprojekt relevanten Konstrukten zu erwarten ist. Der umfassende Literaturkorpus zu Stimmungen legt beispielsweise nahe, dass unsere Stimmung stark von Situation zu Situation schwankt (Penner et al., 1994). Eine Messung zu einem Zeitpunkt führt mit einer hohen Wahrscheinlichkeit zu einem anderen Ergebnis als eine Messung einige Zeit später. Auch viele Konstrukte mit Bezug zur Mediennutzung schwanken erheblich zwischen Situationen (Schnauber-Stockmann, Scharnow, et al., 2025).

Wiederholte Messungen sind sinnvoll, wenn diese Vermutung auch auf das zu untersuchende Konstrukt zutreffen sollte. Bei eher stabilen Merkmalen reicht hingegen eine Querschnittserhebung aus (siehe Abschnitt 1.4). Eine Panelbefragung ist bei intraindividuellen Veränderungen über einen längeren Zeitraum hinweg angebracht, wenn z. B. die Bewältigung von Entwicklungsaufgaben in der Jugend untersucht wird (siehe Abschnitt 1.4).

Anhand von wiederholten Messungen lässt sich feststellen, ob sich ein Konstrukt im Zeitverlauf verändert. Wenn dieser Fall vorliegt, ist die Frage von Interesse, ob in dieser zeitlichen Dynamik Muster zu erkennen sind. Daran knüpft die Frage an, welche Tendenzen sich im Erhebungszeitraum zeigen. Denkbar wäre beispielsweise, dass die Nutzung von informationsorientierten Podcasts im zeitlichen Umfeld von Wahlen zunimmt, Podcasteinsteiger:innen mit zunehmender Nutzung ihre Auswahl stärker kuratieren oder sich spezifische Zyklen zeigen: Eine Person könnte Podcasts z. B. an Werktagen weniger als an Wochenendtagen (wöchentlicher Rhythmus, Wochentags- vs. Wochenendeffekt) oder im Hinblick auf die Tageszeit morgens mehr als abends nutzen (täglicher Rhythmus, z. B. Struckmann & Karnowski, 2016). Die Studie von Schnauber-Stockmann & Naab (2019) zeigt beispielhaft, dass Nutzer:innen auf

3 Typen von Forschungsfragen

eine neue App über die Zeit zunehmend habitueller und mit weniger Selektionsaufwand zugreifen (siehe Studiensteckbrief im Online-Anhang zum Buch).

Es ist unter anderem abhängig von der Dauer der Erhebung, wie gut sich zeitliche Dynamiken untersuchen lassen (vgl. Abschnitt 6.1). Die Erhebungsdauer sollte sich deshalb an theoretischen Annahmen über die Häufigkeit und Volatilität des Phänomens orientieren. Es ist etwa unmittelbar einsichtig, dass ein langer Erhebungszeitraum über mehrere Monate notwendig ist, um saisonale Effekte zu identifizieren. Für In-situ-Studien ist ein langer Erhebungszeitraum allerdings eher untypisch. Interessieren wir uns dagegen für Gefühlsschwankungen während eines Tages, ist ggf. eine Erhebung über einige Tage ausreichend. Neben der Erhebungsdauer bestimmt die Granularität der Erhebung – also die Anzahl der auszufüllenden Protokolle pro Tag und deren zeitlicher Abstand zueinander –, wie fein Schwankungen gemessen werden können (vgl. Abschnitte 6.1 und 6.3).

3.2.2 Zusammenhänge zwischen mehreren situativ variierenden Konstrukten

Nicht immer schwanken Phänomene nach einem zeitlichen Muster. Ihre Varianz kann auch im Zusammenhang mit der Varianz eines anderen, zeitlich nicht determinierten Phänomens stehen. Wir betrachten in diesem Abschnitt zunächst den Zusammenhang mehrerer situativ gemessener Konstrukte (multivariate Zusammenhänge auf situativer Ebene). Wir verbleiben also zunächst auf der Situationsebene. Später werden wir außerdem Forschungsfragen betrachten, bei denen Personen- und Situationsmerkmale interagieren.

Es ist eine Frage des individuellen Forschungsinteresses und der theoretischen Überlegungen, zu welchen anderen Variablen ein Konstrukt in Beziehung gesetzt werden soll. Zum Beispiel könnten wir die Frage klären, ob ein Podcast genutzt wird oder nicht, an welchem Aufenthaltsort (zu Hause, auf dem Weg zur Arbeit, beim Warten auf einen Termin) und wie groß die Aufmerksamkeit gegenüber dem Inhalt ist. Anschließend könnten wir prüfen, ob Paralleltätigkeiten

3.2 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

die Aufmerksamkeit einschränken. Auch Kovariationen mit Kognitionen oder Emotionen sind Ausgangspunkt vieler Forschungsfragen. In der Podcastingstudie interessierte uns beispielsweise der Zusammenhang zwischen narrativer Transportation und Aufmerksamkeit bei der Nutzung von True-Crime-Podcasts. Die Annahme war, dass Personen sich umso besser in das auditive Geschehen hineinversetzen können, je aufmerksamer sie zuhören (siehe Studiensteckbrief im Online-Anhang zum Buch).

Otto et al. (2020) erhoben in einer ereignisabhängigen ESM-Studie z. B. Situationen, in denen die Teilnehmer:innen mit politischen Informationen konfrontiert wurden, und maßen deren emotionale Reaktionen und Aufmerksamkeit. Die Studie zeigte unter anderem, dass Wut, nicht aber Angst mit größerer Aufmerksamkeit gegenüber den Informationen einherging. Karsay et al. (2023) erhoben über fünf Tage dreimal täglich zu festen Uhrzeiten rezipierende bzw. produzierende Nutzungsarten sozialer Medien und damit einhergehendes affektives Wohlbefinden. Posting und Messaging waren positiv und Browsing negativ mit Wohlbefinden assoziiert. Klingelhoef et al. (2024) untersuchten situativ variierende Motivationen, die einen Einfluss darauf hatten, ob Menschen zeitweise auf digitale Medien verzichteten (siehe Studiensteckbrief im Online-Anhang zum Buch).

MERKE

Wir können zeitliche Dynamiken untersuchen, z. B. Muster im Tagesverlauf oder Tendenzen im Erhebungszeitraum, um intraindividuelle Varianz zu erklären. Aber auch andere räumliche und soziale Kontexte oder innere Zustände können situativ variieren und im Zusammenhang mit der Mediennutzung stehen.

In Abschnitt 3.1 zu Personenunterschieden haben wir bereits die Studie von Siebers et al. (2022) (siehe Studiensteckbrief im Online-Anhang zum Buch) und Valkenburg et al. (2021) (siehe Studiensteckbrief im Online-Anhang zum Buch) kennengelernt. Siebers et

3 Typen von Forschungsfragen

al. (2022) fanden eine Between-Person-Korrelation zwischen der Nutzung sozialer Medien und wahrgenommener Ablenkung. Dieser Zusammenhang zeigte sich auch auf situativer Ebene, wenn auch auf schwächerem Niveau: Jugendliche fühlten sich in Situationen, in denen sie soziale Medien nutzten, abgelenkter als in Situationen, in denen sie keine sozialen Medien nutzten. Anders sieht es hingegen bei Valkenburg et al. (2021) im Hinblick auf den Zusammenhang zwischen der Nutzung sozialer Medien und Selbstwertgefühl aus: Während die Autor:innen zeigten, dass Personen mit insgesamt höherer Social-Media-Nutzung ein im Durchschnitt über ihre wiederholten Messungen geringeres Selbstwertgefühl aufwiesen, konnten sie keinen Zusammenhang auf der Ebene der situativ gemessenen Konstrukte feststellen. Dieser Unterschied in den Ergebnissen auf der Personen- und Situationsebene verdeutlicht eindrucksvoll, dass sich interindividuelle Zusammenhänge nicht zwingend auch intraindividuell zeigen müssen.

Die Beispiele belegen, welche hohe Bedeutung es hat, ob Forschungsfragen auf der Personen- oder Situationsebene untersucht werden. Ggf. können sich die Ergebnisse erheblich unterscheiden. Ignoriert man, dass die Daten aus wiederholten Befragungen von denselben Personen stammen, kann es sogar zu einer Variante des ökologischen Fehlschlusses kommen. Dieser entsteht, wenn Daten auf Aggregatniveau ausgewertet werden (wenn z. B. die wiederholten Angaben derselben Personen zu Personenmittelwerten verrechnet werden, mit denen dann auf Personenebene weitergerechnet wird), daraus aber unzulässigerweise auf Zusammenhänge auf der nicht aggregierten Ebene (hier die situative Ebene) geschlossen wird. Die Darstellung des Simpsons Paradoxons in Abschnitt 9.5.2.1 wird diesen Fehlschluss noch einmal veranschaulichen.

3.2.3 Zeitversetzte Zusammenhänge und Autokorrelationen

Mehrere zeitlich variierende Konstrukte schwanken nicht notwendigerweise gleichzeitig. Auch zeitversetzte Variationen können theore-

3.2 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

tisch und praktisch relevant sein. Menschen regulieren z. B. durch die Nutzung von Medien ihre Stimmung in Vorbereitung auf eine später folgende Aufgabe (Knobloch, 2003) oder Auswirkungen von Medien-nutzung treten erst mit Verzögerung ein (Thomas, 2022). Z. B. kann die Nutzung sozialer Medien die wahrgenommene Ablenkung zwei Stunden später voraussagen; in die umgekehrte Richtung, in der das Ablenkungsgefühl die Nutzung voraussagen würde, findet sich jedoch kein nennenswerter Effekt (Siebers et al., 2022) (siehe Studiensteckbrief im Online-Anhang zum Buch). Ein anderes Beispiel zeigt, dass die Motivation, vorübergehend auf digitale Medien zu verzichten, um Ablenkung zu vermeiden, Wohlbefinden zu steigern oder offline prä-senter zu sein, etwa zwei bis drei Stunden später tatsächlich zu etwas gesteigertem Disconnection-Verhalten führen kann (Klingelhofer et al., 2024) (siehe Studiensteckbrief im Online-Anhang zum Buch).

Gelegentlich gibt es auch Forschungsfragen zu autokorrelativen Effekten. Mit Autokorrelation ist die Korrelation einer Variable mit sich selbst zu einem anderen Zeitpunkt gemeint. Wir könnten zum Beispiel die Frage untersuchen, wie die Stimmung einer Person zu einem Zeitpunkt mit ihrer Stimmung kurze Zeit später zusammenhängt.

3.2.4 Fragen zum Zusammenspiel von Personen- und Situationsmerkmalen

Oben haben wir einige Beispiele für situative Schwankungen von Phänomenen vorgestellt, die wir in In-situ-Erhebungen erfassen können. Solche situativen Schwankungen sind aber unter Umständen nicht für alle Personen identisch, sondern es liegen interindividuelle Unterschiede im Hinblick auf die intraindividuelle Varianz vor. Situative Schwankungen von Phänomenen können anhand von Personenmerkmalen erklärt werden (Haupteffekte von Variablen auf der Personenebene) oder Unterschiede zwischen Situationen können mit Unterschieden zwischen Personen zusammenspielen (Interaktionseffekte von Variablen auf Personen- und Situationsebene).

In der Podcastingstudie haben wir folgende Überlegung ange-

3 Typen von Forschungsfragen

stellt (siehe Studiensteckbrief im Online-Anhang zum Buch): Ob sich Hörer:innen in einen wahren Kriminalfall hineinversetzen können, könnte nicht nur abhängig sein von der Aufmerksamkeit, die sie dem Podcast in einer konkreten Nutzungssituation widmen. Das situative Transportationserleben könnte auch davon abhängen, ob sie sich grundsätzlich gut in Geschichten hineinversetzen können (Transportabilität, Dal Cin et al., 2004). Dies wäre ein Beispiel für einen Haupteffekt eines stabilen Personenmerkmals. Weiterhin haben wir oben beispielhaft überlegt, ob Menschen in der Zeit vor und nach Wahlen zunehmend informationsorientierte Podcasts auswählen. Solche Veränderungen über die Zeit (Situationsebene) sind aber eventuell nicht für alle Menschen gleich, sondern möglicherweise treffen sie für Menschen mit hohem Interesse für politische Themen zu, für Menschen mit geringem Interesse hingegen nicht. Hier läge eine Interaktion zwischen dem Personenmerkmal des Politikinteresses und dem Situationsmerkmal des Abstands zum Wahltermin vor.

MERKE

Forschungsfragen nach Haupt- und Interaktionseffekten beschäftigen sich damit, wie situative Veränderungen mit Personenmerkmalen zusammenhängen. Dafür sind wiederholte In-situ-Messungen mit mehreren Personen notwendig.

In-situ-Studien, die sich mit Haupt- und Interaktionseffekten befassen, beschränken sich nicht notwendigerweise auf Merkmale, die den Personen inhärent sind, wie ihre Charaktereigenschaften. Auch andere Variablen können situationsübergreifend stabil bei einer Person wirken und sich zwischen Personen unterscheiden, wie strukturelle Einbindungen (z. B. haben manche Personen ein stabileres familiäres Netzwerk als andere) oder Medienausstattung (z. B. die Qualität der Kopfhörer für die Podcastnutzung). Auch solche Faktoren auf Personenebene können die intraindividuelle Variation erklären.

3.2.5 Klassen-/Typenbildung von Personen und Situationen

Die meisten Fragestellungen auf der Situationsebene in der KMW beziehen sich auf Beschreibungen der Häufigkeit situativer Kontexte, zeitliche Dynamiken oder Zusammenhänge zwischen stabilen und situativ variierenden Konstrukten. Gelegentlich werden aber auch Situationen zu Klassen zusammengefasst. Beispielsweise fragten Karnowski et al. (2024) in einer ESM-Studie junge Erwachsene wiederholt, ob und wenn ja in welchem Rezeptionsmodus (lesen, schauen, hören) und auf welcher Plattform (Webseite, App, Social Media Messenger u. a.) sie Onlinenachrichten konsumiert haben (siehe Studiensteckbrief im Online-Anhang zum Buch). Anhand dieser situativen Merkmale gruppieren die Autorinnen anschließend Nachrichtennutzungssituationen. Im Rahmen einer ereignisabhängigen ESM-Studie identifizieren auch Ernst & Schnauber-Stockmann (2026) Situationsklassen (siehe Studiensteckbrief im Online-Anhang zum Buch). Die Teilnehmer:innen füllten nach Beendigung einer TikTok- oder Instagram-Nutzungsperiode ein ESM-Protokoll aus, in dem sie u. a. angaben, warum sie die jeweilige App geschlossen haben. Auf Basis dieser sogenannten Disengagement Cues (z. B. Ortswechsel, Empfang einer Pushbenachrichtigung, Erreichung des Mediennutzungsziels) konnten die Autorinnen verschiedene Typen von Disengagement-Situationen identifizieren.

Es ist ebenfalls möglich, auf Basis von situativ erhobenen Konstrukten Personenklassen oder -typen zu unterscheiden, etwa anhand der Frage, ob es Personen gibt, bei denen bestimmte Nachrichtennutzungssituationen besonders häufig auftreten.

4 Ethische und rechtliche Aspekte von In-situ-Studien

Wissenschaftliche Forschung muss grundsätzlich den Anforderungen an Professionalität, Integrität, Redlichkeit und Reflexivität entsprechen. Neben rechtlichen Verpflichtungen sind forschungsethische Standards ein Aspekt sog. guter wissenschaftlicher Praxis (vgl. Deutsche Forschungsgemeinschaft, 2025). Im Vordergrund angewandter Forschungsethik steht die Reflexion der eigenen Handlungsentscheidungen als Forscher:in unter Berücksichtigung der Interessen und Rechte der Beteiligten. Es handelt sich daher um einen Prozess der Abwägung gesetzlicher Bestimmungen (z. B. Datenschutzgesetze), standesrechtlicher und institutioneller Normen (verankert z. B. in Ethikkodizes oder den Satzungen lokaler Ethikkommissionen) und individueller ethischer Erwägungen einerseits sowie methodischer Überlegungen und forschungspraktischer Möglichkeiten im konkreten Forschungsprozess andererseits (Franzke et al., 2021; Schlütz & Möhring, 2018; Schlütz & Zillich, 2023). Ziel dieses Abwägungsprozesses ist es, die Studiengüte insgesamt zu optimieren. Dazu bedarf es eines auf den Einzelfall bezogenen Reflexionsprozesses, der systematisch, deliberativ und kollaborativ angelegt ist und der die zentrale Rolle der Ethik für die Erzielung innovativer, aber vor allem auch belastbarer wissenschaftlicher Ergebnisse berücksichtigt (Leonelli et al., 2021). Eine solche kasuistische Vorgehensweise ist auch für In-situ-Studien geeignet und wird daher weiter unten beschrieben.

Ohne grundsätzliche ethische Erwägungen in der (deutschen) KMW zu diskutieren (vgl. dafür z. B. Roehse et al., 2023; Schlütz & Möhring, 2016, 2017), gehen wir im Folgenden auf grundlegende Prinzipien und ethisch-methodische Abwägungsprozesse und ihre Konsequenzen für In-situ-Studien ein (vgl. Capon et al., 2016;

4.1 Leitprinzipien forschungsethischer Reflexion

Kirtley, 2022; Robbins, 2017). Besonders relevant ist in diesem Zusammenhang das Feld der sog. **Internet Research Ethics** (IRE, Locatelli et al., 2023). IRE befasst sich speziell mit digitalen Forschungsansätzen und behandelt Themen wie informierte Einwilligung, Privatsphäre und Anonymität sowie Datensicherheit für Forschungsvorhaben in Bezug auf die Nutzung des Internets (Ess, 2019). Herausforderungen in diesen Bereichen ergeben sich in vielen In-situ-Studien, die sich häufig mit den Phänomenen digitaler Mediennutzung befassen und eine große Fülle ebenso detaillierter wie umfangreicher Daten erfassen. Dies gilt insbesondere, wenn die Befragungsdaten mit digitalen Verhaltensspuren kombiniert werden (Breuer et al., 2021; Struminskaya & Sakshaug, 2023). Wir beginnen die Ausführungen mit handlungsleitenden Prinzipien und deren Abwägung und gehen anschließend auf die Besonderheiten von In-situ-Designs ein. Einzelne Aspekte der forschungsethischen Praxis werden im Laufe des Buches an passenden Stellen ggf. ausführlicher behandelt.

MERKE

Forschungsethik bezieht sich auf die Gestaltung der Beziehungen zwischen den Beteiligten in sämtlichen Phasen des Forschungsprozesses – von der Themenwahl und dem Studiendesign über den Feldzugang, die Datenerhebung und -analyse bis hin zu Fragen der Publikation und Verwertung von Forschungsergebnissen.

4.1 Leitprinzipien forschungsethischer Reflexion

Aktuelle forschungsethische Ansätze, insbesondere aus dem Bereich der IRE (vgl. Zillich et al., 2023), verzichten weitgehend auf fixierte Checklisten. Sie setzen stattdessen auf reflexives Urteilen anhand fallbasierter Abwägungsprozesse (vgl. Schlütz & Möhring, 2025). Ein solches kasuistisches Urteil ist i. d. R. weder eindeutig noch end-

gültig, sondern pluralistisch und reversibel. McKee & Porter (2009) haben einen entsprechenden Ansatz der Entscheidungsfindung entwickelt, bei dem spezifische Leitprinzipien genutzt werden, um fallbasiert konkrete forschungsethische Entscheidungen zu treffen. I. d. R. werden zu diesem Zweck die von Beauchamp & Childress (2019) weiterentwickelten Belmont-Prinzipien – Selbstbestimmung (*respect for autonomy*), Schadensvermeidung (*nonmaleficence*) bzw. Fürsorge (*beneficence*) sowie Gerechtigkeit (*justice*) – zugrunde gelegt.

Die **Selbstbestimmung** bzw. Autonomie der Teilnehmer:innen wird maßgeblich durch eine informierte Einwilligung, den sog. Informed Consent sichergestellt. Hinzu kommt ggf. eine ergänzende Aufklärung im Nachgang der Datenerhebung (das sog. Debriefing). Im Rahmen der Einwilligungsprozedur werden potenzielle Teilnehmer:innen über den Zweck der Studie, ihre Inhalte und das konkrete Vorgehen aufgeklärt, damit sie auf dieser Basis entscheiden können, ob sie teilnehmen und ihre Daten teilen wollen oder nicht. Ergänzend sind Hinweise zu Anonymität – falls es sich nicht um personenbezogene Daten handelt – und Datenschutz obligatorisch. Zu den personenbezogenen Daten gehören gemäß Art. 4 DSGVO alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Identifizierbar ist eine Person z. B. über ihren Namen, aber auch indirekt über Standortdaten, den Bezug zu einer Onlinekennung oder zu einem oder mehreren individuellen Merkmalen. Beim Vorliegen vieler verschiedener Merkmale besteht durch deren Kombination die Möglichkeit, Personen zu deanonymisieren. Auf diesen Aspekt muss insbesondere bei der Veröffentlichung von Daten geachtet werden (vgl. Morehouse et al., 2024).

Die informierte Einwilligung gibt Forschenden das Recht, Teilnehmer:innen in das Forschungsprojekt einzubeziehen, und stellt gleichzeitig sicher, dass deren Rechte geschützt werden. Dazu gehört, dass die Einwilligung freiwillig ist, d. h. die Zustimmung kann auch verweigert oder im Nachhinein widerrufen werden, ohne dass sich daraus negative Konsequenzen ergeben. Zudem muss sie von einer geschäftsfähigen Person erteilt werden, anderenfalls ist ein sog. *proxy consent* notwendig, z. B. bei Kindern durch eine

4.1 Leitprinzipien forschungsethischer Reflexion

erziehungsberechtigte Person. Schließlich müssen die potenziellen Teilnehmer:innen über die Ziele der Studie und die eigene Rolle darin sowie über die Verwendung ihrer Daten ausreichend aufgeklärt werden, damit sie verstehen, was von ihnen erwartet wird. Aus diesem Grund müssen die Informationen präzise, verständlich und vollständig sein (Perrault & McCulloch, 2019; Perrault & Nazione, 2016). Die Einwilligung kann implizit (der Beginn der Befragung wird als Zustimmung gewertet) oder explizit (zu Beginn der Umfrage wird ein entsprechendes Feld angekreuzt) erfolgen. Sie kann prospektiv im Vorfeld der Befragung oder (erneut) retrospektiv z. B. nach einem Debriefing, das über eine mögliche Täuschung aufklärt, eingeholt werden (Verbeke et al., 2023). Welche Vorgehensweise angemessen ist, hängt von der Art der erhobenen Daten ab. Der Begriff der personenbezogenen Daten kann gem. Art. 4, 1 DSGVO recht weit ausgelegt werden. Aus diesem Grund ist es im Kontext von In-situ-Studien empfehlenswert, die Einwilligung prospektiv und explizit einzuholen.

Eine fallbezogene Schaden-Nutzen-Abwägung (Rosnow & Rosenthal, 2014) wird vorgenommen, um den erwarteten Nutzen – für Gesellschaft, Wissenschaft und die Forschenden, aber auch für die Teilnehmer:innen (z. B. durch Incentives, Credit Points für Studierende, Spaß, Selbsteinsicht o. Ä.) – in ein ausgewogenes bzw. vertretbares Verhältnis zu möglichen Schädigungen der Beteiligten zu setzen. Es geht darum, Schaden von den Studienteilnehmer:innen abzuwenden (**Schadensvermeidung**). Neben individuellen Schädigungen sind auch aufgewendete Ressourcen relevant. In diesem Zusammenhang sollte auch berücksichtigt werden, welche negativen Konsequenzen der Verzicht auf die geplante Studie hätte (Rosnow & Rosenthal, 2014).

Das Streben nach **Gerechtigkeit** hat zum einen Konsequenzen für die Stichprobenbildung, die möglichst inklusiv sein sollte, zum anderen betrifft es die Barrierefreiheit von verwendeten Apps, Websites oder Instrumenten, und zwar sowohl auf inhaltlicher (etwa durch verständliche Formulierungen) als auch auf technischer Ebene (Eignung für Screen Reader etc., siehe z. B. Berger et al., 2019;

Bigl et al., 2024). Aus forschungsethischer Sicht richtet sich die Aufmerksamkeit besonders auf sog. vulnerable Zielgruppen. Das Thema ‚Vulnerabilität‘ sollte allerdings im Hinblick auf die Autonomie der betreffenden Personen kritisch hinterfragt werden, um entsprechende Populationen nicht zu bevormunden bzw. automatisch von Forschung auszuschließen (Lahman, 2018; Tiidenberg, 2018). Im Zusammenhang mit dem Prinzip der Gerechtigkeit steht auch die Open-Science-Bewegung (siehe Kapitel 10), die die Sicherung guter wissenschaftlicher Praxis im Blick hat, ohne dabei allerdings explizit ethische Fragestellungen zu berücksichtigen (Dienlin et al., 2021; Fox et al., 2021).

Die o. g. Prinzipien können genutzt werden, um einen Prozess der kasuistischen Entscheidungsfindung im konkreten Anwendungsfall anzuleiten. McKee & Porter (2009) haben einen solchen Ansatz entwickelt, der sich eignet, um im Forschungsprozess auftretende Dilemmata und komplexe ethische Fragestellungen zu reflektieren und angemessene Entscheidungen zu treffen (für konkrete Beispiele unterschiedlicher Studien siehe Eble et al., 2014; Pentzold, 2015; Schlütz & Zillich, 2023). Wir beschreiben die Vorgehensweise anhand der Beispielstudie in Abschnitt 4.2.

MERKE

Folgende Prinzipien leiten forschungsethische Überlegungen an: Die (informationelle) Selbstbestimmung bezieht sich auf die freiwillige und widerrufbare Zustimmung zur Teilnahme an einer Studie bzw. deren Ablehnung. Das Prinzip der Schadensvermeidung impliziert eine gründliche Schaden-Nutzen-Abwägung, die den zu erwartenden Nutzen in ein vertretbares Verhältnis zu möglichen Schädigungen setzt. Das Gerechtigkeitsprinzip geht vom Gleichheitsgrundsatz aus und soll die faire Teilhabe aller am wissenschaftlichen Erkenntnisgewinn sicherstellen.

4.2 Forschungsethische Besonderheiten von In-situ-Studien

Grundsätzlich ist es ein Vorteil von In-situ-Studien, dass sie es durch ihre Verankerung im alltäglichen Lebensumfeld ermöglichen, bestimmte Zusammenhänge (wie z. B. den Einfluss der Mediennutzung auf mentale Krankheiten) zu erforschen, die aus forschungsethischen Gründen nicht induziert, d. h. im Rahmen von Experimentaldesigns manipuliert werden können (Reis, 2012). Allerdings sind In-situ-Studien für die Teilnehmer:innen im Vergleich zu anderen Studiendesigns extrem zeitaufwändig sowie durch die häufigen Befragungen im Tagesablauf anspruchsvoll. Die Teilnahme kann daher durchaus als Belastung im Sinne eines Schadens empfunden werden. Je nach Forschungsfrage und Zielgruppe kann zudem die (in anderen Kontexten häufig als Incentive bewertete) mit der Teilnahme einhergehende Selbsterkenntnis forschungsethisch problematisch sein, wenn z. B. pathologische Verhaltensweisen erforscht werden (Trull & Ebner-Priemer, 2020). Mit Blick auf die in Abschnitt 4.1 vorgestellten Leitprinzipien lassen sich folgende zentrale Aspekte für Forschungsethik bei In-situ-Studien bündeln (vgl. auch Capon et al., 2016; Jacobson et al., 2020; Kirtley, 2022; Nock et al., 2021; Robbins, 2017). Aus forschungspraktischen bzw. -ökonomischen Gründen werden sich nicht immer alle Anforderungen umsetzen lassen, manchmal ist dies auch aus methodischen Gründen nicht erstrebenswert. Aber es lohnt sich in jedem Fall, darüber im Vorfeld der eigenen Studie nachzudenken, um reflektierte Entscheidungen zu treffen. Wir diskutieren diese Aspekte zunächst auf theoretischer Ebene und wenden sie dann beispielhaft an.

(1) Die Selbstbestimmung der Teilnehmer:innen von In-situ-Studien ist vor allem informationeller Natur. Entsprechend ist sie eng verbunden mit Datenschutz und Datensicherheit. Diese werden z. B. sichergestellt durch verschlüsselte Datenübertragung, passwortgesicherte lokale Speicherplattformen und die Nutzung datenschutzkonformer Cloudlösungen. Beim Datenschutz müssen auch die Rechte dritter Personen beachtet werden. Deren Daten können

Bestandteil der situativen Erhebung sein, etwa wenn Teilnehmer:innen Chatnachrichten oder Bilder weitergeben. Ein weiterer Aspekt ist eine mögliche Weiterverwertung der Daten durch Dritte (*data sharing*). Hier müssen z. B. die entsprechenden Zugangs- und Nutzungsregeln der verwendeten Open-Data-Plattform offengelegt werden. Zentrales Instrument, um die Selbstbestimmung sicherzustellen, ist eine ausführliche und verständliche **Einwilligungserklärung** (*Informed Consent*, siehe Abschnitt 4.1), die die Freiwilligkeit der Teilnahme deutlich herausstellt und das Datenhandling (inkl. Open-Access-Regelungen, vgl. Kapitel 10) nachvollziehbar beschreibt. Wichtig ist, dass die Teilnehmer:innen eine aktive Rolle beim Schutz ihrer Privatsphäre einnehmen, also z. B. zu ungelegenen Zeiten die Alarmierung unterdrücken können, auch wenn dies zu Einschränkungen der Stichprobengüte führen kann. Es ist zudem sinnvoll, einen kontinuierlichen und direkten Kontakt zum Forschungsteam zu ermöglichen (z. B. per E-Mail oder SMS), um Rückfragen (z. B. zu Datenschutz oder Belastung) zeitnah beantworten zu können (siehe Abschnitt 8.3.2).

(2) Ein wichtiger Aspekt der Schaden-Nutzen-Abwägung ist die Vermeidung von Schaden durch die **Sicherstellung von Anonymität** bei der Datenerhebung und -auswertung. Entscheidend ist zunächst, dass personenbezogene Daten, die für die Methodik unerlässlich sind (wie z. B. die Mobilnummer für die Signalversendung), nicht in einem Datensatz mit den Befragungsdaten gespeichert und nach Beendigung der Feldzeit so schnell wie möglich gelöscht werden. Zudem gilt das Gebot der Datensparsamkeit, vor allem für die Aufzeichnung von Kontextdaten und ergänzenden Verhaltensspuren. Je weniger Daten erhoben werden, die potenziell zu einer De-Anonymisierung herangezogen werden könnten, umso mehr sind die Anforderungen an die Anonymität gewährleistet (vgl. Morehouse et al., 2024). Ist ein kompletter Verzicht auf bestimmte Daten nicht möglich, sollten sie auf einem möglichst hohen Aggregationsniveau erfasst werden (also z. B. Postleitzahlengebiete statt genaue Adressen), weil das die Identifizierung erschwert. Um den Nutzen für die Teilnehmer:innen (im Verhältnis zu dem ihnen entstehenden Auf-

4.2 Forschungsethische Besonderheiten von In-situ-Studien

wand) zu erhöhen, kann man sie durch ein angemessenes Incentive (z. B. einen Geldbetrag oder die Teilnahme an einem Gewinnspiel) entschädigen (siehe Abschnitt 8.2). Insbesondere bei Studierenden, die häufig mit Credit Points entlohnt werden, gilt es allerdings zu bedenken, dass die Teilnahme dadurch nicht erzwungen werden darf (Scollon et al., 2009).

(3) Gerechtigkeit bezieht sich vorwiegend auf die Zusammenstellung einer möglichst **inklusiven Personenstichprobe** für eine In-situ-Studie. Dieses Vorgehen ist auch aus methodischen Gründen ratsam. Allerdings ist es eine große Herausforderung, weil es ohnehin schwierig ist, Personen für Längsschnittstudien zu gewinnen, zumal für solche, die vermeintlich oder tatsächlich eine umfangreiche Menge unterschiedlicher Daten sammeln. Hinzu kommt, dass je nach Studienkonzeption bestimmte technische Voraussetzungen für die Teilnahme gegeben sein müssen, etwa der Besitz eines Smartphones. Beim Einsatz bestimmter Apps spitzt sich die Problematik noch zu, wenn diese z. B. nur für bestimmte Betriebssysteme geeignet sind. Zusätzlich ist eine stabile, mobile Datenverbindung erforderlich, über die auch nicht allen Personen verfügen (etwa Schüler:innen). Wenn möglich, sollten Geräte oder ein mobiles Datenvolumen für Teilnehmer:innen kostenfrei bereitgestellt werden, um auch ressourcenschwachen Personen die Teilnahme zu ermöglichen. Zudem sollte sowohl bei den Fragebögen der Vorab- und/oder Nachbefragung als auch bei den Protokollen in der In-situ-Phase auf Barrierefreiheit geachtet werden, etwa durch die Optimierung für Screen Reader, die Untertitelung von Videos oder die Formulierung in leichter Sprache (vgl. Bigl et al., 2024).

Verstärkt werden die genannten Herausforderungen, wenn es sich um sensible Fragestellungen wie etwa in der Gesundheitsforschung handelt oder wenn In-situ-Befragungsdaten mit digitalen Verhaltensspuren oder anderen Kontext- oder Paradata kombiniert werden. Damit sind Informationen gemeint, die neben den eigentlichen Daten im Befragungsprozess erhoben werden (vgl. Felderer et al., 2019).

Wichtig ist für alle In-situ-Studien eine umfangreiche Informed-Consent-Prozedur, die auf die jeweilige Zielgruppe abgestimmt ist

und auf die technischen Besonderheiten des Designs eingeht (Conner & Lehman, 2012). Insgesamt kann es sinnvoll bzw. geboten sein, die geplante Studie von einer Ethikkommission begutachten zu lassen, um ein Ethikvotum einzuholen. Ein solches Votum ist inhaltlich sinnvoll, um die forschungsethischen Herausforderungen der eigenen Forschung im Vorfeld zu reflektieren und auf diese Weise die Studienqualität insgesamt zu steigern. Zudem können die Expert:innen aus der Kommission ggf. beratend tätig werden. Häufig besteht zudem die formale Notwendigkeit, weil etliche Fachzeitschriften und Forschungsförderinstitutionen mittlerweile ein positives Ethikvotum voraussetzen (vgl. z. B. Chatfield & Dove, 2023). Für den Prozess sollte entsprechend Zeit eingeplant werden, denn je nach den Gegebenheiten an der eigenen Hochschule und Komplexität der Studie kann die Einholung ein aufwändiges Unterfangen sein.

Auf der Webseite des BMBF-geförderten Projekts FeKoM (Forschungsethik in der Kommunikations- und Medienwissenschaft, <https://www.forschungsethik-kmw.de>) finden sich Handreichungen, Tools und Checklisten zum Thema Forschungsethik, die Anregungen für eigene Studien geben.

Um die genannten Aspekte im Rahmen eines **fallbasierten Abwägungsprozesses** zu illustrieren, greifen wir auf unsere Beispielstudie (siehe Studiensteckbrief im Online-Anhang zum Buch) zurück. Dort haben wir den Einfluss der Situationsbedingungen auf das Transportationserleben beim Hören von True-Crime-Podcasts erforscht (Abbildung 4.1). Im Folgenden gehen wir auf einige Aspekte des Studiendesigns ein, die forschungsethische Relevanz besitzen, und diskutieren Herausforderungen und Lösungsmöglichkeiten.

Die Daten wurden, wie es bei ESM-Studien üblich ist, auf zwei Ebenen erfasst: Individuelle Dispositionen haben wir im Rahmen einer online durchgeführten Vorabbefragung erhoben, während situatives Erleben und Kontextvariablen *in situ* erfasst wurden. Auf beiden Ebenen sind wir standardisiert vorgegangen. Die Zugänglichkeit zu den Daten war insofern beschränkt, als es sich um Primärdaten in Form von Selbstberichten handelt, die per se nicht öffentlich sind. Die Daten zur Nutzung von True-Crime-Podcasts können aus

4.2 Forschungsethische Besonderheiten von In-situ-Studien

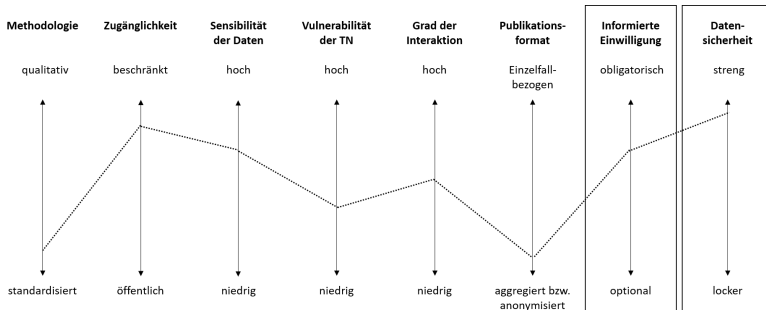


Abbildung 4.1: Forschungsethische Entscheidungsfindung basierend auf McKee und Porter (2009) in Anwendung auf die Beispielstudie

zwei Gründen als sensibel gelten. Zum einen handelt es sich um ein emotionales Thema (Gewalt und Verbrechen), das eine Triggerwirkung haben oder Vulnerabilitätsfurcht auslösen kann. Zum anderen wird eine Fülle detaillierter Informationen über Nutzungssituationen erhoben. Auflösungsgrad und Detailliertheit der Daten machen die Teilnehmer:innen insofern verwundbar, als sie theoretisch eine Identifikation ermöglichen, wenn dem nicht an anderer Stelle durch Datensparsamkeit oder -aggregation wirksam begegnet wird. Konkret wurde technisch sichergestellt, dass die Mobilfunknummern für die Übersendung der In-situ-Fragebögen und die E-Mail-Adressen für das Gewinnspiel separat von den Erhebungsdaten gespeichert und nach Ablauf der Erhebungszeit gelöscht wurden. Der Grad der Interaktion mit den Teilnehmer:innen lag durch die standardisierten, aber schriftlich vorgelegten Fragen im mittleren Bereich. Die erhobenen Daten sind zwar sehr detailliert, aber die Ergebnispublikation ist hoch aggregiert in Form personenübergreifender Kennwerte und daher aus forschungsethischer Sicht unproblematisch. Auf Basis dieser Überlegungen ergab sich, dass für die Studie eine explizite informierte Einwilligung der Studienteilnehmer:innen geboten war. Zu-

4 Ethische und rechtliche Aspekte von In-situ-Studien

dem musste besonders auf Datenschutz geachtet werden, was durch die Administrierung der Studie über den DSGVO-konformen *SoSci-Survey*-Server gewährleistet wurde.

In der Beispielstudie lautete der Informed Consent für den situativen Teil der Erhebung wie folgt:

Vielen Dank, dass Sie auch am zweiten Teil unserer True-Crime-Studie teilnehmen! Der zweite Teil der Untersuchung erstreckt sich über zwei Wochen, und zwar vom 18.09. bis 01.10.2023. Er besteht aus bis zu vier sehr kurzen Umfragen (max. 2–3 Minuten) pro Tag. In dem Teil geht es uns vor allem um die jeweilige Situation, in der Sie Podcasts hören. Sie benötigen für diesen Teil keine gesonderte App. Wir werden Ihnen einfach eine SMS mit einem Befragungslink schicken, der Sie zum Kurzfragebogen leitet. Sie können ihn dann zwischendurch am Handy ausfüllen. Dafür benötigen Sie mobiles Datenvolumen oder eine WLAN-Verbindung. Der Link ist nach dem Versand eine Stunde gültig. In diesem Zeitfenster können Sie die Umfrage ausfüllen. Wenn Sie gerade keine Gelegenheit dazu haben, machen Sie einfach bei der nächsten Umfrage wieder mit. Für das Gelingen der Studie ist es aber wichtig, dass so viele Kurzfragebögen wie möglich beantwortet werden. Wir freuen uns daher, wenn Sie Ihr Smartphone bereithalten und so häufig wie möglich mitmachen – aber natürlich sollen Sie Ihr normales Verhalten nicht aufgrund der Studie verändern, denn gerade das interessiert uns ja. Selbstverständlich ist auch die Teilnahme an diesem Teil freiwillig und kann jederzeit abgebrochen werden. Auch die Kurzfragebögen sind vollständig anonym. Es werden keine zusätzlichen Datenspuren erhoben, wie z. B. Ihr Aufenthaltsort. Die Angaben, die Sie im Rahmen der Befragung freiwillig machen, werden durchgängig vertraulich behandelt und ausschließlich für wissenschaftliche Zwecke genutzt.

4.2 Forschungsethische Besonderheiten von In-situ-Studien

Als kleines Dankeschön für Ihre Mühe verlosen wir unter allen Teilnehmenden der kompletten Studie einen Kopfhörer mit Lärmreduktionsfunktion nach Wahl im Wert von bis zu 245 Euro. Am Ende der Studie haben Sie Gelegenheit, an der Verlosung teilzunehmen.

Bitte geben Sie nun an, ob Sie unter diesen Bedingungen in die Teilnahme am zweiten Teil der Studie einwilligen oder nicht.

Ich habe die Studien- und Datenschutzinformationen gelesen und verstanden. Ich bin mit der Teilnahme am zweiten Teil der o. g. Studie für wissenschaftliche Forschungszwecke einverstanden. Ich versichere außerdem, dass ich 16 Jahre oder älter bin.

- ☐ Ja
- ☐ Nein

Die oben beispielhaft aufgeführten sowie ggf. weitere Entscheidungsdimensionen, wie etwa das Erkenntnisinteresse (vgl. Heise & Schmidt, 2014), müssen für die Spezifika jeder Studie adaptiert und unter Zuhilfenahme der vorgeschlagenen Prinzipien der Selbstbestimmung, Schadensvermeidung und Gerechtigkeit bewertet werden, um konkrete forschungsethische Herausforderungen im Kontext einer eigenen In-situ-Studie zu reflektieren und geeignete Lösungen zu entwickeln.

5 In-situ-Designs in verschiedenen Untersuchungsanlagen

5.1 Kombination von In-situ-Designs mit anderen Erhebungsmethoden

Häufig werden In-situ-Designs mit anderen Erhebungsverfahren kombiniert. Im Folgenden stellen wir die gängigsten Methodenkombinationen vor (siehe Schnauber-Stockmann & Karnowski, 2020).

Sehr gängig ist die Kombination mit einer **quantitativen Vorab- und/oder Nachbefragung**. Die Vorabbefragung eignet sich für die Erhebung stabiler Personenmerkmale. Weiterhin dient sie häufig dem sog. Screening, also der Vorauswahl der Teilnehmer:innen hinsichtlich forschungsrelevanter Merkmale. Beispielsweise kann es das Ziel sein, Personen zu identifizieren, die zumindest zweimal wöchentlich Podcasts hören oder einer bestimmten Altersgruppe angehören. Schließlich können im Rahmen der Vorabbefragung wichtige Studieninformationen vermittelt und Informationsmaterialien sowie Kontaktmöglichkeiten bereitgestellt werden. Im Rahmen der Vorabbefragung wird oft auch der Informed Consent für alle Teile der Studie eingeholt (siehe Abschnitt 4.2). Die Nachbefragung wiederum kann dazu dienen, weitere Personenmerkmale zu erfassen, die Teilnehmer:innen nachträglich über den genauen Zweck der Untersuchung aufzuklären (Debriefing) oder um Nachfragen zur In-situ-Phase zu stellen. Die Abfrage von Kontaktdaten zur Ausschüttung von Incentives findet häufig ebenfalls im Rahmen der Nachbefragung statt. In der Podcastingstudie (siehe Studiensteckbrief im Online-Anhang zum Buch) wurde die In-situ-Phase mit einer Vorab- und einer Nachbefragung kombiniert.

5.1 Kombination von In-situ-Designs mit anderen Erhebungsmethoden

Die Vorabbefragung (in Form einer herkömmlichen Onlineumfrage) enthielt die erste von zwei Einwilligungserklärungen. Zudem wurden dort stabile Personenmerkmale erhoben, z. B. Nutzungsgewohnheiten und allgemeine Bewertungen von True-Crime-Podcasts. Abschließend wurden die Befragten über den Ablauf des In-situ-Teils der Studie informiert und um Einwilligung zur ESM-Phase gebeten. In der Nachbefragung wurden u. a. methodische Merkmale – insbesondere hinsichtlich der Reaktivität – erfragt.

MERKE

Fast immer werden In-situ-Studien mit einer **Vorabbefragung** kombiniert, die u. a. der Rekrutierung, dem Screening, der Messung von Personenmerkmalen und dem Einholen des Informed Consent dient.

In-situ-Methoden bieten sich auch in Kombination mit **qualitativen Befragungen** an. Die *in situ* gesammelten Informationen werden in dieser Kombination z. B. als Prompts für qualitative Interviews genutzt. In der Studie von Pigg et al. (2014) zum Thema Schreibverhalten und Bedeutung von Schreiben im Alltag wurden die Teilnehmer:innen z. B. eine Woche lang fünfmal am Tag per SMS dazu befragt, was sie gerade schreiben oder kurz zuvor geschrieben hatten. Diese Antworten wurden genutzt, um im Rahmen qualitativer Interviews das Schreibverhalten der Teilnehmer:innen näher zu beleuchten.

Schließlich können In-situ-Studien mit **digitalen Verhaltensdaten** kombiniert werden (De Calheiros Velozo et al., 2022; Keusch & Conrad, 2022; siehe zur Verknüpfung von digitalen Verhaltens- und Befragungsdaten allgemein z. B. Stier et al., 2020). Zwei Zugangswege zu digitalen Verhaltensdaten sind im Zusammenhang mit In-situ-Studien zentral: Tracking und Datenspenden. Beide orientieren sich an den Nutzer:innen und erlauben die Sammlung von digitalen Verhaltensdaten auf der individuellen Ebene, d. h., wir können einzelnen Teilnehmer:innen ihre individuellen digitalen Verhaltens-

spuren zuordnen (z. B. ihre Nutzungshäufigkeit von TikTok). Trackingdaten setzen voraus, dass die Teilnehmer:innen ein Programm auf ihrem Gerät installieren, das während der Studienzeit für die Forschungsfrage relevante Daten (und nur diese, Stichwort ‚Datensparsamkeit‘, siehe Abschnitt 4.2) aufzeichnet und an die Forschenden übermittelt (siehe z. B. Christner et al., 2022; Ohme et al., 2024). Die Studie von Aalbers et al. (2022) (siehe Studiensteckbrief im Online-Anhang zum Buch) kombiniert z. B. ESM- mit Trackingdaten.

Im Rahmen von Datenspenden stellen die Teilnehmer:innen ihre digitalen Verhaltensdaten zu Forschungszwecken zur Verfügung, sie ‚spenden‘ sie an die Forschenden. Dabei kann es sich um unterschiedliche Daten handeln, z. B. Data Download Packages von Social-Media-Plattformen (siehe z. B. Carrière et al., 2025; Driel et al., 2022; Ohme et al., 2024) oder selbstgenerierte Daten wie z. B. Bildschirmfotos oder Videos (z. B. Baumgartner et al., 2022). Datenspenden erfreuen sich u. a. aus forschungsethischen Gründen großer Beliebtheit, denn sie beziehen die Nutzer:innen aktiv in den Erhebungsprozess ein und ermöglichen ihnen einen souveränen Umgang mit den eigenen Daten (siehe z. B. Boeschoten et al., 2022).

Ein weiterer verbreiteter Zugang zu digitalen Verhaltensdaten sind sog. Application Programming Interfaces (APIs). Die Forschenden können über eine Softwareschnittstelle auf Plattformdaten zugreifen. Häufig werden darüber aber keine personenbezogenen Daten übertragen, sodass ihr Einsatz für In-situ-Studien eingeschränkt ist. Ernst et al. (2026) haben jedoch im Rahmen einer ESM-Studie über die Spotify-API Daten auf der Individualebene gesammelt. Einen ausführlichen Überblick über die verschiedenen Zugänge zu digitalen Verhaltensdaten (bzgl. der Social-Media-Nutzung) geben Ohme et al. (2024).

Die Kombination von In-situ-Designs, die auf Selbstauskünften beruhen, mit digitalen Verhaltensdaten kann aus mehreren Gründen sinnvoll sein. Zum einen können über das Smartphone getrackte digitale Verhaltensdaten für die automatische Auslösung von Protokollen bei ereignisabhängigen Stichproben dienen (siehe hierzu Abschnitt 6.3.2.1). Zum anderen können digitale

Verhaltensdaten die über die Protokolle erhobenen Selbstauskünfte um Beobachtungsdaten ergänzen. Das kürzer gehaltene Protokoll kann die Teilnehmer:innen entlasten, weil bestimmte Variablen (z. B. digitale Mediennutzung oder Aufenthaltsorte) automatisch erfasst (genauer: beobachtet) und nicht erfragt werden. Die Erfassung digitaler Verhaltensdaten kann zudem auch Probleme von Selbstberichten wie eine verzerrte Wahrnehmung oder mangelnde Erinnerung umgehen. Variablen, die keiner Selbstauskunft bedürfen oder bei denen Selbstauskünfte wenig zuverlässig sind, können mithilfe digitaler Verhaltensdaten erhoben und gemeinsam mit den Selbstauskünften aus den In-situ-Protokollen ausgewertet werden. Aalbers et al. (2022) konnten z. B. zeigen, dass Smartphonennutzung (Log-Daten-Erhebung) mit Prokrastination (Selbstreport) einhergeht (siehe Studiensteckbrief im Online-Anhang zum Buch). Müller et al. (2021) konnten ermitteln, dass Mobilitätsmuster (GPS-Tracking) mit Wohlbefinden (Selbstreport) zusammenhängen. Ernst et al. (2026) nutzten die Spotify-API, um die Nutzung von Musikstreaming zu erfassen. Sie testeten dann Mood-Management-Annahmen, indem sie die Daten mit Selbstauskunftsdaten zur aktuellen Stimmung kombinierten.

Bei der Einbindung von digitalen Verhaltensdaten sind eine Reihe von Aspekten zu beachten:

- Die Anforderungen an die Kompetenzen der Forschenden im Umgang mit den erhobenen Daten (siehe hierzu z. B. Breuer et al., 2020; De Calheiros Velozo et al., 2022; Stier et al., 2020) und an die Forschungsethik erhöhen sich (siehe hierzu z. B. Breuer et al., 2021).
- Die technischen Voraussetzungen müssen gegeben sein, etwa der Zugang zu Trackingssoftware oder zu Tools für Datenspenden (Ohme et al., 2024). Einige kommerzielle ESM-Tools verfügen über entsprechende Trackingoptionen, die meist aber auf einzelne Betriebssysteme, fast immer Android, beschränkt sind. Ein Überblick über vorhandene ESM-Software inklusive Informationen zu Trackingmöglichkeiten haben Wieland et al.

5 In-situ-Designs in verschiedenen Untersuchungsanlagen

(2025) zusammengestellt und online verfügbar gemacht (der Link findet sich im Online-Anhang zum Buch unter Ressourcen). Für die Sammlung von Datenspenden stehen bisher nur wenige Tools zur Verfügung (z. B. Araujo et al., 2022; Boeschoten et al., 2023; Haim et al., 2023; Pfiffner et al., 2024). Teilweise kommen für einzelne Projekte individuell entwickelte Apps und Tools zum Einsatz.

- Wir müssen abwägen, ob die zusätzliche Belastung der Teilnehmer:innen gerechtfertigt ist. Die Erhebung digitaler Verhaltensdaten kann zwar die Länge des Protokolls verkürzen, sie ist aber häufig mit anderen Belastungen für die Teilnehmer:innen verbunden. Im Falle von Tracking z. B. mit kürzeren Akkulaufzeiten von Smartphones, im Falle von Datenspenden mit zusätzlichem Aufwand, etwa bei der Anforderung und dem Hochladen von Data Download Packages, oder auch dem zusätzlichen Tragen von Wearables. Ganz generell verschärfen sich Datenschutzbedenken mit der umfassenderen Preisgabe von Informationen.

Diese Einschränkungen verringern die Teilnahmebereitschaft oft deutlich (De Calheiros Velozo et al., 2022; Hase & Haim, 2024; Keusch et al., 2019; Welbers et al., 2024). Wir empfehlen deshalb eine über dieses Buch hinausgehende intensive Einarbeitung der Forschenden in das Thema (siehe z. B. De Calheiros Velozo et al., 2022; Harari et al., 2016; Keusch et al., 2019; Müller et al., 2021; Ohme et al., 2024; Pfiffner et al., 2024; Welbers et al., 2024).

MERKE

Die Kombination von In-situ-Methoden mit Beobachtungsverfahren, die digitale Verhaltensdaten speichern, ermöglicht eine genauere Erfassung individueller Mediennutzung sowie weiterer Merkmale der Situation, indem Selbstberichte durch objektive Verhaltensmessungen ergänzt und Verzerrungen redu-

ziert werden. Digitale Verhaltensdaten werden vor allem über Tracking oder Datenspenden gewonnen. Beide Verfahren erfordern allerdings umfassende technische, datenschutzrechtliche und ethische Abwägungen.

5.2 Measurement-Burst-Designs

„Einfache“ In-situ-Studien dienen aufgrund der überwiegend kurzen Feldzeit von einigen Tagen vor allem dazu, eher kurzfristige Zusammenhänge und Phänomene zu untersuchen. Es sind aber auch Fragestellungen vorstellbar, bei denen Zusammenhänge oder Phänomene über längere Zeiträume von Interesse sind. Wir könnten uns etwa dafür interessieren, wie sich die Nutzung von politischen Podcasts und deren Einfluss auf Einstellungen in den verschiedenen Phasen eines Wahlkampfes (kein Wahlkampf, Beginn des Wahlkampfes, „heiße Phase“ des Wahlkampfes und direkt nach der Wahl) unterscheiden. Ein weiteres zu untersuchendes Thema könnte die Nutzung von Podcasts zur Stimmungsregulierung in verschiedenen Phasen der Jugend sein.

Für die Untersuchung solcher oder ähnlicher Fragestellungen eignen sich sog. Measurement-Burst-Designs (z. B. Cho et al., 2019; Nestler, 2021; Sliwinski, 2008; Stawski et al., 2015). Measurement Bursts sind eine spezielle Art von Längsschnittdesigns, bei denen mehrere Bursts, also In-situ-Studien, in eine größere Längsschnittstudie eingebettet werden. Dieselben Personen nehmen nicht nur einmal an einer In-situ-Studie teil, sondern die In-situ-Phase wird nach einer gewissen Zeit (z. B. nach einem Monat, einem Jahr) mehrfach wiederholt. Dieses Design ermöglicht sowohl die Untersuchung von kurzfristigen intraindividuellen Zusammenhängen innerhalb der einzelnen Bursts wie jedes übliche In-situ-Design auch. Darüber hinaus können Veränderungen in den Prävalenzen oder Mittelwerten über die Zeit hinweg untersucht werden, z. B. ob in Wahlkampfzeiten mehr politische Podcasts gehört werden als außerhalb oder

ob die Podcastnutzung im Verlauf der Jugend zunimmt. Außerdem können Veränderungen in den intraindividuellen Zusammenhängen zwischen den Bursts untersucht werden, etwa ob politische Podcasts einen größeren Einfluss auf Einstellungen in Wahlkampfzeiten haben als außerhalb oder ob die Stimmungsregulierung mit Podcasts in der späten Jugend erfolgreicher ist als in der frühen (zu Auswertungsmöglichkeiten siehe z. B. Nestler, 2021).

Es muss gut abgewogen werden, ob ein solches Design zum Einsatz kommen sollte, denn es ist deutlich ressourcenintensiver als übliche In-situ-Studien. Damit einher geht eine deutlich höhere Belastung für die Teilnehmer:innen, weshalb mit geringerer Teilnahmebereitschaft und höheren Abbruchquoten zu rechnen ist (Sliwinski, 2008).

5.3 Interventions- und Experimentaldesigns

Wie im einleitenden Kapitel ausgeführt, besteht eine zentrale Idee von In-situ-Studien darin, den Alltag möglichst umfassend abzubilden. Es sind aber auch Fragestellungen bzw. Zielsetzungen denkbar, bei denen es sinnvoll ist, dass die Forschenden in die üblichen Abläufe des Alltags eingreifen.

Zum einen können In-situ-Designs im Rahmen von **Interventionsstudien** eingesetzt werden, also Studien, die zur Entwicklung und Prüfung spezifischer Maßnahmen dienen. Im Mittelpunkt können etwa die Wirksamkeit gesundheitsfördernder Handlungen (z. B. Mikesell et al., 2018) oder die Effekte des digitalen Fastens auf das Wohlbefinden stehen (z. B. Hall et al., 2019). Beispielsweise könnten wir zur Steigerung der Informationsnutzung eine Interventionsstudie planen, in der die Teilnehmer:innen mehrfach im Studienverlauf per SMS informiert werden, dass sie sich zeitnah über ein politisches Thema unterhalten sollen. Mit In-situ-Protokollen könnte anschließend geprüft werden, ob sie deshalb mehr politische Podcasts hören. Denkbar ist auch, dass das Ausfüllen der In-situ-Protokolle selbst Teil der Intervention ist, etwa wenn Personen in den Protokollen

über ihre gesundheitliche Verfassung reflektieren sollen und diese Aufzeichnungen die Aufmerksamkeit für die eigene Gesundheit erhöhen (*ecological momentary interventions*, EMI, Kuppens & Myin-Germeyns, 2022; May et al., 2018)

Zum anderen kann der (strenge) Kausalnachweis wichtig sein. Es handelt sich bei In-situ-Studien zwar um Längsschnittstudien, womit – je nach Auswertungsvorgehen (siehe Abschnitt 9.5.2.7) – zumindest die zeitliche Reihenfolge als ein Merkmal von Kausalität berücksichtigt werden kann (die Ursache muss zeitlich vor der Wirkung liegen). Ein Kausalnachweis ist aber nur im Rahmen eines **Experimentaldesigns** möglich, bei dem neben der zeitlichen Abfolge von Variablen auch mögliche Störvariablen kontrolliert werden (zu In-situ-Studien mit Experimentaldesign siehe Neubauer & Schmiedek, 2020; Schmiedek & Neubauer, 2020). Bei einer Kombination von In-situ-Studien mit experimentellen Designs handelt es sich grundsätzlich um Feldexperimente (zu Feldexperimenten allgemein siehe z. B. T. Koch et al., 2025). Denn durch die In-situ-Abfrage im Alltag der Proband:innen über einen längeren Zeitraum hinweg können Störvariablen nicht (komplett) kontrolliert werden.

Im Rahmen von Experimentalstudien ist sowohl ein Between-Person-Design denkbar – verschiedene Teilnehmer:innen werden in Experimental- und Kontrollgruppen eingeteilt – als auch ein Within-Person-Design, indem im Studienverlauf die Teilnehmer:innen verschiedene Experimentalbedingungen durchlaufen. Häufig sind die o. g. Interventionsstudien ebenfalls experimentell angelegt, um die Wirksamkeit bestimmter Interventionen zu testen. Hall et al. (2019) untersuchten z. B. den Zusammenhang zwischen Social-Media-Fasten und Wohlbefinden. Sie teilten ihre Teilnehmer:innen in fünf Experimentalgruppen ein: kein Social-Media-Fasten, eine, zwei, drei oder vier Wochen Social-Media-Fasten. In Daily Diaries berichteten die Teilnehmer:innen dann über verschiedene Dimensionen des Wohlbefindens. Ein Zusammenhang mit digitalem Fasten konnte jedoch nicht gezeigt werden.

5 In-situ-Designs in verschiedenen Untersuchungsanlagen

MERKE

In-situ-Studien erfassen alltägliches Leben (Kognitionen, Emotionen und Verhalten), können aber durch gezielte Eingriffe – etwa in Form von Interventionen oder experimentellen Designs – erweitert werden.

6 Ziehung von Personen- und Situationsstichproben

Die Stichprobenziehung (das sog. Sampling) in In-situ-Studien bewegt sich immer auf zwei Ebenen: Auf der Personenebene wird, wie wir es von allen Arten von Befragungsstudien kennen, eine Gruppe an Teilnehmer:innen ausgewählt (Personenstichprobe), die häufig stellvertretend für eine größere Gruppe stehen. Auf der Situationsebene wird über einen gewissen Zeitraum hinweg eine Menge an Situationen aus dem Alltag dieser Teilnehmer:innen ausgewählt (Situationsstichprobe). Genauso wie die Stichprobe der Teilnehmer:innen häufig dazu dient, auf eine größere Gruppe an Personen zu schließen, dient die Situationsstichprobe dazu, Einblick in die Vielfalt alltäglicher Situationen im Leben der Teilnehmer:innen zu erhalten.

Während die folgenden Ausführungen zu Stichprobengröße, Erhebungsdichte und -zeitraum sowie für die Personenstichprobe auch für Tagebuchstudien gelten, ist die Ziehung der Situationsstichprobe spezifisch für ESM-Designs. In Tagebuchstudien werden auf der Situationsebene fast immer ganze Tage oder zusammenhängende Tagesabschnitte ausgewählt. ESM-Studien erheben hingegen mehrmals täglich über mehrere Tage hinweg Selbstauskünfte von den Teilnehmer:innen. Damit ziehen solche Studien – wie auch der Name Experience Sampling besagt – eine Stichprobe an Situationen oder Momenten aus dem Alltag der Teilnehmer:innen, um von dieser Stichprobe auf die Gesamtheit des alltäglichen Verhaltens, der Kognitionen und Emotionen zu schließen. Die Teilnehmer:innen werden dann gebeten, wiederholt Auskunft zu aktuellen oder kurz zurückliegenden Situationen zu geben, z. B. zu ihrem Mediennutzungsverhalten in den letzten fünf Minuten.

MERKE

In-situ-Studien erfordern eine Stichprobenziehung auf zwei Ebenen. Zum einen werden auf der Personenebene Teilnehmer:innen ausgewählt, die eine größere Gruppe an Personen repräsentieren (Personenstichprobe). Zum anderen wird auf Situationsebene eine Stichprobe an Situationen aus dem Alltag gezogen, die den Alltag der Teilnehmer:innen repräsentiert (Situationsstichprobe).

6.1 Stichprobengröße, Erhebungsdichte und -zeitraum

Zunächst stellt sich wie bei jeder Studie die Frage nach der benötigten Stichprobengröße. Bei In-situ-Designs müssen wir hierfür beide Ebenen berücksichtigen – die Personen- und die Situationsebene.

Auf der Personenebene wird die Stichprobengröße ausschließlich von der Anzahl der Personen bestimmt, die an der Studie teilnehmen (sollen). Auf der Situationsebene ergibt sich die Stichprobengröße aus mehreren Aspekten: zunächst natürlich aus der Anzahl der Teilnehmer:innen – denn je mehr Personen teilnehmen, desto größer wird auch die Situationsstichprobe. Darüber hinaus ist die Länge der Feldzeit (Erhebungszeitraum) in Kombination mit der Anzahl der auszufüllenden Protokolle pro Tag (Erhebungsdichte) entscheidend. Wir können dieselbe Stichprobengröße mit einer Studie mit einem Tag Feldzeit und 15 Protokollen pro Tag erreichen wie mit einer Studie, die 15 Tage Feldzeit hat und pro Tag nur ein Protokoll (z. B. einer Tagebuchstudie). Deshalb ist es bei In-situ-Studien nicht nur relevant, dass wir die benötigte Stichprobengröße bestimmen (siehe hierzu weiter unten), sondern auch, wie die Situationsstichprobe sich innerhalb von Erhebungstagen und über den gesamten Erhebungszeitraum hinweg verteilt. Diese Entscheidungen sind im Zusammenhang mit In-situ-Studien zentral und hängen vom Forschungsinter-

esse ab: Interessieren wir uns für Phänomene, die im Alltag häufig auftreten, z. B. den Zusammenhang von sozialer Mediennutzung und Wohlbefinden, sind kürzere Erhebungszeiträume ausreichend. Eine hohe Erhebungsdichte mit geringeren Abständen zwischen den einzelnen Messungen ist sinnvoll, wenn wir beabsichtigen, die Veränderung der gemessenen Variable nachzuzeichnen oder zeitversetzte Zusammenhänge zu analysieren (siehe Abschnitt 9.5.2.7). Gehen wir hingegen von einer geringeren Variabilität im Tagesverlauf aus oder von seltener auftretenden Phänomenen, z. B. der Nutzung spezieller Nischeninhalte, dann passt eine geringere Erhebungsdichte und ein längerer Erhebungszeitraum besser zu unserem Phänomen.

MERKE

Sowohl die Länge des Erhebungszeitraums (Feldzeit) als auch die Erhebungsdichte (Anzahl auszufüllender Protokolle pro Tag) sind zentrale Designentscheidungen in In-situ-Studien. Sie sind abhängig von der vermuteten zeitlichen Variabilität der gemessenen Konstrukte und bestimmen die Stichprobengröße auf der Situationsebene entscheidend mit.

Für die Begründung und Bestimmung der angemessenen Stichprobengröße – sowohl auf der Personen- als auch auf der Situations-ebene – gibt es verschiedene Herangehensweisen (ausführlich Lakens, 2022). Der „Königsweg“ ist die A-priori-Teststärkenanalyse (für detaillierte Ausführungen siehe Döring & Bortz, 2016, S. 814ff.). Für die Durchführung einer A-priori-Teststärkenanalyse, die oftmals als Poweranalyse bezeichnet wird, müssen wir vor der Durchführung der Untersuchung u. a. das anzustrebende Signifikanzniveau, die Power (Teststärke), die zu erwartenden Effektgrößen oder die sog. *smallest effect sizes of interest*, also die geringste Effektstärke, die noch als realweltlich bedeutsam angesehen wird (siehe z. B. Lakens, 2022), und das Auswertungsverfahren festlegen. Auf dieser Basis lässt sich der optimale Stichprobenumfang berechnen. Die Festlegung von Teststärke und Signifikanzniveau ist im Allgemeinen kein Problem, weil

sich die KMW an den Konventionen der Sozialwissenschaften orientiert. Üblicherweise werden ein Signifikanzniveau von $\alpha = .05$ und eine Teststärke von $1 - \beta = .80$ gesetzt. Schwieriger ist die Vorabfestlegung der zu erwartenden Effektstärke und je nach Auswertungsverfahren weiterer Parameter. Hier müssen wir je nach konkretem Forschungsinteresse auf ein Zusammenspiel an theoretischen Modellen, vorhandenen empirischen Ergebnissen, eigenen Vorstudien oder Simulationen zurückgreifen.

Während es für gängige Studiendesigns und Auswertungsverfahren standardisierte Tools zur Berechnung der Power gibt (beispielsweise G*Power, Faul et al., 2009), ist die Berechnung der angestrebten Stichprobengröße bei In-situ-Studien komplexer. Durch die Mehrebenenstruktur der erhobenen Daten muss auch die Vorgehensweise einer A-priori-Poweranalyse diesen Gegebenheiten angepasst werden. Konkret müssen die entsprechenden Berechnungen der Tatsache Rechnung tragen, dass komplexere Auswertungsverfahren notwendig sind, die Varianz der abhängigen Variable in (mindestens) zwei Ebenen (Personen- und Situationsebene) zerlegt werden kann und die unabhängigen Variablen sowohl auf der Personen- als auch auf der Situationsebene angesiedelt sein können (siehe Kapitel 9). Empfehlungen für entsprechende Poweranalysen liegen erst seit Kurzem vor, wobei sich simulationsbasierte Vorgehensweisen nicht zuletzt aufgrund ihrer breiteren Anwendbarkeit im Vergleich zu anderen Ansätzen durchzusetzen scheinen (für eine ausführlichere Diskussion siehe Lafit, 2022). Aktuell stehen wenige Anwendungen für entsprechende A-priori-Poweranalysen im Zusammenhang mit spezifischen Auswertungsverfahren zur Verfügung, u. a. *PowerAnalysisIL* (Lafit, 2022) und *SIMR* (Arend & Schäfer, 2019).

Allerdings sind diese Verfahren erstens oft sehr rechenintensiv Tuerlinckx et al. (2026). Zweitens erfordern Simulationen bereits umfassende Vorannahmen über die Struktur der Daten, die oft nicht bekannt sind und dann geschätzt oder vorab simuliert werden müssen. Dieses Vorgehen ist nicht in allen Fällen sinnvoll und ausreichend informiert möglich, um solide Berechnungen durchzuführen. Im Online-Anhang zum Buch unter Ressourcen stellen wir einen

vereinfachten R-Code zur Verfügung, der ohne allzu spezifische Vorannahmen einfache Poweranalysen ermöglicht. Aufgrund der sehr dynamischen Entwicklungen in diesem Feld empfehlen wir, vor Studienbeginn den aktuellen Stand der Entwicklungen zu recherchieren.

Doch selbst wenn wir keine A-priori-Poweranalyse durchführen können oder wollen, ist es wichtig, die Stichprobengröße zu planen und zu begründen. Ein weniger exakter, aber dennoch hilfreicher Zugang kann die Orientierung an entsprechender Methodenzliteratur sein. Speziell für Mehrebenenanalysen mit zwei Ebenen (eines der üblichsten Auswertungsverfahren für In-situ-Daten, siehe Abschnitt 9.5.2) geben Arend & Schäfer (2019) – basierend auf Monte-Carlo-Simulationen – hilfreiche Daumenregeln für typische Forschungsfragen (siehe auch Maas & Hox, 2005; weiterhin Park & Yu, 2018 für *Multilevel Latent Class Analysis*).

Schließlich müssen wir auch immer die zeitlichen und finanziellen Ressourcen der Forschenden sowie die Belastung der Teilnehmer:innen berücksichtigen (zur Integration dieser Parameter im Rahmen von A-priori-Poweranalysen siehe Revol et al., 2024). Entsprechende Erwägungen bei der konkreten Durchführung und Rekrutierung diskutieren wir in den Abschnitten 6.2.2 und 6.3.2.

6.2 Ziehung der Personenstichprobe

6.2.1 Grundlagen

Wie die meisten empirischen Studien in der KMW können auch In-situ-Studien nur in seltenen Fällen alle für das Forschungsinteresse relevanten Personen in die Studie mit einbeziehen, also eine Vollerhebung auf der Personenebene durchführen. Entsprechend müssen sich auch In-situ-Studien üblicherweise auf Teilerhebungen beschränken, also eine Stichprobe an Personen ziehen. Die Stichprobenziehung auf der Personenebene ist dabei nicht spezifisch für In-situ-Studien. Vielmehr erfolgt sie genauso, wie wir dies aus der KMW allgemein (vgl. Brosius et al., 2022) und insbesondere aus Befragungsstudien

(vgl. Möhring & Schlütz, 2025) kennen. Entsprechend werden wir an dieser Stelle nicht detailliert auf die einzelnen Aspekte der Stichprobenziehung eingehen, die sich an anderer Stelle nachlesen lassen. Wir geben lediglich einen kurzen Überblick in Anlehnung an die ausführliche Darstellung in Möhring & Schlütz (2025).

Zunächst stellt sich im Hinblick auf die Auswahl einer Stichprobe die grundlegende Frage, ob die Teilerhebung auch Aussagen über eine größere Gruppe, also über die Grundgesamtheit zulassen soll. Dieser Schritt ist möglich, wenn die Stichprobe repräsentativ für die Grundgesamtheit ist und ein verkleinertes Abbild dieser liefert. Es hängt von der Art der Stichprobenauswahl ab, ob dies tatsächlich zutrifft. In Analogie zu allgemeinen Befragungsstudien kann die Personenstichprobe in In-situ-Studien mithilfe verschiedener Auswahlverfahren gezogen werden. Dabei lässt sich zwischen zufallsgesteuerten und nicht zufallsgesteuerten Auswahlverfahren unterscheiden (vgl. Abbildung 6.1).

Auswahlverfahren Personenstichprobe			
zufallsgesteuert		nicht zufallsgesteuert	
<i>einstufig</i>	<i>mehrstufig</i>	<i>willkürlich</i>	<i>bewusst</i>
• einfache Zufallsauswahl • systematische Zufallsauswahl	• geschichtete Auswahl • Klumpenauswahl • Kombinationsverfahren	Convenience Sample, z. B. • Ad-hoc-Stichproben • Schneeballverfahren	• Quota-Verfahren

Abbildung 6.1: Auswahlverfahren für die Personenstichprobe (Möhring & Schlütz, 2025, S. 28)

6.2.1.1 Zufallsgesteuerte Auswahlverfahren

Ein zufallsgesteuertes Auswahlverfahren (Zufallsstichprobe) liegt vor, wenn alle Elemente der Grundgesamtheit die gleiche mathematisch berechenbare Chance haben, ausgewählt zu werden. Die Auswahl eines Elements darf nicht die Chance eines anderen Elements beeinflussen, in die Stichprobe zu gelangen. Generell lassen nur zufallsgesteuerte Auswahlverfahren Rückschlüsse auf die Grundgesamtheit im Sinne einer Repräsentativität der Stichprobe zu.

MERKE

Repräsentativität hängt vom Auswahlverfahren ab und ist streng genommen nur mit Zufallsstichproben zu erreichen.

Bei einem einstufigen Verfahren findet die Zufallsauswahl in einem Schritt statt. Dies kann als einfache oder als systematische Zufallsauswahl erfolgen. Bei der **einfachen Zufallsauswahl** werden die einzelnen Teilnehmer:innen der Stichprobe im Sinne eines mathematischen Lotterieverfahrens aus der Grundgesamtheit ausgewählt. Für eine Studie zu den Hörgewohnheiten könnten z. B. alle Abonnent:innen eines bestimmten Podcasts die Grundgesamtheit bilden. Die Personenstichprobe für unsere Studie können wir nun mithilfe einer einfachen Zufallsauswahl ziehen, indem wir z. B. aus einer Liste aller Abonnent:innen durch Zufallszahlen die Teilnehmer:innen für unsere Stichprobe auswählen. Eine **systematische Zufallsauswahl** liegt vor, wenn jedes n -te Element aus der Grundgesamtheit in die Stichprobe aufgenommen wird. Aus der bereits erwähnten Liste der Abonnent:innen des Podcasts würden wir beispielsweise jede 25. Person für unsere Stichprobe auswählen. Dieses Vorgehen setzt voraus, dass eine solche Liste vorliegt. Wichtig ist zudem, dass die Anordnung der Grundgesamtheit nicht auf einem für das Untersuchungsinteresse relevanten Merkmal basiert, die Liste also z. B. nicht nach der Nutzungshäufigkeit geordnet ist.

Mehrstufige, zufallsgesteuerte Auswahlverfahren sind im Unterschied zu einstufigen Methoden notwendig, wenn die Grundgesamtheit nicht vollständig erfasst ist und beispielsweise in Form einer Liste vorliegt, was eher die Regel als die Ausnahme sein dürfte. Wie der Name bereits nahelegt, vollzieht sich die Auswahl bei diesen Verfahren nicht in einem Schritt, sondern mithilfe mehrerer hintereinandergeschalteter Auswahlsschritte. Bei der **geschichteten Auswahl** wird die Grundgesamtheit zunächst anhand eines für das Forschungsinteresse relevanten Merkmals in verschiedene Teilgesamtheiten aufgeteilt. Diese Schichten sind in sich möglichst homogen, aber zueinander mit Blick auf das Schichtungsmerkmal heterogen, also z. B. Nutzer:innen, die Podcasts häufig oder wenig nutzen. Aus den verschiedenen Schichten werden die Elemente der Stichprobe dann mithilfe einer einfachen oder systematischen Zufallsauswahl ausgewählt. In unserer Studie zu den Hörgewohnheiten der Abonnent:innen eines bestimmten Podcasts könnten wir die Abonnent:innen zunächst anhand ihrer Hörfrequenz (jede Episode, jede zweite Episode, seltener) schichten und dann aus jeder Schicht eine einfache Zufallsauswahl mithilfe einer Zufallszahl treffen. Geschichtete Stichproben verbessern die Stichprobenqualität vor allem von kleinen Stichproben, weil die Repräsentation wichtiger Merkmale sichergestellt wird. Eine Voraussetzung ist allerdings, dass die Merkmale und deren Verteilung bekannt sind.

Bei der **Klumpenauswahl** sind die einzelnen Klumpen in sich hingegen möglichst heterogen und zueinander homogen. Die Einteilung in Klumpen basiert also auf einem Merkmal, das für das Forschungsinteresse nicht relevant ist, aber idealerweise natürlich vorkommt. Die einzelnen Klumpen kann man sich als verkleinerte Grundgesamtheit vorstellen. Für die Stichprobe werden vollständige Klumpen ausgewählt. Eine häufige Art der Auswahl von Klumpenstichproben ist die Auswahl ganzer Schulklassen. Am Beispiel unserer Studie zur Podcastnutzung wäre die Auswahl der Stichprobe mithilfe einer Klumpenstichprobe beispielsweise in Form einer Studie in mehreren Schulklassen möglich.

6.2.1.2 Nicht zufallsgesteuerte Auswahlverfahren

Bei der nicht zufallsgesteuerten Auswahl wird zwischen bewussten und willkürlichen Auswahlverfahren unterschieden. **Willkürliche Auswahlverfahren** (Convenience Sampling) sind weder systematisch noch repräsentativ. Entsprechend sind Rückschlüsse auf die Grundgesamtheit nicht möglich. Aus stichprobentheoretischer Sicht sind willkürliche Stichproben problematisch, denn Personen, die zur Grundgesamtheit gehören, werden willkürlich in die Stichprobe aufgenommen. Ihre Teilnahme liegt nur im eigenen (sog. Selbstrekrutierung, z. B. durch das Klicken auf gepostete Teilnahmelinks in den sozialen Medien) bzw. im Ermessen der Rekrutierenden. Beispielsweise können in einem Schneeballverfahren Teilnehmer:innen gebeten werden, die Aufforderung zur Studienteilnahme (z. B. den Teilnahmelink) selbst weiterzuleiten.

Bei Convenience Samples lässt sich die Wahrscheinlichkeit nicht angeben, mit der eine bestimmte Person aus der Grundgesamtheit in die Stichprobe gelangt. Dies ist aber eine Voraussetzung für inferenzstatistische Verfahren, die in der Datenauswertung angewendet werden. Trotzdem finden auch willkürliche Auswahlverfahren immer wieder Anwendung, denn sie sind zumindest geeignet, Aussagen über Zusammenhänge innerhalb der Stichprobe zu treffen. In unserer Beispielstudie zur Podcastnutzung haben wir den Link zum Fragebogen auf der Website und den Social-Media-Kanälen beliebter Podcasts gepostet und die Hörer:innen konnten selbst entscheiden, ob sie teilnehmen wollten oder nicht. Es handelt sich also um eine nicht-zufallsgesteuerte, genauer um eine willkürliche Stichprobenziehung bzw. ein Convenience Sample (siehe Studiensteckbrief im Online-Anhang zum Buch).

Neben diesen willkürlichen Auswahlverfahren können die Elemente der Stichprobe bei sog. Quota-Verfahren auch **bewusst** ausgewählt werden. Voraussetzung für ein Quota-Verfahren ist allerdings, dass bestimmte Merkmalsverteilungen der Grundgesamtheit bekannt sind (also z. B. der Anteil der True-Crime-Fans unter den Podcasthörer:innen). Anhand dieser Verteilungen lässt sich gezielt

eine Stichprobe als verkleinertes Abbild der Grundgesamtheit zusammenstellen. Diese Stichprobe kann – zumindest mit Blick auf diese Merkmale – wiederum als repräsentativ für die Grundgesamtheit gelten und lässt entsprechende Rückschlüsse zu.

6.2.2 Praktische Umsetzung und Herausforderungen

Nachdem wir im vorherigen Abschnitt grundsätzliche Vorgehensweisen bei der Ziehung von Personenstichproben beschrieben haben, geht es nun um die praktischen Abwägungen in Bezug auf das Sampling bei In-situ-Studien.

6.2.2.1 Validität und Repräsentativität der Personenstichprobe

Die Stichprobengüte auf der Personenebene bestimmt mit, wie extern valide, d. h. verallgemeinerbar die Ergebnisse einer In-situ-Studie sind (siehe auch Exkurs in Kapitel 2). Es ist eine Grundfrage der Stichprobenziehung, ob eine Stichprobe Rückschlüsse – statistisch gesprochen Inferenzschlüsse – auf die Grundgesamtheit zulässt. Aus strikt inferenzstatistischer Sicht kann dies nur für Stichproben gelten, die auf einer Zufallsauswahl basieren, weil nur in diesem Fall auf die Wahrscheinlichkeitstheorie zurückgegriffen werden kann. Doch selbst unter diesen Bedingungen müssen Fehler berücksichtigt werden, die die Repräsentativität und damit die externe Validität einer Personenstichprobe einschränken (*Total Survey Error*, Biemer, 2010; Groves & Lyberg, 2010; Wolf et al., 2016; vgl. Sen et al., 2021 für ein Anwendungsbeispiel). Zum einen kann die Grundgesamtheit im ersten Schritt nicht richtig bestimmt worden sein, womit auch die zufällige Auswahl der Elemente der Stichprobe fehlerhaft wird. Dieser Fehler wird als Coverage Error bezeichnet. Wenn wir also in unserer Studie zur Podcastnutzung die Abonnent:innen eines bestimmten Podcasts (z. B. eines wöchentlich ausgestrahlten True-Crime-Angebots) als Grundgesamtheit definieren und daraus eine Zufallsauswahl treffen, aber Rückschlüsse auf

alle Podcasthörer:innen – also auf eine andere Grundgesamtheit – ziehen, unterläuft uns dieser Fehler. Zudem können wir bei der Zufallsauswahl an sich Fehler machen (Sampling Error). Beispielsweise könnten wir bei einer systematischen Zufallsauswahl aus den Abonnent:innen eines Podcasts die Liste immer abwechselnd mit Männern und Frauen besetzen. Wählen wir nun jedes zehnte Element von der Liste, haben wir nur Frauen in der Stichprobe und damit einen Samplingfehler verursacht.

Einen Sonderfall bei der Betrachtung der Repräsentativität von Stichproben und damit der Übertragbarkeit von Aussagen auf die Grundgesamtheit sind Stichproben, die bewusst mithilfe eines Quota-Verfahrens zusammengestellt wurden. Diese Stichproben geben zwar ein verkleinertes Abbild der Grundgesamtheit wieder, jedoch nur mit Blick auf die quotierten Merkmale. Fallweise muss diskutiert werden, inwieweit und unter welchen Umständen diese Merkmale Rückschlüsse auf die Grundgesamtheit zulassen (vgl. u. a. Döring & Bortz, 2016).

Zuletzt kann – unabhängig vom Stichprobenverfahren – auch der Response Error die Repräsentativität unserer Stichprobe einschränken. Nicht alle für eine Mitwirkung an unserer Studie ausgewählten Personen werden auch teilnehmen. Es ist ein verbreitetes Problem im Zusammenhang mit Selbstselektion, dass die geplante von der tatsächlich realisierten Stichprobe abweicht und es zu einem Auswahlfehler auf dieser Ebene kommt. Gerade bei für die Teilnehmer:innen aufwändigen Designs wie In-situ-Studien kann sich der Response Error als ein zentrales Problem erweisen. So ist es möglich, dass potenzielle Teilnehmer:innen gar nicht erst bereit sind, an der Studie teilzunehmen. Dieses Problem ist nicht spezifisch für In-situ-Studien und generell unter dem Stichwort Response Rate oder Teilnahmebereitschaft hinreichend erforscht (siehe z. B. Groves & Couper, 1998; Schnell, 1997). Zudem gibt es Personen, die zunächst zwar teilnehmen, dann aber während der Feldphase abbrechen. Dieser Ausfall von Teilnehmer:innen ist typisch für Panelstudien und wird auch als Panelmortalität bezeichnet (Longhi, 2017; Scherer & Naab, 2013). Nichtteilnahme und Abbrüche können systematisch sein und die Er-

gebnisse verzerren, d. h. die Personen, die nicht an der Studie teilnehmen oder die abbrechen, unterscheiden sich von denen, die bis zum Schluss mitmachen (Myin-Germeys & Kuppens, 2022a; Rintala et al., 2019; Vachon et al., 2019). In vielen Fällen steigen die Teilnehmer:innen in einer frühen Phase der Studie, d. h. in den ersten Tagen aus (Stone et al., 1991). Das lässt sich kaum verhindern. Wenn durch eine Vorabbefragung grundlegende Informationen von diesen Teilnehmer:innen vorliegen, kann zumindest nach Ende der Feldzeit geprüft werden, welche systematischen Ausfälle es gab (siehe hierzu auch Abschnitte 6.3.2.2 und 9.2.1).

Die Teilnahmebereitschaft wird von diversen Faktoren auf individueller, gesellschaftlicher und auf Gruppenebene beeinflusst (Groves et al., 1992; Keusch, 2015; vgl. auch Möhring & Schlütz, 2025). Menschen aus kollektivistisch orientierten Kulturen beteiligen sich beispielsweise häufiger an Studien. Gerade für ländervergleichende Studien kann dieser Faktor ins Gewicht fallen. Außerdem sind einige gesellschaftliche Gruppen stärker von Befragungsmüdigkeit (*survey fatigue*) betroffen. Dazu zählt auch die häufig einbezogene Zielgruppe der Studierenden (Porter et al., 2004). Zum anderen hängen Teilnahmebereitschaft und -kontinuität in längsschnittlichen Studien von Merkmalen der Person, z. B. Persönlichkeitscharakteristika und soziodemografischen Faktoren, ab (Cheng et al., 2016; Keusch, 2015). Hektner et al. (2007) berichten, dass Teilnehmer:innen an In-situ-Studien tendenziell eher weiblich, gut organisiert, gewissenhaft und mental gesund sind. Diese Verzerrung mag für manche Fragestellungen unproblematisch sein, für andere kann sich daraus jedoch ein Problem ergeben, wenn es etwa um mentale Gesundheit geht (vgl. Kimhy et al., 2012). Für die Durchführung von mobilen In-situ-Studien sind vor allem der notwendige Zugang zu einem Mobilgerät, mit dem die Studie durchgeführt wird, eine ausreichende Datenverbindung und eine gewisse Affinität zu technischen Geräten sowie ggf. die Bereitschaft ausschlaggebend, eine App für die Datenerhebung zu installieren. Diesbezüglich kann es zu systematischen Ausfällen kommen (Keusch et al., 2019; Struminskaya et al., 2021). Bei einigen Forschungsinteressen gehören Nutzer:innen, die auf Internet und

6.2 Ziehung der Personenstichprobe

Smartphone verzichten, vielleicht nicht zur Grundgesamtheit; in allen anderen Fällen ist aber auch ihr systematischer Ausschluss durch die Wahl der Erhebungsmethode problematisch.

Schließlich spielen Faktoren auf der Ebene des Studiendesigns eine wesentliche Rolle für den Erfolg der Rekrutierung (ausführlich siehe Keusch, 2015; vgl. auch Stone et al., 2023). Aus der Literatur zu Onlinebefragungen ist bekannt, dass Art und Zeitpunkt der Kontaktierung einen Einfluss auf die Teilnahmebereitschaft haben. Beispielsweise werden Einladungen am Wochenende und im Sommer schneller beantwortet. Vorankündigungen (auch offline für Onlineerhebungen) und Erinnerungen sind sinnvoll. Eine Einladung von bekannten Absender:innen oder eine Unterstützung der Studie von vertrauenswürdigen Sponsor:innen erhöhen die Teilnahmebereitschaft. Hinsichtlich personalisierter Formulierungen in der Ansprache und der Länge des Einladungstextes ist die Studienlage uneindeutig.

Auch In-situ-spezifische Aspekte des Studiendesigns wie die Feldzeit, die Anzahl der Signale oder die Länge des Protokolls spielen eine Rolle und können dazu führen, dass Personen gar nicht erst teilnehmen oder die Teilnahme abbrechen (siehe Abschnitt 6.3.2.2). Schließlich ist die Gestaltung des Protokolls relevant für die Teilnahmemotivation und auch Gamificationelementen können zur Steigerung der Teilnahmebereitschaft beitragen (Berkel et al., 2017).

MERKE

Validität und Repräsentativität der Personenstichprobe in In-situ-Studien hängen – neben der Art der Stichprobenziehung – auch von deren Ausschöpfung ab. Hierbei können systematische Ausfälle durch kulturelle, individuelle und durch das Studiendesign bedingte Faktoren auftreten.

6.2.2.2 Umsetzung der Stichprobenziehung auf Personenebene

Eine der größten Herausforderungen bei wiederholten In-situ-Studien ist, dass sie einen hohen Einsatz der Teilnehmer:innen erfordern und die Forschenden deshalb besonders auf deren Kooperationsbereitschaft angewiesen sind. Nach Larson & Csikszentmihalyi (1983) ist eine „research alliance“ (S. 45) mit den Versuchspersonen notwendig, um die Güte der Erhebung zu sichern. Mit einer einmaligen Befragung zu einem häufig selbstbestimmten Zeitpunkt ist es nicht getan, sondern die Teilnehmer:innen müssen über einen längeren vorgegebenen Zeitraum immer wieder teilnehmen – ggf. auch in Situationen, in denen ein Ausfüllen störend sein könnte. Diese empfundene Belastung kann zu Selbstselektivität im Vorfeld der Erhebung, zu Studienabbruch, nicht ausgefüllten Protokollen (geringe Compliance) und mangelnder Datenqualität führen (siehe Abschnitt 6.3.2.2).

Nicht zufallsgesteuerte Stichproben sind häufig forschungsökonomisch der einzige Weg, Teilnehmer:innen für In-situ-Studien zu rekrutieren. Auch wenn diese aus stichprobentheoretischer Sicht problematisch sind und die Übertragbarkeit der Ergebnisse auf die Grundgesamtheit auf der Personenebene nicht mehr gegeben ist, können sie bei gutem Sampling auf der Situationsebene eine pragmatische Lösung für die Umsetzung von In-situ-Studien sein (siehe Abschnitt 6.3).

Wie die Teilnehmer:innen für die erforderliche Stichprobe genau rekrutiert werden, ist besonders von der Größe, Beschaffenheit und Erreichbarkeit der Grundgesamtheit sowie den zur Verfügung stehenden Ressourcen abhängig. Im Universitätskontext werden als einfachster Weg gelegentlich reine Studierendenstichproben über Kurse oder zentrale Pools von Teilnehmer:innen rekrutiert. Dieses Vorgehen kann für Forschungsfragen sinnvoll sein, deren Grundgesamtheit Studierende sind. Anderenfalls sind Verzerrungen der Ergebnisse durch die Merkmale dieser speziellen sozialen Gruppe zu befürchten (Meltzer et al., 2012). Weit verbreitet ist auch die Rekru-

tierung über persönliche Netzwerke, z. B. indem die Einladungen zur Studie per E-Mail oder Messenger versendet werden (z. B. Dogruel & Schnauber-Stockmann, 2020). Weiterhin nutzen Studien Social-Media-Kanäle (z. B. Reinecke & Hofmann, 2016) oder zum Thema der Studie passende Kontaktpunkte oder Multiplikator:innen (z. B. in unserer Beispielstudie, siehe Studiensteckbrief im Online-Anhang zum Buch). Bei speziellen Grundgesamtheiten, z. B. durch einen regionalen Fokus oder ein Nischenthema, kann eine gezielte Werbung über Zeitungsaufrufe oder Plakate hilfreich sein (Beispiele finden sich im Online-Anhang zum Buch unter Ressourcen).

Finanziell gut ausgestattete Forschungsprojekte engagieren kommerzielle Marktforschungsanbieter (z. B. Schnauber-Stockmann & Mangold, 2020), die u. a. Online-Access-Panels anbieten, in denen Personen registriert sind, die sich bereit erklärt haben, häufiger an Studien teilzunehmen. Solche Access-Panels gehen häufig aus Convenience-Stichproben hervor, weil die Teilnehmer:innen selbst entscheiden können, ob sie am Panel teilnehmen möchten und willkürlich darin aufgenommen werden. Zudem ist ihre Entscheidung für die Teilnahme an einer spezifischen Studie wiederum selbstselektiv. Häufig werden Quoten festgelegt, um zumindest zu gewährleisten, dass sich die finalen Teilnehmer:innen in einem vorher definierten Verhältnis über verschiedene soziale Kategorien verteilen und ein Abbild der Grundgesamtheit in den bekannten Merkmalen darstellen, z. B. ein bestimmter Anteil von Frauen und Männern, von verschiedenen Altersgruppen und von Bildungsgraden (siehe Quota-Verfahren, Abschnitt 6.2.1.2). Ein zentraler Kostenfaktor ist, über welche Merkmale definiert wird, ob eine Person in die Grundgesamtheit gehört oder nicht. Je seltener ein Merkmal ist, also je geringer dessen sog. Inzidenz, desto teurer ist das Sample. Beispielsweise ist der Anteil der Deutschen, die gelegentlich Messenger-Dienste nutzen, sehr viel höher als der Anteil der Deutschen, die gelegentlich einen Podcast hören (Kupferschmitt, 2024). Die Suche nach Podcastnutzer:innen ist schwieriger und daher mit höheren Kosten verbunden. Auch ohne Unterstützung von kommerziellen Rekrutierungsanbietern arbeiten

6 Ziehung von Personen- und Situationsstichproben

Forschende gelegentlich mit einer bewussten Auswahl anhand von Quotenstichproben (z. B. Naab, 2013).

MERKE

Zufallsstichproben ist zwar grundsätzlich der Vorzug zu geben, allerdings lassen sie sich häufig forschungspraktisch nicht umsetzen. Deshalb greifen Forschende in der Praxis überwiegend auf nicht zufallsgesteuerte Auswahlverfahren zurück. Auch wenn nicht zufallsgesteuerte Stichproben aus stichprobentheoretischer Sicht problematisch sind, können sie bei gutem Sampling auf der Situationsebene eine pragmatische Lösung für die Umsetzung von In-situ-Studien sein.

6.3 Ziehung der Situationsstichprobe

Neben der im vorhergehenden Abschnitt dargestellten Auswahl der Personen, die an einer In-situ-Studie teilnehmen, müssen im zweiten Schritt auch die konkreten Situationen oder Momente aus dem Alltag dieser Teilnehmer:innen ausgewählt werden, die in die Studie einfließen sollen. Wie im Kapitel zur Personenstichprobe befassen wir uns zunächst mit den Grundlagen, um uns diesen Fragen anzunähern und gehen anschließend auf mögliche Herausforderungen während der Umsetzung ein.

6.3.1 Grundlagen

6.3.1.1 Untersuchungseinheiten auf Situationsebene

Auf der Personenebene ist die Wahl der Untersuchungseinheit so intuitiv, dass wir uns darum im Allgemeinen keine Gedanken machen. Wir wollen Aussagen über Personen machen, also ist die einzelne zu betrachtende Entität, von der wir im Rahmen der Befragung Auskünfte einholen, das Individuum. Es sind auch einzelne Individuen, die mithilfe der im vorhergehenden Abschnitt skizzierten Auswahl-

verfahren als Elemente in die Personenstichprobe eingehen. Auf der Situationsebene ist diese Zuordnung nicht so offensichtlich. Was genau sind hier die einzelnen Elemente, die wir aus dem Alltag auswählen, um Teil der Situationsstichprobe zu werden? Oder anders ausgedrückt: Was ist unsere Untersuchungseinheit auf der Situationsebene? In welche Einheiten zerlegen wir den Alltag, bevor wir daraus eine Stichprobe ziehen können?

Für die Situationsstichprobe in In-situ-Studien schlagen wir in Anlehnung an Schnauber-Stockmann, Bayer, et al. (2025) die folgenden drei Untersuchungseinheiten vor: Momente, Episoden und Segmente. Je nach Untersuchungsinteresse sollte der passende Auflösungsgrad und die entsprechende Untersuchungseinheit auf der Situationsebene gewählt werden.

- **Momente** sind ‚Schnappschüsse‘ aus dem Alltag. Sie dauern nur ein paar Sekunden oder Minuten, ihre Grenzen sind also durch einen spezifischen und kurzen Zeitrahmen („jetzt gerade“, „in diesem Moment“) definiert. Eine umfassende Beschreibung der relevanten Situationsfaktoren (je nach Forschungsinteresse) zu diesem Zeitpunkt liefert ein klares Bild von Momenten des täglichen Lebens. Insbesondere bei wiederholten Messungen bei denselben Personen zu zufällig ausgewählten Momenten können solche Stichproben zur Beschreibung des täglichen Lebens herangezogen werden, um Zusammenhänge zwischen situativen Faktoren und der Mediennutzung zu erkennen.
- **Episoden** konzentrieren sich auf ein bestimmtes Zielverhalten, in der KMW häufig auf ein Mediennutzungsverhalten. Die Grenzen dieser Untersuchungseinheit werden also durch den Beginn und die Beendigung des interessierenden Verhaltens definiert (z. B. die Rezeption einer Podcastepisode). Diese Einheit auf der Situationsebene ermöglicht eine detaillierte Beschreibung eines Zielverhaltens und eignet sich beispielsweise für die Untersuchung von kognitiven Prozessen während der Mediennutzung (z. B. Schnauber-Stockmann et al., 2018).

- **Segmente** erstrecken sich über einen längeren Zeitraum. Die Grenzen der Segmente lassen sich je nach Forschungsinteresse auf unterschiedliche Weise festlegen. Jeder situative Faktor mit eindeutigen Anfangs- und Endpunkten kann zur Segmentierung eines Tages verwendet werden. Typische Beispiele sind (Tages-)Zeit, Ort, primäre Aktivitäten und soziale Konstellationen (siehe z. B. die sog. Day Reconstruction Method von Kahneman et al., 2004). Innerhalb dieser Segmente können die relevanten situativen Faktoren erhoben werden. Diese Untersuchungseinheit ermöglicht es zum Beispiel, die Mediennutzung innerhalb von Tagen zu verorten (z. B. Hoplamazian et al., 2018).

MERKE

Die Untersuchungseinheit bezeichnet diejenige Einheit, in die wir die Grundgesamtheit zerlegen, um daraus eine Stichprobe ziehen zu können. Die Untersuchungseinheit legt fest, welche Entität wir als Merkmalsträger für unser Forschungsinteresse definieren. In In-situ-Studien können die Untersuchungseinheiten auf der Situationsebene Momente, Episoden oder Segmente sein.

Tagebuchstudien sind häufig in Segmente aufgeteilt, während in ESM-Studien überwiegend Momente oder Episoden die Untersuchungseinheit bilden. Bei Tagebuchstudien findet meist keine Stichprobenziehung auf der Situationsebene statt, sondern alle im Untersuchungszeitraum liegenden Segmente werden einbezogen. Im Folgenden gehen wir auf die drei Unterarten von ESM-Studien aus Kapitel 2 – echte ESM-Studien, Quasi-ESM-Studien und ereignisabhängige ESM-Studien – und die damit verbundenen Spezifika der Stichprobenziehung ein.

6.3.1.2 Echte ESM-Studien

Echte ESM-Studien greifen auf das Prinzip der Zufallsauswahl zurück, d. h., die Zeitpunkte der Alarmierung werden zufällig aus der Menge aller möglichen Momente im Alltag der Teilnehmer:innen ‚gezogen‘. Oder anders ausgedrückt: Die Teilnehmer:innen werden zu einem zufälligen Zeitpunkt alarmiert und gebeten, Auskunft über den aktuellen Moment zu geben. Im strikten statistischen Sinne lassen nur derartige Zufallsstichproben einen Repräsentationsschluss auf den Alltag der Teilnehmer:innen zu (vgl. auch Abschnitt 6.2.1.1)

Bei der **einfachen Zufallsauswahl** werden zufällig ein oder mehrere Momente pro Tag aus dem Alltag der Teilnehmer:innen ausgewählt. Bei einer derartigen Auswahl laufen wir – aufgrund des Zufalls – jedoch Gefahr, dass mehrere zufällig ausgewählte Zeitpunkte pro Tag sehr nahe beieinander liegen, dass die Teilnehmer:innen also beispielsweise um 9.59 Uhr, 10.03 Uhr, 10.10 Uhr und 10.19 Uhr alarmiert werden und unser Einblick in den Alltag damit weniger umfassend ist, als wir uns wünschen würden. Datenanalytisch gesprochen würden wir in diesem Fall in unserer Messung nicht die notwendige Varianz erreichen, um einen validen Einblick in den gesamten Alltag der Teilnehmer:innen zu bekommen. Eine **geschichtete Zufallsauswahl** eignet sich, um diesem Problem vorzubeugen (vgl. u. a. Reinecke & Hofmann, 2016). Der Alltag der Teilnehmer:innen wird zunächst in verschiedene Zeitfenster (Schichten) eingeteilt (in der Annahme, dass Tageszeit und Mediennutzung zusammenhängen). Innerhalb der einzelnen Zeitfenster wird dann zufällig ein Moment ausgewählt. Beabsichtigen wir für unsere Studie zur Podcastnutzung, vier Signale pro Tag zwischen 6.00 und 22.00 Uhr zu versenden, so würden wir diesen Zeitraum in vier gleich große Zeitfenster von jeweils vier Stunden unterteilen (6.00 bis 10.00 Uhr, 10.00 bis 14.00 Uhr, 14.00 bis 18.00 Uhr und 18.00 bis 22.00 Uhr) und innerhalb eines jeden Zeitfensters dann zufällig einen Moment auswählen, um eine höhere Varianz zu gewährleisten. Alternativ legen manche Studien Mindestabstände zwischen zwei Signalen fest (z. B. Klingelhoefer et al. (2024), siehe Studiensteckbrief im Online-

Anhang zum Buch).

6.3.1.3 Quasi-ESM-Studien

Quasi-ESM-Studien sind echten ESM-Studien mit Blick auf die Ziehung der Situationsstichprobe weitgehend ähnlich. Auch hier werden die Zeitpunkte der Alarmierung zufällig ausgewählt, überwiegend als einfache oder geschichtete Zufallsauswahl (vgl. Abschnitt 6.3.1.2). Der Unterschied bezieht sich auf die Untersuchungseinheit, die bei echten ESM-Studien der Moment ist. Quasi-ESM-Studien weichen von diesem Merkmal ab, indem das relevante Zeitfenster ausgeweitet wird und z. B. nach den vergangenen 30 Minuten gefragt wird. Die Untersuchungseinheit kann aus einem Segment – bestimmt durch den Zeitraum von 30 Minuten – bestehen: Die Teilnehmenden berichten über ihr Verhalten, ihre Emotionen oder Kognitionen innerhalb dieses Segments. Die Untersuchungseinheit kann auch aus einer Episode bestehen, die innerhalb eines vordefinierten Zeitraums stattgefunden hat. Als Beispiel sei das Hören eines Podcasts in den zwei Stunden vor oder seit dem letzten Signal genannt. Die Teilnehmer:innen werden in diesem Fall gebeten, ihre Antworten auf diese Episode zu beziehen.

Zu bedenken ist, dass aus statischer Sicht keine echte Zufallsauswahl mehr aus dem Alltag getroffen wird. Sinn dieses Vorgehens ist es aber, die Chance der Messung für den Fall zu erhöhen, dass das Untersuchungsobjekt nicht durchgängig im Alltag prävalent ist. Während die ursprünglichen Studien von Csikszentmihalyi et al. (vgl. Abschnitt 1.2) beispielsweise das Flow-Erleben untersuchten, das sich jederzeit im Alltag messen lässt, richtet die KMW ihr Interesse häufig auf Aspekte der Mediennutzung, die nicht in jeder Alltagssituation stattfindet. Eine Quasi-ESM-Studie kann durch Ausweitung des Zugriffszeitraums auf Segmente oder Episoden statt auf Momente als Untersuchungseinheiten helfen, trotzdem für das Untersuchungsinteresse relevantes Verhalten, Emotionen oder Kognitionen zu messen.

Auch in Quasi-ESM-Studien werden die Teilnehmer:innen zu zu-

fälligen Zeitpunkten alarmiert. Die Teilnehmer:innen werden dann instruiert, an einen definierten Zeitraum zu denken und entweder ihre Antworten auf das gesamte Segment oder auf konkrete Episoden in diesem Zeitraum zu beziehen. Klingelhoefer et al. (2024) fragen beispielsweise nach den vergangenen zwei Stunden (siehe Studiensteckbrief im Online-Anhang zum Buch) und in unserer Beispielstudie nach dem gesamten Zeitraum seit dem letzten Signal (siehe Studiensteckbrief im Online-Anhang zum Buch).

6.3.1.4 Ereignisabhängiges ESM

Bei ereignisabhängigen ESM-Studien ist die Untersuchungseinheit auf der Situationsebene immer eine Episode. Hier wählen wir alle für unser Untersuchungsinteresse relevanten Ereignisse aus dem Alltag der Teilnehmer:innen aus. In derartigen Studien ist die Grundgesamtheit nicht der gesamte Alltag der Teilnehmer:innen, sondern nur die Gesamtheit spezifischer Situationen im Alltag, beispielsweise aller Podcastnutzungssituationen.

So könnten die Teilnehmer:innen beispielsweise immer ein Protokoll ausfüllen müssen, wenn sie ihre Podcast-App schließen. Der Vorteil solcher ereignisabhängiger Samplingstrategien ist, dass wir bei jeder Messung wirklich auch Auskünfte in Bezug auf unser Untersuchungsinteresse erhalten. Dadurch minimieren wir den Abdeckungsfehler, weil nur relevante Episoden in die Stichprobe gelangen. Dies gilt vor allem, wenn es möglich ist, das interessierende Ereignis technisch gestützt zu erkennen. Die Teilnehmer:innen können dann automatisch alarmiert werden, damit sie ein Protokoll ausfüllen (siehe hierzu Abschnitt 6.3.2.1). In ihrer Studie zur Beendigung der Mediennutzung *in situ* nutzten Ernst & Schnauber-Stockmann (2026) z. B. eine ESM-App, die die im Rahmen der Studie relevante TikTok- und Instagram-Nutzung tracken konnte (z. B. das Öffnen und Schließen der App) und die Teilnehmer:innen direkt nach dem Schließen der App automatisiert alarmierte (siehe Studiensteckbrief im Online-Anhang zum Buch). Ist diese Möglichkeit nicht gewährleistet, wenn es z. B. um die Nutzung des linearen Fernsehens

geht, dann müssen die Teilnehmer:innen selbst erkennen, wann sie ein Protokoll ausfüllen müssen, was zu fehlenden Episoden führen kann. Schließlich können Teilnehmer:innen die Befragungen zum relevanten (Mediennutzungs-)Verhalten als eine übermäßige Belastung wahrnehmen, wenn es sich um sehr häufig auftretendes Verhalten handelt. Dann bedarf es auch im Rahmen von ereignisabhängigen ESM-Studien einer (Zufalls-)Auswahl an Episoden (siehe z. B. Masur, 2019a). Ernst & Schnauber-Stockmann (2026) begrenzten die tägliche Anzahl an auszufüllenden Protokollen z. B. auf sechs, die zur Sicherstellung der Verteilung über den Tag auf drei Zeitfenster entfielen: Maximal zweimal wurden Protokolle zwischen 0.00 und 08.00 Uhr, zwischen 08.00 und 16.00 Uhr und zwischen 16.00 und 0.00 Uhr ausgelöst (siehe Studiensteckbrief im Online-Anhang zum Buch).

6.3.2 Praktische Umsetzung und Herausforderungen

6.3.2.1 Möglichkeiten der technischen Umsetzung

Die technischen Voraussetzungen für die Durchführung von ESM-Studien haben sich durch die flächendeckende Verbreitung internetfähiger mobiler Endgeräte (insbesondere Smartphones) in den vergangenen Dekaden grundlegend verbessert. Durch die ubiquitäre Verfügbarkeit werden Smartphones heute üblicherweise für die Alarmierung genutzt. Die damit einhergehenden Voraussetzungen für die Studienteilnahme, also der Besitz eines Smartphones und vielfach auch die Übernahme der Kosten für den anfallenden Datenverkehr durch die Teilnehmer:innen sind, aus forschungsethischer Sicht nicht ganz unproblematisch (siehe Abschnitt 4.2). Ein solches Vorgehen ist dennoch weit verbreitet.

Die Alarmierung kann mithilfe einer auf dem Smartphone der Teilnehmer:innen installierten App ausgelöst werden (so z. B. bei Aalbers et al. (2022); Ernst & Schnauber-Stockmann (2026) und Klingelhoefner et al. (2024), siehe Studiensteckbriefe im Online-Anhang zum Buch). Die App dient dann nicht nur der Alarmierung,

sondern gleichzeitig auch dem Ausfüllen des In-situ-Protokolls. Der Vorteil dieser Anwendungen ist, dass teilweise auch weitere Sensordaten des Smartphones (z. B. der Standort) oder Aktivitäten auf dem Gerät (z. B. Öffnen und Schließen von Apps) einbezogen werden können (ausführlich zum Einbezug von Smartphonedaten siehe Harari et al. (2016); vgl. auch Abschnitt 5.1). Diese Daten können dazu genutzt werden, automatisch ereignisbasiert ESM-Protokolle auszulösen (vgl. Abschnitte 6.3.1.4 und 6.3.2.1). So wurden die Teilnehmer:innen in einer Studie von Reichow & Friemel (2020) immer dann alarmiert, wenn sie ihren Standort wechselten. Bei Ernst & Schnauber-Stockmann (2026) (siehe Studiensteckbrief im Online-Anhang zum Buch) wurden die Teilnehmer:innen immer dann alarmiert, nachdem sie TikTok oder Instagram auf ihrem Smartphone schlossen.

Die verfügbaren ESM-Apps unterscheiden sich deutlich in ihren Funktionalitäten. Nicht alle Apps können Sensordaten einbeziehen und die In-situ-Protokolle technisch ereignisbasiert auslösen und die wenigsten funktionieren für alle gängigen Smartphonebetriebssysteme. Wir müssen daher stets prüfen, über welche Funktionen eine App genau verfügt und ob sie den Anforderungen an das Studiendesign genügt. In den vergangenen beiden Dekaden gab es eine Vielzahl wissenschaftlicher und kommerzieller Initiativen, Apps für ESM-Studien zu entwickeln. Viele davon, insbesondere im wissenschaftlichen Kontext, wurden nur wenige Male genutzt und nicht weiter aktualisiert. Hier sei ebenfalls auf den von Wieland et al. (2025) zusammengestellten Überblick über vorhandene ESM-Software verwiesen, der online verfügbar ist (Der Link findet sich im Online-Anhang zum Buch unter Ressourcen).

Neben diesen All-in-one-Lösungen ist auch die Trennung von Alarmierung und In-situ-Protokoll auf technischer Ebene möglich. In diesem Fall werden mittels entsprechender SMS-Portale, via E-Mail oder Messenger-App die individuellen Einladungen zu den In-situ-Protokollen an die Teilnehmer:innen verschickt. Eine entsprechende integrierte Lösung, die insbesondere in der deutschen KMW vielfach Verwendung findet (so z. B. bei Karnowski et al. (2017), Naab et al.

(2019) und Schlütz et al. (2027), siehe Studiensteckbriefe im Online-Anhang zum Buch), bietet das Onlinebefragungstool *SoSci Survey* an (Leiner, 2024). Die Umsetzung über ein derartiges Onlinebefragungstool erfordert allerdings große Sorgfalt, um die Verbindung der erhobenen In-situ-Datensätze mit den Daten auf der Personenebene zu ermöglichen, beispielsweise mit einer eindeutigen Identifikationsnummer. Es empfiehlt sich, diese Verknüpfung vor der Feldphase ausführlich zu testen.

Neben der Möglichkeit der signalbasierten Auslösung, also der externen Alarmierung der Teilnehmer:innen überwiegend mit einem Smartphone, ist auch ein selbstausgelöstes Ausfüllen möglich. Diese Option wird im Allgemeinen nur bei ereignisabhängigen ESM-Studien genutzt. Die Teilnehmer:innen werden gebeten, immer dann, wenn sie ein bestimmtes Verhalten ausführen oder ausgeführt haben, beispielsweise immer nach Ende der Podcastnutzung, ein In-situ-Protokoll auszufüllen. Damit geht allerdings eine gewisse Unschärfe einher, weil die Güte der Stichprobe von der Genauigkeit der Instruktionen, die die Forschenden den Teilnehmer:innen bereitstellen, sowie der Compliance der Teilnehmer:innen abhängt.

MERKE

Hinsichtlich der praktischen Erhebung der Situationsstichprobe besteht zunächst ein Unterschied zwischen signalbasierter Auslösung und Selbstauslösung. Beim signalbasierten Vorgehen werden die Teilnehmer:innen durch ein externes Gerät, in der Regel durch ihr Smartphone, alarmiert, ein In-situ-Protokoll auszufüllen. Die technische Umsetzung ist sowohl mit speziellen Apps als auch mit einem Onlinebefragungstool in Verbindung mit einer Alarmierung via SMS, E-Mail o. ä. möglich. Bei der Selbstauslösung entscheiden die Teilnehmer:innen basierend auf den Instruktionen der Forschenden selbst, wann sie ein Protokoll ausfüllen.

Es hängt maßgeblich vom Forschungsinteresse ab, welcher Weg

für eine Studie geeignet ist. Sollen nur Selbstauskunftsdaten erhoben werden, so ist die Kombination aus Alarmierung per SMS oder E-Mail und einer konventionellen Onlinebefragung, die über das Smartphone aufgerufen werden kann, vollauf ausreichend. Falls über das Smartphone automatisiert erhobene Verhaltensdaten mit ausgewertet werden sollen, empfiehlt sich eine integrierte App-basierte Lösung. Neben den Fragen zu technischen Funktionen müssen bei der Auswahl der Softwarelösung rechtliche bzw. forschungsethische Aspekte beachtet werden, insbesondere eine datenschutzrechtlich einwandfreie Speicherung der Daten. Hier sollte auf die DSGVO-Konformität des jeweiligen Anbieters geachtet werden. Die Ansätze der technischen Umsetzung sind mit unterschiedlich hohen Kosten verbunden, die ebenfalls im Vorfeld bedacht werden müssen.

6.3.2.2 Stichprobengüte

Auch bei der Situationsstichprobe sind mögliche Fehler zu bedenken, die die Repräsentativität der Stichprobe einschränken können. Üblich bei ESM-Studien sind sog. Coverage Errors, die aus pragmatischen Erwägungen resultieren. Die Zeitpunkte der zufälligen Alarmierungen beschränken sich i. d. R. auf die Wachzeiten der Teilnehmer:innen, also beispielsweise auf den Zeitraum zwischen 8.00 und 22.00 Uhr. Diese Einschränkung der verfügbaren Stunden repräsentieren dann die sog. **Auswahlgesamtheit**, was jedoch gelegentlich problematisch ist. Podcasts werden beispielsweise häufig kurz vor dem Einschlafen gehört, also ggf. nach 22.00 Uhr. Wenn wir dann aus einer eingeschränkten Auswahlgesamtheit auf die Podcastnutzung rund um die Uhr und an sieben Tagen die Woche der Teilnehmer:innen schließen würden, läge ein Coverage Error vor. Daneben können auf der Situationsebene Sampling Errors auftreten, also Fehler bei der zufälligen Auswahl der einzelnen Elemente, z. B. wenn Signale nicht zugestellt werden, weil der Akku des Smartphones leer ist oder die Teilnehmer:innen keinen Empfang haben oder – wie häufig bei Studien mit Kindern oder Jugendlichen – das Smartphone nicht genutzt werden darf (vgl. Fecke et al., 2024).

Weiterhin spielt für die Güte der Situationsstichprobe die sog. **Response Latency**, also die Zeit zwischen dem Signal bzw. bei selbstausgelösten Protokollen dem zu berichtenden Ereignis und dem Ausfüllen des Protokolls eine Rolle: Besonders bei zufallsbasierten Stichprobenplänen (siehe Abschnitt 6.3.1) ist der Zeitpunkt der Alarmierung entscheidend für die gezogene Untersuchungseinheit. Er gibt an, welcher Moment mit „jetzt gerade“ gemeint ist oder welches Segment mit diesem Moment endet. Beantworten die Teilnehmer:innen das Protokoll aber zu einem späteren Zeitpunkt, kann diese Verzögerung zu Verzerrungen der Situationsstichprobe führen. Einige Apps und Softwaretools für Befragungen sind deshalb mit der Einschränkung ausgestattet, dass Protokolle nur eine gewisse Zeit nach der Alarmierung (z. B. zehn Minuten, eine halbe Stunde) ausgefüllt werden können.

MERKE

Die Response Latency, der zeitliche Abstand zwischen dem Signal bzw. dem zu berichtenden Ereignis und dem Ausfüllen des Protokolls, ist ein wesentliches Qualitätskriterium für In-situ-Studien. Sie sollte möglichst gering sein.

Schließlich ist auf der Situationsebene mit fehlenden Werten und damit ggf. mit Response Errors zu rechnen. Denn selten füllen Teilnehmer:innen alle Protokolle über den gesamten Zeitraum der Studie gewissenhaft aus. Dieses Thema wird in der Literatur unter dem Begriff der **Compliance** behandelt (siehe z. B. Eisele, Vachon, et al., 2022; Rintala et al., 2019). Im Durchschnitt liegt die Compliance in ESM-Studien bei etwa 70 bis 85 % (Eisele, Kasanova, et al., 2022; Fritz et al., 2024; Jones et al., 2019; Myin-Germeys & Kuppens, 2022a; Stone et al., 2023). Allerdings beziehen sich diese Werte auf die in Publikationen berichteten Complianceraten. Diese Information fehlt aber in etlichen Publikationen oder sie wird auf Basis der Teilnehmer:innen berechnet, die bereits eine vorab festgelegte Mindestcompliance erreicht haben (siehe auch Abschnitt 9.2.1), so-

dass die tatsächliche durchschnittliche Compliancerate ggf. niedriger ausfällt (Fritz et al., 2024). In ESM-Studien mit selbstausgelösten Protokollen lässt sich die Compliance kaum bestimmen.

MERKE

Die Compliance, der Anteil ausgefüllter Protokolle an allen gesendeten Signalen, ist ein wesentliches Qualitätskriterium für ESM-Studien. Sie sollte möglichst hoch sein.

Häufig werden Teilnehmer:innen vor der Datenauswertung ausgeschlossen, wenn sie eine gewisse Mindestcompliance (z. B. 60 %) unterschreiten (siehe auch Abschnitt 9.2.1). Die zugrunde liegende Annahme lautet, dass die Situationsstichprobe dann nicht mehr repräsentativ für den Alltag der Teilnehmer:innen ist und die Ergebnisse somit verzerrt sind. Es ist allerdings abhängig von den Gründen für die **fehlenden Werte**, ob und wenn ja in welchem Ausmaß es zu Verzerrungen kommt. In der Methodenliteratur wird zwischen verschiedenen Arten von fehlenden Werten (sog. Missings) unterschieden. Fehlende Werte können rein zufällig sein (*missing completely at random*, MCAR). Mit MCAR ist gemeint, dass die fehlenden Werte keinen Zusammenhang mit anderen, für die Studie relevanten Variablen haben. Die fehlenden Werte sind quasi eine zufällige Stichprobe aller Werte. Beispielsweise kann es gelegentlich zu technischen Schwierigkeiten kommen, sodass einzelne Protokolle nicht ausgefüllt werden. Weiterhin können fehlende Werte *missing at random* (MAR) sein. Entgegen der Bezeichnung meint MAR, dass die fehlenden Werte nicht rein zufällig sind, aber dass die Faktoren (z. B. Personen- oder Situationsmerkmale), die für die fehlenden Werte verantwortlich sind, bekannt sind und nicht mit der Ausprägung der fehlenden Beobachtung zusammenhängen. Beispielsweise könnten wir feststellen, dass männliche Teilnehmer häufiger Protokolle unserer Podcaststudie auslassen als weibliche Teilnehmerinnen. Spielt das Geschlecht keine Rolle für die zu untersuchenden Zusammenhänge, handelt es sich um MAR-Ausfälle. Schließlich können fehlende Werte systema-

tisch sein (*not missing at random*, NMAR). In diesem Fall hängt das Fehlen direkt mit der Ausprägung des fehlenden Wertes zusammen (Courvoisier et al., 2012; Enders, 2022; Schafer & Graham, 2002; Stone et al., 2023). Beispielsweise könnten gerade Personen, die während der Alarmierung einen Podcast hören, das Protokoll nicht ausfüllen, weil sie nicht beim Hörerlebnis gestört werden möchten.

Während rein zufällige Ausfälle (MCAR) vor allem die Teststärke einer Stichprobe (die sog. Power, siehe Abschnitt 6.1) beeinträchtigen und – je nach Auswertungsmethode (siehe Kapitel 9) – die Analyse der Daten erschweren, können MAR- und vor allem NMAR-Ausfälle die Qualität einer Studie beeinträchtigen. Deshalb ist es wichtig, sich über die Ausfallgründe im Klaren zu sein, um ihnen durch die Gestaltung der Studie entgegenzuwirken und sie bei der Interpretation der Ergebnisse zu berücksichtigen. Gründe für nicht rein zufällige Ausfälle lassen sich in drei Kategorien einteilen: Merkmale der Person, Merkmale der Situation und Studieneigenschaften (siehe z. B. Ono et al., 2019).

Merkmale der Person sind stabile Eigenschaften wie z. B. soziodemografische Merkmale (Alter, Geschlecht) oder Persönlichkeitsmerkmale, die i. d. R. bereits im Rahmen einer Vorabbefragung (siehe Abschnitt 5.1) erhoben werden. Im Anschluss an die Feldzeit kann analysiert werden, ob bestimmte Personenmerkmale mit einem höheren Anteil nicht ausgefüllter Protokolle einhergehen. Zum Beispiel könnte es sein, dass jüngere Teilnehmer:innen im Vergleich zu älteren weniger Protokolle ausfüllen. Über Ergebnisse, die in diese Richtung deuten, berichten z. B. bei Morren et al. (2009), Ono et al. (2019) und Rintala et al. (2019). Solche Auffälligkeiten können dann in einer Publikation der Ergebnisse berichtet werden (siehe Abschnitt 10.3). Es kann sich hierbei um MAR handeln, wenn z. B. das Alter keinen Zusammenhang mit den uns interessierenden Variablen aufweist. Besonders gravierend sind solche Personenunterschiede in der Compliance allerdings, wenn ein direkter Zusammenhang mit dem Inhalt der Studie besteht. Geht es in der Studie z. B. um depressive Verstimmungen (Situationsebene), könnte befürchtet werden, dass Personen mit Depressionen (Personenmerkmal) eher dazu

neigen, Protokolle auszulassen, wenn sie sich gerade besonders deprimiert fühlen, nicht aber Personen, die keine Depressionen haben. Unter diesen Bedingungen hängt nicht nur der Ausfall eines Protokolls von einem Personenmerkmal ab, sondern auch die fehlende Antwort, weil vor allem Personen mit Depression nicht mitmachen, wenn sie sich gerade schlecht fühlen. In diesem Fall läge NMAR vor, was problematisch wäre.

Merkmale der Situation können zeitliche Faktoren, Kontextfaktoren (Ort, soziales Umfeld) oder psychologische States von Personen sein. Am einfachsten zu erkennen sind zeitliche Faktoren in signalbasierten ESM-Studien, aus denen hervorgeht, wann das Protokoll hätte ausgefüllt werden müssen. Bei selbstausgelösten Protokollen ist diese Möglichkeit hingegen nicht gegeben, weil nicht bekannt ist, ob das auslösende Ereignis stattgefunden hat und nur kein Protokoll ausgefüllt wurde. In einigen ESM-Studien zeigen sich entsprechende Ausfallmuster im Tagesverlauf. Häufig werden Protokolle am Morgen nicht beantwortet (z. B. Courvoisier et al., 2012; Rintala et al., 2019). Auch im Wochenverlauf können sich Ausfallmuster zeigen, etwa unterschiedliche Antwortraten zwischen Wochentagen und Wochenenden. Hier gilt: Solche Ausfallmuster können für die Studie wichtig sein, müssen es aber nicht. Wenn man auf diese Weise z. B. immer die erste Nutzungsepisode für einen Podcast am Tag nicht erfasst, wird damit die Abschätzung der Gesamtnutzungsdauer beeinträchtigt (NMAR). Geht es in einer Studie dagegen um den Zusammenhang mit Emotionen und ist dieser Zusammenhang am Morgen und zu anderen Zeiten identisch, beeinträchtigen die Ausfälle nicht die Datenqualität, sondern nur die -quantität (MAR).

Schwieriger wird es bei anderen Merkmalen auf der Situations-ebene, die erhoben worden sein müssen, um ihr Verzerrungspotenzial abschätzen zu können. So ist es möglich, dass Teilnehmer:innen eher zuhause, nicht aber im ÖPNV auf eine Alarmierung reagieren. Das würde die Messung der mobilen Podcastnutzung im Rahmen einer Studie verzerren. Dieser Befund gilt nicht nur für die Frage, ob ein Podcast gehört wurde, sondern auch, ob weitere studienrelevante Merkmale (z. B. das Rezeptionserleben) mit dem Ort der Nutzung

einhergehen. In Bezug auf die psychologischen States einer Person ist z. B. anzunehmen, dass eine Person nicht antwortet, wenn sie besonders gestresst ist. Solche Ausfälle sind problematisch (NMAR), wenn es in der Studie um die Frage geht, ob im gestressten Zustand eher Podcasts genutzt werden.

Während der Zusammenhang zwischen nicht ausgefüllten Protokollen und Merkmalen der Person und der Situation vor allem wichtig ist, um Verzerrungen der Ergebnisse festzustellen und als Limitation anzuerkennen sowie diesen ggf. statistisch entgegenzuwirken, liegen die **Studieneigenschaften** in der Hand der Forschenden und können mitbedacht werden. Als wichtiger Faktor hat sich die Länge der Feldzeit erwiesen. Im Laufe der Feldzeit sinkt – abgesehen von kompletten Studienabbrüchen – der Anteil ausgefüllter Protokolle (Courvoisier et al., 2012; Eisele, Kasanova, et al., 2022; Fuller-Tyszkiewicz et al., 2013; Ono et al., 2019; Rintala et al., 2019). Besonders bei Feldzeiten von mehr als zwei Wochen geht die Antwortrate zurück (Berkel et al., 2019; Stone et al., 1991). Unklar ist die Forschungslage hinsichtlich der Dichte, d. h. der Anzahl der täglichen Signale bzw. auszufüllenden Protokolle. Gerade bei Ereignissen, bei denen unklar ist, wie oft sie im Alltag vorkommen, ist aus Sicht der Forschenden eine hohe Protokolldichte sinnvoll. Es gibt Methodenstudien und Metaanalysen, die nahelegen, dass viele Protokolle sowohl zu einer geringeren (Vachon et al., 2019), als auch zu einer höheren Compliance führen können (Eisele, Kasanova, et al., 2022). Übergreifend lassen sich jedoch keine Zusammenhänge zwischen der Anzahl der auszufüllenden Protokolle und der Compliance belegen (Jones et al., 2019; Morren et al., 2009; Ono et al., 2019; Soyster et al., 2019).

Es zeigen sich aber zeitliche Antwortmuster, insofern als morgendliche Protokolle tendenziell eher ausgelassen werden. Ein möglicher Grund für dieses Verhalten liegt in den individuellen Tagesabläufen der Teilnehmer:innen. Oft sind ESM-Studien auf bestimmte Tagesabschnitte wie typische Wachzeiten beschränkt. Diese Beschränkung ist unproblematisch, wenn die Grundgesamtheit der Situationen, die uns interessieren, nur diese Wachzeiten umfassen (sie-

he Abschnitt 6.3.2.2). Diese müssen aber nicht für alle Personen gleich sein. Wenn es das Forschungsinteresse zulässt, indem keine festen Zeitpunkte notwendig sind und die technische Umsetzbarkeit gewährleistet ist (siehe Abschnitt Kapitel 6.3.2), kann es eine Option sein, die Teilnehmer:innen selbst angeben zu lassen, zu welchen Zeiten sie gewöhnlich aufstehen und zu Bett gehen. Es ist sinnvoll, das Senden der Signale individuell an die verfügbaren Zeiten der Teilnehmer:in anzupassen, um den Anteil ausgefüllter Protokolle zu erhöhen (Rintala et al., 2019; Silvia et al., 2013).

Die Länge des Protokolls spielt für die Compliance ebenfalls eine Rolle. Die Befunde in der Forschung sind zwar nicht eindeutig, in der Tendenz zeigt sich aber, dass kürzere Protokolle zu weniger Ausfällen führen (Eisele, Kasanova, et al., 2022; Morren et al., 2009; keinen Einfluss finden z. B. Ono et al., 2019; Rintala et al., 2019, allerdings sind alle ESM-Protokolle, die in diesen beiden Publikationen betrachtet werden, relativ kurz, sodass hier auch die geringe Varianz in der Protokolllänge den fehlenden Zusammenhang erklären könnte).

MERKE

Ob fehlende Protokolle nur die Power einer Studie beeinträchtigen oder auch zu Verzerrungen in den Ergebnissen führen, hängt davon ab, ob das Nichtausfüllen zufällige oder systematische Ursachen hat. Die Gründe für systematische Ausfälle lassen sich drei Kategorien zuordnen: Merkmale der Person, Merkmale der Situation und Studieneigenschaften. Gerade das zuletzt genannten Merkmal liegt in der Hand der Forschenden. Es sollte daher sorgfältig zwischen der erforderlichen Detailiertheit der Erhebung (Länge der Feldzeit, Anzahl, Umfang und Terminierung der Protokolle) und der Gefahr von Ausfällen abgewogen werden.

7 Konstruktion der Erhebungsinstrumente

Im Hinblick auf die Konstruktion der Erhebungsinstrumente geht es darum, Validität und Reliabilität bei der Messung der Konstrukte sicherzustellen. Wir gehen zweistufig vor und beschreiben zunächst in Abschnitt 7.1, wie wir idealerweise vorgehen sollten, um maximale Validität und Reliabilität sicherzustellen. Anschließend geht es in Abschnitt 7.2 um mögliche Herausforderungen bei der konkreten Umsetzung.

7.1 Grundlagen der Fragebogenkonstruktion

Eine durchdachte Fragebogenkonstruktion ist zentral, um die Validität und Reliabilität des Instruments sicherzustellen. Dazu gehört sowohl die Formulierung der Fragen als auch deren Anordnung im Fragebogen (die sog. Dramaturgie, Möhring & Schlütz, 2025). Der strategische Aufbau ist wichtig, um die Teilnahmebereitschaft über einen gelingenden Einstieg sicherzustellen, die Motivation mit einem geeigneten Spannungsbogen zu erhalten und unerwünschte Wechselwirkungen zwischen den Fragen wie z. B. Kontexteffekte zu vermeiden. Die Konstruktion von Fragebögen für die Personenebene (also für die Vorab- und Nachbefragung, siehe Abschnitt 5.1) unterscheidet sich nicht von herkömmlichen Querschnittsbefragungen. Wir gehen daher nur kurz auf relevante Punkte ein, bevor wir uns mit In-situ-Protokollen und ihren Besonderheiten ausführlicher befassen. Es sei im Wesentlichen auf einschlägige Publikationen zur Befragung verwiesen, die in Buchform (z. B. Faulbaum, 2019; Möhring & Schlütz, 2025; Porst, 2014; Scholl, 2018) und als Aufsätze mit unterschiedlichen Foki vorliegen (etwa Möhring & Schlütz, 2013a;

7.1 Grundlagen der Fragebogenkonstruktion

Möhring & Schlütz, 2013b; Taddicken, 2013; Wagner-Schelewsky & Hering, 2019; Weiß et al., 2019).

Eine ESM-Studie kombiniert häufig herkömmliche Querschnittsbefragungen mit wiederholten situativen Erhebungen (siehe Abschnitt 5.1). In der querschnittlich angelegten Vorabbefragung werden Informationen auf der Personenebene gesammelt, also etwa allgemeine, d. h. situationsübergreifende Nutzungsgewohnheiten, Einstellungen, Persönlichkeitsmerkmale oder soziodemografische Merkmale. Diese Daten werden für die Beschreibung der Stichprobe genutzt, können aber auch als unabhängige Variablen oder Kontrollvariablen in die Analyse eingehen (siehe Kapitel 9). Den Abschluss einer In-situ-Studie bildet häufig eine Nachbefragung, die zur Erhebung von Personenvariablen, zum Debriefing oder für die Ermittlung methodischer Aspekte genutzt werden kann. Für die Podcastingstudie gab es zum Beispiel drei Erhebungsinstrumente (abrufbar im Studiensteckbrief im Online-Anhang zum Buch): Im Rahmen der Vorabbefragung wurden Nutzungsgewohnheiten, -motive und -vorlieben sowie Bewertungen von True-Crime-Podcasts auf der Personenebene erhoben. Zusätzlich wurden Informationen über übliche Hörsituationen wie Kontextvariablen und Transporterleben im Allgemeinen für den Vergleich mit den situativen Daten erfasst. In den ESM-Protokollen wurden das situative Nutzungserleben – inklusive Transportation als State-Variable – und der Hörkontext erhoben. In der Nachbefragung wurden methodische Merkmale (vor allem in Bezug auf die Reaktivität) erfragt und Transportabilität als Trait-Variable gemessen.

7.1.1 Allgemeine Prinzipien der Operationalisierung

Wenn wir einen Fragebogen oder ein Protokoll konstruieren, ist es hilfreich, uns zu vergegenwärtigen, dass nicht einfach Merkmale abgerufen werden, sondern die befragte Person auf Basis der situativ verfügbaren Informationen zielorientiert handelt. Die Person durchläuft einen **kognitiven Prozess**, der aus mehreren Schritten besteht (Schwarz & Oyserman, 2001): Zunächst muss die Frage nach-

7 Konstruktion der Erhebungsinstrumente

vollzogen und ihre Bedeutung entschlüsselt werden. Wenn wir also fragen „Wie häufig hören Sie Podcasts im Allgemeinen?“, überlegt die befragte Person zunächst, was unter den Begriff Podcast fällt (Geht es z. B. auch um Hörbücher, wenn sie auf einer Podcasting-plattform erscheinen?) und wie „im Allgemeinen“ auszulegen ist. Im zweiten Schritt muss sie verfügbare relevante Wissensbestände zum nachgefragten Objekt abrufen, sich also an konkrete Ereignisse erinnern. Auf dieser Basis werden im dritten Schritt Inferenzschlüsse und Schätzungen vorgenommen, weil die erfragte Häufigkeit, Dauer oder Intensität des Fragegegenstands häufig nicht ohne Weiteres abrufbar ist oder keine exakte Quantifizierung vorgenommen werden kann. Für eine standardisierte Befragung muss diese innere Antwort viertens auf ein vorgegebenes Antwortformat (z. B. eine Skala oder Antwortitems) zugeschnitten werden. Zum Beispiel können wir abgestufte Antwortmöglichkeiten vorgeben, etwa „unter einer Stunde/Woche“, „1 bis unter 2 Std./Woche“, „2 bis unter 3 Std./Woche“, „3 bis unter 5 Std./Woche“ und „5 Std./Woche und mehr“ sowie „Ich weiß es nicht“. Die Antwort muss dann in die jeweilige Kategorie eingepasst werden. Schließlich können fünftens weitere Anpassungen notwendig sein, etwa um sich mit der eigenen Antwort einem vermeintlich sozial erwünschten Antwortverhalten anzunähern (Krumal, 2013).

Entscheidend für die Formulierung geeigneter Testfragen und den dazugehörigen Antwortvorgaben ist, den Bezugsrahmen der Forschenden – d. h. die relevanten Dimensionen der Forschungsfrage – in Indikatoren zu überführen, so dass Befragte den o. g. kognitiven Antwortprozess möglichst störungsfrei durchlaufen können. Zu diesem Zweck müssen die als Konstruktindikatoren dienenden Testfragen zunächst systematisch aus den Dimensionen der Forschungsfrage hergeleitet werden, um die Konstruktvalidität sicherzustellen. Mit Blick auf die Standardisierung ist es wichtig, dass der Bezugsrahmen jeder Frage eindeutig ist, d. h., dass die Befragten die Fragen gleich im Sinne des dahinterstehenden Konstrukts verstehen. Nur dann können die Daten der individuellen Befragten sinnvoll zusammengefasst und verallgemeinert werden. Konkret sind bei der Formulierung

7.1 Grundlagen der Fragebogenkonstruktion

der Testfragen neben der Berücksichtigung der Befragungssituation und dem Vorwissen der Befragten drei grundlegende Aspekte zu beachten: **Verständlichkeit, Auskunftsfähigkeit und Auskunftswilligkeit**.

Neben den Fragen müssen auch die Antwortvorgaben bestimmten Kriterien genügen, weil auch sie den kognitiven Antwortprozess beeinflussen. Empirische Erkenntnisse lassen darauf schließen, dass Befragte aus den Vorgaben auf die Bedeutung der Frage schließen und sie als Informationsquelle für ein übliches Verhalten nutzen, an dem sie ihr Verhalten messen (Möhring & Schlütz, 2025). Daher ist es entscheidend, dass die Antwortalternativen das mögliche Spektrum abdecken, den Bezugsrahmen verdeutlichen, trennscharf und eindeutig sind und nicht zuletzt auf die Formulierung der Frage abgestimmt werden. Es ist jedoch zu beachten, dass unterschiedliche Arten von Fragen bzw. Skalen sich sowohl auf die Datengüte als auch auf die Benutzungsfreundlichkeit auswirken können (vgl. z. B. Couper et al., 2006 zum Unterschied von Slidern und Radio-Buttons). Inhaltlich ist es wichtig, die Formulierungen nicht zu extrem zu wählen bzw. eine entsprechende Bandbreite an Antwortvorgaben vorzusehen, damit ausreichend Raum für Variabilität bleibt und Boden- bzw. Deckeneffekte vermieden werden (Eisele, Vachon, et al., 2022). Mit einem Pretest sollte geprüft werden, ob diese Anforderungen erfüllt sind (siehe Abschnitt 8.1).

MERKE

Ziel der Operationalisierung ist es, eine Forschungsfrage anhand einzelner Testfragen valide und reliabel umzusetzen. Dazu muss der kognitive Antwortprozess – Verstehen, Erinnern/Urteilen, Schätzen und Anpassen – optimal unterstützt werden, damit die befragte Person ihn möglichst störungsfrei durchlaufen kann.

Im Rahmen standardisierter Befragungen finden verschiedene **Fragetypen** Anwendung. Offene Fragen ohne vorgegebene

7 Konstruktion der Erhebungsinstrumente

Antwortoptionen sind möglich, sollten aber sparsam eingesetzt werden, weil sie sowohl für die Befragten in der Beantwortung als auch für die Forschenden in der Auswertung vergleichsweise aufwändig sind. In erster Linie werden daher geschlossene Fragen verwendet, denn sie sind valider und reliabler und lassen Vergleiche zwischen Befragten zu. Geschlossene Fragen lassen sich anhand ihres Datenniveaus klassifizieren in Auswahlfragen auf Nominalniveau, Rangordnungsfragen auf Ordinalniveau und Fragen der Intensitätsmessung/Skalen auf (quasi-)metrischem Niveau. Alle Datenniveaus finden eine Verwendung in In-situ-Studien. Es empfiehlt sich, bereits bei der Erstellung des Erhebungsinstruments zu überlegen, welche Analyseverfahren zur Beantwortung der Forschungsfragen herangezogen werden, um das erforderliche Datenniveau der einzelnen Testfragen sicherzustellen.

Häufig werden Skalen eingesetzt, mit deren Hilfe latente Konstrukte wie etwa Transportation und Transportabilität in unserer Beispielstudie messbar gemacht werden (Appel et al., 2015; Dal Cin et al., 2004). Im Allgemeinen bestehen Skalen aus einer Reihe von Items, die im Fragebogen in einer sog. Itembatterie präsentiert und bei der Datenanalyse zu einem Index verrechnet werden. Auf diese Weise lassen sich komplexe Konstrukte oder Phänomene i. d. R. valider abbilden als mit Einzelitems. Bei Skalen kann jedes Item einer Batterie als Messwiederholung verstanden werden, die in ihrer Summe zu einer zuverlässigeren Messung eines Konstrukts führt (Hartmann & Reinecke, 2013). Auf diese Weise wird das Konstrukt erschöpfender repräsentiert und die Abbildungstreue steigt. Konstruktion und Validierung solcher Skalen sind allerdings sehr aufwändig. Es empfiehlt sich daher, nach Möglichkeit auf bestehende Skalen zurückzugreifen.

Die einzelnen Items einer Skala werden mithilfe einer Messskala mit verbalisierten Ankerpunkten (z. B. von 1 „stimme überhaupt nicht zu“ bis 5 „stimme voll und ganz zu“) von den Befragten bewertet. Mit Blick auf eine optimale Messgüte sollte die Anzahl der Skalenpunkte zwischen 4 und 7 liegen (Lozano et al., 2008). Die Einstufung anhand solcher Skalen ist absolut, die resultierenden Da-

ten können daher als quasimetrisch betrachtet und in der Analyse entsprechend verwendet werden, obschon sie streng genommen lediglich ordinales Datenniveau aufweisen (vgl. aber Larroulet Philippi, 2021).

7.1.2 Spezifika des In-situ-Protokolls

Für die Erfassung der Situationsmerkmale in In-situ-Studien wird ein Kurzfragebogen verwendet, der in frühen Experience-Sampling-Studien als *Experience Sampling Formular* (ESF) bezeichnet wurde (Kubey & Csikszentmihalyi, 1990). Wir nutzen den Begriff Protokoll.

MERKE

Wir bezeichnen den in In-situ-Studien verwendeten Kurzfragebogen als Protokoll. Im Protokoll werden die für das Forschungsprojekt relevanten situativen Aspekte erfasst.

Je nach Forschungsinteresse werden auf der Situationsebene beispielsweise folgende Dimensionen erhoben (Hektner et al., 2007; für ein Beispiel siehe Hektner et al., 2007, S. 294–297):

- räumlicher und sozialer Kontext (Wo? Mit wem?),
- Aktivitäten (Was? Z. B. Mediennutzung),
- innere Verfasstheit (Gedanken, Emotionen, Motive etc.) sowie
- weitere individuelle, situativ variable Merkmale

In den Protokollen sind sowohl offene als auch geschlossene Fragen unterschiedlicher Standardisierungsgrade und Datenniveaus möglich. Aus Reliabilitätsgesichtspunkten und um die Befragten nicht zu sehr zu belasten, sind standardisierte, geschlossene Abfrageformen im Allgemeinen sinnvoller. Neben diesen Selbstauskünften können automatisiert weitere Paradata erfasst werden, wie z. B.

7 Konstruktion der Erhebungsinstrumente

der Zeitpunkt des Ausfüllens, um die Response Latency, also die Zeit zwischen Signal und Ausfüllen zu berechnen (siehe Abschnitt 9.2.1). Für die Formulierung von Fragen und Antwortvorgaben im Protokoll müssen die in Abschnitt 7.1.1 vorgestellten Prinzipien der Operationalisierung beachtet werden. Zentral ist vor allem die präzise und knappe Formulierung der Fragen und Antwortvorgaben, damit sie schnell erfasst und auch auf dem Smartphone ohne Probleme bearbeitet werden können (für Beispiele siehe Berkel et al., 2017; siehe auch Eisele et al., 2025).

Je nach In-situ-Design bzw. Samplingstrategie (siehe Kapitel 2 und Abschnitt 6.3) ergeben sich unterschiedliche **zeitliche Bezüge** für die im Protokoll erfassten Merkmale, die gleichzeitig die Untersuchungseinheiten bilden. Wir haben in Abschnitt 6.3.1.1 Momente, Episoden und Segmente differenziert, die jeweils unterschiedliche zeitliche Bezüge variierender Dauer definieren. Es kann z. B. um die momentane Verfasstheit, eine spezifische Mediennutzung in der letzten Stunde bzw. im Verlauf des vergangenen Tages oder die Beschreibung eines längeren Zeitabschnitts gehen. Dieser zeitliche Bezug wird im Rahmen des kognitiven Antwortprozesses berücksichtigt: Je größer der Zeitrahmen ist, umso intensiver muss das Erlebte sein, um als relevant eingestuft und berichtet zu werden (Schwarz, 2007). Fragen wir also z. B. nach Stress in diesem Moment, in der letzten halben Stunde oder am gesamten heutigen Tag, wird die Antwort jeweils anders ausfallen. Echte ESM-Studien („Wie gestresst fühlen Sie sich jetzt?“) sind am besten geeignet, auch geringere Intensitäten zu erfassen. Wichtig für die Formulierung der Protokollfragen ist, dass der Zeitrahmen explizit in der Frage genannt wird (z. B. „jetzt“, „momentan“, „in der letzten halben Stunde“, „seit dem letzten Protokoll“ etc., siehe auch Eisele et al., 2025).

Neben Aspekten der Fragenformulierung spielt für die Konstruktion der Protokolle deren Länge eine entscheidende Rolle. In der Literatur werden zwar sehr unterschiedliche Umfänge von bis zu 100 Items genannt (Eisele, Vachon, et al., 2022), es ist aber plausibel, dass sich eine übermäßige Länge tendenziell negativ auf diverse Qualitätsindikatoren von In-situ-Studien auswirkt (Morren et al., 2009,

7.1 Grundlagen der Fragebogenkonstruktion

siehe Abschnitt 6.3.2.2). Weiterhin kann sich eine übermäßige Ausfülldauer auf die Sorgfalt auswirken, mit der die Teilnehmer:innen das Protokoll ausfüllen (siehe Abschnitt 7.2.2). Die Protokolllänge ist somit ein entscheidender Parameter für die Studienqualität von In-situ-Studien.

Wir müssen nicht nur entscheiden, welche Konstrukte in den Protokollen enthalten sein müssen, sondern auch, wie diese konkret gemessen werden sollen. Die in Abschnitt 7.1.1 beschriebene Verwendung von Skalen mit umfangreichen Itembatterien widerspricht dem Ziel, das Protokoll möglichst kurz zu halten. So lässt sich beispielsweise das Spannungserleben postrezeptiv mit zwölf Items valide messen (Wünsch et al., 2023). Für eine Messung der Spannung während der Nutzung ist eine kürzere Skala sinnvoller, um die Unterbrechung der Mediennutzung kurz und die Belastung der Teilnehmer:innen gering zu halten und so Fehler durch Abbrüche oder Auslassungen zu reduzieren. Eine reduzierte Anzahl von Items bildet ein Phänomen allerdings häufig weniger gut ab. Trotzdem ist es für In-situ-Protokolle i. d. R. unerlässlich, auf gekürzte Skalen, oft auch auf Single-Item-Messungen zurückzugreifen, um das Protokoll kurz zu halten. Es gibt allerdings nur wenige validierte Instrumente speziell für In-situ-Studien. Ausnahmen bilden das *Experience Sampling Method Item Repository* (Kirtley et al., 2021), das online verfügbar ist (der Link findet sich im Online-Anhang zum Buch unter Ressourcen) sowie die Sammlung von Wolfers & Baumgartner (2025), die für zentrale kommunikationswissenschaftliche Konstrukte Single-Item-Messungen validiert haben. Deshalb ist es häufig notwendig und sinnvoll, herkömmliche Kurzskalen zu nutzen (siehe z. B. das Open Access Repositorium für sozial- und verhaltenswissenschaftliche Messinstrumente ZIS von GESIS) oder selbst bestehende Skalen zu kürzen.

Fällt die Entscheidung auf Kurzskalen, dann empfehlen Shrout und Lane (2012), Skalen aus nicht weniger als drei Items zu konstruieren, um die Validität und Reliabilität sicherzustellen. Sehr häufig werden wir in In-situ-Protokollen allerdings auf Single-Item-Messungen zurückgreifen (müssen), die in der Kritik stehen, weil

7 Konstruktion der Erhebungsinstrumente

ihre Validität und Reliabilität als geringer gilt im Vergleich zu Skalen mit mehreren Items (siehe Abschnitt 9.2.2.1.2, Diamantopoulos et al., 2012; Stone et al., 2023; siehe aber Song et al., 2023 für gegenteilige Ergebnisse). Die Festlegung auf Einzelitems erfordert deshalb eine besondere Sorgfalt vonseiten der Forschenden und ist eine zentrale Herausforderung für die Konstruktion von In-situ-Protokollen.

MERKE

Mangelnde Sorgfalt beim Ausfüllen kann die Datenqualität von In-situ-Studien beeinträchtigen: Mit steigender empfundener Belastung, etwa aufgrund langer Protokolle, nimmt die Sorgfalt der Teilnehmer:innen ab. Diese Problem muss gegen die nötige Detailliertheit der Erhebung abgewogen werden. Dafür ist meist der Einsatz von Kurzskalen und Single-Item-Messungen notwendig, um Protokolle in der Länge zu beschränken.

Empfehlenswert sind die Beiträge von Eisele et al. (2025), die das Evaluationsinstrument ESM-Q vorstellen, und von Wolfers & Baumgartner (2025), die Single-Item-Messungen für zentrale kommunikationswissenschaftliche Konstrukte validieren. Beide geben konkrete Handlungsempfehlungen für die Konstruktion und Auswahl von Single-Item-Messungen für In-situ-Studien (siehe auch Tuerlinckx et al., 2026). Wir fassen im Folgenden die Kernaspekte zusammen.

1. **Orientierung an bereits bestehenden Messinstrumenten:** Es bietet sich an, bereits vorhandene validierte Langskalen als Grundlage zu nehmen. Es sollte ein Item ausgewählt werden, das hinreichend stark mit der Gesamtskala korreliert. Allerdings ist zu beachten, dass die meisten Langskalen und damit auch die darin enthaltenen Items für die Messung auf der Personenebene entwickelt wurden. Anhand der folgenden Schritte sollte kritisch geprüft werden, ob sie sich auch für den

Einsatz in In-situ-Protokollen eignen. Unter Umständen sind Adaptionen der Items notwendig.

2. **Definitionsgeleitete Auswahl des Items:** Das Item soll bestmöglich zur Definition des theoretischen Konstrukts passen, also dessen Kern messen. Die Auswahl kann – wenn die Ressourcen vorhanden sind – im Rahmen einer eigenen Vorstudie stattfinden. In diesem Fall können wir Teilnehmer:innen z. B. bestehende Langskalen sowie die Konstruktdefinition vorlegen und sie bitten, die Passung der Items zur Definition zu beurteilen (Colquitt et al., 2019; Matthews et al., 2022; Wolfers & Baumgartner, 2025). Forschende können die Beurteilung aber auch selbst durchführen. Es empfiehlt sich, dass mehrere Forschende zunächst getrennt auswählen und dann über eventuelle Abweichungen diskutieren.
3. **Eignung zur Messung von intraindividuellen Schwankungen:** Wir müssen sorgfältig abwägen, ob das zugrunde liegende Konstrukt situativ schwankt. Auf Basis theoretischer Überlegungen ist der empirische Test auf die situative Variation empfehlenswert, um die situative Sensitivität des Items sicherzustellen. Wir können diesen Test im Rahmen eines Pretests des In-situ-Designs (siehe Abschnitt 8.1) oder im Vorfeld im Rahmen einer Vignettenstudie durchführen. Wolfers & Baumgartner (2025) präsentieren ihren Teilnehmer:innen z. B. verschiedene Vignetten, in denen sie zentrale Situationsmerkmale (z. B. das Stresslevel) manipulieren und testen, ob die Items annahmenkonform beantwortet werden (z. B. höherer Stress in der stressigsten Situation).
4. **Kürze, Verständlichkeit und klarer zeitlicher Bezug:** Aufgrund der begrenzten Länge des Protokolls und des mehrmaligen (täglichen) Ausfüllens ist es im In-situ-Kontext entscheidend, auf Kürze, Verständlichkeit und Klarheit (z. B. keine doppelten Verneinungen) der Items zu achten. Am Beispiel von Emotionsmessungen zeigt sich z. B., dass es sinnvoller ist,

7 Konstruktion der Erhebungsinstrumente

nur Adjektive (z. B. „fröhlich“) anstelle von Aussagesätzen zu präsentieren (z. B. „Ich bin fröhlich“). Zudem muss – wie bereits in Abschnitt 7.1.2 erläutert – der zeitliche Bezug eindeutig sein. Soll das Item z. B. für „jetzt gerade“ oder die vergangene halbe Stunde beantwortet werden?

5. **Zusammenhänge mit verwandten Konstrukten:** Wie auch bei der Validierung von Langskalen müssen wir bei der Auswahl und Validierung von Single-Item-Messungen auf deren annahmekonformen Korrelationen mit verwandten Konstrukten achten. Eine Prüfung ist im Rahmen eines Pretests möglich, in dem idealerweise nicht nur das Single-Item, sondern – wenn vorhanden – die Langskala, aus der es entstammt, erhoben und mit anderen Konstrukten korreliert werden. In diesem Fall sollte das Single-Item nicht (deutlich) schlechter abschneiden als die Langskala (siehe zu empirischen Ergebnissen Song et al., 2023; Wolfers & Baumgartner, 2025).

Wenn keine bereits validierten Single-Item-Messungen vorliegen – was aktuell noch häufig der Fall ist –, testen und entwickeln Forschende ihre Messinstrumente idealerweise im Rahmen von Vorstudien (zum Einsatz qualitativer Verfahren für die Validitätssicherung im In-situ-Kontext siehe Schorlepp et al., 2025). Ist diese Möglichkeit aufgrund von knappen Ressourcen nicht gegeben, sollte dennoch bei der Auswahl auf zentrale Aspekte wie die Passung zur Kerndefinition, die Orientierung an bereits etablierten Langskalen und die Itemformulierung (Kürze, Verständlichkeit, zeitlicher Bezug) geachtet werden.

MERKE

Die Auswahl von Kurzskalen und Single-Item-Messungen bedarf großer Sorgfalt. Idealerweise testen und entwickeln Forschende ihre Messinstrumente im Rahmen von Vorstudien.

7.2 Herausforderungen für die praktische Umsetzung

Auch wenn wir unser Protokoll für die Erhebung der Situationsvariablen bestmöglich erstellt haben, kann es zu Einschränkungen der Datenqualität kommen (Bolger et al., 2003; Eisele, Kasanova, et al., 2022). Die Gründe hierfür teilen sich in zwei Gruppen: Reaktivität der Methode und mangelnde Sorgfalt vonseiten der Teilnehmer:innen. Es ist wichtig, die dahinter liegenden Mechanismen zu verstehen, um anhand von Designentscheidungen nach Möglichkeit gegensteuern zu können, das verbleibende Verzerrungspotenzial zu erkennen und daraus ggf. Entscheidungen für die Datenaufbereitung und -auswertung abzuleiten (siehe Kapitel 9).

7.2.1 Reaktivität

Reaktivität ist ein unerwünschter Effekt empirischer Forschung, der die Validität von Studien einschränkt. Es geht um das Phänomen, dass ein Messvorgang die soziale Realität beeinflusst, statt sie lediglich zu dokumentieren (Barta et al., 2012; König et al., 2022; Scholl, 2013). Damit kann die Erfassung des zu messenden Konstrukts beeinträchtigt oder sogar verfälscht werden. Reaktivität ist bei nahezu jedem Erhebungsverfahren gegeben, dessen sich die Teilnehmer:innen bewusst sind. Die Frage ist daher eher, wie ausgeprägt die Reaktivität bei In-situ-Studien im Vergleich zu anderen Verfahren ist.

MERKE

Reaktivität (in In-situ-Studien) bedeutet, dass Befragte aufgrund ihrer Studienteilnahme ihr (Antwort-)Verhalten ändern.

Wenn eine Person etwa mit fortschreitender Feldzeit einer In-situ-Studie angibt, weniger Podcasts zu hören, könnte es sein, dass sie wirklich weniger hört, z. B. weil ihr durch die wiederkehrende Befragung klar wurde, wie viel Zeit sie dafür aufgewendet hat. Bei einem

entsprechend veränderten Verhalten greifen Mechanismen wie Sensibilisierung für das Thema der Studie, Steigerung der Salienz und eine daran anschließende intensivere **Selbstbeobachtung**. Konkret werden sich Nutzer:innen des eigenen Verhaltens durch die wiederholte Abfrage bewusst und reagieren mit einer Anpassung, die ohne die Messung vermutlich nicht eingetreten wäre (Barta et al., 2012; Conner & Reid, 2012; Reynolds et al., 2016). Bei Interventionsstudien ist dieser Effekt sogar beabsichtigt, um diese stetige Selbstbeobachtung zu initiieren (vgl. etwa Korotitsch & Nelson-Gray, 1999; Kuppens & Myin-Germeys, 2022; May et al., 2018, siehe Abschnitt 5.3).

Es gibt empirische Hinweise, dass Reaktivität in In-situ-Studien zwar auftritt (im Überblick für gesundheitsbezogene Verhaltensweisen siehe König et al., 2022), aber eher schwach ausgeprägt ist und insbesondere sensible Themen und Verhaltensweisen (z. B. suchtartiges oder gesundheitsschädliches Verhalten), außergewöhnliche Lebenslagen (z. B. stressige Phasen) oder spezielle Zielgruppen (z. B. Menschen mit Depression) betrifft (Barta et al., 2012; Conner & Reid, 2012; McCarthy et al., 2015; Merrilees et al., 2008; Myin-Germeys & Kuppens, 2022a; Reynolds et al., 2016). Zudem deuten einige Studien darauf hin, dass die Art der Stichprobenziehung auf der Situationsebene einen Einfluss ausübt: Reaktivität ist bei planbar auftretenden Protokollen (z. B. Intervalltagebüchern, siehe Kapitel 2) höher als bei zufälligen Samplingplänen (Barta et al., 2012; Myin-Germeys & Kuppens, 2022a).

Neben diesem Effekt erhöhter Selbstwahrnehmung kann sich Reaktivität auch dahingehend äußern, dass sich das individuelle **Antwortverhalten** im Laufe der Studie verändert. Bestimmte Verhaltensweisen werden vielleicht noch ausgeführt, aber nicht mehr berichtet. Ein Grund hierfür ist sozial erwünschtes Antwortverhalten der Teilnehmer:innen, um ein positives Bild von sich selbst zu präsentieren (Bäckström & Björklund, 2021; Barta et al., 2012), indem vermeintlich negativ zu bewertende Angaben vermieden (*underreporting*) oder umgekehrt wünschenswerte Angaben stärker betont werden (*overreporting*). Ein weiterer Grund für die Anpassung des Antwortverhaltens basiert auf einem Lerneffekt: Teilnehmer:innen

7.2 Herausforderungen für die praktische Umsetzung

stellen im Laufe der Zeit vielleicht fest, dass das Ausfüllen der Protokolle deutlich länger dauert, wenn ein relevantes Verhalten aufgetreten ist. Sie merken, dass es sich lohnt, ausweichend zu antworten, um Folgefragen zu vermeiden und die Erhebung abzukürzen. Dieser Effekt lässt sich vermeiden, wenn alle Protokollversionen in etwa gleich lang sind. In diesem Fall besteht jedoch ggf. das Risiko, dass das Protokoll von den Teilnehmer:innen als unnötig lang wahrgenommen wird, weil sie kein interessierendes Ereignis berichten können.

Eine weitere Form von Reaktivität zeigt sich in sog. **Skalenverankerungen** bzw. Response Shifts (Barta et al., 2012; Csikszentmihalyi & Larson, 1987; Moskowitz et al., 2009; Schwartz et al., 2006; Stone et al., 2023). Dieses Phänomen bezieht sich vor allem auf Abfragen anhand von Antwortskalen (z. B. von 1 „trifft überhaupt nicht zu“ bis 5 „trifft voll und ganz zu“). Der Effekt zeigt sich in einer Kalibrierung des Antwortverhaltens: Nach einigen Antworten verändert sich die Bedeutung, die die Teilnehmer:innen den Skalenabstufungen zuschreiben. Eine Teilnehmerin schätzt z. B. die eigene Stimmung („Ich bin gut gelaunt“) in einem der ersten Protokolle als sehr gut ein und vergibt eine 5. Danach tritt aber eine Situation ein, in der sie noch besser gelaunt ist. Dadurch wird ihr bewusst, dass ihre Stimmung, die sie zuvor mit 5 bewertet hatte, eher eine 3 oder 4 war. Tritt zukünftig also eine der ersten Bewertung vergleichbare Situation ein, wird sie keine 5, sondern eine 3 oder 4 vergeben, was aus der rekalierten Sicht einer realistischeren Einschätzung entspricht (*initial elevation bias*, Shrout et al., 2018). Im besten Fall werden die Antworten über die Zeit genauer und verlässlicher (Csikszentmihalyi & Larson, 1987; Reynolds et al., 2016). Problematisch sind lediglich die ersten Messungen, bevor die Kalibrierung stattgefunden hat. Aus Studienplanungssicht kann sinnvoll sein, eine **Trainingsphase** zur Messkalibrierung einzuplanen, die anschließend nicht in die eigentliche Auswertung eingeht. Die Folge ist allerdings eine längere Feldzeit, was aus Belastungsgründen wiederum nur eingeschränkt zu empfehlen ist. Hier muss je nach Thema abgewogen werden, ob eine Trainingsphase notwendig und möglich ist. Eine intensive Schulung der Teilnehmer:innen im Vorfeld ist aber auf jeden Fall vorteilhaft

(siehe Abschnitt 8.3).

7.2.2 Mangelnde Qualität einzelner Antworten

Aufgrund der hohen empfundenen Belastung durch In-situ-Studien kann es passieren, dass Teilnehmer:innen nicht in allen Situationen mit der gebotenen Sorgfalt antworten und einzelne Testfragen oder Items, die nicht als Pflichtfelder gekennzeichnet sind, im Protokoll auslassen. Dieses Thema ist unter dem Stichwort Item Nonresponse für Befragungsstudien allgemein umfassend erforscht (z. B. Roßmann, 2017; Yan & Curtin, 2010). Tendenziell entscheiden sich Teilnehmer:innen allerdings eher komplett gegen eine Teilnahme, brechen die Studie ganz ab oder füllen nicht alle Protokolle aus (Unit Nonresponse), als dass sie mangelnde Sorgfalt walten lassen (Fuller-Tyszkiewicz et al., 2013). Zur Korrektur von Item Nonresponse stehen – falls notwendig und gewünscht – statistische Methoden zur Imputation zur Verfügung (siehe Abschnitt 9.2.2.2). Es kann jedoch auch der Fall eintreten, dass die Teilnehmer:innen den in Abschnitt 7.1.1 beschriebenen kognitiven Antwortprozess abkürzen, um möglichst zügig zu einer aus ihrer Sicht ausreichend passenden Antwort zu kommen. Diese Heuristik wird als ‘Satisficing’ bezeichnet (Krosnick, 1999; C. Roberts et al., 2019). Andere Befragte antworten schlicht nachlässig, etwa weil sie abgelenkt oder grundsätzlich unmotiviert sind (sog. Careless Responding, Barta et al., 2012; Jaso et al., 2022).

MERKE

Die mangelhafte Qualität einzelner Antworten kann auf einer Heuristik beruhen, die den kognitiven Antwortprozess abkürzt (Satisficing), oder schlicht auf Nachlässigkeit und Desinteresse zurückzuführen sein (Careless Responding).

Prinzipiell wird Satisficing durch eine eingeschränkte oder im Verlauf der Feldphase nachlassende Motivation, mangelnde kogni-

7.2 Herausforderungen für die praktische Umsetzung

ve Fähigkeiten oder eine zu hohe Schwierigkeit der Aufgabe begünstigt (Roßmann, 2017; Sturgis & Brunton-Smith, 2023). Ebenso wie beim Careless Responding sind die Daten fehlerhaft und bestehende Zusammenhänge bleiben ggf. unerkannt. Es gibt im Kontext von Querschnittsbefragungen Empfehlungen für den Umgang mit diesen Phänomenen (siehe z. B. Meade & Craig, 2012; C. Roberts et al., 2019; Roßmann, 2017). In In-situ-Studien kann insbesondere Careless Responding verschiedene Formen annehmen (Berkel et al., 2019; Eisele, Kasanova, et al., 2022; Fuller-Tyszkiewicz et al., 2013; Reynolds et al., 2016). Zum einen kann es sich darin zeigen, dass eine Person zufällig antwortet oder immer die gleichen Antworten auswählt (zu ähnlichen Phänomenen wie *response sets* oder *straightlining* siehe Kim et al., 2019). Anders als bei Querschnittsbefragungen kann sich dieses Verhalten über die Feldzeit hinweg verändern. Am Anfang ist die Sorgfalt groß, aufgrund der Belastung sinkt jedoch im Laufe der Zeit die Kooperationsbereitschaft. Entsprechende Hinweise finden Berkel et al. (2019) für Feldzeiten von über zwei Wochen. Eisele, Kasanova, et al. (2022) konnten hingegen keine veränderte Kooperationsbereitschaft während der Feldzeit nachweisen.

Ein häufig verwendeter Indikator, um solches Antwortverhalten zu identifizieren, ist die **Ausfülldauer**. Ausgangspunkt ist die Annahme, dass eine kürzere Antwortzeit mit geringerer Datenqualität einhergeht, weil die Teilnehmer:innen nicht sorgfältig lesen bzw. antworten (Roßmann, 2017). Es ist umstritten, ob diese Annahme gleichermaßen für In-situ-Studien gilt. Die Protokolle sind ohnehin überwiegend sehr kurz, sodass nur eine geringe Varianz in der Ausfülldauer besteht und damit besonders kurze Antwortzeiten schwer zu identifizieren sind. Zudem gibt es im In-situ-Kontext Hinweise, dass lange Antwortzeiten bei bestimmten Fragetypen (speziell Erinnerungsaufgaben/Recall) zu einer geringeren Datenqualität führen, weil Befragte während des Antwortens abgelenkt waren und anderen Tätigkeiten nachgegangen sind (Berkel et al., 2019; siehe aber auch Jaso et al., 2022).

Schließlich kann sich Careless Responding auch in einem längeren zeitlichen Abstand zwischen dem Signal und dem Ausfüllen des Pro-

7 Konstruktion der Erhebungsinstrumente

tokolls ausdrücken (Reynolds et al., 2016). Mit steigender Response Latency steigt auch die Gefahr einer verzerrten Situationsstichprobe (siehe Abschnitt 6.3.2.2).

8 Durchführung einer In-situ-Studie

8.1 Pretest

Unabhängig davon, wie sorgfältig die Erhebungsinstrumente entwickelt, evaluiert und am Schreibtisch mithilfe theoretischer Verfahren wie ESM-Q (Eisele et al., 2025) evaluiert wurden, ist es immer empfehlenswert, diese sowie alle weiteren Aspekte einer In-situ-Studie im Rahmen eines sog. Pretests vor ihrem Einsatz im Feld gründlich zu überprüfen (vgl. ausführlich Möhring & Schlütz, 2025, siehe Abschnitt 6.2). Es ist sinnvoll, sowohl methodisch versierte Personen als auch Angehörige der angestrebten Zielgruppe einzubeziehen.

MERKE

Im Pretest wird die inhaltliche und technische Eignung der Erhebungsinstrumente überprüft, bevor sie ins Feld gehen.

Im Pretest sollten die Instrumente, also die Fragebögen der Vorab- und Nachbefragung sowie das In-situ-Protokoll auf Verständlichkeit der Fragen, Länge etc. getestet werden. Kognitive Interviews eignen sich beispielsweise, um das korrekte inhaltliche Verständnis der Protokollfragen sicherzustellen (vgl. Beatty & Willis, 2007). Diese Form des Pretests ist vergleichbar mit der Methode des lauten Denkens (ggf. ergänzt um Verständnisanfragen): Beim Ausfüllen des Fragebogens bzw. Protokolls werden die Pretester:innen gebeten, ihre Gedanken zu artikulieren. Auf diese Weise werden Informationen gesammelt, die dazu dienen, die Qualität der Antworten zu bewerten und festzustellen, ob die Frage die intendierten Informationen liefert (für Beispiele siehe Eisele, Vachon, et al., 2022). Ein solcher Pretest basiert auf

dem in Abschnitt 7.1.1 vorgestellten kognitiven Antwortprozess – Verstehen, Erinnern/Urteilen, Schätzen und Anpassen. Er zielt darauf ab, diesen Prozess abzubilden und ggf. damit verbundene Probleme zu identifizieren, um die Reliabilität und Validität der Antworten sicherzustellen. Eine weniger aufwändige Form des Pretests besteht darin, Testpersonen zu bitten, die Fragebögen und Protokolle auszufüllen und ihre Anmerkungen an die Forschenden zu übermitteln. Software für Onlinebefragungen bietet dafür üblicherweise eine Pretestfassung mit offenen Kommentarfeldern an (wie z. B. *SoSci Survey*, Leiner, 2024). Außerdem ist für eine In-situ-Studie der reibungslose technische Ablauf zentral, also etwa der korrekte Versand der Signale, die Verbindung der Datensätze aus den verschiedenen Studienteilen, die Funktionsfähigkeit der eingesetzten technischen Lösung, z. B. der App etc. Diese Aspekte sollten zentraler Teil des Pretests sein.

8.2 Incentivierung

Viele Studien bieten ihren Teilnehmer:innen Belohnungen, also Incentives an, um die Teilnahmebereitschaft zu steigern und die Selbstselektivität zu reduzieren. Darüber hinaus sollen Incentives zum vollständigen und sorgfältigen Ausfüllen motivieren und damit die Datenqualität erhöhen. Wir können zwischen monetären und nicht monetären Belohnungen unterscheiden. Nicht-monetäre Incentives können etwa Spenden an Wohltätigkeitsorganisationen sein oder der Zugang zu Studienergebnissen. Bei Siebers et al. (2024) (siehe Studiensteckbrief im Online-Anhang zum Buch) konnten die Teilnehmer:innen beispielsweise ihr Ausfüllverhalten über eine interaktive Website in Echtzeit beobachten.

Der Einfluss von Incentives wurde bereits intensiv untersucht (vgl. Keusch, 2015; Singer, 2018). Empirische Ergebnisse zeigen, dass monetäre Incentives die Teilnahmebereitschaft allgemein sowie in In-situ-Studien im Besonderen erhöhen (Keusch et al., 2019; Kibuchi et al., 2020). Nicht-monetäre Incentives haben sich als weniger effektiv

erwiesen (Göritz & Neumann, 2016; im Überblick Keusch, 2015; zu gegenteiligen Ergebnissen aber Edwards et al., 2023).

Monetäre Incentives können vorab oder nachträglich und dann in Abhängigkeit von der tatsächlichen Teilnahme und dem regelmäßigen Ausfüllen der Protokolle ausgezahlt werden (siehe Abschnitt 6.3.2.2). Die Studienlage zeigt allerdings konsistent, dass grundsätzlich gewährte und vorab ausgezahlte Incentives an alle Teilnehmer:innen wirksamer sind als solche, die nachträglich und abhängig von der Teilnahme gewährt werden. Je nach Rekrutierungsstrategie sind voraussetzungsfreie Belohnungen bei Onlinebefragungen allerdings kaum umzusetzen. Daher ist die nachträgliche, an die Compliance geknüpfte Incentivierung bei In-situ-Studien deutlich verbreiteter. Alternativ erweist sich teilweise auch eine Kombination aus vorab und nachträglich ausgezahlten Belohnungen als erfolgreich (Göritz, 2008; Patrick et al., 2013).

Hinsichtlich der **Höhe der Incentivierung** können bereits kleine Beträge bei einfachen Befragungen erfolgreich sein (z. B. 2 US-\$, Gajic et al., 2012), wobei mit steigender Summe auch die Antwortrate steigt (Singer, 2018). In einer Metaanalyse von ESM-Studien im medizinischen Bereich zeigen Vachon et al. (2019), dass die Höhe des Incentives zudem positiv mit der Compliance korreliert. Dabei bieten sich höhere Beträge (im Überblick Singer, 2018) und mehrere Belohnungen über die Panelwellen hinweg an (Jäckle & Lynn, 2008). Siebers et al. (2024) zahlten ihren Teilnehmer:innen z. B. 0,30 Euro für jedes ausgefüllte Protokoll (siehe Studiensteckbrief im Online-Anhang zum Buch).

Das Budget reicht häufig nicht aus, alle Teilnehmer:innen – unabhängig davon, ob und wenn ja wie gut sie die Studie durchgehalten haben – zu belohnen. Daher werden häufig Gutscheine oder Sachpreise **verlost** (Zhu et al., 2018). Wertigere Lotteriepreise steigern die Teilnahmebereitschaft mehr als niedrigere, auch bei geringeren Gewinnchancen (Gajic et al., 2012). Auch in eine Verlosung können alle Teilnehmer:innen einbezogen werden oder nur solche, die regelmäßig Protokolle ausgefüllt haben. In der Studie von Siebers et al. (2024) nahmen diejenigen, die alle sechs Protokolle eines Tages ausgefüllt

8 Durchführung einer In-situ-Studie

hatten – zusätzlich zu anderen Maßnahmen –, an einer täglichen Lotterie teil, bei der 4 x 25,00 Euro verlost wurden (siehe Studiensteckbrief im Online-Anhang zum Buch). Karnowski et al. (2017) verlost vier Gutscheine im Wert von 50,00 Euro unter allen, die einen Vorabfragebogen und über acht Tage einmal täglich einen Kurzfragebogen ausfüllten (siehe Studiensteckbrief im Online-Anhang zum Buch, für weitere Beispiele siehe Berkel et al. (2017)). Im Rahmen unserer Beispielstudie wurde unter allen Teilnehmer:innen des In-situ-Teils ein Kopfhörer mit Noise-Cancelling-Funktion verlost. Eine Mindestanzahl an beantworteten Protokollen war nicht Voraussetzung, um an der Verlosung teilzunehmen (siehe Studiensteckbrief im Online-Anhang zum Buch). Insgesamt sollten für die Höhe der Incentivierung neben dem verfügbaren Budget die Länge des Erhebungszeitraums und der einzelnen Protokolle, ihre Komplexität sowie die Häufigkeit der Signale berücksichtigt werden. Eisele, Vachon, et al. (2022) verweisen darauf, dass viele Ethikkommissionen eine Orientierung der Belohnung am Aufwand der Beteiligung erwarten. Dennoch arbeiten einige Studien aufgrund eingeschränkter Budgets auch ohne monetäre Anreize (z. B. Dogruel & Schnauber-Stockmann, 2020; Naab, 2013).

MERKE

Für die Steigerung der Teilnahmebereitschaft erweisen sich vor allem monetäre Incentives als effektiv. Deren Höhe sollte sich am Aufwand der Beteiligung orientieren. Obwohl vorab ausbezahlte Incentives erfolgsversprechender sind, wird eine Belohnung – oft aufgrund von Budgetbegrenzungen – i. d. R. nachträglich ausgezahlt und an eine bestimmte Compliance geknüpft.

8.3 Schulung, Feldphase und Debriefing

In-situ-Studien erstrecken sich in der Regel über einen längeren Zeitraum und sind komplex in ihrer inhaltlichen und technischen Ausgestaltung. Die Anforderungen an die Teilnehmer:innen sind also insgesamt hoch. Es bedarf einer Reihe von Bemühungen, die *research alliance* (Larson & Csikszentmihalyi, 1983, siehe Abschnitt 6.2.2.2) aufzubauen und die Teilnahmebereitschaft und -motivation sicherzustellen. Unterstützende Maßnahmen sind – neben Incentives – die Bereitstellung geeigneter Informationen, eine Schulung im Vorfeld der Studie, eine enge Betreuung während der Feldphase sowie ein Debriefing.

8.3.1 Informationen im Vorfeld der Studie

Aus forschungsethischen Gründen (vgl. Kapitel 4), aber auch um einen reibungslosen Ablauf einer In-situ-Studie sicherzustellen, ist es notwendig, dass wir die Teilnehmer:innen im Vorfeld genau über den Ablauf der Studie und über die an sie gestellten Erwartungen (z. B. die Anzahl auszufüllender Protokolle) informieren. Es bietet sich an, im Vorfeld der Studie eine entsprechende Veranstaltung einzuplanen, bei der Studieninformationen in einem persönlichen Gespräch und/oder einem ausführlichen, aber einfach verständlichen schriftlichen Briefing weitergegeben werden. Das Briefing zur Studie von Aalbers et al. (2022) fand etwa im Klassenraum der jugendlichen Teilnehmer:innen statt (siehe Studiensteckbrief im Online-Anhang zum Buch). Dabei unterstützten die Forschenden bei der Installation der Studienapp und erläuterten den Ablauf der ESM-Phase. Auch Erklärvideos sind denkbar. Im Online-Anhang zum Buch unter Ressourcen finden sich einige Beispiele für Erklärvideos. Ziel ist, die Teilnehmer:innen zu motivieren und ihnen zentrale Aspekte der Vorgehensweise zu verdeutlichen (Häufigkeit der Alarmierung, Wiederholungscharakter, Unmittelbarkeit des Ausfüllens) und Protokolleinträge zu üben. Vor allem können wir kommunizieren, dass täglich mehrere identische Fragebögen auszufüllen sind, und das Vorgehen

im Falle eines Problems (Kontaktmöglichkeit) und die Bewältigung besonderer Situationen (z. B. Fragen von anderen Personen) besprechen. Schließlich ist eine solche Schulung eine passende Gelegenheit, Feedback von den Teilnehmer:innen einzuholen. Ein solches Vorgehen kann die fehlerfreie Beteiligung verbessern, Abbrüche vermeiden sowie Compliance und Teilnahmemotivation steigern (Palmier-Claus et al., 2011; Stone et al., 2023; Wen et al., 2021).

Über ein Manual, eine Studienwebseite oder in der verwendeten Erhebungsassp können wir Informationen bereitstellen, auf die die Teilnehmer:innen jederzeit zurückgreifen können. Die konkreten Informationen sind abhängig vom Studieninteresse und -verlauf. Wichtig sind folgende **Teilnahmehinweise** (Formulierungsvorschläge für Instruktionen, die an die eigene Studie angepasst werden können, finden sich z. B. bei Eisele, Vachon, et al., 2022):

- technische Voraussetzungen (z. B. bestimmte Betriebssysteme, Notwendigkeit einer mobilen Datenverbindung oder WLAN-Verbindung)
- technische Einstellungen: die verwendete App, ihre Installation und Bedienung, notwendige Einstellungen am eigenen Gerät (z. B. Benachrichtigungen und Berechtigungen)
- Aufbau der Erhebung: Vorabbefragung, In-situ-Phase, Nachbefragung etc.
- Beginn und Ende der Erhebung
- zentrale Konstrukte und ihre Messung: Es empfiehlt sich, das Protokoll durchzusprechen und die Operationalisierungen zu erklären, etwa auf welchen Zeitraum sich das Protokoll bezieht (z. B. „jetzt gerade“ vs. „in der letzten Stunde“)
- Art der Alarmierung (z. B. per App, SMS, E-Mail)
- Häufigkeit und Zeitraum der Signale: Ggf. ist es angebracht, nicht die genaue Anzahl und die konkreten Zeiten anzugeben;

8.3 Schulung, Feldphase und Debriefing

zentral ist bei selbstausgelösten ereignisbasierten ESM-Studien die genaue Definition des Ereignisses und wie das Protokoll aufgerufen werden kann

- Handlungsanweisungen zur Alltagsgestaltung: Der Alltag soll möglichst nicht angepasst und das Smartphone immer bereitgehalten werden
- Handlungsanweisungen über die erwartete Compliance: z. B. wie lange nach dem Signal ein Protokoll ausgefüllt werden kann (maximale Response Latency); was zu tun ist, wenn ein Signal versäumt wurde oder zwei aufeinanderfolgende Signale bereitstehen; ob und wie ein Protokoll nachgeholt werden kann
- vollständige Informationen zur Datenerhebung, -speicherung und -verarbeitung (siehe Kapitel 4)
- Informationen zur Incentivierung
- Kontaktdaten für Rückfragen

In der Podcastingstudie wurden die Teilnehmer:innen folgendermaßen instruiert (siehe Studiensteckbrief im Online-Anhang zum Buch):

Der zweite Teil der Untersuchung erstreckt sich über zwei Wochen, und zwar vom 18.09. bis 01.10.2023. Er besteht aus bis zu vier sehr kurzen Umfragen (max. 2–3 Minuten) pro Tag. In dem Teil geht es uns vor allem um die jeweilige Situation, in der Sie Podcasts hören.

Sie benötigen für diesen Teil keine gesonderte App. Wir werden Ihnen einfach eine SMS mit einem Befragungslink schicken, der Sie zum Kurzfragebogen leitet. Sie können ihn dann zwischendurch am Handy ausfüllen. Dafür benötigen Sie mobiles Datenvolumen oder eine WLAN-Verbindung. Der Link ist nach dem Versand eine Stunde gültig. In diesem Zeitfens-

ter können Sie die Umfrage ausfüllen. Wenn Sie gerade keine Gelegenheit dazu haben, machen Sie einfach bei der nächsten Umfrage wieder mit. Für das Gelingen der Studie ist es aber wichtig, dass so viele Kurzfragebögen wie möglich beantwortet werden. Wir freuen uns daher, wenn Sie Ihr Smartphone bereithalten und so häufig wie möglich mitmachen – aber natürlich sollen Sie ihr „normales“ Verhalten nicht aufgrund der Studie verändern, denn gerade das interessiert uns ja.

Selbstverständlich ist auch die Teilnahme an diesem Teil freiwillig und kann jederzeit abgebrochen werden. Auch die Kurzfragebögen sind vollständig anonym. Es werden keine zusätzlichen Datenspuren erhoben (wie z. B. Ihr Aufenthaltsort). Die Angaben, die Sie im Rahmen der Befragung freiwillig machen, werden durchgängig vertraulich behandelt und ausschließlich für wissenschaftliche Zwecke genutzt.

Als kleines Dankeschön für Ihre Mühe verlosen wir unter allen Teilnehmer:innen der kompletten Studie einen Kopfhörer mit Lärmreduktionsfunktion nach Wahl (im Wert von bis zu 245 Euro). Am Ende der Studie haben Sie Gelegenheit, an der Verlosung teilzunehmen.

8.3.2 Betreuung während der Feldphase und Debriefing

Auch während der Feldphase ist es ratsam, kontinuierlich Kontakt mit den Teilnehmer:innen zu halten, um ihre Teilnahmebereitschaft etwa durch gelegentliche motivierende Kurznachrichten aufrecht zu erhalten. Umgekehrt sollten auch die Teilnehmer:innen die Möglichkeit haben, Kontakt zu den Forschenden aufzunehmen, z. B. per SMS, E-Mail oder über die eingesetzte App. Wenn die individuelle Compliance kontinuierlich überwacht wird, können auch solche Teilnehmer:innen gezielt kontaktiert werden, die eine geringe Beteiligung zeigen (Silvia et al., 2013). Zusätzlich ist es sinnvoll, dass eine Person aus dem Team an der Studie teilnimmt, um eventuell auftretende technische Fehler (z. B. fehlende Alarmierungen) während der

Feldphase rasch zu erkennen und zu beheben. Diese Daten müssen natürlich vor der Analyse aus dem Datensatz entfernt werden.

Zum Abschluss der Befragung ist ein Debriefing empfehlenswert. Eine solche Nachbesprechung kann aus forschungsethischen Gründen angezeigt sein, um die Teilnehmer:innen nachträglich über den genauen Studienzweck aufzuklären, wenn dies im Vorfeld der Studie aus methodischen Gründen nicht möglich oder nicht sinnvoll war (vgl. Verbeke et al., 2023). Weiterhin können im Rahmen des Debriefings die Incentives ausgeschüttet und die notwendigen Kontaktdaten dafür erhoben werden.

Ein Debriefing kann auch genutzt werden, um zu ermitteln, welche Erfahrungen die Teilnehmer:innen mit der Studienteilnahme gemacht haben. Es ist sinnvoll, am Ende der Erhebung standardisiert Feedback einzuholen, um die Zuverlässigkeit der Ergebnisse abschätzen zu können. Beispielsweise können die Befragten reflektieren, inwiefern der Erhebungszeitraum ihrem gewöhnlichen Alltag entsprach, ob besondere Ereignisse aufgetreten waren und wie sie mit den Protokollen zurechtkamen. Solche Einschätzungen können helfen, methodische Limitationen wie etwa Reaktivitätseffekte abzuschätzen (vgl. auch Eisele, Vachon, et al., 2022). Im Rahmen unserer Beispielstudie wurden verschiedene Items bewertet, um Aspekte der Studie wie externe Validität, Careless Responding oder Reaktivität besser beurteilen zu können (siehe Studiensteckbrief im Online-Anhang zum Buch, siehe auch Ernst et al. (2026) für ähnliche Formulierungen):

- Aufgrund der Studienteilnahme hatte ich Sorgen um private Informationen, die ich online preisgegeben habe.
- Aufgrund der Studienteilnahme hat sich die Art und Weise verändert, wie ich Podcasts höre.
- Die letzte Woche war typisch für meinen momentanen Alltag.

8 Durchführung einer In-situ-Studie

- Die Studie hat sich gut in den Alltag eingepasst und hat kaum gestört.
- Es gab viele Momente, wo ich den Fragebogen nicht ausfüllen konnte oder wollte.
- Die kurzen Befragungen haben mich beim Hören eines Podcasts gestört.
- Es war anstrengend, die Fragen immer wieder zu beantworten.
- Ich habe die Kurzfragebögen in der Regel gewissenhaft ausgefüllt.

9 Auswertung einer In-situ-Studie

Im Folgenden gehen wir auf das Datenmanagement und die Auswertung von In-situ-Daten ein. Die zentralen Schritte sind im Online-Anhang zum Buch in R (R Core Team, 2025) anhand unserer Beispielstudie durchgeführt und kommentiert. Der Code soll den Leser:innen dieses Buches helfen, die technische Komponente des Datenmanagements und der Datenauswertung nachzuvollziehen. Er bietet außerdem eine Ressource, die die Leser:innen für eigene Datensätze adaptieren können. Wir haben uns für R entschieden, weil R ein freies und kostenlos zugängliches und durch seine Erweiterungsmöglichkeiten in Form von Paketen sehr flexibel einsetzbares Tool ist. R findet zunehmend Verbreitung in der KMW und eignet sich besonders für den Umgang mit komplexeren Datenstrukturen und für fortgeschrittene Auswertungsverfahren, wie sie im Zusammenhang mit In-situ-Daten häufig vorliegen. Weder dieses Buch noch der Online-Anhang zum Buch können allerdings in die statistischen Grundlagen aller vorgestellten Verfahren und in ihre Verwendung von R einführen, sondern konzentrieren sich auf den datenanalytischen Umgang mit In-situ-Daten. Grundlegende Einführungen in R finden sich in Buchform z. B. bei Field (2026), Gehrau et al. (2022), Sauer (2019) und Wickham et al. (2023) sowie im Lehrveranstaltungsskript von Schnauber-Stockmann et al. (2024).

Leser:innen finden im Handbuch von Myin-Germeyns & Kuppens (2022b) zwei Kapitel zur Aufbereitung und Auswertung von In-situ-Daten: Viechtbauer (2022a) gibt einen ausführlichen Überblick über grundlegende Auswertungsmethoden und Lafit (2022) widmet sich den fortgeschrittenen Methoden. Eine detaillierte Literaturübersicht zu (vor allem fortgeschrittenen) Auswertungen von In-situ-Daten findet sich auch im Anhang von Trull & Ebner-Priemer (2020). Zudem legen Bolger & Laurenceau (2013) einen Fokus auf die Auswertung

von In-situ-Daten und stellen grundlegende sowie fortgeschrittene Auswertungsverfahren übersichtlich vor. Revelle & Wilt (2019) gehen dezidiert auf die Auswertung von In-situ-Daten mit R ein.

9.1 Struktur von In-situ-Daten

Die Daten von In-situ-Studien weisen eine **Mehrebenenstruktur** auf. Auf der unteren Ebene liegen die einzelnen Messzeitpunkte (also die einzelnen Protokolle), und auf der oberen Ebene die Personen. Die Messzeitpunkte sind in Personen geschachtelt, d. h., mehrere Messzeitpunkte stammen von derselben Person. Wenn wir nun ein Merkmal wiederholt messen und die Ergebnisse variieren, sie unterscheiden sich also zwischen den Messungen, dann können zwei Ursachen (Messfehler ausgenommen) dafür maßgeblich sein: (1) Personen unterscheiden sich und (2) Situationen unterscheiden sich. Personenunterschiede sagen aus, dass eine Person z. B. generell besser gelaunt oder affiner für Podcasts ist als eine andere. Dabei beschreibt die **interindividuelle Varianz** die Unterschiede zwischen Personen. In Abbildung 9.1 sehen wir bei der ersten und zweiten Person von oben nur Personenunterschiede: Eine hört nie, die andere immer Podcasts. Situationsunterschiede zeigen Veränderungen auf, die eine Person im Zeitverlauf durchläuft: Sie ist z. B. am Montagmorgen schlechter gelaunt als am Dienstagabend und hört unter der Woche mehr Podcasts als am Wochenende. Diese **intraindividuelle Varianz** bildet die Unterschiede innerhalb einer Person in unterschiedlichen Situationen ab. In Abbildung 9.1 sehen wir bei der dritten und vierten Person Beispiele für intraindividuelle Varianz: Die beiden Personen hören in einigen Situationen Podcasts, in anderen nicht. Zusätzlich erkennen wir auch Personenunterschiede, denn wir sehen, dass die Person in der dritte Zeile seltener Podcasts hört als die Person in der vierten Zeile. Selten zeigt sich ein Verhalten so einheitlich (nicht) wie bei den Personen in der ersten und zweiten Zeile; überwiegend werden wir – wie bei den Personen in der dritten und vierten Zeile – beide Varianzquellen finden. Liegen mehr als drei Messzeitpunkte

9.1 Struktur von In-situ-Daten

vor, ist es statistisch möglich, interindividuelle und intraindividuelle Varianz voneinander zu unterscheiden (siehe Abschnitt 9.5.2.2).

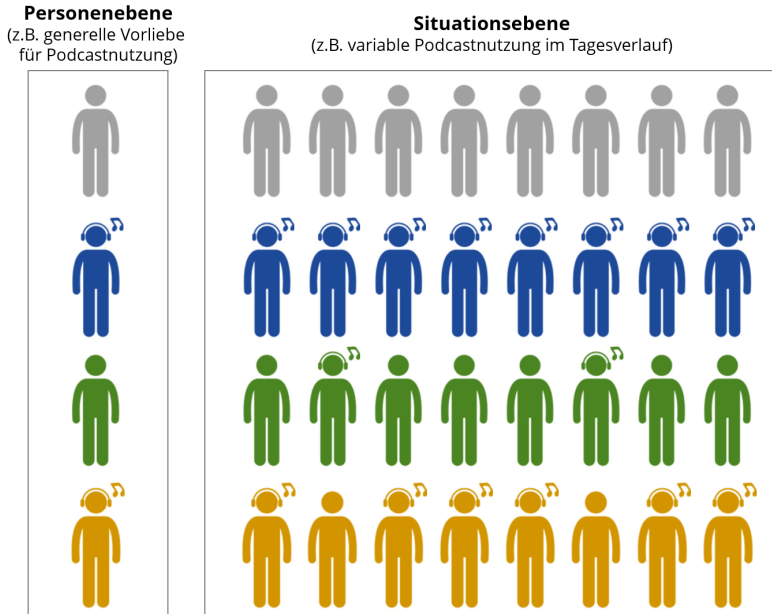


Abbildung 9.1: Personen- und situationsabhängige Varianz (inter-/intraindividuell)

Merke

In-situ-Daten haben eine Mehrebenenstruktur: Messzeitpunkte sind in Personen geschachtelt, sodass Unterschiede sowohl zwischen Personen (interindividuelle Varianz, *between-person variance*) als auch innerhalb einer Person zwischen verschiedenen Messsituationen (intraindividuelle Varianz, *within-person*

variance) auftreten können.

Grundsätzlich enthält der Datensatz einer In-situ-Studie immer Variablen auf der Situationsebene. Einerseits handelt es sich um zeitbezogene **Variablen, die den Messzeitpunkt eindeutig identifizieren** und beschreiben. Sie geben etwa den Tag der Erhebung an (z. B. Studientag 1-8), um welches Protokoll des Tages es sich handelt (z. B. 1-4 bei 4 Protokollen am Tag), die Uhrzeit usw. Andererseits gibt es auf der Situationsebene **Variablen, die Antworten auf die Fragen des Protokolls enthalten** (siehe auch Trull & Ebner-Priemer, 2020). Zudem benötigt jeder In-situ-Datensatz eine eindeutige **Identifikationsnummer (ID)** für jede teilnehmende Person, um die einzelnen Messzeitpunkte den Teilnehmer:innen zuzuordnen zu können. Sehr oft sind zudem **Personenvariablen** von Interesse, die wir im Rahmen einer Vorab- oder Nachbefragung erheben. Diese Variablen sind zeitstabil, d. h. für jeden Messzeitpunkt einer Person identisch. Sie sind häufig in einem eigenen Datensatz enthalten, der dieselbe ID für eine Person ausweisen muss wie der Situationsdatensatz der entsprechenden Person. Anhand dieser ID können die Personenvariablen dann zugespielt werden (siehe hierzu Abschnitte 6.3.2 und 8.1).

Merke

In-situ-Datensätze enthalten immer Variablen auf der Situationsebene (z. B. Zeitpunkte, Antworten). Diese Variablen können durch Personenvariablen ergänzt werden, die aus einer Vorab- oder Nachbefragung stammen. Deren Messung findet nur einmal statt, weil sie als intraindividuell stabil angenommen werden.

Je nach Studiendesign können auch noch weitere Ebenen oberhalb der Person (z. B. Familien, Peergroups, Paare) relevant sein oder es sind Variablen notwendig, die verschiedene Studienphasen

kennzeichnen (z. B. Pre-/Post-Intervention, Welle des Measurement-Bursts o. ä.). Auf diese Aspekte gehen wir im Folgenden nicht näher ein (siehe z. B. Bolger & Laurenceau, 2013; Trull & Ebner-Priemer, 2020 und Exkurs im Abschnitt 9.5.2.1).

Die Daten können im Datensatz in zwei Formaten vorliegen, im sog. **Long- oder Wide-Format**. Im Long-Format bildet jeder Messzeitpunkt für jede Person eine Zeile und es gibt jeweils nur eine Spalte pro Variable auf der Situationsebene sowie je eine Spalte für die zeitstabilen Personenvariablen (siehe Tabelle 9.1). Die Ausprägungen dieser Personenvariablen sind für alle Zeilen derselben Person identisch, denn diese Personenvariablen verändern sich nicht über die Messzeitpunkte hinweg. Im Wide-Format stellt jede Zeile eine:n Teilnehmer:in dar. Für jede Situationsvariable pro Messzeitpunkt (z. B. Messzeitpunkt t_1 , Messzeitpunkt t_2 etc.) gibt es eine Spalte, zudem noch je eine Spalte für jede zeitstabile Personenvariable (siehe Tabelle 9.2).

Je nach Auswertungsverfahren ist das Wide- oder Long-Format notwendig. Die häufiger benötigte Datenstruktur ist aber das Long-Format. Die Umstrukturierung von Datensätzen vom Wide- in das Long-Format und umgekehrt ist in R und anderen Programmen möglich siehe Kapitel A im Online-Anhang zum Buch.

Merke

In-situ-Datensätze können im Long- oder Wide-Format vorliegen. Im Long-Format gibt es pro Messzeitpunkt und Person eine Zeile, während im Wide-Format alle Messzeitpunkte einer Person in einer Zeile stehen. Für die meisten Analysen wird das Long-Format benötigt.

Tabelle 9.1: Beispieldaten im Long-Format

Person	Messzeitpunkt	Situationsvariablen		Personenvariablen	
		Podcast	Stimmung	Geschlecht	Alter
1	1	ja	5	weiblich	23
1	2	nein	3	weiblich	23
2	1	ja	2	weiblich	40
2	2	ja	1	weiblich	40
3	1	nein	5	männlich	21
3	2	nein	2	männlich	21
4	1	nein	1	weiblich	32
4	2	ja	4	weiblich	32

Tabelle 9.2: Beispieldaten im Wide-Format

Person	Situationsvariablen				Personenvariablen	
	t1		t2		Geschlecht	Alter
	Podcast	Stimmung	Podcast	Stimmung		
1	ja	5	nein	3	weiblich	23
2	ja	2	ja	4	weiblich	40
3	nein	5	nein	2	männlich	21
4	nein	1	ja	3	weiblich	32

9.2 Datenmanagement

9.2.1 Dateninspektion, -prüfung und -bereinigung

Vor Beginn der eigentlichen Analyse ist es sinnvoll, den Datensatz zu inspizieren, um einen ersten Eindruck von den Daten zu erhalten. Dieser Schritt dient der Plausibilitäts- und Vollständigkeitsprüfung sowie der Prüfung von möglichen systematischen Verzerrungen und kann für die Datenbereinigung wichtige Parameter liefern.

Im Folgenden geben wir einen Überblick über mögliche Überprüfungen mit dem Schwerpunkt auf der Inspektion des In-situ-Datensatzes. Natürlich müssen aber auch die Daten einer Vorab- oder Nachbefragung geprüft und ggf. bereinigt werden. Nicht alle der folgenden Schritte sind immer notwendig, dies hängt u. a. vom Design der Studie ab. Die Schritte sollen einen Eindruck davon vermitteln, wie wir uns mit den Daten vertraut machen und wie wir die Daten bereinigen können (siehe auch Revol et al., 2024).

1. Schritt – Grundlegende Checks

Fallzahlen: Wie viele Fälle hat der Datensatz auf der Personen- und Situationsebene? Wie viele Personen haben sich an der Studie beteiligt und wie viele Messzeitpunkte liegen von allen Personen vor?

Ausfülldauer: Wie lange dauerte das Ausfüllen der Protokolle durchschnittlich?

Vollständigkeit der Variablen: Sind alle benötigten Variablen vorhanden?

- Hierfür bietet es sich an, eine Liste aller im Datensatz vorhandenen Variablen zu erstellen und auf Vollständigkeit zu prüfen.
- Wir empfehlen, die Variablen so zu benennen, dass wir leichter den Überblick behalten. Die Vorgehensweise bei der Benennung von Variablen orientiert sich an persönlichen Präferenzen. Sinnvoll ist in jedem Fall, zusammengehörige Variablen gleich zu benennen (wurde beispielsweise die narrative Transportation anhand von sechs Items erhoben, könnten die Variablen `narr_trans_01`, `narr_trans_02` usw. benannt werden).

Sinnhaftes Ausfüllen: Sind alle Variablen ausgefüllt, die ausgefüllt sein müssen, und sind alle Zellen leer, in denen kein Wert stehen sollte (z. B. bei Filtern)?

- Diese Fragen können wir über Häufigkeitsauszählungen prüfen.
- Hierbei erkennen wir auch, ob es bei bestimmten Variablen vie-

9 Auswertung einer In-situ-Studie

le fehlende Werte gibt. Falls ja, sollten wir den Umgang mit fehlenden Werten überprüfen (siehe Abschnitt 9.2.2.2).

Verteilung der Variablen: Sind alle Werte plausibel? Wie verteilen sich die Variablen?

- Gibt es nur Werte, die im Wertebereich der jeweiligen Variablen liegen? Enthalten offene Felder unplausible Angaben (z. B. eine ungewöhnliche Angabe bei der Nutzungsdauer)?
- Sind fehlende Werte als solche zu erkennen? In R müssen diese Werte mit NA (not available) gekennzeichnet sein. Haben wir im Protokoll z. B. für „weiß nicht/keine Angabe“ eine -9 vergeben, müssen wir vor den Auswertungen diesen Wert noch als NA codieren.
- Wie verteilen sich die Variablen? Hier bietet sich bei metrischen Variablen ein Blick z. B. auf Mittelwert, Standardabweichung, Minimum und Maximum an, bei kategorialen Variablen sind Häufigkeitsauszählungen sinnvoll. Auch eine Visualisierung der Ergebnisse kann hilfreich sein.

2. Schritt – Zentrale Indikatoren für die Teilnahmequalität

Abbruchquote: Wie viele Teilnehmer:innen haben die Studie abgebrochen?

- Es ist möglich, dass Teilnehmer:innen zwar die Vorabbefragung ausfüllen, aber die In-situ-Phase gar nicht anfangen oder sehr früh abbrechen (also nur am ersten Tag teilnehmen oder nur wenige Protokolle ausfüllen). Um die Quote zu berechnen, vergleichen wir die Anzahl an Personen, die beabsichtigten, an der In-situ-Studie teilzunehmen (also die Vorbefragung ausgefüllt und ihr Einverständnis für die Teilnahme gegeben haben), mit denen, die die In-situ-Phase tatsächlich begonnen und/oder bis zu einem bestimmten Tag mitgemacht oder eine bestimmte Anzahl an Protokollen ausgefüllt haben.

- Uch die Teilnahme an einer Nachbefragung kann als (zusätzliches) Kriterium für den Studienabschluss gelten.

Compliance: Wie viele Protokolle haben die einzelnen Teilnehmer:innen ausgefüllt?

- Der Anteil ausgefüllter Protokolle in Bezug auf alle gesendeten Signale wird als Compliance bezeichnet. Die Compliance gilt als wesentliches Qualitätskriterium einer In-situ-Studie: Je höher die Compliance ist, desto unwahrscheinlicher ist es, dass die Situationsstichprobe verzerrt ist. Bei selbstausgelösten Protokollen ist sie kaum zu bestimmen.
- Je nach Struktur enthält der Datensatz auch die Protokolle, bei denen die Teilnehmer:innen zwar zur Befragung aufgefordert wurden, indem sie bei signalbasierten Designs ein Signal erhalten haben, das Protokoll aber nicht ausgefüllt haben. Dann können wir die Compliance aus dem Verhältnis von ausgefüllten zu nicht ausgefüllten Protokolle berechnen und entsprechende Muster anhand ausgefüllter bzw. nicht ausgefüllter Protokolle identifizieren. Zudem ist zu erkennen, ob die Signale wie geplant an die Teilnehmer:innen gesendet wurden, ob also der Stichprobenplan auf Situationsebene erfolgreich war oder ob es (technische) Ausfälle gab. Anschließend ist es sinnvoll, eine Filtervariable anzulegen, aus der hervorgeht, ob es sich bei einem Fall um ein ausgefülltes Protokoll handelt oder nicht (z. B. die Variable **abgeschlossen** mit **ja** für ein vollständiges Protokoll und **nein** für ein nicht ausgefülltes Protokoll, also ein nicht beantwortetes Signal). Enthält der Datensatz nur die ausgefüllten Protokolle, lässt sich die Compliance dennoch berechnen, wenn bekannt ist, wie viele Signale die Teilnehmer:innen erhalten haben. Diesen Wert können wir ins Verhältnis zu den von ihnen ausgefüllten Protokollen setzen.

Response Latency: Wie lang ist die Zeitspanne zwischen Alarmierung und Ausfüllen des Protokolls durchschnittlich?

- Die Response Latency kann nur bei signalbasierten In-situ-Studien bestimmt werden. Zudem hängt es von der verwendeten App oder dem verwendeten Tool ab, ob die notwendigen Informationen für die Berechnung vorliegen.

3. Schritt – Prüfung möglicher systematischer Verzerrungen und Umgang damit

Einzelne Situationen

Ausfülldauer: Gibt es einzelne Protokolle, die besonders schnell oder langsam beantwortet wurden?

- In Bezug auf In-situ-Studien ist die Frage umstritten, ob die Antwortzeit – wie in klassischen Befragungen – ein Qualitätskriterium ist (siehe Abschnitt 7.2.2). Dennoch kann es sinnvoll sein, dass wir uns anschauen, ob es auffällige Protokolle gibt.

Response Latency: Gibt es einzelne Protokolle, bei denen die Ausfüllverzögerung besonders lang ist?

- Mit einer langen Response Latency geht eine Verzerrung der Situationsstichprobe einher (siehe Abschnitt 6.3.2.2). Aus diesem Grund empfiehlt es sich, die Response Latency der einzelnen Protokolle zu analysieren und ggf. anschließend danach zu bereinigen. Häufig sind bereits das Design und die Programmierung der Studie darauf ausgelegt, dass Protokolle nur eine gewisse Zeit ab Alarmierung (z. B. zehn Minuten, eine halbe Stunde) ausgefüllt werden können. Anderenfalls kann die Response Latency nachträglich zur Bereinigung herangezogen und einzelne Protokolle mit einer langen Response Latency können aus der Analyse ausgeschlossen werden. Es gibt keine allgemeingültigen Grenzwerte, wichtig ist, die Entscheidung sinnvoll zu begründen und vorab festzulegen (siehe Abschnitt 10.2). Eine weitere Möglichkeit ist, die Response Latency als Kontrollvariable in den Auswertungen zu berücksichtigen (wie z. B. bei Naab et al. (2019), siehe Studiensteckbrief im Online-Anhang zum Buch).

Personenunterschiede

Studienabbruch: Unterscheiden sich die Personen, die die In-situ-Studie erfolgreich abgeschlossen haben, von den Abbrecher:innen?

- Wenn die Abbrecher:innen zumindest die Vorabbefragung ausgefüllt haben, können wir uns anhand der dort gegebenen Antworten einen Eindruck davon verschaffen, ob sie sich in relevanten Personenmerkmalen von den Teilnehmer:innen unterscheiden, die die Studie abgeschlossen haben. Diese Befunde können später bei der Ergebnisinterpretation berücksichtigt werden.

Ausfülldauer: Gibt es Personen mit besonders kurzen oder langen Antwortzeiten?

- Wir können uns die Antwortzeiten nicht nur für die einzelnen Protokolle auf der Situationsebene anschauen, sondern auch die Frage prüfen, ob es Personen gibt, die häufig besonders schnell oder langsam geantwortet haben. Eine schnelle Beantwortung könnte darauf hindeuten, dass sie die Gesamtstudie nicht ernst genommen und generell nicht sorgfältig teilgenommen haben (Careless Responding, siehe Abschnitt 7.2.2).

Response Latency: Gibt es Personen mit generell besonders langen Ausfüllverzögerungen?

- Analog zur Antwortzeit können wir untersuchen, ob es Personen gibt, die generell immer erst sehr lange nach der Alarmierung antworten, und auch dieses Verhalten als eine Form des Careless Responding werten.

Compliance: Gibt es Personen mit einer besonders niedrigen Compliance?

- Die individuelle Compliance pro Teilnehmer:in kann ebenfalls für die Datenbereinigung herangezogen werden. Es ist üblich, Teilnehmer:innen, die bestimmte Minimalanforderungen, d. h.

eine gewisse Mindestanzahl an ausgefüllten Protokollen, nicht erfüllen, aus der Analyse auszuschließen, weil davon auszugehen ist, dass ihre Situationsstichprobe zu verzerrt ist (Soyster et al., 2019). Es gibt keine festen Vorgaben, welcher Schwellenwert hier angemessen ist. Dies ist von Studie zu Studie abzuwägen. Ggf. ist es auch sinnvoll, dass wir uns lediglich anhand von Personenmerkmalen einen Eindruck verschaffen, welche Personen eine niedrige Compliance aufweisen, und diesen Befund kritisch bei der Interpretation der Ergebnisse berücksichtigen. Denn mit dem Ausschluss von Personen mit geringer Compliance geht wiederum eine Verzerrung der Personenstichprobe einher.

Zeitverläufe

Compliance: Gibt es zeitliche Muster in der Compliance?

- Wir können prüfen, ob die Compliance über die Feldzeit hinweg abnimmt. Dieser Befund ist i. d. R. gut dokumentiert (siehe Abschnitt 6.3.2.2). Ein moderates Sinken der Compliance ist also zu erwarten. Von Interesse sind zudem zeitliche Muster, wenn z. B. eine geringere Compliance am Wochenende im Vergleich zu den Werktagen oder Muster im Tagesverlauf vorliegen. Diese Muster können bei der späteren Dateninterpretation helfen (siehe z. B. Karnowski et al., 2019). Ermittelte Ergebnisse am Wochenende oder am Morgen können weniger verlässlich sein, wenn die Compliance niedrig ist.

Response Latency: Gibt es zeitliche Muster in der Ausfüllverzögerung?

- Zeitliche Muster, z. B. eine länger werdende Response Latency im Verlauf der Feldzeit, können als Rückgang der Motivation gedeutet werden und für eine sinkende Datenqualität sprechen (Reynolds et al., 2016). Ein Bewusstsein hierfür hilft bei der Interpretation der Daten.

Verteilung der Variablen: Gibt es zeitliche Muster im Hinblick

auf die Verteilung der Variablen?

- Die Verteilung der Variablen im Zeitverlauf kann Muster aufweisen, die auf **Reaktivität** hindeuten (siehe Abschnitt 7.2.1). Solche zeitlichen Muster werden getrennt nach Teilnehmer:innen analysiert, denn nur auf dieser Ebene ist Reaktivität erkennbar. Wir müssen hierbei beachten, dass die Ursachen für gefundene Muster sowohl Reaktivität als auch Veränderungen aus anderen Gründen sein können. Typische Muster und ihre mögliche Bedeutung sind:
 - Sinkende Werte bei Verhaltensangaben zeigen, dass entweder das Verhalten seltener ausgeführt oder das Antwortverhalten angepasst wurde.
 - Sinkende Werte bei Skalenfragen wie z. B. Stimmungsmessungen deuten eventuell auf Response Shifts hin, die durch einen *initial elevation bias* ausgelöst wurden (Reynolds et al., 2016; Shrout et al., 2018). Bei einem ausgeprägten Rückgang empfiehlt es sich eventuell, den ersten Tag auszuschließen. Zumindest ist eine Sensitivitätsanalyse empfehlenswert, indem der erste Tag einmal in die Auswertung einbezogen und einmal davon ausgeschlossen wird. Dies zeigt eventuelle Unterschiede in den Ergebnissen auf.
- Weiterhin können wir Muster über die Zeit identifizieren, die auf **mangelnde Sorgfalt** hinweisen. Auch in diesem Fall gilt: Es handelt sich lediglich um Hinweise auf Methodeneffekte, ursächlich können auch tatsächliche Veränderungen während der Feldzeit sein. Das typischste Muster ist eine über die Feldzeit geringer werdende Varianz in den Antworten. Diese kann als Hinweis gewertet werden, dass die Teilnehmer:innen zunehmend dieselbe Antwort ankreuzen (Fuller-Tyszkiewicz et al., 2013). Genauso kann eine über die Feldzeit größer werdende Varianz ein Indiz für mangelnde Sorgfalt sein und auf zufälliges Ankreuzen hindeuten. In diesem Fall bietet es sich an, Teilnehmer:innen zu identifizieren, bei denen diese Antworttendenzen

sehr ausgeprägt sind. In gut begründbaren Fällen können diese von der Auswertung ausgeschlossen werden.

Merke

In-situ-Daten sollten vor der Analyse überprüft werden hinsichtlich (1) grundlegender Verteilungen wie z. B. Fallzahlen, Ausfülldauer und Verteilung der Variablen, (2) zentraler Indikatoren der Teilnahmequalität wie Anzahl ausgefüllter Protokolle (Compliance) oder Ausfüllverzögerung (Response Latency) sowie (3) möglicher systematischer Verzerrungen auf der Situations- und Personenebene oder über die Zeit.

Falls die Datenprüfungen einen Grund zur Datenbereinigung liefern, müssen wir diese Entscheidungen gut abwägen und begründen. Eine geringe Compliance ist vor allem problematisch, wenn die Ausfälle systematisch sind. Zeitliche Antwortmuster können auch auf tatsächliche zeitliche Dynamiken zurückgehen, z. B. eine verringerte Mediennutzung infolge eines wärmeren Sommerwetters oder einer nachlassenden Aufmerksamkeit gegenüber politischen Themen, weil ein politisches Ereignis an Aktualität verliert. Idealerweise legen wir die geplanten Prüfkriterien und Bereinigungs-schritte schon vor der Datenerhebung oder Sichtung der Daten fest (siehe Abschnitt 10.2). Eine zu großzügige Bereinigung beeinträchtigt – wenn nicht vorher eingeplant – die Power einer Studie (siehe Abschnitt 6.1). Entscheidender ist aber, dass hierdurch systematische Verzerrungen erst ausgelöst werden können. Wenn wir z. B. Teilnehmer:innen ausschließen, die überwiegend dieselben Antworten geben, schließen wir entweder typische Durchklicker:innen aus, also Fälle von mangelnder Sorgfalt, oder wir entfernen Personen aus der Studie, die stabile Kognitionen, Emotionen oder Verhalten aufweisen, was durchaus möglich sein kann. Deshalb ist es sinnvoll, Fälle – Teilnehmer:innen oder einzelne Messzeitpunkte –, die wir ausschließen wollen, nicht aus dem Datensatz zu löschen, sondern entsprechend zu kennzeichnen und nur für die Auswertung auszuschließen. Anschließend können wir **Sensiti-**

vitätsanalysen einmal mit und einmal ohne die bereinigten Fälle durchführen, um Veränderungen in den Ergebnissen erkennen und die abweichenden Ergebnisse in die Interpretation einbeziehen zu können. Dieses Vorgehen ist aus forschungsethischer Sicht vorzuziehen, denn zwischen angemessener Datenbereinigung und illegitimer Datenfälschung verläuft ein schmaler Grat. Es ist daher zudem wichtig, das Vorgehen vollständig und transparent zu dokumentieren.

9.2.2 Datenaufbereitung

Nachdem wir uns grundlegend mit dem Datensatz vertraut gemacht und die Bereinigungsverfahren durchgeführt und protokolliert haben, geht es im nächsten Schritt um die Datenaufbereitung, also um die Vorbereitung der Daten für die eigentliche Auswertung. Im Folgenden befassen wir uns zunächst mit den Besonderheiten im In-situ-Kontext: mit der Reliabilitätsprüfung von Skalen und dem Umgang mit fehlenden Werten. Weitere spezifische Aufbereitungsschritte hängen von den genauen Analysen ab und werden in den dazugehörigen Abschnitten beschrieben (siehe Abschnitte 9.3 bis 9.5).

9.2.2.1 Reliabilitätsprüfung

Die Reliabilitätsprüfung, also die Frage, ob die Messung verlässlich ist, ist im In-situ-Kontext aus zwei Gründen herausfordernder als in herkömmlichen Befragungsstudien. Erstens besteht die Absicht, die Last für die Teilnehmer:innen so gering wie möglich zu halten. Aus diesem Grund greifen wir i. d. R. auf Single-Item-Messungen zurück (siehe Abschnitt 7.1.2). Die üblichen Verfahren der Reliabilitätsprüfung von Skalen, wie z. B. Omega (ω) oder Cronbachs Alpha (α) (siehe Abschnitt 9.2.2.1.3), sind unter diesen Bedingungen nicht möglich. Zweitens sind bei Skalenabfragen, also bei der Messung eines Konstrukts mit mehreren Items, die Verfahren zur Berechnung der Reliabilität im In-situ-Kontext noch nicht etabliert und statistisch komplexer (Stone et al., 2023; Trull & Ebner-Priemer, 2020). Im Folgenden gehen wir deshalb mehrstufig vor: Zunächst erklären

wir die grundlegende Problematik und Logik der Reliabilitätsprüfung im In-situ-Kontext. Anschließend gehen wir auf Möglichkeiten der Reliabilitätsprüfung von Single-Item-Messungen ein. Schließlich stellen wir State-of-the-Art-Verfahren zur Berechnung der Reliabilität von Skalen bei Mehrebenenendaten vor und zeigen zusätzlich alternative Wege auf, die zumindest Hinweise auf die Reliabilität von In-situ-Skalen geben können.

9.2.2.1.1 Grundlagen

Bei der Reliabilitätsprüfung geht es immer um das Verhältnis zwischen der Varianz in den Antworten, die inhaltlich bedeutsam ist, und der Varianz in den Antworten, die auf zufällige Messfehler zurückgeht. Systematische Messfehler sind dagegen eine Frage der Validität. Zwei Messungen sollen also unter gleichen Bedingungen auch das gleiche Ergebnis liefern.

Ein Beispiel soll zu einem besseren Verständnis betragen: Wenn wir eine Woche lang täglich unsere Körpergröße messen, werden die Werte immer ein klein wenig schwanken, je nachdem wie genau wir auf das Maßband schauen, es anlegen, wie aufrecht wir stehen etc. Aber die Werte verteilen sich alle um den wahren Wert herum, weil wir gleich groß bleiben. Jede Abweichung vom wahren Wert ist in diesem Fall also rein auf Messfehler zurückzuführen. Dieses Beispiel beschreibt die Retest-Reliabilität – wir führen denselben Test mehrfach durch und jedes Mal sollte sich (fast) dasselbe Ergebnis zeigen. Im Rahmen einer Studie misst aber in der Regel nicht nur eine Person die Körpergröße, sondern viele Menschen werden mit dieser Aufgabe betraut. Daraus ergeben sich zwei Quellen der Varianz: Erstens die inhaltlich bedeutsame Varianz – Personen sind unterschiedlich groß, sie variieren in ihrer wahren Größe. Zweitens die Fehlervarianz, also die Abweichungen in den Messungen, die darauf zurückgehen, dass nicht genau gemessen wurde, weil z. B. das Maßband falsch angelegt wurde.

Es gibt bei Personenvariablen einen einzigen wahren Wert pro Person, um den die einzelnen Messungen einer Person streuen. Alle

Abweichungen der einzelnen Messungen einer Person vom wahren Wert sind zufällige Messfehler. Eine inhaltlich bedeutsame Varianz gibt es nur zwischen Personen – die eine ist größer (wahrer Wert höher), die andere kleiner (wahrer Wert niedriger). Die Reliabilität ergibt sich dann aus dem Verhältnis dieser inhaltlich bedeutsamen Varianz zur Gesamtvarianz (inhaltlich bedeutsame Varianz + Fehlervarianz): Je geringer die Fehlervarianz ist, desto reliabler ist die Messung.

Nun fällt aber auf, dass sich diese Logik nicht nahtlos auf Variablen aus In-situ-Studien übertragen lässt. Denn im In-situ-Kontext sind Merkmale relevant, die nicht stabil sind, also nicht nur zwischen, sondern auch innerhalb einer Person im Zeitverlauf variieren. Wir messen nicht die Größe, die weitgehend stabil bleibt, sondern zum Beispiel die Stimmung, die im Tages- und Wochenverlauf schwankt. Diese (valide) Dynamik ist nicht ohne Weiteres von Messfehlern zu unterscheiden (vgl. auch Schuurman & Hamaker, 2019; Stone et al., 2023). Das Grundprinzip bleibt in diesem Fall zwar zunächst gleich: Wir interessieren uns für das Verhältnis zwischen der inhaltlich bedeutsamen Varianz und der Gesamtvarianz. Bei Merkmalen, die innerhalb einer Person variieren, ist die Varianz innerhalb einer Person aber nicht nur die Fehlervarianz, sondern enthält auch den inhaltlich bedeutsamen Teil, d. h., die Stimmung einer Person ist jetzt anders als vor drei Stunden. Entscheidend für die Reliabilitätsprüfung ist es daher, zwischen den Schwankungen in den Antworten zu unterscheiden, die eine echte Veränderung zeigen, und denen, die nur auf den Messfehler zurückgehen. Es geht darum, wie gut unser Messinstrument eine echte Veränderung auf der Situationsebene messen kann. Nur wenn eine Messung auf der Within-Person-Ebene reliabel misst – also ein hinreichend großer Teil der Varianz inhaltlich bedeutsam ist –, können wir davon ausgehen, dass Veränderungen in der Variable zwischen den Messzeitpunkten auch tatsächliche Veränderungen und keine Messfehler sind (Cranford et al., 2006; Neubauer & Schmiedek, 2020). Nur wenn wir z. B. die Stimmung reliabel messen, können wir Stimmungsschwankungen einer Person auch als solche interpretieren und nicht als zufällige Abweichungen durch eine

ungenauere Messung.

Merke

Reliabilität bedeutet, dass eine Messung konsistent und verlässlich ist – also möglichst wenig von zufälligen Messfehlern beeinflusst wird. Bei In-situ-Studien ist es entscheidend, tatsächlich bestehende intraindividuelle Schwankungen von Messfehlern abzugrenzen.

Eine weitere Besonderheit von In-situ-Messungen besteht darin, dass möglichst kurz sein sollten, weswegen häufig Single-Item-Messungen oder sehr kurze Skalen eingesetzt werden (siehe Abschnitt 7.1.2). Im Folgenden gehen wir daher auf die Reliabilitätsprüfung sowohl bei Single-Item-Messungen als auch bei Skalen ein.

9.2.2.1.2 Single-Item-Messungen

Reliabilitätsprüfungen sind bei Single-Item-Messungen bisher kaum etabliert. Basierend auf einer Variante der klassischen Testtheorie (siehe Abschnitt 9.2.2.1.3) befassen sich Dejonckheere et al. (2022) mit der Möglichkeit, über Test-Retest-Verfahren die Reliabilität von Single-Item-Messungen zu bestimmen, und nutzen Mehrebenenmodelle (siehe Abschnitt 9.5.2). Sie schlagen vor, entweder gelegentlich – also nicht in jedem Protokoll – die Single-Item-Messung im selben Protokoll zu wiederholen. So wird sichergestellt, dass der Retest, d. h. die Wiederholung des Items, unter denselben situativen Bedingungen wie der Test stattfindet. Alternativ schlagen sie vor, gelegentlich sehr kurze Zeitabstände (max. 20 min) zwischen den Protokollen zu lassen – unter der Annahme, dass sich das Konstrukt (z. B. die Stimmung) nicht so schnell ändert und auch dies als valider Retest gesehen werden kann. Beide Verfahren zeigten bei der Messung der Stimmung ähnliche Ergebnisse. In Abwägung der Last für die Teilnehmer:innen und dem Erkenntnisgewinn zur Reliabilität der verwendeten Single-

Item-Messungen muss entschieden werden, ob die Verfahren im Rahmen einer Studie möglich sind.

Exkurs: Vektorautoregressive Modelle

Schuurman und Hamaker (2019) stellen eine Prüfmöglichkeit vor, die allerdings auf dynamische VAR(1)-Modelle (= Vektorautoregressive Modelle) beschränkt ist (siehe hierzu Exkurs im Abschnitt 9.5.3).

9.2.2.1.3 Skalen

Außer mit Single-Item-Messungen werden Konstrukte in In-situ-Befragungen teilweise anhand mehrerer Items einer Skala gemessen. Für die weitere Auswertung müssen diese Einzelwerte zusammengefasst werden. Das Vorgehen entspricht herkömmlichen Befragungsstudien; es können z. B. Summen- oder Mittelwertindizes gebildet werden, sodass für jede Person und jeden Messzeitpunkt ein Indexwert entsteht. Gehen wir davon aus, dass die Items ein dahinterliegendes gemeinsames Konstrukt abbilden (z. B. narratives Transportationserleben), dann empfiehlt sich die Prüfung, ob diese Annahme stimmt. Hierfür müssen wir die Reliabilität der Skala berechnen. Üblich ist hier die Berechnung der internen Konsistenz als Reliabilitätsmaß. Diese funktioniert grundsätzlich nach dem gleichen Prinzip wie die Retest-Reliabilität: Die einzelnen Items bilden die wiederholten Messungen desselben Merkmals für eine Person ab. Ihre gemeinsame Varianz – die Konsistenz, mit der sie beantwortet werden – im Vergleich zur Gesamtvarianz gibt an, wie gut die Skala das Konstrukt misst. Daraus lässt sich die Schlussfolgerung ableiten: Wenn alle Items perfekt korrelieren, d. h., eine Person bei allen Items denselben Wert angibt, dann liegt keine Fehlervarianz vor, weil die Messung reliabel ist.

Die Berechnung der internen Konsistenz ist bei In-situ-Daten jedoch schwieriger: Die Skala wird nicht nur einmal pro Person erhoben, sondern mehrfach. Dadurch entstehen zwei Ebenen, auf de-

nen wir Reliabilität prüfen können oder genauer ausgedrückt die wir bei der Berechnung berücksichtigen müssen: die Personenebene (Between-Person-Reliabilität) und die Situationsebene (Within-Person-Reliabilität). Dieses Vorgehen basiert auf der in Kapitel 2 beschriebenen Grundlogik von In-situ-Daten, die die Möglichkeit eröffnet, die Personen- von der Situationsvarianz zu trennen.

Die Within-Person-Reliabilität gibt an, wie zuverlässig eine Messung den intraindividuell schwankenden Anteil des Konstrukts messen kann. Die Between-Person-Reliabilität bemisst hingegen den stabilen Anteil eines Konstruktes. Viele Merkmale haben nicht nur einen situativen Charakter, sondern auch eine Komponente, die stabil ist – diesen Teil bildet die Varianz zwischen Personen ab. Die Between-Person-Reliabilität gibt also an, wie gut Unterschiede (= Varianz) zwischen den Personen über alle Messzeitpunkte hinweg mit der Messung erfasst werden können, wie konsistent also die Personenmittelwerte sind. Welche der beiden Reliabilitäten für die eigene Forschung relevanter ist, hängt von der Forschungsfrage ab. Bei Forschungsfragen auf der Personenebene (siehe Abschnitt 3.1) ist die Between-Person-Reliabilität und bei Forschungsfragen auf der Situationsebene die Within-Person-Reliabilität zentral (siehe Abschnitt 3.2).

Es gibt bisher keinen Standard, wie sich die Reliabilität von Skalen bei In-situ-Daten bestimmen lässt, aber verschiedene Ansätze, die zumindest grundlegende Ähnlichkeiten aufweisen: Sie betrachten alle die interne Konsistenz, basieren auf der Zerlegung der Varianz in verschiedene Komponenten und trennen zwischen Between- und Within-Person-Reliabilität oder genauer gesagt kontrollieren für die Abhängigkeit der Daten durch ihre Mehrebenenstruktur (im Überblick z. B. auch Alphen et al., 2022; Bolger & Laurenceau, 2013; Castro-Alvarez et al., 2026; Eisele, Vachon, et al., 2022; Trull & Ebner-Priemer, 2020). Sie unterscheiden sich allerdings im Hinblick auf ihren messtheoretischen Zugang. So beschreibt z. B. Nezlek (2017) eine Vorgehensweise, mit der auf Basis des Standardoutputs von Mehrebenenanalysen (Residualvarianzen) Reliabilität berechnet wird (zu Mehrebenenanalysen siehe Abschnitt 9.5.2, zur Kritik an

diesem Ansatz siehe z. B. Eisele, Vachon, et al., 2022). Darüber hinaus finden sich in der Literatur vor allem zwei Zugänge: die Generalisierbarkeitstheorie als Erweiterung der klassischen Testtheorie (Bolger & Laurenceau, 2013; Cranford et al., 2006; Revelle & Wilt, 2019; Shrout & Lane, 2012) und die konfirmatorische Mehrebenenfaktorenanalyse (z. B. Alphen et al., 2022; Geldhof et al., 2014; Lai, 2021).

Die **Generalisierbarkeitstheorie** unterteilt die Gesamtvarianz in die drei Komponenten Zeit, Item (beides Situationsebene) und Person, mit denen verschiedene Between- und Within-Person-Reliabilitätsmaße bestimmt werden können. Weitere hilfreiche Details zu diesem Verfahren finden sich bei Bolger und Laurenceau (2013) und Cranford et al. (2006). In R bietet das weitverbreitete Paket `psych` eine Funktion zur Berechnung der Reliabilitäten auf Basis der Generalisierbarkeitstheorie an (Revelle, 2024). Ihre Schwäche liegt in den relativ strengen Anwendungsvoraussetzungen, die selten erfüllt sind. Es müssen beispielsweise gleiche Ladungen aller Items auf den Faktor und gleiche Fehlervarianzen vorliegen, d. h., alle Items müssen gleich valide und reliabel sein.

Die **konfirmatorische Mehrebenenfaktorenanalyse** (*multi-level confirmatory factor analysis*, MCFA) umgeht diese strengen Voraussetzungen. Eine solche Analyse prüft, ob oder wie gut mehrere Items ein gemeinsames, zugrunde liegendes latentes Konstrukt (= Faktor) messen, indem bei einer MCFA die Mehrebenenstruktur der Daten berücksichtigt wird. MCFAs sind statistisch anspruchsvoll. Weiterführende Informationen finden sich bei Geldhof et al. (2014). Aber auch ohne Detailkenntnisse zu MCFAs gibt es nutzungsfreundliche Möglichkeiten, diese für die Prüfung der Reliabilität von In-situ-Messungen einzusetzen, etwa in Form von R-Paketen. Eine Sammlung von entsprechenden Paketen findet sich im Online-Anhang zum Buch unter Ressourcen.

Auf Basis von MCFAs können ebenenspezifische Reliabilitätskennzahlen für die interne Konsistenz berechnet werden. Am bekanntesten ist Cronbachs Alpha (Cronbach, 1951). Cronbachs Alpha hat allerdings ebenfalls strenge Voraussetzungen, die selten erfüllt sind

(vor allem gleiche Ladungen aller Items auf den Faktor). Aus diesem Grund empfehlen wir Omega (McDonald, 1999), das diese Voraussetzung nicht hat (siehe hierzu auch Goodboy & Martin, 2020; Peters, 2018). Die Werte für Cronbachs Alpha und Omega reichen von 0 bis 1, wobei näher an 1 liegende Werte auf eine bessere Reliabilität der Skala hinweisen. Cronbachs Alpha und Omega können als der Anteil der Varianz in den Skalenwerten interpretiert werden, der auf das Konstrukt, das allen Items zugrunde liegt, zurückzuführen ist. Je höher der Wert ist, desto kleiner ist die Fehlervarianz und desto reliabler die Messung (Cronbach, 1951; McDonald, 1999).

Welche Reliabilitätswerte als ausreichend, gut, ungenügend etc. angesehen werden können, ist nicht generell zu beantworten (Alphen et al., 2022; Nezlek, 2017). Bereits die häufig referierte Grenze von 0.7 ist nicht ohne Kritik (siehe z. B. Field, 2026). Zudem verschärft sich das Problem, wenn in In-situ-Protokollen nur sehr kurze Skalen eingesetzt werden. In diesem Fall ist es unwahrscheinlicher, einen hohen Reliabilitätswert zu erreichen, weil klassische Reliabilitätsmaße wie Cronbachs Alpha sensibel für die Anzahl der Items sind (Calamia, 2019; Nezlek, 2017). Wichtig ist eine Abwägung zwischen der Last für die Teilnehmer:innen und der Reliabilität.

Merke

Im Hinblick auf die Reliabilität von In-situ-Daten unterscheiden wir zwischen der Between-Person-Reliabilität und der Within-Person-Reliabilität. Für die Berechnung bieten sich vor allem konfirmatorische Mehrebenenfaktorenanalysen an.

Exkurs: Fortgeschrittene Reliabilitätstests

Schließlich sei noch auf zwei weitere in diesem Zusammenhang sinnvolle, aber anspruchsvolle Vorgehensweisen hingewiesen: Erstens gibt es Ansätze, die Reliabilität personenspezifisch,

d. h. für jede Person einzeln zu bestimmen (Hu et al., 2016). Der Vorteil dieses Vorgehens ist, dass wir Personen identifizieren können, die sehr niedrige Reliabilitäten aufweisen, und diese Information für die Datenbereinigung oder als Kontrollvariable in der Auswertung berücksichtigen können. Zweitens gibt es erweiterte Prüfmöglichkeiten für Faktorstrukturen, also Fragen zu den Konstrukten, die den gemessenen Items zugrunde liegen: Wie viele sind es? Sind es auf der Personen- und Situationsebene gleich viele Items? Sind die Strukturen zeitlich invariant, bleibt ihre Bedeutung über die verschiedenen Messzeitpunkte hinweg also gleich (Neubauer et al., 2020; Vogelsmeier et al., 2022; Wright et al., 2015)?

Die vorangegangenen Ausführungen zeigen zur Bestimmung der Reliabilität im In-situ-Kontext Best-Practice-Wege auf, die die geschachtelte Datenstruktur berücksichtigen. Sie sind allerdings an vielen Stellen sehr voraussetzungsreich, noch nicht allen Forschenden vertraut und werden in der KMW (bisher) selten eingesetzt und berichtet. In vielen Fällen wird – wenn Reliabilitätswerte berichtet werden – auf Cronbachs Alpha in seiner üblichen, nicht die Mehrebenenstruktur berücksichtigenden Variante zurückgegriffen. Im Folgenden stellen wir die von Forschenden überwiegend verwendeten Vorgehensweisen vor und gehen kurz auf ihre Vor- und Nachteile ein (siehe auch Calamia, 2019; Nezlek, 2017). Ein entscheidender Nachteil ist, dass alle alternativen Optionen gerade bei kleineren Fallzahlen auf der Situationsebene nicht immer für jede Person oder jeden Messzeitpunkt berechenbar sind. Es kommt durchaus vor, dass einzelne Variablen keine Varianz aufweisen und die Berechnung von Cronbachs Alpha deshalb fehlschlägt oder negative Werte auftreten, die nicht im Wertebereich von Cronbachs Alpha liegen.

- *Option 1 (nicht empfehlenswert): Eine Berechnung der Reliabilität über alle Personen und Messzeitpunkte hinweg ist möglich, wenn die Daten im Long-Format vorliegen (siehe Abschnitt*

9.1). Diese Variante vermischt Unterschiede zwischen Personen mit Unterschieden innerhalb von Personen, die nötige Unabhängigkeit der einzelnen Fälle ist nicht gegeben. Jede Person steuert so viele Antworten pro Item bei, wie sie Protokolle ausgefüllt hat. Durch die häufig sehr ähnlichen Antwortstile von Personen wird die Reliabilität sehr wahrscheinlich überschätzt. Die Ergebnisse können nicht sinnvoll interpretiert werden; diese Vorgehensweise sollte daher nicht angewendet werden.

- *Option 2: Die Berechnung der Reliabilität findet für jede Person getrennt, aber über alle Messzeitpunkte hinweg statt* und anschließend folgt z. B. eine Mittelung der Reliabilitätswerte. Im Vergleich zu Option 1 findet keine Vermischung von Unterschieden zwischen und innerhalb von Personen statt. Die Mehrebenenstruktur wird durch die Trennung der Daten nach Personen aufgelöst. Entgegen der in einigen Quellen gegen diese Methode als Kritik vorgebrachte Annahme, dass die Mehrebenenstruktur bei dieser Berechnung ignoriert würde, kann dieses Vorgehen als vertretbare Annäherung angesehen werden, das zumindest einen deskriptiven Wert hat. Denn letztlich wird jede Person getrennt für sich modelliert, jedes In-situ-Protokoll ist eine Ziehung aus den Situationen einer Person. Die gesamte Varianz in den Werten ist intraindividuell, die Reliabilität bezieht sich lediglich auf die Within-Person-Komponente. Eine Voraussetzung ist jedoch, dass nicht zusätzlich von Zeiteffekten und Autokorrelationen ausgegangen wird.
- *Option 3: Für die Berechnung der Reliabilität für nur einen Messzeitpunkt* wird ähnlich wie bei Option 2 die Abhängigkeit der Daten aufgehoben, indem die Messzeitpunkte getrennt betrachtet werden und jede Person pro Messzeitpunkt nur einmal in die Reliabilitätsberechnung einbezogen wird. Ein mögliches Vorgehen ist, z. B. nur den ersten und den letzten Messzeitpunkt auszuwählen sowie ggf. einen Messzeitpunkt in der Mitte der Feldzeit. Indem wir alle zwei oder drei Messzeitpunkte berichten und vergleichen, erhalten wir einen Eindruck über die

Reliabilität. Es ist auch möglich, die Reliabilitätsberechnung für jeden Messzeitpunkt durchzuführen und anschließend alle Reliabilitätswerte zu mitteln, um nur eine Kennzahl zu erhalten und berichten zu können. Dieses Vorgehen ist jedoch nicht fehlerfrei, weil es die Logik der Auswahl der Messzeitpunkte nicht berücksichtigt (Nezlek, 2017) und nicht immer die Information vorhanden ist, um welchen Messzeitpunkt es sich handelt. Wir wissen z. B. je nach verwendeter App oder Software bei einem signalbasierten Design nur, um das wievielte ausgefüllte Protokoll einer Person es sich handelt. Diese Information bildet nicht zwingend den Messzeitpunkt ab, weil nicht immer alle Signale zu ausgefüllten Protokollen führen. Das Vorgehen kann dennoch als vertretbare Annäherung gesehen werden und hat somit einen deskriptiven Wert. Es entspricht am ehesten einer Analyse der Between-Person-Reliabilität.

- *Option 4: Bei Berechnung der Reliabilität für die aggregierten Mittelwerte* werden zuerst für jedes Item die Personenmittelwerte über alle Messzeitpunkte hinweg berechnet und diese neuen Variablen für die Reliabilitätsprüfung herangezogen. Dieses Vorgehen hat den Vorteil, dass die Unabhängigkeit der Daten nicht ignoriert wird, sodass jede Person nur einmal in die Berechnung einbezogen wird. Es kann als vertretbare Annäherung an die Between-Person-Reliabilität gesehen werden und hat somit deskriptiven Wert.

9.2.2.2 Umgang mit fehlenden Werten

In Abschnitt 9.2.1 haben wir uns damit befasst, ob es aufgrund nicht ausgefüllter Protokolle zum Ausschluss von Personen kommen sollte. Aber auch die Protokolle der Personen, die in der finalen Stichprobe enthalten sind, können fehlende Werte (Missings) aufweisen. Zum einen gibt es kaum Personen, die alle Protokolle ausfüllen. Somit fehlen für einige Messzeitpunkte alle Werte einzelner Personen (Unit Nonresponse). Zum anderen kommt es vor, dass innerhalb eines Pro-

tokolls einzelne Fragen nicht beantwortet werden (Item Nonresponse). Es hängt vor allem von der Auswertungsmethode ab, wie wir mit fehlenden Werten umgehen und ob sie problematisch sind. Viele gängige Auswertungsverfahren, z. B. Mehrebenenanalysen, können mit fehlenden Werten umgehen (siehe Abschnitt 9.5.2). Bei anderen Auswertungsverfahren können fehlende Werte hingegen zu deutlichen Einbrüchen der Fallzahlen führen, sodass wir über Möglichkeiten der **Imputation** oder konkret gesagt des modellbasierten Umgangs mit fehlenden Werten nachdenken müssen. Dies ist weitgehend unproblematisch für rein zufällig fehlende Werte (MCAR, siehe Abschnitt 6.3.2.2), möglich für Missing at Random (MAR) und nicht ohne Weiteres möglich für systematische Ausfälle (NMAR). Deshalb ist es notwendig, dass die Ausfallmuster identifiziert (siehe auch Abschnitt 9.2.1) und bei der Dateninterpretation berücksichtigt werden, um ggf. die passende Methode im Umgang mit fehlenden Werten zu wählen.

Zunächst stellt sich die Frage, ob wir überhaupt imputieren, also fehlende Werte durch plausible Schätzungen ersetzen sollten. Daran schließt sich die Frage an, wie wir imputieren können. Bisher gibt es keinen Konsens unter In-situ-Forschenden, wie (und ob überhaupt) Imputation bei In-situ-Daten sinnvoll und möglich ist. In-situ-Daten zielen auf die Messung von intraindividuell variablen und sich im Alltag stetig verändernden Faktoren ab. Daher erscheint es kaum möglich, fehlende Werte aus vorherigen oder späteren Antworten zu schätzen. Dennoch gibt es hierzu einige Ansätze, die in der In-situ-Forschung allerdings noch nicht etabliert sind. Eine Auseinandersetzung mit Möglichkeiten und Grenzen der Imputation für In-situ-Daten und Mehrebenenstrukturen findet sich im Überblick bei Fritz et al. (2024) und Stone et al. (2023) sowie z. B. bei Cursio et al. (2019), Grund et al. (2018), Hox et al. (2017), Ji et al. (2018), Lin et al. (2018) und Slipetz et al. (2023).

9.3 Grundlegende Auswertungen

Im Methoden- und ggf. im Ergebnisteil einer In-situ-Publikation werden üblicherweise – wie bei allen empirischen Publikationen – grundlegende Ergebnisse auf Basis des finalen, bereinigten Datensatzes vorgestellt. Dazu gehören Informationen über die Stichprobenzusammensetzung (z. B. Alter, Geschlecht und Bildungsgrad), die Fallzahlen auf der Personen- und Situationsebene sowie die Abbruchquote und die Compliance. Üblicherweise berichten wir auch die Deskriptivangaben zu den zentralen Variablen und bei Skalen die Reliabilitätswerte (siehe Abschnitt 9.2.2.1.3). Weiterhin ist es sinnvoll, die bivariaten Korrelationen aller (metrischen) Variablen zu berichten. Bei einer Mehrebenenstruktur unterscheiden wir zwischen den intraindividuellen (*within-person*) und den interindividuellen (*between-person*) Korrelationen (Trull & Ebner-Priemer, 2020). Auch wenn wir selbst nur Forschungsfragen auf einer der beiden Ebenen haben, sind diese Informationen z. B. für Metaanalysen wichtig und geben generell einen ersten Überblick über die Datenstruktur und Zusammenhänge. Ggf. eignet sich für den Bericht ein Anhang (siehe auch Abschnitt 10.3). .

9.4 Unterschiede zwischen Personen: Fragen nach interindividueller Varianz

Bei Forschungsfragen auf der Personenebene nutzen wir die In-situ-Daten vor allem für eine valide Erfassung von Informationen (z. B. Nutzungshäufigkeit und -dauer, Naab et al. (2019), siehe Studiensteckbrief im Online-Anhang zum Buch). Dafür müssen wir die In-situ-Daten vor der Auswertung aggregieren, d. h., die Messwerte einer Person über mehrere Situationen zu einem Wert für die Person zusammenfassen. Es hängt von der Forschungsfrage ab, welche Variablen wir wie aggregieren (siehe Abschnitt 3.1). Typische **Aggregationsvarianten** sind :

- Aufsummieren von Variablen über die Feldzeit, z. B. Anzahl

der True-Crime-Nutzungsepisoden

- Berechnung der Durchschnittswerte bei metrischen Variablen, z. B. der durchschnittlichen Stimmung (*person mean*, siehe auch Abschnitt 9.5.2.3)
- Bestimmung der intraindividuellen Varianz, z. B. um zu prüfen, ob es Personen gibt, die überwiegend in der gleichen Stimmung oder erheblichen Stimmungsschwankungen ausgesetzt sind, ob Personen dieselben Gratifikationen suchen oder sich dieses Verhalten je nach Situation unterscheidet oder ob Nutzungsepisoden häufig gleich lang oder jedes Mal unterschiedlich lang sind

Bei allen drei Varianten ist es möglich, weitere Informationen einzubeziehen. Wir können nicht nur einen Wert für die Anzahl der Podcastnutzungsepisoden pro Person berechnen, sondern – wenn es das Forschungsinteresse erfordert – einen Wert für die Nutzung zu Hause und einen zweiten Wert für die Nutzung unterwegs, wenn der Ort auf der Situationsebene erhoben wurde.

Nun können wir mit der eigentlichen Auswertung beginnen. Auf Basis der aggregierten Werte können wir z. B. die Frage beantworten, wie oft, wie lange und wie aufmerksam die Teilnehmer:innen im Durchschnitt Podcasts hören oder wie sie durchschnittlich gelaunt sind. Im Hinblick auf die interindividuelle Varianz können wir zudem feststellen, ob manche Personen häufiger über Stimmungsschwankungen berichten als andere.

Die Einteilung der Nutzer:innen von Podcasts in **Klassen/Typen** – also eine Gruppierung nach ihrer Ähnlichkeit – ist auf der Personenebene möglich. Beispielsweise können in einer ereignisabhängigen In-situ-Studie jeweils der zeitliche, räumliche und soziale Kontext der Podcastnutzung und genutzte Podcastinhalte erfragt, auf der Personenebene aggregiert und anhand einer latenten Klassenanalyse oder einer Clusteranalyse verschiedene Nutzer:innentypen unterschieden werden. Beide Verfahren sind nicht spezifisch für In-situ-Studien und werden daher nicht näher

vorgestellt (siehe z. B. Hagenaars & Mccutcheon, 2003; Hennig et al., 2015; Vermunt & Magidson, 2004). Abschnitt 9.5.3 befasst sich ebenfalls mit der Klassen-/Typenbildung von Personen, bezieht aber die Möglichkeit mit ein, gleichzeitig Situationen einzuteilen und somit die Datenstruktur von In-situ-Studien zu nutzen.

Anhand der aggregierten Werte können wir auch **Unterschiede und Zusammenhänge** anhand der aus der klassischen Querschnittsbefragung bekannten Analysemethoden berechnen, die hier nicht im Detail vorgestellt werden, weil sie nicht spezifisch für In-situ-Studien sind (siehe z. B. Field, 2026; Gehrau et al., 2022; Sauer, 2019; Wickham et al., 2023). Für die Identifizierung von Mittelwertunterschieden stellt sich z. B. die Frage, ob sich Berufstätige von Nichtberufstätigen in ihrer Podcastnutzung unterscheiden (*t*-Test oder Varianzanalyse) oder ob ein Zusammenhang zwischen Persönlichkeitsmerkmalen wie der Transportabilität und der Nutzungsdauer besteht (Korrelation oder Regression). Den Einfluss mehrerer Variablen können wir mit einer multiplen Regression berechnen, indem wir z. B. den Einfluss von Transportabilität und Berufstätigkeit auf die durchschnittliche Nutzungsdauer untersuchen.

Insgesamt können sowohl Personenmerkmale, die im Rahmen einer Vorab- oder Nachbefragung erhoben wurden, als auch verschiedene in situ erhobene und dann auf Personenebene aggregierte Variablen in Zusammenhang gesetzt werden.

Die vorgestellten Auswertungsverfahren, die auf den aggregierten In-situ-Variablen beruhen, sind der einfachste Weg, Forschungsfragen zu beantworten, die sich ausschließlich auf die Personenebene beziehen. Auch Auswertungen auf der Situationsebene bieten die Möglichkeit, Forschungsfragen auf der Personenebene (mit) zu beantworten. Bei Forschungsfragen auf beiden Ebenen empfiehlt sich auf jeden Fall das im Folgenden beschriebene Vorgehen.

9.5 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

Die große Stärke von In-situ-Designs besteht darin, Fragen zur intraindividuellen Varianz zu beantworten. Zunächst gehen wir auf deskriptive Auswertungen auf der Situationsebene ein. Anschließend stellen wir die Grundlagen von Unterschieds- und Zusammenhangsauswertungen auf der Situationsebene vor, denn aufgrund der besonderen Datenstruktur benötigen wir andere statistische Verfahren als bei üblichen querschnittlichen Befragungen. Der Schwerpunkt liegt auf der häufigsten Auswertungsmethode, die zur Analyse von In-situ-Daten eingesetzt wird: der **Mehrebenenmodellierung**. Die grundlegende Logik von Mehrebenenmodellen lässt sich auf viele andere Auswertungsmethoden im Kontext von In-situ-Daten übertragen. Dies gilt sowohl für die in Abschnitt 9.5.3 vorgestellte Methode zur Einteilung in Klassen/Typen als auch für weitere (fortgeschrittene) Auswertungsmethoden (siehe Exkurs in Abschnitt 9.5.3).

9.5.1 Deskriptive Auswertungen

Auf der Situationsebene können wir uns für deskriptive Auswertungen interessieren. So können wir per Häufigkeitsauszählung auswerten, wie oft bestimmte Kontexte, z. B. bestimmte Orte oder Tätigkeiten, vorkommen. Von Interesse kann auch die situative Variation von Konstrukten sein. Für jede:n Teilnehmer:in können wir etwa explorieren, wie die Nutzungsdauer einzelner Podcastnutzungsepisoden im Tagesverlauf variiert, um festzustellen, ob sie sich gleichmäßig über den Tag verteilt oder zu bestimmten Tages- oder Wochenzeiten verstärkt Podcasts gehört werden (zu weiteren Auswertungsmöglichkeiten zeitlicher Dynamiken siehe Abschnitt 9.5.2.7).

In vielen Fällen interessieren wir uns (zusätzlich) für die Zusammenhänge zwischen den Konstrukten. Hierfür bieten sich vor allem mehrebenenanalytische Verfahren an, die wir im Folgenden vorstellen.

Exkurs: Fixed-Effects-Modelle als Alternative zu Mehrebenenanalysen

Liegt der Schwerpunkt ausschließlich auf Within-Person-Zusammenhängen, können wir auch sogenannte Fixed-Effects-Modelle schätzen, wie sie z. B. in der Politikwissenschaft und den Wirtschaftswissenschaften für Paneldaten üblich sind. Es gibt zwei Ansätze: Das sog. De-Meaning, das der Zentrierung am Personenmittelwert entspricht (siehe Abschnitt 9.5.2.3) und das *Least Squares Dummy Variable Model* (LSDVM), bei dem die Personen-ID als Kovariate ins Modell aufgenommen wird. Damit werden automatisch alle beobachteten und unbeobachteten Unterschiede zwischen Personen herausgerechnet, um den reinen Within-Person-Zusammenhang zu berechnen. Allerdings haben diese Modelle den Nachteil, dass wir keine Personenvariablen mehr hinzufügen können, weil diese Varianz schon herausgerechnet worden ist (siehe Abschnitt 9.5.2.4 und 9.5.2.6). Weiterführende Hinweise zu Fixed-Effects-Modellen finden sich z. B. bei Allison (2009).

9.5.2 Mehrebenenanalysen zur Auswertung von In-situ-Daten

9.5.2.1 Grundlagen

Die Daten aus In-situ-Studien haben eine Mehrebenenstruktur. Die einzelnen Messzeitpunkte sind nicht unabhängig voneinander, sondern in Personen geschachtelt, d. h., mehrere Messzeitpunkte stammen von derselben Person (siehe Kapitel 2 und Abschnitt 9.1). Diese Datenstruktur verletzt die Annahme vieler gängiger statistischer Auswertungsverfahren, nach der Messungen unabhängig voneinander sein müssen. Damit gehen falsche Parameterschätzungen und unzuverlässige Ergebnisse einher (siehe z. B. Gelman & Hill, 2006; Langer, 2009). Ignorieren wir, dass die Daten von denselben Personen

stammen, dann kann es zu einer Variante des sog. ökologischen Fehlschlusses kommen (z. B. im Extremfall zum Simpson-Paradoxon): Zusammenhänge, die zwischen Personen gefunden werden, gelten nicht zwingend auch innerhalb von Personen. Abbildung 9.2 verdeutlicht diesen Zusammenhang an einem fiktiven Beispiel. Die orangene Linie auf der linken Seite bildet die Regressionsgerade ab, die wir schätzen würden, wenn wir nicht beachten, dass die Messungen in Personen geschachtelt sind. Alle Messzeitpunkte haben die gleiche Farbe, unabhängig davon, dass sie von verschiedenen Personen stammen. Es sieht nach einem positiven Zusammenhang zwischen Stimmung und Mediennutzung aus. Beachten wir aber, dass die Messungen in Personen geschachtelt sind – nun haben alle zur selben Person gehörenden Messungen eine eigene Farbe – dann erkennen wir, dass dieser Eindruck falsch ist. Während im Aggregat also der Eindruck entsteht, dass eine bessere Stimmung mit höherer Mediennutzung einhergeht, gilt auf der Ebene einzelner Personen tatsächlich, dass eine bessere Stimmung mit weniger Mediennutzung einhergeht. Jede farbige Linie steht für die Regressionsgerade einer einzelnen Person.

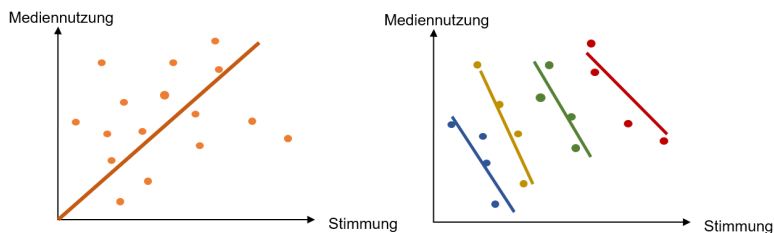


Abbildung 9.2: Fiktives Beispiel eines Simpson-Paradoxons

Merke

Die Daten aus In-situ-Studien weisen eine Mehrebenenstruktur auf, weil mehrere Messzeitpunkte von denselben Personen stammen. Diese Bedingung verletzt die Annahme der Unabhängigkeit der Messwerte voneinander, die für herkömmliche statistische Verfahren gilt und deren Verletzung dort zu verzerrten Ergebnissen führen kann.

Zudem besteht eine Besonderheit von In-situ-Daten darin, dass selten alle Teilnehmer:innen dieselbe Anzahl an Messzeitpunkten beisteuern – aus unterschiedlichen Gründen, von Non-Compliance bis hin zu inhaltlich relevanten Gründen. Diese Aussage gilt z. B. für ereignisabhängige In-situ-Designs, wenn das Ereignis bei mehreren Personen unterschiedlich oft auftritt (siehe Abschnitt 6.3.1.4). Gängige Auswertungsverfahren für wiederholte Messungen (z. B. Varianzanalyse mit Messwiederholung) können damit nicht umgehen.

Mehrebenenmodelle (in der Literatur bezeichnet als *multilevel model*, z. B. Gelman & Hill (2006), *mixed effect model*, z. B. Viechtbauer (2022a), oder *hierarchical linear model*, z. B. Raudenbush & Bryk (2002)) berücksichtigen bei der Auswertung ebendiese hierarchische Struktur der Daten, also die Schachtelung von Situationen (Ebene 1) in Personen (Ebene 2), und können damit umgehen, dass nicht jede:r Teilnehmer:in gleichviele Messzeitpunkte beisteuert.

In der Literatur werden Mehrebenenmodelle oft im Kontext anderer geschachtelter Datenstrukturen diskutiert, in denen Personen die Ebene 1 bilden und übergeordnete Einheiten die Ebene 2, z. B. Schüler:innen (Ebene 1) geschachtelt in Schulklassen (Ebene 2). Dementsprechend wird teilweise von der Ebene 1 als Personenebene gesprochen und von Ebene 2 als Kontext- oder Gruppenebene. Im Fall von In-situ-Daten bilden aber die Situationen die unterste Ebene, die wiederum in Personen geschachtelt sind. Diese Zuordnung sollte uns beim Lesen von Methodenliteratur zur Mehrebenenanalyse also nicht verwirren.

Exkurs: Anzahl und Struktur der Ebenen

Mehrebenenmodelle können auch drei oder mehr Ebenen umfassen, z. B. könnten wir – je nach Forschungsfrage – annehmen, dass Situationen (Ebene 1) in Personen (Ebene 2) und diese wiederum in Schulklassen (Ebene 3) geschachtelt sind. Weiterhin gibt es kreuzklassifizierte Modelle, in denen die Ebenen nicht hierarchisch ineinander geschachtelt sind. Bei Schnauber-Stockmann & Mangold (2020) sind die Messzeitpunkte in Personen und gleichzeitig in Zeitabschnitten am Tag geschachtelt. Ein Messzeitpunkt kann zum einen dem Tagesabschnitt, aus dem er stammt, und zum anderen der Person, von der er stammt, zugeordnet werden. Die Tagesabschnitte sind aber nicht in den Personen geschachtelt – sie gelten für alle Personen gleichermaßen. Die beiden oberen Ebenen liegen also quer zueinander. In diesem Lehrbuch fokussieren wir uns allerdings auf den üblichen Fall, die Schachtelung von Situationen (Ebene 1) in Personen (Ebene 2).

Mehrebenenmodelle können für lineare und nicht lineare Zusammenhänge und mit metrischen, ordinalen und nominalen abhängigen Variablen geschätzt werden. Im Folgenden gehen wir auf die zentralen Aspekte von Mehrebenenmodellen ein, die zum grundlegenden Verständnis beitragen und Möglichkeiten von mehrebenenanalytischen Auswertungen von In-situ-Daten aufzeigen sollen. Wichtig ist zu betonen, dass wir keine umfassende Einführung in die Mehrebenenanalyse geben – empfehlenswerte Lehrbücher mit sozialwissenschaftlichem Fokus sind z. B. Gelman und Hill (2006), Hox et al. (2017), Langer (2009), Raudenbush und Bryk (2002), Sand und Braun (2025) sowie Snijders und Bosker (2012).

Exkurs: Arten der Schätzung

Die Parameter in Mehrebenenmodellen (Regressionskoeffizienten, Varianzkomponenten) können mit verschiedenen Verfahren geschätzt werden. Am häufigsten wird die *Maximum-Likelihood*-Methode eingesetzt, auf der auch die Modelle basieren, die im Anhang geschätzt werden. Es ist aber ebenso möglich, andere Methoden zu verwenden, vor allem bayesianische Methoden. Eine Einführung geben z. B. Hox et al. (2017) sowie Raudenbush & Bryk (2002). Ein Anwendungsbeispiel findet sich bei Link (2026).

Im Folgenden geht es zunächst um die grundlegende Frage, wie viel Varianz in einer Variable überhaupt durch die Situations- und Personenebene erklärbar ist. Im nächsten Schritt folgt die Aufbereitung der Variablen für die Auswertung in Mehrebenenmodellen – genauer: die Zentrierung. Anschließend befassen wir uns mit den Zusammenhängen zwischen zeitlich variierenden und stabilen Faktoren (interindividuelle Unterschiede) sowie zwischen mehreren zeitlich Faktoren (intraindividuelle Unterschiede). Abschließend schauen wir auf das Zusammenspiel von Personen- und Situationsmerkmalen.

Mehrebenenmodelle können in einigen Statistikprogrammen geschätzt werden – so verfügen z. B. SPSS, Mplus oder Stata über entsprechende Funktionen. Weiterhin gibt es spezielle, jedoch kostenpflichtige Software, z. B. HLM (Raudenbush et al., 2019). Im Online-Anhang zum Buch verwenden wir eines der gängigsten Pakete *lme4* (Bates et al., 2015) in R (R Core Team, 2025) zur Schätzung der Mehrebenenmodelle.

9.5.2.2 Zerlegung in Personen- und Situationsvarianz

Bei der Auswertung von In-situ-Daten stellt sich zunächst die Frage: Variiert das Merkmal, für das wir uns interessieren, überhaupt

auf der Situationsebene oder ist es lediglich bei verschiedenen Personen unterschiedlich ausgeprägt? Im Kontext der Beispielstudie im Online-Anhang zum Buch geht es etwa darum, ob die narrative Transportation vor allem deshalb variiert, weil die Teilnehmer:innen generell unterschiedlich narrativ transportiert werden – die einen neigen eher dazu, sich in Geschichten hineinzusetzen, die anderen weniger (wir sprechen dann von Transportabilität) – oder ob die narrative Transportation vor allem deshalb variiert, weil jede:r Teilnehmer:in je nach Situation mal stärker, mal schwächer involviert ist.

Das Verhältnis der unterschiedlichen Varianzquellen wird in Mehrebenenmodellen durch den Intraklassenkorrelationskoeffizienten (*Intraclass Correlation Coefficient*, $ICC = \text{Ebene-2-Varianz} / \text{Gesamtvarianz}$) ausgedrückt. Ein hoher ICC bedeutet, dass die Variable vor allem zwischen Personen variiert und die Situationsebene weniger relevant ist. Ein niedriger ICC deutet hingegen darauf hin, dass vor allem die Situation (und Interaktionen zwischen Situation und Person) relevant ist und nur eine geringe interindividuelle Varianz vorliegt. Ein ICC von 0 zeigt an, dass die Schachtelung der Situationen (Ebene-1-Einheiten) in Personen (Ebene-2-Einheiten), also die hierarchische Datenstruktur, (statistisch) irrelevant ist. In diesem Fall könnten klassische Auswertungsverfahren eingesetzt werden (z. B. die Regressionsanalyse). Dies wird bei In-situ-Daten aber vermutlich nie der Fall sein (für $ICCs$ in Studien zur Mediennutzung siehe Schnauber-Stockmann, Scharkow, et al., 2025). In der Literatur ist nicht einheitlich festgelegt, ab welchem ICC eine Mehrebenenanalyse zwingend notwendig ist. Generell lautet die Empfehlung, bei einem von 0 abweichenden ICC – auch wenn dieser verhältnismäßig klein ist – eine Mehrebenenanalyse durchzuführen, weil die Berücksichtigung der Mehrebenenstruktur nie zu schlechteren Schätzungen führt (z. B. Hox et al., 2017; Langer, 2009; Viechtbauer, 2022a). Der ICC wird für das sogenannte Nullmodell berechnet – ein Mehrebenenmodell ohne Einbeziehung von Prädiktorvariablen.

Es muss berücksichtigt werden, dass der ICC lediglich angibt,

9.5 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

welcher Anteil der Varianz einer Variablen sich maximal durch situative oder Personenfaktoren erklären lässt. Die Kennzahl macht keine Aussage darüber, ob dies durch die im späteren Verlauf einbezogenen Variablen (siehe Abschnitte 9.5.2.4 bis 9.5.2.6) tatsächlich (vollständig) gelingt.

Merke

Der *ICC* gibt an, wie viel Varianz einer Variable auf der Personenebene und im Umkehrschluss also nicht auf der Situationsebene liegt. Er sagt aber nichts darüber aus, ob diese Varianzanteile auf den beiden Ebenen durch die zur Erklärung herangezogenen Variablen erklärt werden können.

9.5.2.3 Zentrierung

Wollen wir In-situ-Daten mehrebenenanalytisch auswerten, also Variablen zur Erklärung der abhängigen Variablen in das Modell mit aufnehmen, müssen wir uns in den meisten Fällen mit dem Thema Zentrierung auseinandersetzen (siehe z. B. Enders & Tofighi, 2007; Hox et al., 2017; Langer, 2009; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Dieses Thema ist aus der klassischen Regressionsanalyse bekannt. Die Zentrierung dient dazu, den Intercept und die Interaktionsterme interpretierbar zu machen. In der klassischen Regressionsanalyse zentrieren wir, indem wir den Gesamtmittelwert einer Variable über alle Personen berechnen und diesen vom Wert jeder Person abziehen. Ein positiver Wert bedeutet, dass eine Person über dem Durchschnitt aller Personen liegt, ein negativer, dass sie darunter liegt. Die Zusammenhänge zwischen den Variablen bleiben jedoch unverändert.

Bei Mehrebenenanalysen gibt es dazu ein Pendant: die **Zentrierung am Gesamtmittelwert über alle Personen und Messzeitpunkte** hinweg (Zentrierung am *grand mean*). Für Situationsvariablen bedeutet ein positiver Wert, dass eine Person zu einem Messzeitpunkt über dem Durchschnitt aller Personen an allen Mess-

zeitpunkten liegt, ein negativer, dass sie darunter liegt. Für Personenvariablen ist nur die Zentrierung am Gesamtmittelwert möglich, denn diese Variablen variieren nicht zwischen den Messzeitpunkten. Ein positiver Wert bedeutet, dass die Person über dem Durchschnitt aller Teilnehmer:innen liegt, ein negativer darunter. Auch hier gilt: Die grundlegenden Zusammenhänge bleiben unverändert.

Die zweite Form ist die **Zentrierung am Personenmittelwert** (Zentrierung am *person mean*). Für jede Person wird gesondert ein individueller Mittelwert über alle Messzeitpunkte hinweg errechnet und dieser Mittelwert von jedem einzelnen Messzeitpunkt derselben Person abgezogen. Dann liegt der Wert bei 0, wenn die Person z. B. so gelaunt ist, wie sie es im Durchschnitt ist. Der Wert hat ein positives Vorzeichen, wenn die Person besser als ihr Durchschnitt gelaunt ist, und ein negatives Vorzeichen, wenn sie schlechter als ihr Durchschnitt gelaunt ist. Die Referenzgröße ist also nicht mehr, ob eine Person zu einem Messzeitpunkt besser oder schlechter als alle anderen Personen im Durchschnitt gelaunt ist, sondern referenziert auf die Person selbst als Bezugsgröße. Diese Form der Zentrierung verändert auch die Interpretation der Zusammenhänge in den Analysen, denn sie separiert die Between-Person-Varianz und enthält allein die Within-Person-Varianz. Sie repräsentiert also Veränderungen innerhalb des Individuums, alle Unterschiede zwischen Personen sind herausgerechnet. Oft ist diese Form der Zentrierung sinnvoll, wenn wir uns für intraindividuelle Zusammenhänge interessieren, also Zusammenhänge zwischen den Variablen auf der Situationsebene. Diese können wir so in ihrer Reinform anschauen (Enders & Tofighi, 2007; Viechtbauer, 2022a siehe auch Abschnitt 9.5.2.5).

Diese Form wird in der Literatur oft Zentrierung am Gruppenmittelwert (*group mean*) genannt, weil Mehrebenenanalysen häufig für die Auswertung von Personen geschachtelt in Gruppen (z. B. Schüler:innen in Klassen) verwendet werden. Die Personenebene ist bei In-situ-Daten die obere Ebene, weshalb wir vom Personen- und nicht vom Gruppenmittelwert sprechen.

Merke

Es stehen zwei Formen der Zentrierung zur Verfügung: die Zentrierung am Gesamtmittelwert (*grand mean*) und am Personenmittelwert (*person mean*). Die zuletzt genannte Form ermöglicht eine Fokussierung auf die bei In-situ-Studien im Vordergrund stehenden intraindividuellen Veränderungen und ist deshalb häufig die sinnvolle Wahl.

In Kombination mit der Zentrierung am Personenmittelwert kommt auch dem Personenmittelwert selbst eine Bedeutung zu. Während die am Personenmittelwert zentrierte Variable die intra-individuelle Veränderung auf der Situationsebene anzeigt, kann der Personenmittelwert selbst als Trait-Komponente aufgefasst werden und zeigt interindividuelle Unterschiede in der Variable an. Deshalb wird dieser Personenmittelwert häufig als eigenständiger Prädiktor in Analysen einbezogen. Weitere Erläuterungen hierzu und ein konkretes Beispiel für Within-Between-Modelle finden sich in Abschnitt 9.5.2.5.

9.5.2.4 Zusammenhänge zwischen stabilen und situativ variierenden Konstrukten

Die Mehrebenenanalyse ermöglicht es, Personenunterschiede aufzuzeigen, indem wir Personenmerkmale als unabhängige Variablen in die Auswertung einbeziehen. Die berücksichtigten Variablen auf Personenebene sollten am Gesamtmittelwert zentriert sein (siehe Abschnitt 9.5.2.3). So können wir z. B. feststellen, ob ein Zusammenhang zwischen dem Alter (einmalig gemessen) oder der Transportabilität (einmalig gemessen) und der narrativen Transportation (in situ gemessen) besteht. Einbezogene Personenmerkmale werden in Mehrebenenanalysen den fixen Effekten (*fixed effects*) zugeordnet (siehe Abschnitt 9.5.2.5) und gelten für alle Personen gleichermaßen – was auch gar nicht anders möglich ist, weil sie nicht innerhalb von Personen variieren.

Ein Vorteil gegenüber der Aggregation der In-situ-Variablen und der anschließenden Auswertung von Zusammenhängen auf der Personenebene (siehe Abschnitt 9.4) ist, dass die Mehrebenenanalyse berücksichtigt, wie viele Messzeitpunkte von einer Person vorliegen und in die Analyse einbezogen werden. Personen, von denen wir mehr Messzeitpunkte haben, erhalten ein höheres Gewicht, weil sie für die Ergebnisse wichtiger sind als die Personen, von denen weniger Messzeitpunkte vorliegen. Dieses Vorgehen ist plausibel, weil wir bei mehr Fällen i. d. R. von einer genaueren Messung ausgehen können.

Aber nicht nur einmalig gemessene Personenmerkmale können berücksichtigt werden; auch auf der Personenebene aggregierte, vorher in situ gemessene Werte können sinnvoll sein. So ist z. B. die Frage von Interesse, ob Personen, die generell häufig Podcasts hören (aggregierte Nutzungshäufigkeit, z.B. durch Summenbildung über alle Nutzungsepisoden einer Person) eher narrativ transportiert werden als solche, die seltener Podcasts hören. Wie an anderer Stelle ausgeführt (siehe Abschnitt 3.1), ist es sinnvoll, die Nutzungshäufigkeit *in situ* zu erheben und anschließend zu aggregieren. Durch die Aggregation wird die Variable zur Personenvariable und wir können sie – wie die anderen Personenvariablen auch – als fixen Effekt in die Mehrebenenanalyse einbeziehen.

9.5.2.5 Zusammenhänge zwischen mehreren situativ variierenden Konstrukten

Weiterhin können wir Zusammenhänge zwischen Variablen auf der Situationsebene untersuchen, z. B. ob es einen Zusammenhang zwischen narrativer Transportation und Paralleltätigkeit gibt. Eine zentrale Besonderheit von Mehrebenenanalysen ist die Unterscheidung zwischen fixen Effekten (*fixed effects*) und Varianzkomponenten (*random effects* oder *random slopes*). Der fixe Effekt der Paralleltätigkeit ist über alle Personen (also Ebene-2-Einheiten) hinweg gleich, d. h., dass die Paralleltätigkeit für alle Personen denselben Effekt auf die narrative Transportation hat –

9.5 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

für alle Teilnehmer:innen gilt etwa: Wenn sie keiner Paralleltätigkeit nachgehen, sind sie stärker narrativ transportiert. Es unterscheiden sich zwischen den Personen also nur die Intercepts, nicht aber die Steigung der Geraden. Personen gehen im Schnitt mehr oder weniger häufig einer Paralleltätigkeit nach (*random intercept*), aber Veränderungen der Paralleltätigkeit haben für alle denselben Einfluss auf die narrative Transportation (die Steigung der Geraden ist gleich; Random-Intercept-Modell siehe Abbildung 9.3 links).

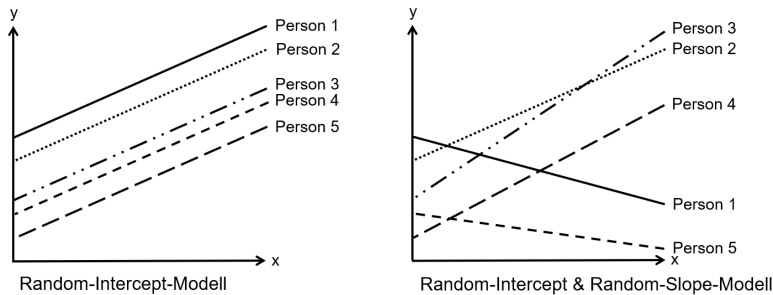


Abbildung 9.3: Random-Intercept- sowie Random-Intercept- & Random-Slope-Modell (Schnauber, 2017)

Doch diese Annahme ist nicht immer realistisch. Es ist möglich, dass einige Personen stärker narrativ transportiert werden, wenn sie einer Paralleltätigkeit nachgehen, etwa einer kognitiv wenig anspruchsvollen Hausarbeit, bei anderen dieser Effekt eintritt, wenn sie keiner Paralleltätigkeit nachgehen, und bei wieder anderen Personen kein Zusammenhang besteht. Der Zusammenhang zwischen Paralleltätigkeit und narrativer Transportation kann von Person zu Person variieren. Die Mehrebenenanalyse kann dies durch Varianzkomponenten abbilden (siehe hierzu Neubauer & Schmiedek, 2020). Wir können also zulassen, dass die unabhängige Variable auf der Situationsebene bei verschiedenen Personen einen unterschiedlichen Einfluss auf die abhängige Variable hat. Die Slopes, also die Stei-

gerung der Geraden, können sich unter diesen Bedingungen auch unterscheiden (Random-Intercept- & Random-Slope-Modell, siehe Abbildung 9.3 rechts).

Merke

Wir können in Mehrebenenmodellen fixe Effekte und Varianzkomponenten schätzen. Fixe Effekte modellieren, dass die unabhängige Variable die abhängige Variable über alle Personen hinweg gleich beeinflusst. Varianzkomponenten modellieren, dass die unabhängige Variable einen personenspezifisch unterschiedlichen Einfluss auf die abhängige Variable ausübt.

Doch was bedeutet es nun, wenn sich der Zusammenhang zwischen zwei Variablen zwischen Personen unterscheidet, die Geradensteigungen also variieren? Hilfreich kann es sein, die **Form der variierenden Slopes** zu explorieren, indem wir z. B. den Anteil der Slopes über 0 (also mit positiver Steigung) und unter 0 (also mit negativer Steigung) oder das Intervall analysieren, in dem 90 oder 95 % der Slopes liegen (für ein Beispiel siehe Schnauber, 2017). Die Befunde vermitteln einen Eindruck davon, ob vor allem die Stärke des Zusammenhangs variiert, wenn z. B. die meisten Slopes positiv sind. Es können auch entgegengesetzte Zusammenhänge vorliegen, wenn es sowohl positive als auch negative Slopes gibt.

Wie in Abschnitt 9.5.2.3 eingeführt, ist es sinnvoll, die unabhängigen Variablen auf der Situationsebene zu zentrieren. Die abhängige Variable wird dagegen nicht zentriert. Während die Zentrierung am Gesamtmittelwert die Interpretation der Zusammenhänge nicht verändert, ist dies für die Zentrierung am Personenmittelwert (also der Abweichung jedes Messzeitpunkts einer Person von ihrem individuellen Durchschnittswert über alle Situationen hinweg) der Fall. Denn diese Form der Zentrierung separiert die Between-Person-Varianz und enthält allein die Within-Person-Varianz. Sie bereinigt die Variable darum, dass es Personen gibt, die z. B. generell aufmerksamer sind als andere, und fokussiert allein auf die Frage, ob eine Person

9.5 Unterschiede zwischen Situationen: Fragen nach der intraindividuellen Varianz

gerade aufmerksamer oder unaufmerksamer ist als sonst. Sie repräsentiert also allein Veränderungen innerhalb des Individuums, indem alle Unterschiede zwischen den Personen herausgerechnet werden. Es hängt auch von der Forschungsfrage ab, welche der beiden Zentrierungsformen hilfreich ist (Enders & Tofighi, 2007; Viechtbauer, 2022a, 2022b). Die Zentrierung am Personenmittelwert ist sinnvoller, wenn intraindividuelle Zusammenhänge von Interesse sind. Die Interpretation lautet in diesem Fall: Wenn eine Person aktuell aufmerksamer ist als sonst (als ihr eigener Durchschnitt), dann erlebt sie ein höheres Maß an narrativer Transportation.

Möglich ist auch, dass wir uns für beides gleichzeitig interessieren: Wir können z. B. auswerten, ob zum einen ein intraindividueller Zusammenhang zwischen der Podcastnutzungsdauer und der narrativen Transportation besteht und ob sich zum anderen Personen, die generell unterschiedlich lange Podcasts hören, in ihrer narrativen Transportation unterscheiden. Diese Befunde lassen sich mit der Mehrebenenanalyse ebenfalls abbilden, indem wir den Personenmittelwert selbst als eigene Variable zusätzlich zur am Personenmittelwert zentrierten Variable in die Auswertung einführen. Der Koeffizient der zentrierten Variablen zeigt den intraindividuellen Zusammenhang an: Wenn eine Person aktuell länger Podcast hört als im eigenen Durchschnitt, dann ist sie in dieser Situation stärker narrativ transportiert. Den interindividuellen Zusammenhang erhalten wir hingegen durch den Koeffizienten des Personenmittelwerts: Personen, die generell länger Podcast hören als andere, sind in dieser Situation stärker narrativ transportiert. Solche **Within-Between-Modelle** sind wirkmächtige Auswertungsmöglichkeiten von In-situ-Daten. Sie erlauben die klare Trennung und gleichzeitige Betrachtung von Between- und Within-Person-Zusammenhängen.

Neben den im Mittelpunkt des Forschungsinteresses stehenden Zusammenhängen beziehen wir häufig auch Kontrollvariablen in unsere Modelle ein. Neben inhaltlich-theoretischen Kontrollvariablen kann es sinnvoll sein, auch methodische Kontrollvariablen wie z. B. den Studenttag, die Tageszeit oder die Response Latency einzubeziehen (z. B. Naab et al. (2019), siehe Studiensteckbrief im Online-

Anhang zum Buch). Welche Kontrollvariablen wir einbeziehen, sollte theoretisch oder methodisch begründet und idealerweise im Vorfeld festgelegt sein, weil das Einbeziehen von Kontrollvariablen die Ergebnisse verändern kann.

Exkurs: Moderations- und Mediationsanalyse

Mehrebenenanalysen erlauben, Moderations- und Mediationsannahmen zu testen. Vorsicht ist jedoch im Hinblick auf Mediationsannahmen geboten, da diese an einige Voraussetzungen gebunden sind, u. a. an eine klare zeitliche Reihenfolge (unabhängige Variable – Mediator – abhängige Variable), die in In-situ-Studien nicht immer gegeben ist (für eine Diskussion über die Angemessenheit von Mediationsanalysen allgemein siehe z. B. Coenen, 2022; Kline, 2015; Pieters, 2017). Moderationsanalysen können sowohl innerhalb einer Ebene als auch zwischen den Ebenen durchgeführt werden (siehe auch Abschnitt 9.5.2.6). So könnten wir annehmen, dass die Nutzungsdauer nur die narrative Transportation beeinflusst, wenn die Person zu Hause ist, nicht aber, wenn sie unterwegs ist. Zur Prüfung dieser Annahme nehmen wir in unser Modell den Interaktionsterm zwischen Nutzungsdauer und Ort auf, indem wir die beiden Variablen multiplizieren.

9.5.2.6 Zusammenspiel von Personen- und Situationsmerkmalen

Schließlich können wir uns für das Zusammenspiel von Personen- und Situationsmerkmalen interessieren. Haben wir z. B. festgestellt, dass für verschiedene Personen die Ausübung von Paralleltätigkeiten unterschiedlich mit narrativer Transportation zusammenhängt (Random Slopes, siehe Abschnitt 9.5.2.5), können wir nach Personenmerkmalen suchen, die dies erklären. Es liegt z. B. die Annahme nahe, dass bei Personen mit einer höheren Transportabilität (Perso-

nenmerkmal) das Hören eines Podcasts ohne eine Paralleltätigkeit mit höherer narrativer Transportation zusammenhängt. Bei Personen mit einer niedrigeren Transportabilität könnte der Zusammenhang umgekehrt sein: Sie könnten stärker narrativ transportiert sein, wenn sie während der Podcastnutzung nebenher einer anderen Tätigkeit nachgehen. Technisch handelt es sich hierbei um Interaktionen zwischen Variablen in Form einer Moderation (siehe Exkurs in Abschnitt 9.5.2.5). Da wir uns hier für die Interaktion einer Variable auf Personen- mit einer auf Situationsebene interessieren, spricht man von einer **Cross-Level-Interaktion**. Da Cross-Level-Interaktionen für alle Personen gleichermaßen gelten, handelt es sich hierbei wieder um fixe Effekte.

9.5.2.7 Zeitliche Dynamiken, zeitversetzte Zusammenhänge und Autokorrelationen

In diesem Kapitel befassen wir uns mit der Zeit als Variable im Mehrebenenmodell und mit zeitversetzten Zusammenhängen zwischen Variablen. Schließlich betrachten wir noch das Thema Autokorrelation, also die Frage, inwieweit die Ausprägung einer Variable durch ihre Ausprägung zu einem vorherigen Zeitpunkt (mit-)bestimmt wird.

Zunächst ist die Frage von Interesse, wie **zeitliche Veränderungen** eines situativ variierenden Merkmals untersucht werden können. Da Zeit ebenfalls nichts anderes als ein situativ variierendes Merkmal ist, können wir eine Variable, die die Zeit misst (siehe Abschnitt 9.5.2.5), in ein Mehrebenenmodell einführen und z. B. das Zusammenspiel mit Personenvariablen testen. Es kann z. B. die Annahme zugrunde liegen, dass sich zeitliche Nutzungsmuster zwischen Berufstätigen und Nichtberufstätigen unterscheiden. Im Folgenden werden einige Beispiele vorgestellt, was solche Variablen beinhalten können.

Wir könnten annehmen, dass die Stimmung im Studienverlauf schlechter wird, weil die Teilnehmer:innen durch die Studie angestrengt sind. Dieser Befund würde sich in einem negativen Zusam-

menhang zwischen Stimmung und Studientag zeigen. Oder wir könnten prüfen, ob die Podcastnutzungsdauer am Wochenende länger ist als unter der Woche. Für diesen Zweck würden wir eine Zeitvariable konzipieren, die jedem Messzeitpunkt zuweist, ob er am Wochenende oder unter der Woche liegt.

Neben linearen Zusammenhängen ist denkbar, dass andere zeitliche Muster zu erwarten sind. Wir könnten z. B. für die Nutzungsdauer im Tagesverlauf (unabhängig vom jeweiligen Wochentag) annehmen, dass sie nachts gering ist, in den Morgenstunden ansteigt und am Mittag am längsten ist, bevor sie wieder abfällt, also dem Verlauf einer umkehrten U-Form entspricht. Eine sinnvolle Zeitvariable könnte z. B. von 0 bis 23 (Stunden des Tages) verlaufen und mit einer polynomialen Transformation der Zeitvariable modelliert werden (z. B. Bortz & Schuster, 2010, S. 199–200; für eine Anwendung mit RTR-Daten, die in ihrer Struktur In-situ-Daten ähnlich sind, siehe Bachl, 2014).

Merke

Zeit kann als situativ variierende Variable in einem Mehrebenenmodell betrachtet werden.

Kommen wir nun zu **zeitversetzten Zusammenhängen** zwischen Variablen: Der Zusammenhang zwischen zwei Situationsvariablen, die zum selben Messzeitpunkt erhoben wurden, sagt statistisch nichts über die Kausalität des Zusammenhangs aus. Für Aussagen über die Wirkrichtung (zur Kausalitätsthematik allgemein siehe aber z. B. Rohrer & Murayama, 2023) müssen wir Ursache und Wirkung in eine zeitliche Reihenfolge bringen, die das Ergebnis theoretischer Überlegungen ist. Als Beispiel sei der Zusammenhang zwischen Stimmung und Nutzungsdauer genannt. Für die Untersuchung stimmungabhängiger Medienauswahlprozesse können wir z. B. die Frage prüfen, ob Personen, die zu einem bestimmten Zeitpunkt schlecht gelaunt waren, zu einem späteren Zeitpunkt länger Podcasts hören. Dieses Verhalten könnte ein Zeichen für Mood-Management sein. Bei

einer Untersuchung über die Wirkung von Podcasts auf die Stimmung stellt sich die Frage, ob sich die Stimmung zum nächsten Zeitpunkt nach der Podcastnutzung verändert hat.

Die Auswertung folgt dem in Abschnitt 9.5.2.5 beschriebenen Schema. Der Unterschied liegt in der Datenaufbereitung, in der wir die zeitversetzten Variablen anspielen – wie z. B. den Stimmungswert vom vorherigen Messzeitpunkt. Der Messzeitpunkt muss im Datensatz in derselben Zeile stehen wie die Werte des folgenden Messzeitpunkts. Dabei stellen sich einige Fragen:

- Wie lange dauern die Intervalle zwischen den Protokollen? Der Abstand zwischen den beiden Messzeitpunkten sollte nicht zu lang sein, wenn wir von einem Zusammenhang zwischen Stimmung und Podcastnutzungsdauer im Sinne des Mood-Managements ausgehen. Die schlechte Laune am Morgen erklärt vermutlich nicht die Podcastnutzung am Nachmittag. Bereits während der Planung der Studie sollte bedacht werden, welcher (maximale) zeitliche Abstand zwischen den Protokollen angemessen und sinnvoll ist (siehe Abschnitt 6.1).
- Umgekehrt sind auch Fragestellungen möglich, die einen eher größeren Abstand vorsehen. Dann ist zu überlegen, ob wir z. B. nicht den Wert des vorherigen Protokolls anspielen, sondern eines noch früheren.
- Was ist, wenn der Wert, den wir anspielen wollen, fehlt? Es wird in etlichen Fällen fehlende Werte geben, weil nicht alle Teilnehmer:innen die Protokolle lückenlos beantworten (siehe Abschnitt 6.3.2.2). Für diese Fälle sollte gut begründet werden, ob ein anderer Wert angespielt werden kann (z. B. vom vorherigen oder folgenden Protokoll) oder ob eine Imputation sinnvoll ist (siehe Abschnitt 9.2.2.2). Alternativ fehlt dieser Fall für die Analyse.
- Ist es plausibel, dass sich über Nacht Effekte zeigen, indem z. B. die Stimmung am Vorabend die Podcastnutzung am nächsten

Tag erklärt? In Studien zu Mood-Management und auch in vielen anderen Fällen ist diese Plausibilität vermutlich nicht gegeben. Es ist daher sinnvoll, für zeitversetzte Zusammenhänge den ersten Messzeitpunkt des Tages auszuschließen.

In allen Fällen geht es um Fragen des theoretisch plausiblen und empirisch erfassbaren oder erfassten Zeitabstands, der eine Wirkungsannahme plausibel macht (siehe hierzu Fritz et al., 2024; Neubauer & Schmiedek, 2020; Viechtbauer, 2022a). Wir müssen also theoretische Argumente abwägen sowie konzeptuelle und methodische Entscheidungen idealerweise im Rahmen einer Präregistrierung festhalten (siehe Abschnitt 10.2).

Merke

Bei einer Modellierung zeitversetzter Zusammenhänge zwischen Variablen müssen wir bei der Planung der Studie sowie der Datenaufbereitung auf theoretisch plausible Zeitabstände zwischen den Messzeitpunkten achten.

Neben zeitversetzten Zusammenhängen zwischen verschiedenen Variablen kann es sinnvoll sein, den Zusammenhang einer Variable mit sich selbst einzubeziehen. Es ist jedoch statistisch problematisch, diese **Autokorrelation**, d. h. die Korrelation einer Variable mit sich selbst zu einem anderen Zeitpunkt, in Mehrebenenmodelle zu integrieren (siehe hierzu z. B. Allison, 2015). In diesem Fall sind andere Formen der Modellierung sinnvoll (siehe Exkurs in Abschnitt 9.5.3).

9.5.3 Klassen-/Typenbildung von Situationen und Personen

Bisher standen vor allem Auswertungsverfahren im Mittelpunkt, die einer Regressionslogik folgen, also auf die Aufdeckung von Zusammenhängen zwischen einzelnen Variablen ausgelegt sind. Die Klassen-/Typenbildung auf Basis von In-situ-Daten gehören

statistisch zu den fortgeschrittenen Auswertungsverfahren. Bei Fragestellungen, die darauf abzielen, Situationen sinnvoll zusammenzufassen, kann es hilfreich sein, sich mit der *Multilevel Latent Class Analysis* (MLCA) zu befassen. Die MLCA basiert ebenfalls auf der Logik von Mehrebenenanalysen. Allerdings sind die abhängigen Variablen latent, d. h., sie werden nicht gemessen, sondern aus den vorhandenen Daten geschätzt. Sie geben die Anzahl und Zugehörigkeit zu verschiedenen Klassen an. Die MLCA ermöglicht es, gleichzeitig Situationen und Personen einzuteilen bzw. bei der Identifikation der Situationsklassen die Mehrebenenstruktur der Daten zu berücksichtigen. Es wird davon ausgegangen, dass Situationen auf der Situationsebene zu Klassen zusammengefasst werden können und es auf der Personenebene Personenklassen gibt, die sich in der Verteilung der Situationsklassen unterscheiden (Lukočienė et al., 2010). Hilfreiche Einstiege in die MLCA bieten z. B. Lukočienė et al. (2010), Vermunt (2008) sowie Vermunt & Magidson (2004). Lyrvall et al. (o. J.) stellen das R-Paket `multilevLCA` vor, mit dem MLCAs in R durchgeführt werden können ().

Karnowski et al. (2024) identifizieren etwa mithilfe einer MLCA mobile Nachrichtenrepertoires. Die Teilnehmer:innen berichteten über ihre mobile Nachrichtennutzung und gaben in den ESM-Protokollen an, über welchen Zugang (z. B. Nachrichtenapp, soziale Medien) und in welchem Modus (z. B. lesen, schauen, hören) sie Nachrichten konsumierten. Es zeigten sich fünf Situationsklassen: *Direct App Access*, *Listening and Watching*, *Social Media/Textual*, *Social Media/Visual* und *Direct Website Access* (siehe Studiensteckbrief im Online-Anhang zum Buch). Ein weiteres Anwendungsbeispiel in der KMW findet sich bei Ernst & Schnauber-Stockmann (2026). Die Teilnehmer:innen berichteten in einer ereignisabhängigen ESM-Studie jeweils nach dem Schließen von Instagram oder TikTok, warum sie – mit Vorgabe verschiedener Schließgründe – die Nutzungsepisode beendet hatten. Die Autorinnen identifizierten fünf Situationstypen: *Non-selfdetermined Disengagement*, *Self-determined Disengagement*, *Priority Shift Disengagement*, *Temporary Disengagement* und „*Mindless*“ *Disengagement* (siehe Studiensteckbrief im

Online-Anhang zum Buch).

Exkurs: Fortgeschrittene Auswertungsverfahren

In-situ-Designs sind zwar keine neuen Erhebungsmethoden (siehe Abschnitt 1.2), ihre Anwendung nimmt aber seit einiger Zeit in der KMW und auch in anderen Fächern wie der Psychologie oder Medizin einen immer höheren Stellenwert ein. Immer mehr und teilweise auch immer komplexere Fragestellungen sollen mit In-situ-Daten beantwortet werden. Die Anforderungen an Validität und Reliabilität sowie deren Prüfbarkeit wachsen ebenfalls. Deshalb werden fortlaufend neue Auswertungsverfahren entwickelt bzw. für In-situ-Daten adaptiert. Im Folgenden soll ein kurzer Überblick – ohne Anspruch auf Vollständigkeit – über solche fortgeschrittenen Auswertungsverfahren gegeben werden. Ziel ist es, den Leser:innen weitere Anwendungsbeispiele und Ansatzpunkte für eine eigenständige Recherche zu liefern. Der Schwerpunkt liegt auf Verfahren, die Möglichkeiten zur dynamischen Modellierung geben, also berücksichtigen, dass die erhobenen In-situ-Daten in zeitlichem Bezug zueinander stehen und sich gegenseitig beeinflussen.

Ein auch in der KMW etwas häufiger eingesetztes Auswertungsverfahren ist das *Dynamic Structural Equation Modeling* (z. B. McNeish & Hamaker, 2020). Diese besondere Form von Strukturgleichungsmodellen berücksichtigt die Mehrebenenstruktur der Daten und ermöglicht somit die gleichzeitige Auswertung von Zusammenhängen auf Personen- und Situationsebene, kombiniert mit Zeitreihenmodellierung. Anwendungsbeispiele finden sich z.B. bei Aalbers et al. (2022) (siehe Studiensteckbrief im Online-Anhang zum Buch), Verbeij et al. (2022) und Valkenburg et al. (2021) (siehe Studiensteckbrief im Online-Anhang zum Buch).

Weiterentwicklungen wie *Hierarchical Continuous Time Dynamic Structural Equation Models* (z. B. Driver & Voelke,

2018) sind für In-situ-Daten geeigneter, weil sie mit ungleichen Abständen zwischen den einzelnen Messzeitpunkten – wie bei In-situ-Studien üblich – besser umgehen können. Hier liegen erste Anwendungsbeispiele mit einem kommunikationswissenschaftlichen Bezug vor (Klingelhoef et al., 2026; Marciano et al., 2022). Weiterhin gibt es über die in diesem Buch vorgestellten Möglichkeiten der Untersuchung von zeitlichen Zusammenhängen hinausgehende komplexere Modellierungsverfahren, die ebenfalls die Möglichkeit bieten, die Mehrebenenstruktur der Daten zu berücksichtigen. Beispiele sind latente Wachstumskurvenmodelle (Muthén, 2002), vektorautoregressive Modelle (Lafit, 2022; Schuurman & Hamaker, 2019; Tuerlinckx et al., 2026) oder latente Markov-Modelle (Bartolucci et al., 2012; Vermunt et al., 1999).

10 Open Science und Reporting

10.1 Einführung

Die grundlegende Idee von Open Science besteht darin, wissenschaftliche Vorgehensweisen und Ergebnisse allgemein verfügbar zu machen und dadurch Transparenz und wissenschaftliche Qualität zu sichern. Open Science soll Wissens- und Datenasymmetrien abbauen (Dienlin et al., 2021; Nosek et al., 2015) und im Idealfall Gerechtigkeit und Inklusion fördern (Fox et al., 2021). Open Science bezieht sich auf sämtliche Phasen des Forschungsprozesses. Entsprechend sind die Open-Science-Bemühungen vielfältig und umfassen u. a. Daten (Open Data), Studienmaterialien (Open Materials), Publikationen (Open Access), Software (Open Source), Lehrmaterialien (Open Educational Resources, OER), Evaluationsprozesse (Open Peer Review) und Infrastrukturen (z. B. das Open Science Framework OSF). Unter Open Science fällt auch die Präregistrierung von Studien (i. S. v. *open reproducible research*), auf die wir in Abschnitt 10.2 eingehen.

Merke

Das Ziel von Open Science ist es, Forschung und ihre Ergebnisse allgemein zugänglich zu machen, um wissenschaftliche Qualität zu sichern und Gerechtigkeit und Inklusion zu fördern.

Ein Ursprung der aktuellen Open-Science-Bewegung in der KMW liegt in der **Replikationskrise** der Psychologie, da zentrale Befunde nicht in nachfolgenden Studien repliziert werden konnten (Open Science Collaboration, 2015). Dieses Problem hatte mehrere

(vermeintliche) Ursachen: Die Dokumentation des Vorgehens in den Originalstudien war für eine Replikation häufig nicht ausführlich genug. Replizierbar ist ein Ergebnis, wenn gilt: „It can be obtained with other random samples drawn from a multidimensional space that captures the most important facets of the research design“ (Asendorpf et al., 2013, S. 109). Es geht also darum, dass eine Studie, die demselben Vorgehen folgt, zum gleichen Ergebnis kommen sollte. Im Sinne von Open Science ist die Veröffentlichung aller Materialien (z. B. Erhebungsinstrumente und experimentelle Manipulationen) sinnvoll, um dieser Anforderung gerecht zu werden. Auch das ausführliche Reporting, also die genaue Dokumentation von methodischen und datenanalytischen Vorgehensweisen in Forschungsberichten und Publikationen, spielt hierfür eine zentrale Rolle (siehe z. B. Roekel et al., 2019; Schnauber-Stockmann & Karnowski, 2020; Stone & Shiffman, 2002; Trull & Ebner-Priemer, 2020). Reportingstandards erleichtern die Transparenz, indem sie Forschende dazu anregen, das Design ihrer Studie, inklusive forschungsethischer Überlegungen, den Analyseplan und die Ergebnisse in der Veröffentlichung vollständig und nachvollziehbar zu beschreiben (Kirtley et al., 2021).

Außerdem besteht die Möglichkeit, dass fragwürdige Forschungspraktiken (sog. Questionable Research Practices, vgl. Matthes et al., 2015) zum Einsatz gekommen sind, etwa das sog. *p-hacking* oder *HARKing* (*Hypothesizing after Results Are Known*). Ersteres umfasst z. B. die Änderung der Auswertungsstrategie und der in die Auswertung einbezogenen (Kontroll-)Variablen oder Fälle, bis die gewünschten hypothesenbestätigenden Ergebnisse herauskommen ($p < .05$ als Kriterium für Signifikanz, deshalb *p-hacking*). Letzteres meint, dass Hypothesen nach Auswertung der Daten an die Ergebnisse angepasst werden, weil die ursprünglichen Hypothesen nicht durch die Ergebnisse bestätigt werden. Beide Vorgehensweisen haben ihren Ursprung auch in der Publikationskultur – veröffentlicht werden häufig nur Studien, die zu signifikanten Ergebnissen gelangen. Studien mit Nullergebnissen, also mit nicht signifikanten Resultaten, werden nicht veröffentlicht und verschwinden gewissermaßen in der Schub-

lade (daher der Begriff *file drawer problem*). So kann der Eindruck entstehen, dass sich ein Zusammenhang in allen Studien bestätigt, wenn nur diese publiziert werden (*publication bias*). Dieses Problem besteht auch in der KMW (Sun & Pan, 2020). Um dem zu begegnen, ist ein zentraler Bestandteil von Open Science die Präregistrierung, also die Formulierung und Veröffentlichung von Hypothesen und Forschungsfragen, des Forschungsdesigns, der Datenmanagement- und Analysepläne bereits im Vorfeld der Datenerhebung oder -analyse. Im folgenden Abschnitt gehen wir auf Präregistrierung und Reportingstandards für In-situ-Studien sowie die Möglichkeiten und Grenzen der Veröffentlichung von Materialien und Daten sowie des Codes ein.

10.2 Präregistrierung

In-situ-Studien haben deutlich mehr Entscheidungspfade als klassische Befragungsstudien; es gibt daher (noch) keine etablierten methodischen Standards wie bei anderen Erhebungsmethoden. Aus der Open-Science-Perspektive – zur Sicherstellung von Replizierbarkeit und Transparenz sowie zur Verhinderung von fragwürdigen Forschungspraktiken – ist es daher sehr wichtig, alle Schritte vorab gründlich zu planen und zu dokumentieren. Präregistrierungen werden üblicherweise öffentlich verfügbar gemacht, z. B. im OSF.

Die Podcastingstudie wurde im Vorfeld der Datenerhebung im OSF präregistriert. Die Präregistrierung und alle Materialien (z. B. Fragebögen), die Daten, der Auswertungscode und eine Ergänzung, die die Abweichungen von der Präregistrierung dokumentiert, sind im im Studiensteckbrief im Online-Anhang zum Buch verfügbar .

Es gibt für Open-Science-Praktiken (Präregistrierung, Metadaten etc.) bisher noch wenige Vorlagen für In-situ-Studien (eine Ausnahme ist z. B. Kirtley et al., 2021). Bei Langener et al. (2024) findet sich für die Präregistrierung von mobilen digitalen Verhaltensdaten eine Vorlage, die hilfreich sein kann, wenn im Rahmen der Studie solche Daten miterhoben werden. Im Folgenden gehen wir auf die

zentralen Aspekte ein, die dokumentiert werden sollten und in diesem Buch als notwendige Bestandteile für In-situ-Studien beschrieben wurden. Wir möchten aber noch einmal ausdrücklich betonen, wie wichtig eine *umfassende* Dokumentation ist, die dazu dient, die Freiheitsgrade der Forschenden (siehe hierzu Nosek et al., 2018; Wicherts et al., 2016) offenzulegen oder einzuschränken.

Unter Ressourcen haben wir im Online-Anhang zum Buch entsprechende Templates und auch Beispielpräregistrierungen zusammengestellt. Beispiele für Präregistrierungen finden sich auch in einigen unserer Steckbriefstudien im Online-Anhang zum Buch.

Bei der Präregistrierung werden die zu prüfenden Hypothesen und die zu beantwortenden Forschungsfragen im Vorfeld festgelegt und öffentlich kommuniziert. Die Präregistrierung gilt auch für Erhebung, Aufbereitung und Auswertung der Daten (Wicherts et al., 2016). Gerade bei Aufbereitung und Auswertung können an vielen Stellen alternative, gleichermaßen legitime Optionen gewählt werden, die ggf. zu unterschiedlichen Ergebnissen führen würden. So müssen wir uns z. B. bei der Compliance entscheiden, ab welchem Grenzwert Teilnehmer:innen aus der Analyse ausgeschlossen werden. Denn der Ein- und Ausschluss von Fällen kann das Ergebnis verändern. Es gibt hier keine festen Regeln (siehe Abschnitte 6.3.2.2 und 9.2.1), wichtig ist, dass die Entscheidung über mögliche Grenzwerte im Vorfeld getroffen wird. Weiterhin können wir zwischen verschiedenen Auswertungsmethoden wählen, von denen ggf. einige – aber andere nicht – signifikante Ergebnisse erbringen. In der Präregistrierung legen wir das Vorgehen deshalb vorab fest, um zu verhindern, dass Hypothesen oder Auswertungsstrategien nachträglich ‚geschönt‘ werden, um signifikante Ergebnisse zu erzielen. Solche Forschungspraktiken gelten als fragwürdig, weil dadurch valide gewonnene Nullergebnisse nicht sichtbar werden, obwohl sie einen wichtigen Beitrag zum wissenschaftlichen Erkenntnisgewinn leisten können.

In Tabelle 10.1 sind die zentralen Informationen aufgeführt, die in einer Präregistrierung festgehalten werden sollten. Diese decken sich in großen Teilen mit den Reportingstandards (siehe Abschnitt 10.3).

Eine Präregistrierung ist mit Arbeitsaufwand vor dem eigentlichen Studienstart verbunden. Dieser Arbeitsaufwand fällt allerdings ohnehin an, z. B. wenn im Rahmen eines Ethikvotums die Vorgehensweise vorab offengelegt werden muss (siehe Abschnitt 4.2) oder während der Datenaufbereitung und -auswertung Entscheidungen zu treffen sind. Es erfordert also ohnehin einen zeitlichen Aufwand, zentrale Aspekte einer Studie möglichst umfänglich zu durchdenken und zu planen. Im Rahmen einer Präregistrierung hat dieses planvolle Vorgehen einige erhebliche Vorteile: Bereits vor der eigentlichen Datenerhebung zeigt sich, ob alle Aspekte berücksichtigt und die benötigten Variablen auch gemessen werden. Umgekehrt fällt uns vielleicht auf, dass bestimmte Konstrukte gar nicht notwendig sind und ein Verzicht auf deren Erhebung der Datensparsamkeit dient. Zudem kann die Situation eintreten, dass die Auswertungsmethode, die für eine Studie am besten geeignet ist, bestimmte Anforderungen an das Studiendesign stellt, das entsprechend angepasst werden muss. Diese vorausschauenden Maßnahmen verhindern, dass uns unüberlegte Entscheidungen später wieder einholen.

Dennoch wird es fast immer Aspekte geben, die nicht wie geplant laufen. Wichtig ist deshalb, eine Präregistrierung als Plan zu begreifen, aber nicht als ‚Gefängnis‘ (DeHaven, 2017). Abweichungen sind immer möglich, sollten aber begründet und transparent dokumentiert werden, etwa in einem (Online-)Anhang (siehe z. B. die Checkliste bei Wicherts et al., 2016; siehe auch Kirtley et al., 2021; Lakens, 2024).

Es ist möglich, über die Präregistrierung hinausgehende Auswertungen von unerwarteten Ergebnissen zu berichten, die als explorativ gekennzeichnet sein sollten. Es wird zwischen **explorativen** und **konfirmatorischen Auswertungen** unterschieden. Konfirmatorische Auswertungen umfassen die Hypothesen und Forschungsfragen sowie die dazugehörigen Auswertungsschritte, die wir in der Präregistrierung festgehalten haben. Explorative Auswertungen umfassen alle zusätzlichen, nicht präregistrierten Auswertungen und dienen dazu, zusätzliche Erkenntnisse aus den Daten zu gewinnen, die über die ursprünglich festgelegten Ziele hinausgehen. Sie können neue und

10.3 Reporting-Standards und Veröffentlichung von Studienmaterialien und -daten

unvorhergesehene Erkenntnisse liefern, die in künftigen Forschungsvorhaben (konfirmatorisch) vertieft untersucht werden können.

Merke

Präregistrierungen dokumentieren alle Hypothesen, Forschungsfragen und methodischen Schritte vor dem Studienstart. Damit werden die Freiheitsgrade der Forschenden nach der Datenerhebung begrenzt, um fragwürdige Forschungspraktiken zu verhindern und valide Nullergebnisse sichtbar zu machen. Präregistrierungen helfen den Forschenden, ihre Studie im Vorfeld detailliert zu planen, wobei begründete und nachvollziehbare Planabweichungen möglich und manchmal unvermeidlich sind.

10.3 Reporting-Standards und Veröffentlichung von Studienmaterialien und -daten

Die ausführliche Dokumentation methodischer und datenanalytischer Vorgehensweisen in Forschungsberichten und Publikationen spielt für die Transparenz und damit auch die Replizierbarkeit (falls auch Daten geteilt werden, zudem die Reproduzierbarkeit) von empirischen Studien eine entscheidende Rolle. Wir empfehlen daher, das Design der eigenen Studie, den Analyseplan und die Ergebnisse in der Veröffentlichung vollständig zu beschreiben (Kirtley et al., 2021).

Es gibt bereits einige Vorschläge für Reportingstandards für In-situ-Studien in der Literatur (z. B. May et al., 2018; Roekel et al., 2019; Schnauber-Stockmann & Karnowski, 2020; Stone et al., 2023; Trull & Ebner-Priemer, 2020). Im Wesentlichen beinhalten diese Vorschläge ähnliche Aspekte wie die Präregistrierung, denn auch in der Forschungsarbeit selbst sollte das Vorgehen genau dokumentiert sein. Hinzu kommen Informationen, die auf den tatsächlichen Studienergebnissen basieren – etwa Fallzahlen auf der Personen- und Situations-

onsebene, die Compliance und Reliabilitätswerte von Skalen. Neben dem Reporting in der Publikation selbst ist es aus Open-Science-Sicht wünschenswert, in einem Repository (z. B. OSF) zusätzlich alle Materialien (z. B. Protokolle, Briefingunterlagen, Vorabfragebogen) sowie die anonymisierten Daten und den Analysecode zu teilen, um der **Reproduzierbarkeit** von Ergebnissen gerecht zu werden. Ein Ergebnis ist reproduzierbar, wenn andere Forschende es mit einer erneuten Analyse der ursprünglichen Daten rekonstruieren können: „Researcher B [...] obtains exactly the same results (e. g., statistics and parameter estimates) that were originally reported by Researcher A [...] from A’s data when following the same [data analysis]“ (Asendorpf et al., 2013, S. 109).

Rechtliche und ethische Hürden können einer Veröffentlichung der Daten im Wege stehen. Diese betreffen insbesondere den Datenschutz bzw. das Recht auf informationelle Selbstbestimmung. Wir müssen also den zu veröffentlichenden Datensatz kritisch daraufhin prüfen, ob er Personendaten enthält und diese ggf. löschen oder anonymisieren. Dies gilt besonders bei so detailreichen und teilweise sehr umfangreichen Datensätzen wie sie im Rahmen von In-situ-Studien entstehen. Bei einer Anonymisierung muss daher auch das Risiko einer nachträglichen Deanonymisierung berücksichtigt werden. Im Zweifelsfall sollte von einer Veröffentlichung der Daten abgesehen werden (siehe auch Fox et al., 2021). Zudem ist es wichtig, auf eine solche Veröffentlichung und die daraus möglicherweise resultierende Zweitverwertung der Daten im Informed Consent hinzuweisen und das explizite Einverständnis der Teilnehmer:innen einzuholen (siehe Abschnitt 4.2, siehe hierzu allgemein Morehouse et al., 2024).

Merke

Die Veröffentlichung der Daten und des Analysecodes einer Studie ermöglicht die Reproduzierbarkeit der Befunde. Vor der Veröffentlichung von Daten sollte eine sorgfältige daten-

10.3 Reporting-Standards und Veröffentlichung von Studienmaterialien und -daten

schutzrechtliche Prüfung stattfinden und die Einwilligung der Teilnehmer:innen vorliegen. Denn auch anonymisierte Daten können sensibel sein – je detailreicher sie sind, desto größer ist das Risiko der Deanonymisierung.

Tabelle 10.1: Inhalte einer Präregistrierung und der Reportingstandards

	Präregistrierung	Reporting
Forschungsfragen und Hypothesen	Auflistung der Forschungsfragen und Hypothesen (siehe Kapitel 3).	
Studiendesign und Feldzeit	Beschreibung des Gesamtsettings der Studie – ggf. Kombination mit anderen Methoden (siehe Abschnitt 5.1).	Zusätzlich: tatsächliche Response Latencies berichten (siehe Abschnitt 9.2.1).
	Design der In-situ-Studie möglichst genau beschreiben und begründen, z. B. Dauer der Feldzeit, Form der Protokollauflösung, erlaubte Zeit zwischen Signal und Ausfüllen; siehe Kapitel 2, Abschnitte 6.1 und 6.3.	
Technische Aspekte	Für die Datenerhebung genutzte Geräte, z. B. nur Android-Geräte.	ggf. in der Feldzeit aufgetretene Probleme, z. B. Smartphones, bzw. Betriebssysteme,
	Genutzte Software bzw. Anbieter für die Datenerhebung, z. B. Sosci-Survey, und Art der Signalsendung und des Ausfüllens, z. B. per SMS, per Notifikation, webbasiert oder appbasiert; siehe Abschnitt 6.3.2.1.	
Teilnehmer:innen	Beschreibung der Zielgruppe.	
Fallzahlen	Beschreibung der Stichprobenziehung und Rekrutierung (siehe Abschnitt 6.2).	Fallzahl auf der Personenebene bei Studienstart und final; Reporting der Ausschlussgründe; zu Bereinigung siehe auch weiter unten sowie Abschnitt 9.2.1.
	angestrebte Fallzahl auf der Personenebene.	

10.3 Reporting-Standards und Veröffentlichung von Studienmaterialien und -daten

	Präregistrierung	Reporting
	Begründung der Stichprobengröße auf der Personen- und Situationsebene, z. B. A-priori-Power-Analyse;* siehe Abschnitt 6.1. Anzahl der geplanten Signale pro Tag oder angenommene Anzahl bei ereignisabhängigem ESM.	Anzahl der gesendeten Signale pro Tag; tatsächlich ausgefüllte Protokolle, die darauf basierende Compliance, ggf. Angaben zu zeitlichen Mustern in der Compliance; siehe Abschnitt 9.2.1.
Protokoll	Anzahl Fragen/Items pro Protokoll. erwartete Antwortzeit pro Protokoll. Dokumentation aller Fragen/Items und Antwortskalen mit Angabe der Quelle, aus der die Fragen/Items stammen; siehe Kapitel 7; Informationen zur geplanten Berechnung von Skalen/Indices und sonstigen geplanten Transformationen; geplante Reliabilitätsmaße; Vorgehen, falls Reliabilitätswerte nicht ausreichend sind etc.; siehe Abschnitt 9.2.2.1.	durchschnittliche Antwortzeit pro Protokoll. Dokumentation aller Fragen/Items und Antwortskalen;* Informationen zur Berechnung von Skalen/Indices und sonstigen Transformatoren, inklusive psychometrischer Informationen, z. B. Reliabilitäten, <i>M</i> , <i>SD</i> etc., und Angabe der Quelle, aus der die Fragen/Items stammen; siehe Kapitel 7, Abschnitte 9.2.2.1 und 9.5.1.
Briefing und Motivation der Teilnehmer:innen	Beschreibung der Schulung von Teilnehmer:innen*; siehe Abschnitt 8.3.	
Ethikvotum und Informed Consent	Beschreibung aller Maßnahmen zur Steigerung der Teilnahme-Motivation, z. B. Incentives; siehe Abschnitt 8.2. falls vorhanden: Informationen zum Ethikvotum; siehe Abschnitt 4.2. Informationen zum forschungsethischen Vorgehen, z. B. Informed Consent, und zum Datenschutz*; siehe Abschnitt 4.1.	
Datenaufbereitung	Vorgehen bei der Datenbereinigung: Ausschluss von Personen, z. B. Mindestcompliance oder Careless Responding bei X Protokollen...; Ausschluss von Protokollen; siehe Abschnitte 7.2 und 9.2.1.	

	Präregistrierung	Reporting
	Umgang mit fehlenden Werten, z. B. Imputationsverfahren; siehe Abschnitt 9.2.2.2.	
	Vorbereitungen für die Auswertung: Art der Zentrierung, Anspiel von zeitversetzten Variablen etc.; siehe Kapitel 9.	
Auswertung		Bericht der bivariaten Korrelationen, within und between, z. B. für Metaanalysen,* siehe Abschnitt 9.3.
	Auswertungsplan: Vorgehen so genau und replizierbar wie möglich beschreiben; darauf eingehen, was gemacht wird, wenn Voraussetzungen nicht erfüllt werden, z. B. Verteilung, Nichtkonvergenz der Modelle etc.; z. B. welche einfacheren Modelle dann berechnet werden; idealerweise Präregistrierung des Auswertungscodes; siehe Kapitel 9.	finaler Auswertungsplan; ggf. auf Probleme eingehen; siehe Kapitel 9.
	Kontrollvariablen mit Begründung; siehe Abschnitt 9.5.2.5.	
	Inferenzkriterien, z. B. p -Wert.	
	Angabe der verwendeten Software, z. B. R, inklusive der verwendeten Pakete; siehe Kapitel 9.	

Anmerkung. In Anlehnung an Schnauber-Stockmann & Karnowski (2020), Stone & Shiffman (2002) und Trull & Ebner-Priemer (2020). Einige Aspekte können beim Reporting in der Publikation auch aus Platzgründen in einen Online-Anhang ausgegliedert werden – vor allem die mit * gekennzeichneten bieten sich hierfür an.

11 Checkliste

Diese knappe Checkliste für eine In-situ-Studie ist als praktische Hilfestellung gedacht, um sich vor Studienbeginn alle wichtigen Aspekte noch einmal zu vergegenwärtigen und auf möglicherweise offene Punkte aufmerksam zu werden. Welche Aspekte relevant für die jeweilige Studie sind und welche Details zu bedenken sind, ist in den einzelnen Kapiteln des Buches ausführlich erläutert.

Wir haben die Checkliste in vier Abschnitte eingeteilt: Unter **Gesamtplanung** gehen wir auf die den gesamten Forschungsprozess begleitenden Themen *zeitliche und finanzielle Rahmenbedingungen*, *Ethik* und *Präregistrierung* sowie *Open Science* ein. Unter **Studienplanung** fallen die Konzeption des *Studiendesigns* und des *In-situ-Protokolls* sowie die der *Datenaufbereitung und Auswertung*, weil letztere zentral für die Bestimmung der Stichprobengröße ist und ggf. auch im Rahmen einer Präregistrierung vor Studienstart festgehalten wird. Zudem geht es um die *Stichprobenplanung* auf Personen- und Situationsebene, die Frage nach *zusätzlichen Messungen*, z. B. der Vorab- und Nachbefragung, sowie die *technische Umsetzung*. Unter **Durchführung** fassen wir den *Pretest*, die *Teilnehmer:innenschulung und Incentivierung* sowie die wichtigsten Aspekte während der *Feldzeit* zusammen. Schließlich gehen wir auf **Datenhandling und Auswertung** ein.

Gesamtplanung

Überblick über zeitliche und finanzielle Rahmenbedingungen

- ✓ Haben wir den zeitlichen Aufwand und Ablauf im Blick?
Dies betrifft z. B. ...

11 Checkliste

- mögliche Vorstudien, z. B. zur Skalen- oder Itemvalidierung
 - ggf. die Präregistrierung
 - ggf. das Einholen eines Ethikvotums
 - die technische Umsetzung (Programmierung)
 - den Pretest
 - die Rekrutierung einer ausreichend großen Personenstichprobe
 - die Feldzeit und ggf. Nachbefragung
- ✓ Haben wir alle anfallenden Kosten im Blick?
Hierunter fallen Kosten für ...
- Rekrutierung (Druck von Flyern, Bezahlung eines Panelanbieters o. ä.)
 - ggf. Incentivierung
 - Technik (App-Nutzung, SMS-Versand)

Ethik

- ✓ Erheben wir personenbezogene Daten, bspw. auch zur Auszahlung von Incentives, und fällt unsere Studie somit unter die DSGVO?
- Wenn ja, ist eine Abstimmung mit der:dem Datenschutzbeauftragten notwendig?
- ✓ Benötigen wir ein Ethikvotum?
- Wenn ja: Wo können wir es beantragen und wie? Welche Informationen über die Studie müssen dafür vorliegen und wie viel Zeit müssen wir dafür einplanen?
- ✓ Was müssen wir bei der Einwilligungserklärung, also dem Informed Consent, bedenken?
Zentrale Aspekte sind z. B. ...
- Datenschutz (z. B. Anonymität, Widerspruchsrecht, verantwortliche Person) und Datensicherheit (z. B. Übermittlung und Zugang)

- Art und Umfang der Datenverwertung (z. B. für Sekundäranalysen)
- Informationen über das Thema, das technische Vorgehen und die Dauer bzw. den Aufwand der jeweiligen In-situ-Studie
- Aufklärung über die Freiwilligkeit der Teilnahme
- geeignete Präsentation der Informationen

Präregistrierung und Open Science

- ✓ Wollen wir unsere Studie präregistrieren?
 - Wenn ja, wo wollen wir die Präregistrierung einreichen, z. B. im OSF?
- ✓ Wollen wir unsere Materialien, z. B. In-situ-Protokolle, Datensätze und unseren Auswertungscode öffentlich zugänglich machen?
Wenn ja, müssen wir ...
 - die Sensibilität unserer Daten abwägen
 - entscheiden, wo wir unsere Materialien, Datensätze und unseren Auswertungscode verfügbar machen wollen, z. B. im OSF

Studienplanung

Studiendesign

- ✓ Welche Art von In-situ-Studie passt am besten zu unseren Forschungsfragen und Hypothesen?

In-situ-Protokoll

- ✓ Wird aus den Frage- und/oder Itemformulierungen klar, auf welchen Zeitraum sich die Antwort beziehen soll, also z. B. den aktuellen Moment, die zurückliegende halbe Stunde o. ä.?

11 Checkliste

- ✓ Was wollen wir alles im Protokoll messen?
Dabei ist zu beachten ...
 - ob es bereits geeignete Kurzskaleten oder Single Items gibt oder wir solche selbst entwickeln müssen (ggf. Skalen-/Item-Pretest einplanen)
 - dass unser Protokoll nicht zu ausführlich ausfällt
 - dass wir bei Filterführung im Protokoll ggf. darauf achten, dass alle Protokollvarianten ähnlich lang sind, um Reaktivitätseffekte zu vermeiden

Datenaufbereitungs- und Auswertungsplanung

- ✓ Wie wollen wir unsere Daten bereinigen?
Darunter fällt z. B. ...
 - der Ausschluss einzelner Protokolle bei zu großer Response Latency
 - der Ausschluss von Teilnehmer:innen bei zu geringer Compliance
- ✓ Mit welchen statistischen Verfahren wollen wir unsere Hypothesen prüfen bzw. unsere Forschungsfragen beantworten?
 - dabei müssen wir die Komplexität der Daten berücksichtigen

Stichprobenplanung

Umfang

- ✓ Welche Fallzahlen auf Personen- und Situationsebene brauchen wir, um die geplanten Analysen vornehmen zu können?

Personenstichprobe

- ✓ Wer ist unsere Grundgesamtheit und was ist unsere Stichprobenziehungsstrategie?

- ✓ Wie gestalten wir die Rekrutierung?
Darunter fallen z. B. ...
 - Überlegungen zu möglichen Kontaktwegen, z. B. Mailinglisten
 - ggf. die Gestaltung von Flyern, Social-Media-Posts o. ä.

Situationsstichprobe

- ✓ Wie viele Tage läuft unsere Studie?
- ✓ Ist eine „Testphase“ geplant, also z. B. ein Tag, der nicht in die Auswertungen eingeschlossen wird?
- ✓ Wie lange nach Signalempfang kann ein Protokoll ausgefüllt werden?
- ✓ Haben wir die Teilnehmer:innenbelastung und unser Forschungsinteresse gut gegeneinander abgewogen?

Bei signalbasierten Designs:

- ✓ Wie viele Signale erhalten die Teilnehmer:innen pro Tag?
- ✓ Legen wir ein Zeitfenster fest, in dem die Signale versendet werden, z. B. von 8–22 Uhr?
- ✓ Legen wir Mindestabstände zwischen Signalen fest?
- ✓ Müssen wir bei den Abständen zwischen zwei Signalen etwas für die spätere Auswertung beachten, z. B. dass diese nicht zu klein oder zu groß ausfallen?

Bei ereignisabhängigen Designs:

- ✓ Legen wir eine Obergrenze an auszufüllenden Protokollen pro Tag oder Zeitabschnitt fest, z. B. bei Ereignissen, die sehr häufig auftreten?
- ✓ Können wir das Protokoll automatisch auslösen lassen, z. B. über eine App?

11 Checkliste

- ✓ Wenn nein: Haben wir das Ereignis gut genug beschrieben, dass die Teilnehmer:innen wissen, wann sie ein Protokoll ausfüllen müssen? Wie können die Teilnehmer:innen auf das Protokoll zugreifen, wenn sie selbstausgelöst ausfüllen müssen?

Zusätzliche Messungen

- ✓ Was müssen wir in einer Vorabbefragung abfragen?
Das betrifft z. B. ...
 - Screeningfragen, die nach Zielgruppenzugehörigkeit fragen
 - unabhängige sowie Kontrollvariablen auf Personenebene
 - die Einholung des Informed Consent
- ✓ Brauchen wir eine Nachbefragung?

Wenn ja:

- Wann soll diese stattfinden? Direkt im Anschluss oder mit Zeitverzögerung?
 - Welche weiteren unabhängigen und Kontrollvariablen auf Personenebene müssen wir hier erheben?
 - Wollen wir nach der Studienerfahrung fragen?
- ✓ Wollen wir zusätzliche Daten erheben, z. B. mittels Tracking?

Technische Umsetzung

- ✓ Setzen wir eine spezielle App ein?

Wenn ja:

- Welche App? Ist sie kostenpflichtig?
- Sind wir mit ihrer Funktionsfähigkeit ausreichend gut vertraut?
- Gehen mit der App Anforderungen an die technische Ausstattung der Teilnehmer:innen einher?
- Haben diese Voraussetzungen evtl. einen Einfluss auf die Stichprobe, etwa durch den Ausschluss eines bestimmten Betriebs-

systems?

Wenn nein:

- Welches andere Vorgehen wählen wir, um die Teilnehmer:innen zu alarmieren und ihnen die Protokolle zugänglich zu machen, z. B. SMS-Versand?

Durchführung

Pretest

- ✓ Funktioniert unser Studiensetup technisch einwandfrei?
Hierzu gehören v. a. ...
 - die Signalversendung, wenn nicht selbstausgelöst, und damit verbunden der Aufruf des In-situ-Protokolls
 - die Verknüpfbarkeit der einzelnen Studienteile, z. B. Vorabbefragung und In-situ-Protokolle
 - die „Übergabe“ von der Vorabbefragung zum In-situ-Teil, z. B. Onboarding in eine App oder Abfrage von Mobilfunknummer und SMS-Versand
 - die Filterführung
- ✓ Ist unser In-situ-Protokoll, und mögliche andere Studienteile, fehlerfrei und verständlich?

Schulung

- ✓ Was muss in die Schulungsunterlagen für die Teilnehmer:innen aufgenommen werden?
- ✓ Wie stellen wir die Verfügbarkeit der Schulungsunterlagen sicher?
 - z. B. Versand per E-Mail, Abrufbarkeit über eine Projektwebseite

11 Checkliste

- ✓ Benötigen wir eine persönliche, virtuelle oder Präsenz-Schulung oder reichen schriftliche Unterlagen bzw. Videos o. ä. aus?
 - Wenn wir eine persönliche Schulung benötigen: wo und wie können wir diese durchführen?

Incentivierung und Teilnehmer:innenmotivation

- ✓ Wie incentivieren wir unsere Teilnehmer:innen?
 - monetär oder anderweitig?
- ✓ Wie können wir die Incentives verteilen?
 - z. B. an alle Teilnehmer:innen, die eine Mindest-Compliance erfüllt haben oder in Form einer Verlosung
- ✓ Welche weiteren Maßnahmen zur Steigerung der Compliance planen wir?
 - z. B. Reminder, Motivationsnachrichten

Feldzeit

- ✓ Wie stellen wir sicher, dass unsere Teilnehmer:innen uns während der Feldzeit erreichen?
 - z. B. über eine Studien-E-Mail oder die eingesetzte App
- ✓ Wie können wir den Studienverlauf im Auge behalten, also Feldmonitoring betreiben, d. h. kontrollieren, wie viele Teilnehmer:innen noch teilnehmen, wie hoch die Compliance ist etc.?
 - z. B. durch Teilnahme eines Team-Mitglieds an der Studie

Datenhandling und Auswertung

Datenhandling

- ✓ Haben wir sichergestellt, dass nicht benötigte, v. a. personenbezogene Daten, z. B. zur Incentivierung, zeitnah gelöscht werden?
- ✓ Sehen alle Daten vollständig und plausibel aus? (Dateninspektion)
- ✓ Haben wir alle Datenbereinigungsschritte wie geplant durchgeführt?
 - Wenn nicht, sollten wir Abweichungen von unserem Plan dokumentieren und begründen, v. a. wenn die Studie präregistriert wurde

Auswertung

- ✓ Haben wir die Stichprobe transparent beschrieben? (deskriptive Auswertungen)
- ✓ Haben wir alle hypothesenprüfenden bzw. forschungsfragenbeantwortenden Analysen wie geplant durchgeführt?
 - Wenn nicht, sollten wir Abweichungen von unserem Plan dokumentieren und begründen, v. a. wenn die Studie präregistriert wurde
- ✓ Ist es sinnvoll, weitere explorative Auswertungen durchzuführen?

Literaturverzeichnis

- Aalbers, G., Abeele, M. M. P. vanden, Hendrickson, A. T., Marez, L. de, & Keijsers, L. (2022). Caught in the moment: Are there person-specific associations between momentary procrastination and passively measured smartphone use? *Mobile Media & Communication*, 10(1), 115–135. <https://doi.org/10.1177/2050157921993896>
- Allison, P. D. (2009). *Fixed effects regression models*. SAGE Publications, Inc.
- Allison, P. D. (2015). *Don't put lagged dependent variables in mixed models*. <https://statisticalhorizons.com/lagged-dependent-variables/>
- Alphen, T. van, Jak, S., Jansen in de Wal, J., Schuitema, J., & Peetsma, T. (2022). Determining reliability of daily measures: An illustration with data on teacher stress. *Applied Measurement in Education*, 35(1), 63–79. <https://doi.org/10.1080/08957347.2022.2034822>
- Appel, M., Gnambs, T., Richter, T., & Green, M. C. (2015). The Transportation Scale–Short Form (TS–SF). *Media Psychology*, 18(2), 243–266. <https://doi.org/10.1080/15213269.2014.987400>
- Araujo, T., Ausloos, J., Atteveldt, W. van, Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., Velde, B. van de, Vreese, C. de, & Welbers, K. (2022). OSD2F: An Open-Source Data Donation Framework. *Computational Communication Research*, 4(2), 372–387. <https://doi.org/10.5117/CCR2022.2.001.ARAU>

- Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, 24(1), 1–19. <https://doi.org/10.1037/met0000195>
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2), 108–119. <https://doi.org/10.1002/per.1919>
- Bachl, M. (2014). *Analyse rezeptionsbegleitend gemessener Kandidatenbewertungen in TV-Duellen*. epubli GmbH.
- Bäckström, M., & Björklund, F. (2021). Socially desirable responding in experience sampling: Consequences for personality research. *Journal of Individual Differences*, 42(4), 157–166. <https://doi.org/10.1027/1614-0001/a000342>
- Barta, W. D., Tennen, H., & Litt, M. D. (2012). Measurement reactivity in diary research. In M. R. Mehl & T. S. Conner (Hrsg.), *Handbook of Research Methods for Studying Daily Life* (S. 108–123). The Guilford Press.
- Bartolucci, F., Pennoni, F., & Farcomeni, A. (2012). *Latent Markov models for longitudinal data*. CRC Press, Taylor & Francis Group.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Baumgartner, S. E., Sumter, S. R., Petkevič, V., & Wiradhany, W. (2022). A novel iOS data donation approach: automatic processing, compliance, and reactivity in a longitudinal study. *Social Science Computer Review*, 41(4), 1456–1472. <https://doi.org/10.>

1177/08944393211071068

- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311. <https://doi.org/10.1093/poq/nfm006>
- Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8. Aufl.). Oxford University Press.
- Berger, U., Fehlinger, M., Mühleck, J., Wick, K., & Schwager, S. (2019). Inklusive Forschung: Validierung der Skala zur Allgemeinen Selbstwirksamkeitserwartung (SWE) in Leichter Sprache an einer Stichprobe von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf. *PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, 69(09/10), 398–406. <https://doi.org/10.1055/a-0831-2270>
- Berkel, N. van, Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys*, 50(6), 1–40. <https://doi.org/10.1145/3123988>
- Berkel, N. van, Goncalves, J., Lovén, L., Ferreira, D., Hosio, S., & Kostakos, V. (2019). Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies*, 125, 118–128. <https://doi.org/https://doi.org/10.1016/j.ijhcs.2018.12.002>
- Beyens, I., Pouwels, J. L., Van Driel, I. I., Keijsers, L., & Valkenburg, P. M. (2024). Social media use and adolescents' well-being: developing a typology of person-specific effect patterns. *Communication Research*, 51(6), 691–716. <https://doi.org/10.1177/00936502211038196>
- Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5), 817–848. <https://doi.org/10.1093/poq/nfn006>

[//doi.org/10.1093/poq/nfq058](https://doi.org/10.1093/poq/nfq058)

- Bigl, B., Gognelashvili, K., & Gehrau, V. (2024). Forschungsethik, Barrierefreiheit und inklusive Forschung in der Kommunikationswissenschaft. *Publizistik*, 69(3), 299–332. <https://doi.org/10.1007/s11616-024-00854-7>
- Blair, E., & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research*, 14(2), 280. <https://doi.org/10.1086/209112>
- Boase, J., & Ling, R. (2013). Measuring mobile phone use: self-report versus log data. *Journal of Computer-Mediated Communication*, 18(4), 508–519. <https://doi.org/10.1111/jcc4.12021>
- Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, 4(2), 388–423. <https://doi.org/10.5117/CCR2022.2.002.BOES>
- Boeschoten, L., De Schipper, N. C., Mendrik, A. M., Van Der Veen, E., Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, 8(90), 5596. <https://doi.org/10.21105/joss.05596>
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54(1), 579–616. <https://doi.org/10.1146/annurev.psych.54.101601.145030>
- Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: an introduction to diary and experience sampling research*. Guilford Press.
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozial-*

wissenschaftler (7. Aufl.). Springer.

- Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, 236(4798), 157–161. <https://doi.org/10.1126/science.3563494>
- Breuer, J., Baghal, T. A., Sloan, L., Bishop, L., Kondyli, D., & Linnardis, A. (2021). Informed consent for linking survey and social media data - Differences between platforms and data types. *IAS-SIST Quarterly*, 45(1). <https://doi.org/10.29173/iq988>
- Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, 22(11), 2058–2080. <https://doi.org/10.1177/1461444820924622>
- Brosius, H.-B., Haas, A., & Unkel, J. (2022). *Methoden der empirischen Kommunikationsforschung: Eine Einführung*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-34195-4>
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217. <https://doi.org/10.1037/h0047470>
- Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly*, 55(1), 50–79. <https://doi.org/10.1086/269241>
- Calamia, M. (2019). Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychological Assessment*, 31(3), 285–291. <https://doi.org/10.1037/pas0000599>

- Capon, H., Hall, W., Fry, C., & Carter, A. (2016). Realising the technological promise of smartphones in addiction research and treatment: An ethical review. *International Journal of Drug Policy*, 36, 47–57. <https://doi.org/10.1016/j.drugpo.2016.05.013>
- Carrière, T. C., Boeschoten, L., Struminskaya, B., Janssen, H. L., De Schipper, N. C., & Araujo, T. (2025). Best practices for studies using digital data donation. *Quality & Quantity*, 59(Suppl 1), 389–412. <https://doi.org/10.1007/s11135-024-01983-x>
- Castro-Alvarez, S., Bringmann, L. F., Back, J., & Liu, S. (2026). The many reliabilities of psychological dynamics: An overview of statistical approaches to estimate the internal consistency reliability of intensive longitudinal data. *Psychological Methods*, 31(2), 281–296. <https://doi.org/10.1037/met0000778>
- Chang, L., & Krosnick, J. A. (2003). Measuring the frequency of regular behaviors: Comparing the “typical week” to the “past week”. *Sociological Methodology*, 33(1), 55–80. <https://doi.org/10.1111/j.0081-1750.2003.t01-1-00127.x>
- Chan-Olmsted, S., & Wang, R. (2022). Understanding podcast users: Consumption motives and behaviors. *New Media & Society*, 24(3), 684–704. <https://doi.org/10.1177/1461444820963776>
- Chatfield, K., & Dove, E. (2023). Clarifying our policy on requiring ethics approval in submitted manuscripts. *Research Ethics*, 19(2), 103–106. <https://doi.org/10.1177/17470161231161402>
- Cheng, A., Zamarro, G., & Orriens, B. (2016). Personality as a predictor of unit nonresponse in panel data: An analysis of an internet-based survey. *SSRN Electronic Journal*, 49(3), 672–698. <https://doi.org/10.2139/ssrn.2852614>
- Cho, G., Pasquini, Giancarlo., & Scott, S. B. (2019). Measurement

- burst designs in lifespan developmental research. In F. C. Worrell (Hg.), *Oxford Research Encyclopedia of Psychology*. Oxford Academic. <https://doi.org/10.1093/acrefore/9780190236557.013.348>
- Christner, C., Urman, A., Adam, S., & Maier, M. (2022). Automated tracking approaches for studying online media use: A critical review and recommendations. *Communication Methods and Measures*, 16(2), 79–95. <https://doi.org/10.1080/19312458.2021.1907841>
- Coenen, L. (2022). The indirect effect is omitted variable bias. A cautionary note on the theoretical interpretation of products-of-coefficients in mediation analyses. *European Journal of Communication*, 37(6), 679–688. <https://doi.org/10.1177/02673231221082244>
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, 104(10), 1243–1265. <https://doi.org/10.1037/apl0000406>
- Conner, T. S., & Lehman, B. J. (2012). Getting started: Launching a study in daily life. In M. R. Mehl & T. S. Conner (Hrsg.), *Handbook of Research Methods for Studying Daily Life*. (S. 89–107). The Guilford Press.
- Conner, T. S., & Reid, K. A. (2012). Effects of intensive mobile happiness reporting in daily life. *Social Psychological and Personality Science*, 3(3), 315–323. <https://doi.org/10.1177/1948550611419677>
- Couper, M. P., Tourangeau, R., Conrad, F. G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A web experiment. *Social Science Computer Review*, 24(2), 227–245. <https://doi.org/10.1177/0898010106287111>

//doi.org/10.1177/0894439305281503

- Courvoisier, D. S., Eid, M., & Lischetzke, T. (2012). Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics. *Psychological Assessment*, 24(3), 713–720. <https://doi.org/10.1037/a0026733>
- Craig, C. M., Brooks, M. E., & Bichard, S. (2023). Podcasting on purpose: Exploring motivations for podcast use among young adults. *International Journal of Listening*, 37(1), 39–48. <https://doi.org/10.1080/10904018.2021.1913063>
- Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, 32(7), 917–929. <https://doi.org/10.1177/0146167206287721>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Csikszentmihalyi, M., & Kubey, R. (1981). Television and the rest of life: A systematic comparison of subjective experience. *Public Opinion Quarterly*, 45(3), 317–328. <https://doi.org/10.1086/268667>
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease*, 175(9), 526–536. <https://doi.org/10.1097/00005053-198709000-00004>
- Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In M. Csikszentmihalyi (Hg.), *Flow and the foundations of positive psychology* (S. 35–54). Sprin-

- ger Netherlands. https://doi.org/10.1007/978-94-017-9088-8_3
- Cursio, J. F., Mermelstein, R. J., & Hedeker, D. (2019). Latent trait shared-parameter mixed models for missing ecological momentary assessment data. *Statistics in Medicine*, 38(4), 660–673. <https://doi.org/10.1002/sim.7989>
- Dal Cin, S., Zanna, M. P., & Fong, G. T. (2004). Narrative persuasion and overcoming resistance. In E. S. Knowles & J. A. Linn (Hrsg.), *Resistance and Persuasion* (S. 175–191). Lawrence Erlbaum Associates.
- De Calheiros Velozo, J., Niemeijer, K., & Vaessen, T. (2022). Passive sensing. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 251–262). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- DeHaven, A. (2017). *Preregistration: A plan, not a prison*. <https://www.cos.io/blog/preregistration-plan-not-prison>
- Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment*, 34(12), 1138–1154. <https://doi.org/10.1037/pas0001178>
- Desjarlais, M. (2024). Subtle momentary effects of social media experiences: An experience sampling study of posting and social comparisons on connectedness and self-esteem. *Journal of Computer-Mediated Communication*, 29(3), zmae004. <https://doi.org/10.1093/jcmc/zmae004>
- Deutsche Forschungsgemeinschaft. (2025). *Leitlinien zur Sicherung*

- guter wissenschaftlicher Praxis: Kodex*. <https://doi.org/10.5281/ZENODO.14281892>
- Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, 40(3), 434–449. <https://doi.org/10.1007/s11747-011-0300-3>
- Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., ... Vreese, C. de. (2021). An agenda for open science in communication. *Journal of Communication*, 71(1), 1–26. <https://doi.org/10.1093/joc/jqz052>
- Dogruel, L., & Schnauber-Stockmann, A. (2020). What determines instant messaging communication? Examining the impact of person- and situation-level factors on IM responsiveness. *Mobile Media & Communication*, 9(2), 210–228. <https://doi.org/10.1177/2050157920943926>
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-41089-5>
- Driel, I. I. van, Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, 16(4), 266–282. <https://doi.org/10.1080/19312458.2022.2109608>
- Driver, C. C., & Voelkle, M. C. (2018). Hierarchical Bayesian continuous time dynamic modeling. *Psychological Methods*, 23(4), 774–799. <https://doi.org/10.1037/met0000168>

- Eble, M., Ziegele, M., & Jürgens, P. (2014). Forschung in geschlossenen Plattformen des Social Web. In M. Welker, M. Taddicken, J.-H. Schmidt, & N. Jakob (Hrsg.), *Handbuch Online-Forschung* (S. 123–149). von Halem.
- Edwards, P. J., Roberts, I., Clarke, M. J., DiGuseppi, C., Woolf, B., & Perkins, C. (2023). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews*, 2023(11). <https://doi.org/10.1002/14651858.MR000008.pub5>
- Eisele, G., Hiekkaranta, A., Kunkels, Y. K., Rot, M. A. H., Van Ballegooijen, W., Bartels, S. L., Bastiaansen, J. A., Beymer, P. N., Bylsma, L. M., Carpenter, R. W., Ellison, W. D., Fisher, A. J., Forkmann, T., Frumkin, M. R., Fulford, D., Naragon-Gainey, K., Greene, T., Heininga, V. E., Jones, A., ... Kirtley, O. J. (2025). ESM-Q: A consensus-based quality assessment tool for experience sampling method items. *Behavior Research Methods*, 57(4), 124. <https://doi.org/10.3758/s13428-025-02626-1>
- Eisele, G., Kasanova, Z., & Houben, M. (2022). Questionnaire design and evaluation. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 71–90). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, 29(2), 136–151. <https://doi.org/10.1177/1073191120957102>
- Enders, C. K. (2022). *Applied missing data analysis* (2. Aufl.). The

Guilford Press.

- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138. <https://doi.org/10.1037/1082-989X.12.2.121>
- Ernst, A., Dietrich, F., Rohr, B., Reinecke, L., & Scharkow, M. (2026). Revisiting the digital jukebox in daily life: applying Mood Management Theory to algorithmically curated music streaming environments. *Human Communication Research*, hqaf030. <https://doi.org/10.1093/hcr/hqaf030>
- Ernst, A., & Schnauber-Stockmann, A. (2026). Won't stop 'til you get enough? Determinants of disengaging from mobile media apps in daily life. *Communication Research*, 53(5), 659–686. <https://doi.org/10.1177/00936502251378582>
- Ess, C. M. (2019). Internet research ethics and social media. In R. Iphofen (Hg.), *Handbook of Research Ethics and Scientific Integrity*. Springer International Publishing. https://doi.org/10.1007/978-3-319-76040-7_12-1
- Eveland, W. P. Jr., & Morey, A. C. (2010). Challenges and opportunities of panel designs. In E. P. Bucy & R. L. Holbert (Hrsg.), *Sourcebook for Political Communication Research: Methods, Measures, and Analytical Techniques* (S. 19–33). Routledge.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Faulbaum, F. (2019). *Methodische Grundlagen der Umfrageforschung*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-319-76040-7_12-1

1007/978-3-531-93278-1

- Fecke, M., Fehr, A., & Schlütz, D. (2024). Die Mobile Experience Sampling Methode (MESM) in der Kinder- und Jugendmedienforschung: Chancen und Herausforderungen. *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, 60, 1–21. <https://doi.org/10.21240/mpaed/60/2024.04.21.X>
- Felderer, B., Birg, A., & Kreuter, F. (2019). Paradata. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 413–422). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_29
- Field, A. (2026). *Discovering Statistics Using R and RStudio* (2. Aufl.). SAGE Publications Ltd.
- Fox, J., Pearce, K. E., Massanari, A. L., Riles, J. M., Szulc, L., Ranjit, Y. S., Trevisan, F., Soriano, C. R. R., Vitak, J., Arora, P., Ahn, S. J. (Grace), Alper, M., Gambino, A., Gonzalez, C., Lynch, T., Williamson, L. D., & L. Gonzales, A. (2021). Open science, closed doors? Countering marginalization through an agenda for ethical, inclusive research in communication. *Journal of Communication*, 71(5), 764–784. <https://doi.org/10.1093/joc/jqab029>
- Franzke, A. S., Muis, I., & Schäfer, M. T. (2021). Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands. *Ethics and Information Technology*, 23(3), 551–567. <https://doi.org/10.1007/s10676-020-09577-5>
- Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., & Bringmann, L. F. (2024). So you want to do ESM? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Sci-*

- ence, 7(3). <https://doi.org/10.1177/25152459241267912>
- Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., & Mills, J. (2013). Does the burden of the experience sampling method undermine data quality in state body image research? *Body Image*, 10(4), 607–613. <https://doi.org/10.1016/j.bodyim.2013.06.003>
- Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2018). Experience sampling methods: a discussion of critical trends and considerations for scholarly advancement. *Organizational Research Methods*, 22(4), 969–1006. <https://doi.org/10.1177/1094428118802626>
- Gajic, A., Cameron, D., & Hurley, J. (2012). The cost-effectiveness of cash versus lottery incentives for a web-based, stated-preference community survey. *The European Journal of Health Economics*, 13(6), 789–799. <https://doi.org/10.1007/s10198-011-0332-0>
- Gehrau, V., Maubach, K., & Fujarski, S. (2022). *Einfache Datenauswertung mit R: Eine Einführung in uni- und bivariate Statistik sowie Datendarstellung mit RStudio und R Markdown*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-34285-2>
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19(1), 72–91. <https://doi.org/10.1037/a0032138>
- Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press.
- Goodboy, A. K., & Martin, M. M. (2020). Omega over alpha for reliability estimation of unidimensional communication measures.

- Annals of the International Communication Association*, 44(4), 422–439. <https://doi.org/10.1080/23808985.2020.1846135>
- Göritz, A. S. (2008). The long-term effect of material incentives on participation in online panels. *Field Methods*, 20(3), 211–225. <https://doi.org/10.1177/1525822X08317069>
- Göritz, A. S., & Neumann, B. P. (2016). The longitudinal effects of incentives on response quantity in online panels. *Translational Issues in Psychological Science*, 2(2), 163–173. <https://doi.org/10.1037/tps0000071>
- Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56(4), 475. <https://doi.org/10.1086/269338>
- Groves, R. M., & Couper, M. (1998). *Nonresponse in Household Interview Surveys*. Wiley.
- Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879. <https://doi.org/10.1093/poq/nfq065>
- 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. <https://doi.org/10.1177/1094428117703686>
- Ha, L., Hu, X., Fang, L., Henize, S., Park, S., Stana, A., & Zhang, X. (2015). Use of survey research in top mass communication journals 2001–2010 and the total survey error paradigm. *Review of Communication*, 15(1), 39–59. <https://doi.org/10.1080/15358593.2015.1014401>
- Hagenaars, J., & Mccutcheon, A. (2003). Applied latent class analy-

- sis models. *Canadian Journal of Sociology / Cahiers canadiens de sociologie*, 28(4), 584–586. <https://doi.org/10.2307/3341848>
- Haim, M., Leiner, D., & Hase, V. (2023). Integrating data donations in online surveys. *Medien & Kommunikationswissenschaft*, 71(1-2), 130–137. <https://doi.org/10.5771/1615-634X-2023-1-2-130>
- Hall, J. A., Kearney, M. W., & Xing, C. (2019). Two tests of social displacement through social media use. *Information, Communication & Society*, 22(10), 1396–1413. <https://doi.org/10.1080/1369118X.2018.1430162>
- Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, 11(6), 838–854. <https://doi.org/10.1177/1745691616650285>
- Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, 18, 83–90. <https://doi.org/10.1016/j.cobeha.2017.07.018>
- Hartmann, T., & Reinecke, L. (2013). Skalenkonstruktion in der Kommunikationswissenschaft. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 41–60). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18776-1_3
- Hase, V., & Haim, M. (2024). Can we get rid of bias? Mitigating systematic error in data donation studies through survey design strategies. *Computational Communication Research*, 6(2), 1. <https://doi.org/10.5117/ccr2024.2.2.hase>
- Heise, N., & Schmidt, J.-H. (2014). Ehtik der Online-Forschung. In

- M. Welker, M. Taddicken, J.-H. Schmidt, & N. Jakob (Hrsg.), *Handbuch Online-Forschung* (S. 519–539). von Halem.
- Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience Sampling Method: Measuring the Quality of Everyday Life*. Sage Publications.
- Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of Cluster Analysis*. CRC Press.
- Hoplamazian, G. J., Dimmick, J., Ramirez, A., & Feaster, J. (2018). Capturing mobility: The time–space diary as a method for assessing media use niches. *Mobile Media & Communication*, 6(1), 127–145. <https://doi.org/10.1177/2050157917731484>
- Hox, J. J., Moerbeek, M., & Schoot, R. van de. (2017). *Multilevel Analysis: Techniques and Applications* (3. Aufl.). Routledge. <https://doi.org/10.4324/9781315650982>
- Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., Neale, M. C., Sisk, C. L., & Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling: a Multidisciplinary Journal*, 23(4), 532–543. <https://doi.org/10.1080/10705511.2016.1148605>
- Hyers, L. (2018). *Diary Methods* (Bd. 1). Oxford University Press. <https://doi.org/10.1093/oso/9780190256692.001.0001>
- Jäckle, A., & Lynn, P. (2008). Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias. *Survey Methodology*, 34(1), 105–117.
- Jacobson, N. C., Bentley, K. H., Walton, A., Wang, S. B., Fortgang, R. G., Millner, A. J., Coombs, G., Rodman, A. M., & Coppersmith, D. D. L. (2020). Ethical dilemmas posed by mo-

- bile health and machine learning in psychiatry research. *Bulletin of the World Health Organization*, 98(4), 270–276. <https://doi.org/10.2471/BLT.19.237107>
- Jaso, B. A., Kraus, N. I., & Heller, A. S. (2022). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*, 27(6), 958–981. <https://doi.org/10.1037/met0000312>
- Ji, L., Chow, S.-M., Schermerhorn, A. C., Jacobson, N. C., & Cummings, E. M. (2018). Handling missing data in the modeling of intensive longitudinal data. *Structural Equation Modeling: a Multidisciplinary Journal*, 25(5), 715–736. <https://doi.org/10.1080/10705511.2017.1417046>
- Johannes, N., Meier, A., Reinecke, L., Ehlert, S., Setiawan, D. N., Walasek, N., Dienlin, T., Buijzen, M., & Veling, H. (2021). The relationship between online vigilance and affective well-being in everyday life: Combining smartphone logging with experience sampling. *Media Psychology*, 24(5), 581–605. <https://doi.org/10.1080/15213269.2020.1768122>
- Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, 114(4), 609–619. <https://doi.org/10.1111/add.14503>
- Kahlor, L., Dunwoody, S., Griffin, R. J., Neuwirth, K., & Giese, J. (2003). Studying heuristic-systematic processing of risk communication. *Risk Analysis*, 23(2), 355–368. <https://doi.org/10.1111/1539-6924.00314>
- Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., &

- Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702), 1776–1780. <https://doi.org/10.1126/science.1103572>
- Karnowski, V., Knop-Huelss, K., & Olbermann, Z. (2024). Mobile news access, mobile news repertoires, and users' tendency to talk about the news – an experience sampling study on mobile news consumption. *Online Media and Global Communication*, 3(2), 213–234. <https://doi.org/10.1515/omgc-2023-0051>
- Karnowski, V., Kämpel, A. S., Leonhard, L., & Leiner, D. J. (2017). From incidental news exposure to news engagement: How perceptions of the news post and news usage patterns influence engagement with news articles encountered on Facebook. *Computers in Human Behavior*, 76, 42–50. <https://doi.org/10.1016/j.chb.2017.06.041>
- Karnowski, V., Naab, T. K., & Schlütz, D. (2019). On the challenges of measuring mobile social media use: Explaining differences between data from surveys and mobile experience sampling. In C. Peter, T. K. Naab, & R. Kühne (Hrsg.), *Measuring Media Use and Exposure – Recent Challenges and Developments* (S. 59–89). von Halem.
- Karnowski, V., & Von Pape, T. (2009). MobileTV im Alltag der Nutzer: Ergebnisse einer dreimonatigen Panelstudie. In J. Krone (Hg.), *Fernsehen im Wandel. Mobile TV & IPTV in Deutschland und Österreich* (S. 241–257). Nomos. <https://doi.org/10.5771/9783845217901-241>
- Karsay, K., Matthes, J., Schmuck, D., & Ecklebe, S. (2023). Messaging, posting, and browsing: A mobile experience sampling study investigating youth's social media use, affective well-being, and loneliness. *Social Science Computer Review*, 41(4), 1493–1513. <https://doi.org/10.1177/08944393211058308>

- Keusch, F. (2015). Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. *Management Review Quarterly*, 65(3), 183–216. <https://doi.org/10.1007/s11301-014-0111-y>
- Keusch, F., & Conrad, F. G. (2022). Using smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and Methodology*, 10(4), 863–885. <https://doi.org/10.1093/jssam/smab035>
- Keusch, F., & Kreuter, F. (2021). Digital trace data. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Hrsg.), *Handbook of Computational Social Science, Volume 1* (1. Aufl., S. 100–118). Routledge. <https://doi.org/10.4324/9781003024583-8>
- 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. <https://doi.org/10.1093/poq/nfz007>
- Kibuchi, E., Sturgis, P., Durrant, G. B., & Maslovskaya, O. (2020). Do interviewers moderate the effect of monetary incentives on response rates in household interview surveys? *Journal of Survey Statistics and Methodology*, 8(2), 264–284. <https://doi.org/10.1093/jssam/smy026>
- Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, 37(2), 214–233. <https://doi.org/10.1177/0894439317752406>
- Kimhy, D., Vakhrusheva, J., Jobson-Ahmed, L., Tarrier, N., Malaspina, D., & Gross, J. J. (2012). Emotion awareness and regulation in individuals with schizophrenia: Implications for so-

- cial functioning. *Psychiatry Research*, 200(2-3), 193–201. <https://doi.org/10.1016/j.psychres.2012.05.029>
- Kirtley, O. J. (2022). Ethical issues in experience sampling method research. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 91–101). The center for Research on Experience sampling; Ambulatory methods Leuven (REAL).
- Kirtley, O. J., Lafit, G., Achterhof, R., Hiekkaranta, A. P., & Myin-Germeys, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Advances in Methods and Practices in Psychological Science*, 4(1). <https://doi.org/10.1177/2515245920924686>
- Kline, R. B. (2015). The Mediation Myth. *Basic and Applied Social Psychology*, 37(4), 202–213. <https://doi.org/10.1080/01973533.2015.1049349>
- Klingelhoef, J., Gilbert, A., Adrian, C., & Meier, A. (2026). Possible futures all at once: time frame and time lag in short-term longitudinal media effects research on well-being. *Journal of Communication*, 76(1), 78–91. <https://doi.org/10.1093/joc/jqaf037>
- Klingelhoef, J., Gilbert, A., & Meier, A. (2024). Momentary motivations for digital disconnection: An experience sampling study. *Journal of Computer-Mediated Communication*, 29(5), zmae013. <https://doi.org/10.1093/jcmc/zmae013>
- Knobloch, S. (2003). Mood adjustment via mass communication. *Journal of Communication*, 53(2), 233–250. <https://doi.org/10.1111/j.1460-2466.2003.tb02588.x>
- Koch, C., Frei, N. K., & Herrmann, I. (2021). Das Smartphone-

- Tagebuch in der Mediennutzungsforschung. Ein qualitativer Methodenansatz unter der Lupe. *Medien & Kommunikationswissenschaft*, 69(2), 299–319. <https://doi.org/10.5771/1615-634X-2021-2-299>
- Koch, T., Seeger, C., & Müller, P. (2025). *Das Experiment in der Kommunikations- und Medienwissenschaft*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-49131-4>
- König, L. M., Allmeta, A., Christlein, N., Van Emmenis, M., & Sutton, S. (2022). A systematic review and meta-analysis of studies of reactivity to digital in-the-moment measurement of health behaviour. *Health Psychology Review*, 16(4), 551–575. <https://doi.org/10.1080/17437199.2022.2047096>
- Korotitsch, W. J., & Nelson-Gray, R. O. (1999). An overview of self-monitoring research in assessment and treatment. *Psychological Assessment*, 11(4), 415–425. <https://doi.org/10.1037/1040-3590.11.4.415>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236. <https://doi.org/10.1002/acp.2350050305>
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50(1), 537–567. <https://doi.org/10.1146/annurev.psych.50.1.537>
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, 47(4), 2025–2047. <https://doi.org/10.1007/s11135-011-9640-9>
- Kubey, R., & Csikszentmihalyi, M. (1990). *Television and the Quality of Life: How Viewing Shapes Everyday Experience*. Erlbaum.

- Kubey, R., Larson, R., & Csikszentmihalyi, M. (1996). Experience sampling method applications to communication research questions. *Journal of Communication*, 46(2), 99–120. <https://doi.org/10.1111/j.1460-2466.1996.tb01476.x>
- Kupferschmitt, T. (2024). ARD/ZDF-Medienstudie 2024: Trends bei Video- und Audioplattformen. *Media Perspektiven*, (26), 1–9.
- Kuppens, P., & Myin-Germeys, I. (2022). Research questions that can be answered with ESM research. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 21–32). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Lafit, G. (2022). Non-normal, higherlevel, and VAR(1) models for the analysis of ESM data. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 185–216). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Lafit, G., Artner, R., & Ceulemans, E. (2024). Enabling analytical power calculations for multilevel models with autocorrelated errors through deriving and approximating the precision matrix. *Behavior Research Methods*, 56(7), 8105–8131. <https://doi.org/10.3758/s13428-024-02435-y>
- Lahman, M. (2018). *Ethics in Social Science Research: Becoming Culturally Responsive*. SAGE.
- Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, 26(1), 90–102. <https://doi.org/10.1037/met0000287>

- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), 33267. <https://doi.org/10.1525/collabra.33267>
- Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, 10(1). <https://doi.org/10.1525/collabra.117094>
- Langener, A. M., Siepe, B. S., Elsherif, M., Niemeijer, K., Andresen, P. K., Akre, S., Bringmann, L. F., Cohen, Z. D., Choukas, N. R., Drexler, K., Fassi, L., Green, J., Hoffmann, T., Jagesar, R. R., Kas, M. J. H., Kurten, S., Schoedel, R., Stulp, G., Turner, G., & Jacobson, N. C. (2024). A template and tutorial for preregistering studies using passive smartphone measures. *Behavior Research Methods*, 56(8), 8289–8307. <https://doi.org/10.3758/s13428-024-02474-5>
- Langer, W. (2009). *Mehrebenenanalyse: Eine Einführung für Forschung und Praxis* (2. Aufl.). VS Verl. für Sozialwiss.
- Larroulet Philippi, C. (2021). On measurement scales: Neither ordinal nor interval? *Philosophy of Science*, 88(5), 929–939. <https://doi.org/10.1086/714873>
- Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. In Harry T. Reis (Hg.), *Naturalistic Approaches to Studying Social Interaction* (S. 41–56). Jossey-Bass.
- Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The People's Choice: How The Voter Makes up his Mind in a Presidential Campaign*. Columbia University Press.
- Lee, C., Hornik, R., & Hennessy, M. (2008). The reliability and stability of general media exposure measures. *Communication Methods and Measures*, 2(1-2), 6–22. <https://doi.org/10.1080/19312450802063024>

- Leiner, D. (2024). *SoSci Survey (Version 3.5.01)*. <https://www.soscisurvey.de>
- Leonelli, S., Lovell, R., Wheeler, B. W., Fleming, L., & Williams, H. (2021). From FAIR data to fair data use: Methodological data fairness in health-related social media research. *Big Data & Society*, 8(1). <https://doi.org/10.1177/20539517211010310>
- Lewin, K. (1939). Field theory and experiment in social psychology: Concepts and methods. *American Journal of Sociology*, 44(6), 868–896. <https://doi.org/10.1086/218177>
- Lin, X., Mermelstein, R., & Hedeker, D. (2018). A shared parameter location scale mixed effect model for EMA data subject to informative missing. *Health Services and Outcomes Research Methodology*, 18(4), 227–243. <https://doi.org/10.1007/s10742-018-0184-5>
- Link, E. (2026). Single episodes of health information seeking, scanning, and avoidance: Findings of an experience sampling methods study of German residents suffering from acute or chronic illness. *Communication Research*, 53(2), 151–178. <https://doi.org/10.1177/00936502251328717>
- Locatelli, E., Franzke, A. S., & Ess, C. M. (2023). Introduction. Over twenty years of internet research ethics: key concepts and future challenges. *Comunicazioni Sociali*, (2), 131–138. <https://publicatt.unicatt.it/handle/10807/270782>
- Longhi, S. (2017). Panel research methods. In J. Matthes, C. S. Davis, & R. F. Potter (Hrsg.), *The International Encyclopedia of Communication Research Methods*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118901731.iecrm0179>
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the

- number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73–79. <https://doi.org/10.1027/1614-2241.4.2.73>
- Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, 40(1), 247–283. <https://doi.org/10.1111/j.1467-9531.2010.01231.x>
- Lynn, P. (2009). Methods for longitudinal surveys. In P. Lynn (Hg.), *Methodology of Longitudinal Surveys* (1. Aufl., S. 1–19). Wiley. <https://doi.org/10.1002/9780470743874.ch1>
- Lyrvall, J., Di Mari, R., Bakk, Z., Oser, J., & Kuha, J. (o. J.). *multi-levelLCA: An R package for single-level and multilevel latent class analysis with covariates*. <https://doi.org/10.48550/ARXIV.2305.07276>
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multi-level modeling. *Methodology*, 1(3), 86–92. <https://doi.org/https://doi.org/10.1027/1614-2241.1.3.86>
- Marciano, L., Driver, C. C., Schulz, P. J., & Camerini, A.-L. (2022). Dynamics of adolescents’ smartphone use and well-being are positive but ephemeral. *Scientific Reports*, 12(1), 1316. <https://doi.org/10.1038/s41598-022-05291-y>
- Masur, P. K. (2019a). Capturing situational dynamics and processes: Strengths and pitfalls of the experience sampling method. In P. Müller, S. Geiß, C. Schemer, T. K. Naab, & C. Peter (Hrsg.), *Dynamische Prozesse der öffentlichen Kommunikation. Methodische Herausforderungen* (S. 174–212). von Halem.
- Masur, P. K. (2019b). *Situational Privacy and Self-Disclosure: Com-*

- munication Processes in Online Environments*. Springer Berlin Heidelberg.
- Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. (2015). Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013. *Communication Methods and Measures*, 9(4), 193–207. <https://doi.org/10.1080/19312458.2015.1096334>
- Matthews, R. A., Pineault, L., & Hong, Y.-H. (2022). Normalizing the use of single-item measures: Validation of the single-item compendium for organizational psychology. *Journal of Business and Psychology*, 37(4), 639–673. <https://doi.org/10.1007/s10869-022-09813-3>
- May, M., Junghaenel, D. U., Ono, M., Stone, A. A., & Schneider, S. (2018). Ecological Momentary Assessment methodology in chronic pain research: A systematic review. *The Journal of Pain: Official Journal of the American Pain Society*, 19(7), 699–716. <https://doi.org/10.1016/j.jpain.2018.01.006>
- McCarthy, D. E., Minami, H., Yeh, V. M., & Bold, K. W. (2015). An experimental investigation of reactivity to ecological momentary assessment frequency among adults trying to quit smoking. *Addiction*, 110(10), 1549–1560. <https://doi.org/10.1111/add.12996>
- McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Psychology Press. <https://doi.org/10.4324/9781410601087>
- McKee, H. A., & Porter, J. E. (2009). *The Ethics of Internet Research: A Rhetorical, Case-based Process*. Peter Lang.
- McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, 25(5), 610–635. <https://doi.org/>

10.1037/met0000250

- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. <https://doi.org/10.1037/a0028085>
- Mehl, M. R., & Conner, T. S. (Hrsg.). (2012). *Handbook of Research Methods for Studying Daily Life*. The Guilford Press.
- Meier, A. (2022). Studying problems, not problematic usage: Do mobile checking habits increase procrastination and decrease well-being? *Mobile Media & Communication*, 10(2), 272–293. <https://doi.org/10.1177/20501579211029326>
- Meltzer, C. E., Naab, T., & Daschmann, G. (2012). All student samples differ: On participant selection in communication science. *Communication Methods and Measures*, 6(4), 251–262. <https://doi.org/10.1080/19312458.2012.732625>
- Merrilees, C. E., Goeke-Morey, M., & Mark Cummings, E. (2008). Do event-contingent diaries about marital conflict change marital interactions? *Behaviour Research and Therapy*, 46(2), 253–262. <https://doi.org/10.1016/j.brat.2007.11.007>
- MeTag / MeSoftware. (o. J.). Abgerufen <https://mesoftware.org/index.php/metag/>
- Mikesell, L., Marti, F. A., Guzmán, J. R., McCreary, M., & Zima, B. (2018). Affordances of mHealth technology and the structuring of clinic communication. *Journal of Applied Communication Research*, 46(3), 323–347. <https://doi.org/10.1080/00909882.2018.1465195>
- Möhring, W., & Schlütz, D. (Hrsg.). (2013a). *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft*.

- Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-531-18776-1>
- Möhring, W., & Schlütz, D. (2013b). Standardisierte Befragung – Messmethodik und Designs in der Medienwirkungsforschung. In W. Schweiger & A. Fahr (Hrsg.), *Handbuch Medienwirkungsforschung* (S. 565–579). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18967-3_30
- Möhring, W., & Schlütz, D. (2025). *Die Befragung in der Medien- und Kommunikationswissenschaft: Eine praxisorientierte Einführung*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-44191-3>
- Morehouse, K. N., Kurdi, B., & Nosek, B. A. (2024). Responsible data sharing: Identifying and remedying possible re-identification of human participants. *American Psychologist*, 80(6), 928–941. <https://doi.org/10.1037/amp0001346>
- Morren, M., Dulmen, S. van, Ouwerkerk, J., & Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: A systematic review. *European Journal of Pain*, 13(4), 354–365. <https://doi.org/10.1016/j.ejpain.2008.05.010>
- Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology / Psychologie canadienne*, 50(3), 131. <https://doi.org/10.1037/a0016625>
- Müller, S., Bayer, J., Ross, M. Q., Mount, J., Stachl, C., Harari, G. M., Chang, Y.-J., & Le, H. T. K. (2021). *Analyzing GPS data for psychological research: A tutorial*. <https://osf.io/rekuw/>
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29(1), 81–117. <https://doi.org/10.2333/bhmk.29.81>

- Myin-Germeys, I., & Kuppens, P. (2022a). Experience sampling methods, and introduction. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 7–19). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Myin-Germeys, I., & Kuppens, P. (Hrsg.). (2022b). *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies*. The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Naab, T. K. (2013). *Gewohnheiten und Rituale der Fernsehnutzung: theoretische Konzeption und methodische Perspektiven* (1. Aufl.). Nomos.
- Naab, T. K., Karnowski, V., & Schlütz, D. (2019). Reporting mobile social media use: How survey and experience sampling measures differ. *Communication Methods and Measures*, 13(2), 126–147. <https://doi.org/10.1080/19312458.2018.1555799>
- Nestler, S. (2021). Modeling intraindividual variability in growth with measurement burst designs. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(1), 28–39. <https://doi.org/10.1080/10705511.2020.1757455>
- Neubauer, A. B., & Schmiedek, F. (2020). Studying within-person variation and within-person couplings in intensive longitudinal data: lessons learned and to be learned. *Gerontology*, 66(4), 332–339. <https://doi.org/10.1159/000507993>
- Neubauer, A. B., Voelke, M. C., Voss, A., & Mertens, U. K. (2020). Estimating reliability of within-person couplings in a multilevel framework. *Journal of Personality Assessment*, 102(1), 10–21.

<https://doi.org/10.1080/00223891.2018.1521418>

- Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, 69, 149–155. <https://doi.org/10.1016/j.jrp.2016.06.020>
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>
- Nock, M. K., Kleiman, E. M., Abraham, M., Bentley, K. H., Brent, D. A., Buonopane, R. J., Castro-Ramirez, F., Cha, C. B., Dempsey, W., Draper, J., Glenn, C. R., Harkavy-Friedman, J., Hollander, M. R., Huffman, J. C., Lee, H. I. S., Millner, A. J., Mou, D., Onnela, J.-P., Picard, R. W., ... Pearson, J. L. (2021). Consensus statement on ethical & safety practices for conducting digital monitoring studies with people at risk of suicide and related behaviors. *Psychiatric Research and Clinical Practice*, 3(2), 57–66. <https://doi.org/10.1176/appi.prcp.20200029>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2024). Digital trace data collection for

- social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, 18(2), 124–141. <https://doi.org/10.1080/19312458.2023.2181319>
- Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. *Journal of Medical Internet Research*, 21(2), e11398. <https://doi.org/10.2196/11398>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Otto, L. P., Thomas, F., Maier, M., & Ottenstein, C. (2020). Only one moment in time? Investigating the dynamic relationship of emotions and attention toward political information with mobile experience sampling. *Communication Research*, 47(8), 1131–1154. <https://doi.org/10.1177/0093650219872392>
- Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., Lewis, S. W., & Dunn, G. (2011). Experience sampling research in individuals with mental illness: Reflections and guidance. *Acta Psychiatrica Scandinavica*, 123(1), 12–20. <https://doi.org/10.1111/j.1600-0447.2010.01596.x>
- Park, J., & Yu, H.-T. (2018). Recommendations on the sample sizes for multilevel latent class models. *Educational and Psychological Measurement*, 78(5), 737–761. <https://doi.org/10.1177/0013164417719111>
- Patrick, M. E., Singer, E., Boyd, C. J., Cranford, J. A., & McCabe, S. E. (2013). Incentives for college student participation in web-based substance use surveys. *Addictive Behaviors*, 38(3), 1710–1714. <https://doi.org/10.1016/j.addbeh.2012.08.007>

- Penner, L. A., Shiffman, S., Paty, J. A., & Fritzsche, B. A. (1994). Individual differences in intraperson variability in mood. *Journal of Personality and Social Psychology*, 66(4), 712–721. <https://doi.org/10.1037/0022-3514.66.4.712>
- Pentzold, C. (2015). Forschungsethische Prämissen und Problemfelder teilnehmenden Beobachtens auf Online-Plattformen. In A. Maireder, J. Ausserhofer, C. Schumann, & M. Taddicken (Hrsg.), *Digitale Methoden in der Kommunikationswissenschaft* (S. 61–85). FU Berlin. <https://doi.org/https://doi.org/10.17174/dcr.v2.4>
- Perks, L. G., & Turner, J. S. (2019). Podcasts and productivity: A qualitative uses and gratifications study. *Mass Communication and Society*, 22(1), 96–116. <https://doi.org/10.1080/15205436.2018.1490434>
- Perrault, E. K., & McCulloch, S. P. (2019). Concise consent forms appreciated – still not comprehended: Applying revised common rule guidelines in online studies. *Journal of Empirical Research on Human Research Ethics*, 14(4), 299–306. <https://doi.org/10.1177/1556264619853453>
- Perrault, E. K., & Nazione, S. A. (2016). Informed consent – uninformed participants: Shortcomings of online social science consent forms and recommendations for improvement. *Journal of Empirical Research on Human Research Ethics*, 11(3), 274–280. <https://doi.org/10.1177/1556264616654610>
- Peters, G.-J. (2018). *The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality*. <https://doi.org/10.31234/osf.io/h47fv>
- Pfiffner, N., Witlox, P., & Friemel, T. N. (2024). Data donation

- module: A web application for collecting and enriching data donations. *Computational Communication Research*, 6(2), 1. <https://doi.org/10.5117/CCR2024.2.4.PFIF>
- Pieters, R. (2017). Meaningful mediation analysis: Plausible causal inference and informative communication. *Journal of Consumer Research*, 44(3), 692–716. <https://doi.org/10.1093/jcr/ucx081>
- Pigg, S., Grabill, J. T., Brunk-Chavez, B., Moore, J. L., Rosinski, P., & Curran, P. G. (2014). Ubiquitous writing, technologies, and the social practice of literacies of coordination. *Written Communication*, 31(1), 91–117. <https://doi.org/10.1177/0741088313514023>
- Porst, R. (2014). *Fragebogen: Ein Arbeitsbuch*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-02118-4>
- Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121), 63–73. <https://doi.org/10.1002/ir.101>
- Prior, M. (2009). Improving media effects research through better measurement of news exposure. *The Journal of Politics*, 71(3), 893–908. <https://doi.org/10.1017/s0022381609090781>
- R Core Team. (2025). *R: A language and environment for statistical computing*. <https://www.R-project.org/>
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1), 203–212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2. Aufl.). Sage

Publications.

- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2019). *HLM 8 for Windows*. Scientific Software International, Inc.
- Reichow, D., & Friemel, T. N. (2020). Mobile communication, social presence, and perceived security on public transport. *Mobile Media & Communication*, 8(2), 268–292. <https://doi.org/https://doi.org/10.1177/2050157919878759>
- Reinecke, L., & Hofmann, W. (2016). Slacking off or winding down? An experience sampling study on the drivers and consequences of media use for recovery versus procrastination. *Human Communication Research*, 42(3), 441–461. <https://doi.org/10.1111/hcre.12082>
- Reis, H. T. (2012). Why researchers should think "real-world": A conceptual rationale. In M. R. Mehl & T. S. Conner (Hrsg.), *Handbook of Research Methods for Studying Daily Life*. (S. 3–21). The Guilford Press.
- Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying everyday experience. In H. T. Reis & C. M. Judd (Hrsg.), *Handbook of Research Methods in Social and Personality Psychology* (S. 190–222). Cambridge University Press.
- Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. 10.32614/CRAN.package.psych
- Revelle, W., & Wilt, J. (2019). Analyzing dynamic data: A tutorial. *Personality and Individual Differences*, 136, 38–51. <https://doi.org/10.1016/j.paid.2017.08.020>
- Revol, J., Carlier, C., Lafit, G., Verhees, M., Sels, L., & Ceulemans,

- E. (2024). Preprocessing experience-sampling-method data: A step-by-step framework, tutorial website, R package, and reporting templates. *Advances in Methods and Practices in Psychological Science*, 7(4). <https://doi.org/10.1177/25152459241256609>
- Reynolds, B. M., Robles, T. F., & Repetti, R. L. (2016). Measurement reactivity and fatigue effects in daily diary research with families. *Developmental Psychology*, 52, 442–456. <https://doi.org/10.1037/dev0000081>
- Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment*, 31, 226–235. <https://doi.org/10.1037/pas0000662>
- Robbins, M. L. (2017). Practical suggestions for legal and ethical concerns with social environment sampling methods. *Social Psychological and Personality Science*, 8(5), 573–580. <https://doi.org/10.1177/1948550617699253>
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1–25. <https://doi.org/10.1037/0033-2909.132.1.1>
- Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, 83(3), 598–626. <https://doi.org/10.1093/poq/nfz035>
- Roehse, E.-M., Möhring, W., Zillich, A. F., Schlütz, D., & Link, E. (2023). Forschungsethische Praxis in der Kommunikations- und Medienwissenschaft – ein Einblick in die Sicht der Forschenden. *Publizistik*, 68(4), 459–489. <https://doi.org/10.1007/s11616-023-00820-9>

- Roekel, E. van, Keijsers, L., & Chung, J. M. (2019). A review of current ambulatory assessment studies in adolescent samples and practical recommendations. *Journal of Research on Adolescence*, 29(3), 560–577. <https://doi.org/10.1111/jora.12471>
- Rohrer, J. M., & Murayama, K. (2023). These are not the effects you are looking for: Causality and the within-/between-persons distinction in longitudinal data analysis. *Advances in Methods and Practices in Psychological Science*, 6(1). <https://doi.org/10.1177/25152459221140842>
- Rosnow, R. L., & Rosenthal, R. (2014). *Ethical Principles in Data Analysis: An Overview*. Routledge. <https://doi.org/10.4324/9780203840023.ch3>
- Roßmann, J. (2017). *Satisficing in Befragungen: Theorie, Messung und Erklärung*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-16668-7>
- Samuel-Azran, T., Laor, T., & Tal, D. (2019). Who listens to podcasts, and why? The Israeli case. *Online Information Review*, 43(4), 482–495. <https://doi.org/10.1108/OIR-04-2017-0119>
- Sand, M., & Braun, M. (2025). *Mehrebenenanalyse: Eine anwendungsorientierte Einführung mit R*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-47710-3>
- Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren, modellieren und kommunizieren*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-21587-3>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>

- Scherer, H., & Naab, T. K. (2013). Messen im Zeitverlauf. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 103–123). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18776-1_6
- Scherer, H., & Schlütz, D. (2002). Gratifikation à la minute: Die zeitnahe Erfassung von Gratifikationen. In P. Rössler, S. Kubisch, & V. Gehrau (Hrsg.), *Empirische Perspektiven der Rezeptionsforschung* (S. 133–151). Reinhard Fischer.
- Schlütz, D. (2002). *Bildschirmspiele und ihre Faszination: Zuwendungsmotive, Gratifikationen und Erleben interaktiver Medienangebote*. R. Fischer.
- Schlütz, D., & Möhring, W. (2016). Kommunikationswissenschaftliche Forschungsethik – Sonntagsworte, Selbstzweck, Notwendigkeit? *Medien & Kommunikationswissenschaft*, 64(4), 483–496. <https://doi.org/10.5771/1615-634X-2016-4-483>
- Schlütz, D., & Möhring, W. (2017). Das Rad rollt! Eine Antwort und weiterführende Überlegungen zu dem Diskussionsbeitrag von Nele Heise in diesem Heft. *Medien & Kommunikationswissenschaft*, 65(4), 779–783. <https://doi.org/10.5771/1615-634X-2017-4-779>
- Schlütz, D., & Möhring, W. (2018). Between the devil and the deep blue sea: Negotiating ethics and method in communication research practice. *Studies in Communication and Media*, 7(1), 31–58. <https://doi.org/10.5771/2192-4007-2018-1-31>
- Schlütz, D., & Möhring, W. (2025). Forschungsethische Aspekte und Herausforderungen von Online-Kommunikation. In W. Schweiger, K. Beck, & V. Karnowski (Hrsg.), *Handbuch Online-Kommunikation* (S. 1–20). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-18017-1_38-1

- Schlütz, D., & Scherer, H. (2001). Der Einsatz der 'Experience Sampling Method' in der Medienwissenschaft. *Zeitschrift für Medienpsychologie*, 13 (N.F. 1)(3), 146–149.
- Schlütz, D., Schnauber-Stockmann, A., Karnowski, V., & Naab, T. K. (2027). Experiencing podcasts in situ: How the reception situation affects transportation. *Studies in Communication and Media*.
- Schlütz, D., & Zillich, A. F. (2023). Forschungsethik und wissenschaftliche Integrität: Herausforderungen und Chancen für Forschung in und mit digitalen Medien. In S. Stollfuß, L. Niebling, & F. Raczkowski (Hrsg.), *Handbuch Digitale Medien und Methoden* (S. 1–18). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-36629-2_8-1
- Schmiedek, F., & Neubauer, A. B. (2020). Experiments in the wild: Introducing the within-person encouragement design. *Multivariate Behavioral Research*, 55(2), 256–276. <https://doi.org/10.1080/00273171.2019.1627660>
- Schnauber, A. (2017). *Medienselektion im Alltag: die Rolle von Gewohnheiten im Selektionsprozess*. Springer VS.
- Schnauber-Stockmann, A., Bayer, J. B., Harari, G. M., & Karnowski, V. (2025). The situation in media and communication research. *Communication Theory*, 35(1), 25–36. <https://doi.org/10.1093/ct/qtae021>
- Schnauber-Stockmann, A., & Karnowski, V. (2020). Mobile devices as tools for media and communication research: A scoping review on collecting self-report data in repeated measurement designs. *Communication Methods and Measures*, 14(3), 145–164. <https://doi.org/10.1080/19312458.2020.1784402>

- Schnauber-Stockmann, A., & Mangold, F. (2020). Day-to-day routines of media platform use in the digital age: A structuration perspective. *Communication Monographs*, 87(4), 464–483. <https://doi.org/10.1080/03637751.2020.1758336>
- Schnauber-Stockmann, A., Meier, A., & Reinecke, L. (2018). Procrastination out of habit? The role of impulsive versus reflective media selection in procrastinatory media use. *Media Psychology*, 21(4), 640–668. <https://doi.org/10.1080/15213269.2018.1476156>
- Schnauber-Stockmann, A., & Naab, T. K. (2019). The process of forming a mobile media habit: Results of a longitudinal study in a real-world setting. *Media Psychology*, 22(5), 714–742. <https://doi.org/10.1080/15213269.2018.1513850>
- Schnauber-Stockmann, A., Scharnow, M., Karnowski, V., Naab, T. K., Schlütz, D., & Pressmann, P. (2025). Distinguishing person-specific from situation-specific variation in media use: A meta-analysis. *Communication Research*, 52(2), 143–172. <https://doi.org/10.1177/00936502241262664>
- Schnauber-Stockmann, A., Scharnow, M., Winkler, Y., Jost, P., & Denner, N. (2024). *Datenanalyse mit R*. <https://stats.ifp.uni-mainz.de/ba-datenanalyse/>
- Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Leske + Budrich.
- Scholl, A. (2013). Reaktivität im Forschungsprozess. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 79–99). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18776-1_5
- Scholl, A. (2018). *Die Befragung* (4. Aufl.). UVK Verlagsgesellschaft

mbH. <https://doi.org/10.36198/9783838549989>

- Schorrlepp, L., Stadel, M., Bringmann, L. F., Hesselink, M., & Maciejewski, D. (2025). Utilizing qualitative methods to detect validity issues in clinical experience sampling methodology (ESM). *Psychological Assessment*, 37(11), 599–613. <https://doi.org/10.1037/pas0001380>
- Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, 24(1), 70–91. <https://doi.org/10.1037/met0000188>
- Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A. G., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Quality of Life Research*, 15(9), 1533–1550. <https://doi.org/10.1007/s11136-006-0025-9>
- Schwarz, N. (2007). Retrospective and concurrent self-reports: The rationale for real-time data capture. In A. A. Stone, S. Shiffman, A. A. Atienza, & L. Nebeling (Hrsg.), *The Science of Real-Time Data Capture: Self-Reports in Health Research* (S. 11–26). Oxford University Press.
- Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, 22(2), 127–160. <https://doi.org/10.1177/109821400102200202>
- Scollon, C. N., Prieto, C.-K., & Diener, E. (2009). Experience sampling: Promises and pitfalls, strength and weaknesses. In A. C. Michalos & E. Diener (Hrsg.), *Assessing Well-Being* (Bd. 39, S. 157–180). Springer Netherlands. https://doi.org/10.1007/978-90-481-2354-4_8

- Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1), 399–422. <https://doi.org/10.1093/poq/nfab018>
- Shalev, D., Naab, T. K., & Tsfati, Y. (2025). Self-reporting news use in situ and in retrospect. *International Journal of Public Opinion Research*, 37(2), edaf027. <https://doi.org/10.1093/ijpor/edaf027>
- Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Hrsg.), *Handbook of Research Methods for Studying Daily Life* (S. 302–320). The Guilford Press.
- Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavel, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences*, 115(1), E15–E23. <https://doi.org/10.1073/pnas.1712277115>
- Siebers, T., Beyens, I., Pouwels, J. L., & Valkenburg, P. M. (2022). Social media and distraction: An experience sampling study among adolescents. *Media Psychology*, 25(3), 343–366. <https://doi.org/10.1080/15213269.2021.1959350>
- Siebers, T., Beyens, I., & Valkenburg, P. M. (2024). The effects of fragmented and sticky smartphone use on distraction and task delay. *Mobile Media & Communication*, 12(1), 45–70. <https://doi.org/10.1177/20501579231193941>
- Silvia, P. J., Kwapil, T. R., Eddington, K. M., & Brown, L. H. (2013). Missed beeps and missing data: Dispositional and situational predictors of nonresponse in experience sampling research. *Social Science Computer Review*, 31(4), 471–481. <https://doi.org/10.1177/0894439313479902>

- Singer, E. (2018). The use and effects of incentives in surveys. In D. L. Vannette & J. A. Krosnick (Hrsg.), *The Palgrave Handbook of Survey Research* (S. 63–70). Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_9
- Slipetz, L. R., Falk, A., & Henry, T. R. (2023). *Missing data in discrete time state-space modeling of ecological momentary assessment data: A Monte-Carlo study of imputation methods*. arXiv. <http://arxiv.org/abs/2301.13144>
- Sliwinski, M. J. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass*, 2(1), 245–261. <https://doi.org/10.1111/j.1751-9004.2007.00043.x>
- Smyth, J. M., Jones, D. R., Wen, C. K. F., Matera, F. T., Schneider, S., & Stone, A. (2021). Influence of ecological momentary assessment study design features on reported willingness to participate and perceptions of potential research studies: An experimental study. *BMJ Open*, 11(7), e049154. <https://doi.org/10.1136/bmjopen-2021-049154>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2. Aufl.). Sage.
- Song, J., Howe, E., Oltmanns, J. R., & Fisher, A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment*, 30(5), 1662–1671. <https://doi.org/10.1177/10731911221113563>
- Soyster, P. D., Bosley, H. G., Reeves, J. W., Altman, A. D., & Fisher, A. J. (2019). Evidence for the feasibility of person-specific ecological momentary assessment across diverse populations and study designs. *Journal for Person-Oriented Research*, 5(2), 53–64. <https://doi.org/10.17505/jpor.2019.06>

- Stawski, R. S., MacDonald, S. W. S., & Sliwinski, M. J. (2015). Measurement burst design. In S. K. Whitbourne (Hg.), *The Encyclopedia of Adulthood and Aging* (S. 1–5). John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118521373.wbeaa313>
- Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5), 503–516. <https://doi.org/10.1177/0894439319843669>
- Stone, A. A., Kessler, R. C., & Haythomthwatte, J. A. (1991). Measuring daily events and experiences: Decisions for the researcher. *Journal of Personality*, 59(3), 575–607. <https://doi.org/10.1111/j.1467-6494.1991.tb00260.x>
- Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology*, 19(1), 107–131. <https://doi.org/10.1146/annurev-clinpsy-080921-083128>
- Stone, A. A., & Shiffman, S. (2002). Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine*, 24(3), 236–243. https://doi.org/10.1207/S15324796ABM2403_09
- Struckmann, S., & Karnowski, V. (2016). News consumption in a changing media ecology: An MESM-study on mobile news. *Telecommunications & Informatics*, 33(2), 309–319. <https://doi.org/10.1016/j.tele.2015.08.012>
- Struminskaya, B., & Sakshaug, J. W. (2023). Ethical considerations for augmenting surveys with auxiliary data sources. *Public Opinion Quarterly*, 87(S1), 619–633. <https://doi.org/10.1093/poq/nfad030>

- Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2021). Understanding willingness to share smartphone-sensor data. *Public Opinion Quarterly*, 84(3), 725–759. <https://doi.org/10.1093/poq/nfaa044>
- Sturgis, P., & Brunton-Smith, I. (2023). Personality and survey satisficing. *Public Opinion Quarterly*, 87(3), 689–718. <https://doi.org/10.1093/poq/nfad036>
- Sun, Y., & Pan, Z. (2020). Not published is not perished: Addressing publication bias in meta-analytic studies in communication. *Human Communication Research*, 46(2-3), 300–321. <https://doi.org/10.1093/hcr/hqz015>
- Taddicken, M. (2013). Online-Befragung. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (S. 201–217). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-531-18776-1_11
- Thomas, F. (2022). A methodological framework for analyzing the appearance and duration of media effects. *Journal of Communication*, 72(3), 401–428. <https://doi.org/10.1093/joc/jqac013>
- Tiidenberg, K. (2018). Ethics in digital research. In U. Flick (Hg.), *The SAGE Handbook of Qualitative Data Collection* (S. 466–479). SAGE Publications Ltd. <https://doi.org/10.4135/9781526416070.n30>
- Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, 129, 56–63. <https://doi.org/10.1037/abn0000473>
- Tuerlinckx, F., Kuppens, P., Ariens, S., Cloos, L., Dejonckheere, E., Lafit, G., Niemeijer, K., Revol, J., Schat, E., Schreuder, M.,

- Vanhasbroeck, N., & Ceulemans, E. (2026). New developments in experience sampling methodology. *British Journal of Mathematical and Statistical Psychology*, 79, 46–65. <https://doi.org/10.1111/bmsp.12398>
- Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and retention with the experience sampling method over the continuum of severe mental disorders: Meta-analysis and recommendations. *Journal of Medical Internet Research*, 21(12), e14475. <https://doi.org/10.2196/14475>
- Valkenburg, P., Beyens, I., Pouwels, J. L., Driel, I. I. van, & Keijsers, L. (2021). Social media use and adolescents' self-esteem: Heading for a person-specific media effects paradigm. *Journal of Communication*, 71(1), 56–78. <https://doi.org/10.1093/joc/jqaa039>
- Verbeij, T., Pouwels, J. L., Beyens, I., & Valkenburg, P. M. (2022). Experience sampling self-reports of social media use have comparable predictive validity to digital trace measures. *Scientific Reports*, 12(1), 7611. <https://doi.org/10.1038/s41598-022-11510-3>
- Verbeke, K., Krawczyk, T., Baeyens, D., Piasecki, J., & Borry, P. (2023). Informed consent and debriefing when deceiving participants: A systematic review of research ethics guidelines. *Journal of Empirical Research on Human Research Ethics*, 18(3), 118–133. <https://doi.org/10.1177/15562646231173477>
- Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17(1), 33–51. <https://doi.org/10.1177/0962280207081238>
- Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent Markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24(2), 179–207. <https://doi.org/10.2307/1165200>

- Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. Lewis-Beck, A. Bryman, & T. Futing Liao (Hrsg.), *The SAGE Encyclopedia of Social Science Research Methods* (S. 550–553). Sage Publications, Inc. <https://doi.org/10.4135/9781412950589.n472>
- Viechtbauer, W. (2022a). Statistical methods for ESM data. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 153–184). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Viechtbauer, W. (2022b). Structuring, checking, and preparing ESM data. In I. Myin-Germeys & P. Kuppens (Hrsg.), *The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies* (S. 137–152). The Center for Research on Experience Sampling; Ambulatory Methods Leuven (REAL).
- Vogelsmeier, L. V. D. E., Vermunt, J. K., & De Roover, K. (2022). How to explore within-person and between-person measurement model differences in intensive longitudinal data with the R package *lmfa*. *Behavior Research Methods*, 55, 2387–2422. <https://doi.org/10.3758/s13428-022-01898-1>
- Wagner, A., Schwarzenegger, C., & Gentzel, P. (2022). Recordings of digital media life: Advancing (qualitative) media diaries as a method. *Studies in Communication Sciences*, 22(1), 53–68. <https://doi.org/10.24434/j.scoms.2022.01.3066>
- Wagner-Schelewsky, P., & Hering, L. (2019). Online-Befragung. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 787–800). Springer Fachmedien. https://doi.org/10.1007/978-3-658-21308-4_54

- Weiß, B., Silber, H., Struminskaya, B., & Durrant, G. (2019). Mobile Befragungen. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 801–812). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-21308-4_55
- Welbers, K., Loecherbach, F., Lin, Z., & Trilling, D. (2024). Anything you would like to share: Evaluating a data donation application in a survey and field study. *Computational Communication Research*, 6(2), 1. <https://doi.org/10.5117/CCR2024.2.5.WELB>
- Wen, C. K. F., Junghaenel, D. U., Newman, D. B., Schneider, S., Mendez, M., Goldstein, S. E., Velasco, S., Smyth, J. M., & Stone, A. A. (2021). The effect of training on participant adherence with a reporting time frame for momentary subjective experiences in ecological momentary assessment: Cognitive interview study. *JMIR Formative Research*, 5(5), e28007. <https://doi.org/10.2196/28007>
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, 1832. <https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01832>
- Wickham, H., Cetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science* (2. Aufl.). <https://r4ds.hadley.nz/>
- Wieland, M., Herbertz, A., Krone, J., & Otto, L. (2025). *ESM tool collection*. <https://doi.org/osf.io/um87t/>
- Wilhelm, P., Perrez, M., & Pawlik, K. (2012). Conducting research in daily life: A historical review. In M. R. Mehl & T. S. Conner (Hrsg.), *Handbook of Research Methods for Studying Daily Life*

(S. 62–86). The Guilford Press.

- Wolf, C., Joye, D., Smith, T., Fu, Y., Lyberg, L. E., & Weisberg, H. F. (2016). Total survey error: A paradigm for survey methodology. In *The SAGE Handbook of Survey Methodology* (S. 27–40). SAGE Publications Ltd. <https://doi.org/10.4135/9781473957893.n3>
- Wolfers, L., & Baumgartner, S. (2025). *Short but still valid: Validating single-item measures for key communication science constructs for experience sampling research*. https://doi.org/10.31234/osf.io/zkwxa_v1
- Wright, A. G. C., Beltz, A. M., Gates, K. M., Molenaar, P. C. M., & Simms, L. J. (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. *Frontiers in Psychology*, 6, 1914. <https://doi.org/10.3389/fpsyg.2015.01914>
- Wünsch, C., Briselat, T., Reitmeier, S., Keil, J., & Czichon, M. (2023). Skala Spannungserleben (SSE). Entwicklung einer Skala zur postrezeptiven Erfassung von Spannungserleben während der Rezeption narrativer Inhalte. *Studies in Communication and Media*, 12(2-3), 115–146. <https://doi.org/10.5771/2192-4007-2023-2-3-115>
- Yan, T., & Curtin, R. (2010). The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Research*, 22(4), 535–551. <https://doi.org/10.1093/ijpor/edq037>
- Zhu, J. J. H., Chen, H., Peng, T.-Q., Liu, X. F., & Dai, H. (2018). How to measure sessions of mobile phone use? Quantification, evaluation, and applications. *Mobile Media & Communication*, 6(2), 215–232. <https://doi.org/10.1177/2050157917748351>

- Zillich, A. F., Schlütz, D., Roehse, E.-M., Link, E., & Möhring, W. (2023). Practicing internet research ethics: Challenges and solutions from a German perspective. *Comunicazioni Sociali*, (2), 191–202. https://doi.org/10.26350/001200_000185
- Zimmermann, F., & Kohring, M. (2020). Mistrust, disinforming news, and vote choice: A panel survey on the origins and consequences of believing disinformation in the 2017 German parliamentary election. *Political Communication*, 37(2), 215–237. <https://doi.org/10.1080/10584609.2019.1686095>