

Bausteine Forschungsdatenmanagement

Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern

Insa Bechertⁱ Pablo Christmannⁱⁱ Andreas Frankenⁱⁱⁱ
Fabio Franzese^{iv} Maarten Koomen^v Kristina Krell^{vi}
Tanja Kunz^{vii} Senta-Melissa Pflüger^{viii} Percy Scheller^{ix}
Maximilian Trommer^x Christina von Rotz^{xi}

2023

Zitiervorschlag

Bechert, Insa, Pablo Christmann, Andreas Franken, Fabio Franzese, Maarten Koomen, Kristina Krell, Tanja Kunz, Senta-Melissa Pflüger, Percy Scheller, Maximilian Trommer und Christina von Rotz. Plausibilisierung: Eine Handreichung zum Umgang mit inkonsistenten Daten in den Sozialwissenschaften. *Bausteine Forschungsdatenmanagement. Empfehlungen und Erfahrungsberichte für die Praxis von Forschungsdatenmanagerinnen und -managern* Nr. 1/2023: S.1-84. DOI: [10.17192/bfdm.2023.1.8413](https://doi.org/10.17192/bfdm.2023.1.8413).

Dieser Beitrag steht unter einer
[Creative Commons Namensnennung 4.0 International Lizenz \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/).

ⁱGESIS - Leibniz-Institut für Sozialwissenschaften/ISSP. ORCID: [0000-0002-3455-9748](https://orcid.org/0000-0002-3455-9748)

ⁱⁱGESIS - Leibniz-Institut für Sozialwissenschaften/FReDA. ORCID: [0000-0003-0458-9572](https://orcid.org/0000-0003-0458-9572)

ⁱⁱⁱDeutsches Institut für Wirtschaftsforschung/SOEP. ORCID: [0000-0002-5882-0285](https://orcid.org/0000-0002-5882-0285)

^{iv}SHARE BERLIN Institute/SHARE. ORCID: [0000-0002-2947-933X](https://orcid.org/0000-0002-2947-933X)

^vUniversität Bern/TREE. ORCID: [0000-0002-0375-5607](https://orcid.org/0000-0002-0375-5607)

^{vi}Universität Bielefeld/TwinLife. ORCID: [0000-0003-0878-4842](https://orcid.org/0000-0003-0878-4842)

^{vii}GESIS - Leibniz-Institut für Sozialwissenschaften/FReDA. ORCID: [0000-0001-8460-2583](https://orcid.org/0000-0001-8460-2583)

^{viii}SHARE BERLIN Institute/SHARE. ORCID: [0000-0001-8158-2206](https://orcid.org/0000-0001-8158-2206)

^{ix}Weizenbaum-Institut. ORCID: [0000-0002-5791-2938](https://orcid.org/0000-0002-5791-2938)

^xDeutsches Zentrum für Hochschul- und Wissenschaftsforschung. ORCID: [0000-0002-1331-4768](https://orcid.org/0000-0002-1331-4768)

^{xi}Universität Bern/TREE. ORCID: [0000-0002-4456-3310](https://orcid.org/0000-0002-4456-3310)

Abstract

Die vorliegende Handreichung zeigt, dass in quantitativen, sozialwissenschaftlichen Erhebungen Unplausibilitäten entstehen können, die möglicherweise zu Auswertungsproblemen und Fehlanalysen führen. Ziel der Handreichung ist, die möglichen Probleme zu systematisieren und konkrete Hilfestellungen bei der Identifikation, dem Umgang und der Vermeidung von Unplausibilitäten zu geben. Außerdem werden Dokumentationsempfehlungen ausgesprochen.

Diese Handreichung richtet sich dabei in erster Linie an Datenmanager:innen oder Datenproduzent:innen von Erhebungen, die bei der Studienplanung und Datenaufbereitung mit Plausibilitätsproblemen in Berührung kommen. Aber auch für Studierende und Forscher:innen könnte sie eine Hilfestellung darstellen.

Gegliedert ist die Handreichung wie folgt: Nach einer Einleitung und der inhaltlichen Einbettung des Themas in den Forschungsprozess folgt eine allgemeine Einführung in den Plausibilisierungsprozess. Dabei wird zunächst diskutiert, ob Plausibilisierung überhaupt zwangsläufig stattfinden muss oder sollte. Im Anschluss wird ein Überblick über die Möglichkeiten gegeben, mit denen Plausibilitätsprobleme zunächst identifiziert und dann behandelt werden können. Diese Punkte werden allgemeingültig formuliert und beziehen sich nicht auf spezielle Arten von Plausibilitätsproblemen. Diese werden anhand der spezifischen Use Cases (UC) in den Folgekapiteln diskutiert. Die Plausibilitätsprobleme und Beispiele werden weitgehend unabhängig voneinander behandelt, sodass die Leser:innen sich auf die für ihre Studie bzw. Forschungsinteressen relevanten Inhalte konzentrieren oder die Handreichung als eine Art Nachschlagewerk verwenden können.

Inhaltsverzeichnis

1	Einleitung	5
2	Inhaltliche Einbettung in den Forschungsprozess	8
2.1	Total Survey Error	8
2.2	Forschungsdatenzyklus	10
3	Der Plausibilisierungsprozess	12
3.1	Wo können Fehler in Datensätzen auftreten?	12
3.1.1	Variablenspezifische Fehler	12
3.1.2	Fallspezifische Fehler	13
3.1.3	Erhebungsübergreifende Fehler	13
3.1.4	Interviewer:innenbezogene Fehler	13
3.2	Auf welchen Wegen lassen sich fehlerhafte Werte identifizieren?	14
3.2.1	Numerische Variablen	15
3.2.2	Strings	16
3.3	Umgang mit fehlerhaften Werten	17
3.3.1	Keine Änderung vornehmen, ohne Markierung (Flag)	17
3.3.2	Keine Änderung vornehmen, mit Markierung (Flag)	18
3.3.3	Änderung auf unwidersprüchliche „valide“ Werte	18
3.3.4	Änderung auf fehlende Werte (Missings)	19
3.4	Entscheidungskriterien zum Umgang mit unplausiblen Werten	19
3.5	Dokumentation des Umgangs mit unplausiblen Werten	21
4	Filterfehler	23
4.1	Beschreibung	23
4.2	Ursachen	24
4.2.1	In der Studienplanungsphase	24
4.2.2	In der Datenerhebungsphase	25
4.3	Identifikation	26
4.4	Umgang	29
4.5	Dokumentation	31
4.6	Vermeidung	32
5	Wildcodes	34
5.1	Beschreibung	34
5.2	Ursachen	37
5.2.1	In der Studienplanungsphase	37
5.2.2	In der Datenerhebungsphase	37
5.2.3	In der Datenaufbereitungsphase	38
5.3	Identifikation	40
5.3.1	In der Studienplanungsphase	40
5.3.2	In der Datenaufbereitungsphase	40

5.4	Umgang	41
5.5	Dokumentation	45
5.6	Vermeidung von Wildcodes	48
6	Inkonsistenzen bei der Verknüpfung mehrerer Datensätze/-quellen	50
6.1	Beschreibung	50
6.2	Ursachen	52
6.2.1	In der Studienplanungsphase	52
6.2.2	In der Datenerhebungsphase	52
6.2.3	In der Datenaufbereitungsphase	55
6.3	Umgang	57
6.4	Dokumentation	60
6.5	Vermeidung	61
6.5.1	In der Studienplanungsphase	61
7	Inkonsistenzen in längsschnittlichen Episodendaten	64
7.1	Beschreibung	64
7.1.1	Was sind Episoden?	64
7.1.2	Wie werden Episodendaten erhoben?	66
7.1.3	Arten von widersprüchlichen und fehlenden Angaben in Episodendaten	68
7.2	Ursachen von Widersprüchen und Lücken innerhalb Episoden	69
7.2.1	In der Studienplanungsphase	69
7.2.2	In der Datenerhebungsphase	72
7.3	Identifikation	73
7.4	Umgang	75
7.5	Dokumentation	76
7.6	Vermeidung	77
7.6.1	Dependent Interview Design	79
7.6.2	Widerspruchsmodul	79
7.6.3	Gestaltung der Stimuli	80
7.6.4	Grobe Matching-Kriterien	81
7.6.5	Schulung der Interviewer:innen	81
8	Resümee und Ausblick	82

1 Einleitung

Andreas Franken, Kristina Krell, Percy Scheller

Die vorliegende Handreichung beschreibt, wie in quantitativen Erhebungen unplausible Werte entstehen können, die zu Auswertungsproblemen und fehlerhaften Analysen führen. Die Handreichung systematisiert die möglichen Probleme und beschreibt Lösungsansätze sowie Dokumentationsanforderungen.

Unplausibilitäten entstehen bei der Durchführung von Befragungen fast zwangsläufig. Die möglichen Umgangsweisen damit sind jedoch vielfältig und unter anderem abhängig davon, wie schwerwiegend die Problematik ist und wie umfangreich ein Eingriff in die Daten wäre, um sie zu beseitigen. Die unplausiblen Werte können im statistischen Rauschen untergehen, können aber auch einen gravierenden Einfluss auf die Analyseergebnisse haben, wenn sie z. B. Subgruppen-Betrachtungen verfälschen. In der Regel besteht das Ziel von Plausibilisierungsprozeduren darin, Fehler in den Daten und somit auch der darauf basierenden Auswertung zu minimieren.

Die Autor:innengruppe dieser Handreichung ist aus dem Pre-Workshop „Datenproduktion“ 2018 im Vorfeld des „12. Workshops der Panelstudien im deutschsprachigen Raum“ in Bamberg heraus entstanden. Dort wurde festgestellt, dass Reflexionen und ein Austausch zum Thema Datenplausibilisierung bisher unzureichend sind. Die Gruppe besteht aus Praktiker:innen aus mehreren etablierten Studien und Institutionen¹ mit dem Interesse, teilweise seit Jahrzehnten praktizierte Routinen der Datenaufbereitung kritisch zu reflektieren und einen gemeinsamen Referenzrahmen zu erarbeiten.

Das Anliegen der Autor:innen ist folglich, eine Handreichung zu erstellen, auf die sich bei der Bearbeitung oder Dokumentation von Plausibilitätsproblemen und der Beantragung von notwendigen Ressourcen bezogen werden kann. Im Sinne einer Bestandsaufnahme der in den großen deutschsprachigen sozialwissenschaftlichen Studien verbreiteten Umgangstechniken lebt die Handreichung von zahlreichen Best-Practice-Beispielen, diskutiert alternative Vorgehensweisen bei der Plausibilisierung und erläutert deren Konsequenzen. Sie richtet sich in erster Linie an Datenproduzent:innen und Datenmanager:innen. Auch für Studierende und etablierte Forschende kann sie Hilfestellung oder Referenztext sein.

In Kapitel 2 findet zunächst eine inhaltliche Einbettung der Plausibilisierungsvorgänge in den Forschungsprozess statt. Dabei wird verdeutlicht, an welchen Stellen des Forschungsdatenzyklus – von der Studienplanung bis hin zur Datenaufbereitung (auch engl. data processing bzw. editing) – Plausibilitätsprobleme auftreten können und wie sich diese im Total-Survey-Error-Schema verorten lassen.

¹Mitarbeiter:innen der folgenden Studien und Institutionen sind an dem Papier beteiligt: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW), GESIS - Leibniz-Institut für Sozialwissenschaften, Sozio-oekonomisches Panel (SOEP), Survey of Health, Ageing and Retirement in Europe (SHARE), TwinLife, Transitionen von der Erstausbildung ins Erwerbsleben (TREE).

Kapitel 3 bietet eine allgemeine Einführung in den Plausibilisierungsprozess. Hierbei wird zunächst grundsätzlich die Sinnhaftigkeit von Plausibilisierung diskutiert, bevor im Anschluss die Arten der Identifikation von und des Umgangs mit Plausibilitätsproblemen vorgestellt werden.

Beginnend mit Filterfehler-Problem, die wohl jede quantitative Befragung betreffen, und endend mit nur bei der Erhebung von Episoden auftretenden Problemen, behandeln die Kapitel 4 bis 7 schließlich konkrete Arten von Plausibilitätsfehlern und ihre Besonderheiten.

Im abschließenden Resümee wird zusammengefasst, mit welchen Plausibilitätsproblemen bei der Aufbereitung sozialwissenschaftlicher Daten gerechnet werden muss, mit welchen möglichen Verfahren ihnen begegnet werden kann und welcher Bearbeitungsaufwand deshalb bereits im Vorfeld der einer Studie eingeplant werden sollte.

Die Handreichung schließt mit einem Ausblick auf die steigende Bedeutung der Thematik in der wissenschaftlichen Gemeinschaft und bei den Forschungsförder:innen. Zudem werden mögliche, zukünftige Ergänzungen der Publikation um verwandte Inhalte angesprochen.

Konkret beschrieben und diskutiert werden in dieser Handreichung die folgenden Plausibilitätsprobleme: Filterfehler, Wildcodes, Inkonsistenzen bei der Verknüpfung mehrerer Datensätze sowie Inkonsistenzen in längsschnittlichen Episodendaten. Alle Use Cases folgen der gleichen, grundlegenden Kapitelstruktur:

Die Use Cases beginnen mit einer kurzen **Beschreibung** des jeweiligen Plausibilitätsproblems. Da sich die Definitionen verschiedener Arten von Plausibilitätsproblemen teilweise überschneiden, findet in diesem Abschnitt eine klare Abgrenzung der besprochenen Plausibilitätsprobleme statt.

Im darauffolgenden Abschnitt werden mögliche **Ursachen** der Probleme beleuchtet. Damit für die Leser:innen möglichst gut nachvollziehbar ist, wie und warum diese Probleme entstehen, orientiert sich dieser Abschnitt am Forschungsdatenzyklus – sprich, er zeigt auf, in welchen Phasen der Entstehung einer Studie es zu Unplausibilitäten kommen kann (vgl. Kapitel 2 „Inhaltliche Einbettung in den Forschungsprozess“). Üblicherweise sind dies die Studienplanungsphase, die Datenerhebungsphase oder die Datenaufbereitungsphase.

Anschließend wird diskutiert, welche Möglichkeiten der **Identifikation** dieser Plausibilitätsprobleme bestehen. Da die Möglichkeiten in der Regel stark von der Art der Studie, der Erhebungsweise und der Datenstruktur abhängen, wird in diesem Abschnitt mit konkreten Beispielen aus den in dieser Handreichung besprochenen Studien gearbeitet. Die detaillierte Darstellung der Problematiken anhand bestimmter Studien soll die Leser:innen dabei unterstützen zu ermitteln, welche Anwendungsmethode für die eigene Studie am sinnvollsten ist.

Der darauffolgende Abschnitt **Umgang** mit den Plausibilitätsproblemen schließt unmittelbar an die Identifikation an. Auch die Umgangsprozedere hängen stark von der

Art der betroffenen Studie ab. Entsprechend ist auch hier das Ziel, den Leser:innen anhand konkreter Studienbeispiele zu erläutern, welche Möglichkeiten in Betracht gezogen werden können und welche konkreten Problematiken sich aus der Plausibilisierung ergeben könnten.

Die **Dokumentation** von identifizierten (und ggf. auch korrigierten) Unplausibilitäten ist grundlegend für saubere wissenschaftliche Arbeit. Deshalb zeigt der entsprechende Abschnitt auf, in welcher Form (von marginal bis sehr umfangreich, innerhalb der Daten oder im Rahmen zusätzlicher Dokumente) den Datennutzer:innen dargelegt werden kann, ob und wie mit welchen Plausibilitätsproblemen in der Studie umgegangen wird.

Zuletzt widmen sich die Kapitel den möglichen Strategien zur **Vermeidung** der besprochenen Plausibilitätsprobleme. Da auch dieser Abschnitt eine starke Praxisorientierung hat und die Leser:innen dabei unterstützen soll, etwaige Probleme bereits im Vorfeld oder während der Durchführung der Studie zu vermeiden, ist er wiederum entlang des in Kapitel 2 „Inhaltliche Einbettung in den Forschungsprozess“ beschriebenen Forschungsdatenzyklus strukturiert. Auch hier sind es in der Regel die Phasen der Studienplanung, der Datenerhebung und der Datenaufbereitung, in denen die Maßnahmen zur Vermeidung von Plausibilitätsproblemen getroffen werden.

2 Inhaltliche Einbettung in den Forschungsprozess

Insa Bechert

Personen, die an einer Umfrage teilnehmen, bemühen sich in der Regel um eine konsistente Selbstdarstellung in ihrem Antwortverhalten. Trotzdem kommt es immer wieder zu Inkonsistenzen in den Daten. Wie diese Handreichung anhand zahlreicher Beispiele ausführlich beleuchtet, gibt es unterschiedliche Gründe, die zu einer Inkonsistenz der Angaben und damit zu unplausiblen Werten in den verschiedenen Datentypen führen können. Panelstudien leiden darunter, dass Angaben von Befragungspersonen zwischen den Wellen auf unplausible Weise abweichen. Wenn bei Querschnittsdaten beispielsweise nach Einstellungen zu verschiedenen Themen gefragt wird, ist es nicht selten der Fall, dass einzelne Angaben oder Antworten auf verschiedene Fragen im Verhältnis zueinander unplausibel erscheinen. Ein Grund hierfür mag sein, dass Individuen in ihren Einstellungen nun einmal nicht immer konsistent sind. Impulse können zu spontanen Meinungsänderungen führen und im Fall der Panelstudien mögen im Laufe der Zeit sich verändernde oder verblassende Erinnerungen zu inkonsistenten Angaben führen (siehe auch Kapitel 6 zu Inkonsistenzen bei der Verknüpfung verschiedener Datenquellen und Kapitel 7 zu Episodendaten). Nichtsdestotrotz liegt den allermeisten Fällen inkonsistenter Angaben mindestens ein Missverständnis zugrunde. Daher können sowohl die unplausiblen Angaben als auch der Versuch diese zu beheben, zu Fehlern in den Daten führen.

2.1 Total Survey Error

Eine Möglichkeit der systematischen Erfassung dieser potenziellen Fehlerquellen bietet der konzeptionelle Rahmen des „Total Survey Error“ (TSE). Verschiedene, unterschiedlich umfassende TSE-Schemata sind in der Vergangenheit entwickelt worden um potenzielle Fehler im Verlauf eines Umfrageprozesses möglichst umfassend zu klassifizieren und ihre Kausalitäten zu beschreiben. Grundlegende Schemata finden sich in Groves 1989², Groves et al. 2004³, Weisberg 2005⁴ oder Biemer et al. 2017⁵. Ein Überblick über die Entwicklung des TSE in Publikationen von 1944 bis ins 21. Jahrhundert findet sich in Groves und Lyberg 2010⁶. In der Regel liegt dem TSE-Konzept

²Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

³Groves, Robert, Floyd Fowler, Mick Couper, Eleanor Singer und Roger Tourangeau. 2004. *Survey Methodology*. New York: Wiley.

⁴Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.

⁵Biemer, Paul P., Edith Desirée de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker und Brady T. West, Hrsg. 2017. *Total survey error in practice*. Hoboken, New Jersey: Wiley.

⁶Groves, R. M. und L. Lyberg. 2010. „Total Survey Error: Past, Present, and Future“. *Public Opinion Quarterly* 74 (5): 849–79. <https://doi.org/10.1093/poq/nfq065>.

eine Trennung von Messfehlern und Fehlern bei der Repräsentation – das heißt, der Validität der Befragungsstichprobe – zugrunde. Je nach Anwendungsbereich wurden über die Jahre zu verschiedenen Datentypen passende TSE-Strukturen entwickelt. Abbildung 1 zeigt ein grundlegendes TSE-Schema nach Groves et al. (2004), strukturiert anhand der klassischen Teilung nach Mess- und Repräsentationskomponenten.

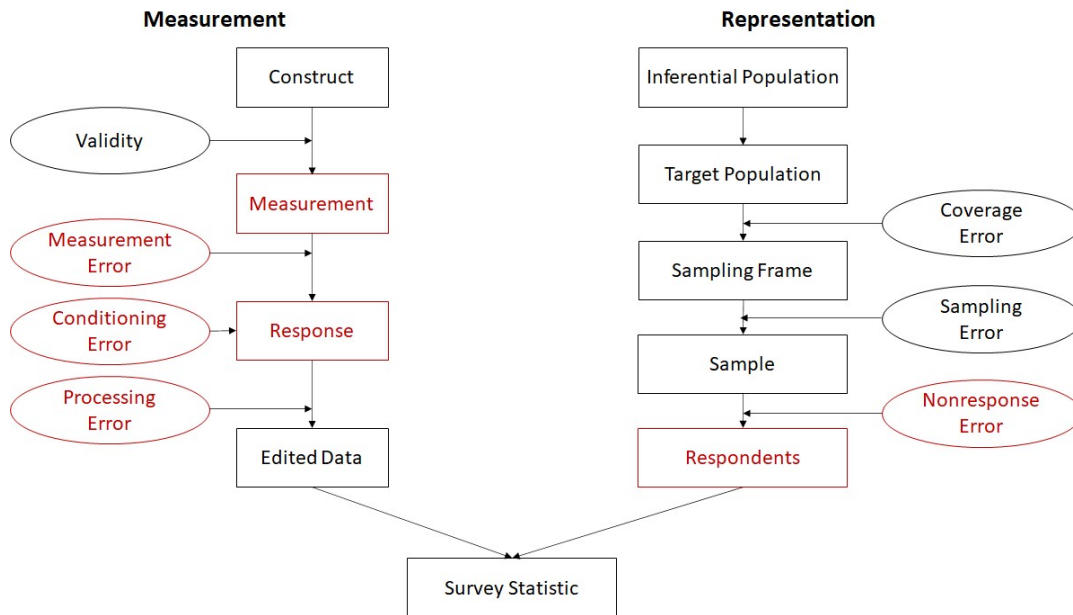


Abbildung 1: Total Survey Error Schema nach Groves et al. (2004), farbliche Hervorhebung und Einfügung des *conditioning error* durch die Autor:innen

Es gibt einige potenzielle Fehlerquellen, die im Rahmen der Datenerhebung und Datenaufbereitung die Frage nach Plausibilisierung aufkommen lassen. In Abbildung 1 sind sie rot dargestellt. Die meisten von ihnen lassen sich dem Bereich der Messung zuordnen. Eine dieser Fehlerquellen wird mit dem Begriff *conditioning error*, also „Voraussetzungsfehler“, beschrieben.⁷ Voraussetzungsfehler treten nur bei Paneldaten, nicht bei Querschnittsdaten auf, da eine vorherige Welle die „Voraussetzung“ für diese Fehler der aktuellen Welle ist. Die Erfahrungen der Befragten mit vorherigen Befragungen des Panels beeinflussen, in welcher Weise die aktuellen Fragen beantwortet werden. So versuchen Befragte unter Umständen eine Befragung abzukürzen, indem sie keine Angaben machen, die, wie sie bereits wissen, zu Folgefragen führen, oder sie machen falsche Angaben in dem Bemühen konsistent mit früheren Angaben zu sein. In dem TSE-Schema in Abbildung 1 sind diese Fehler der Messung, genauer dem Antwortverhalten *Response*, zuzuordnen. Auch bei Querschnittsdaten kann das Bemühen der

⁷Vgl. Neter, John und Joseph Waksberg. 1964. „A Study of Response Errors in Expenditures Data from Household Interviews“. *Journal of the American Statistical Association* 59 (305): 18–55. <https://doi.org/10.1080/01621459.1964.10480699>; Weisberg, *Total Survey Error Approach* und Smith, T. W. 2011. „Refining the Total Survey Error Perspective“. *International Journal of Public Opinion Research* 23 (4): 464–84. <https://doi.org/10.1093/ijpor/edq052>.

Befragten um Konsistenz über Antworten hinweg zu Messfehlern führen. Daher liegt auch in diesem Fall der Fehler im Bereich der Messung bei der *Response*. Ein Beispiel für unplausible Werte aufgrund der Befragungsstichprobe, also im *Representation*-Strang des TSE, ist die versehentliche Befragung einer falschen Person in einer Panel-Folgewelle. Demographische Angaben der Befragungsperson können naturgemäß nicht übereinstimmen.

Die nachfolgenden Kapitel beschreiben im Detail weitere Fälle von Fehlerquellen, die erfahrungsgemäß zu unplausiblen Werten führen können. Sie lassen sich wie folgt im TSE-Schema verorten:

Filterfehler (Kapitel 4) können einerseits in einem fehlerhaften Fragebogendesign begründet liegen, wenn z. B. ein Webfragebogen Befragte gemäß den vorherigen Angaben zur falschen Frage weiterleitet. Andererseits geschieht es auch, dass Interviewer:innen oder Befragte fälschlicherweise Fragen überspringen bzw. Fragen beantworten (lassen), die nicht hätten beantwortet werden sollen oder müssen. Auch in diesen Fällen liegt der Fehler in der Messung und kann im TSE-Schema dem *Measurement Error* zugeordnet werden. Auch die Plausibilität von **Episodendaten** (*spells*, Kapitel 7) leidet unter Messfehlern. Ein fehlerhaftes Fragebogendesign oder Fehlinterpretationen von Angaben der Befragten können auch hier Datenfehler auslösen. Episodendaten, die über mehrere Erhebungswellen erfasst werden, stellen für Befragte mitunter eine Herausforderung dar, weil dort in Bezug auf eine einzelne Episode mehrfache Erinnerungsprozesse (*recalls*) ablaufen müssen.

Niemand ist frei von Fehlern – auch Datenmanager:innen nicht. **Wildcodes** (Kapitel 5) und **Inkonsistenzen bei der Verknüpfung von Datensätzen** (Kapitel 6) sind Beispiele von Fehlern, die unter anderem bei der Datenaufbereitung, d. h. nach der Erhebungsphase entstehen können. Im TSE-Schema sind diese Art Fehler als Verarbeitungsfehler *Processing Error* verortet. Im Falle von Wildcode-Fehlern in der Datenaufbereitungsphase wurden Antworten von Befragten versehentlich einem Wert außerhalb eines vorher festgelegten Wertebereichs zugeordnet. Treten technische Fehler bei der Zusammenführung von Datensätzen auf, beispielsweise weil einer Befragungsperson versehentlich die falsche Personen-ID zugewiesen wurde, passen die Fälle in der Datenmatrix nicht mehr zueinander und die Angaben erscheinen naturgemäß unplausibel. Darüber hinaus können auch bewusst getroffene Entscheidungen im Umgang mit unplausiblen Werten Verarbeitungsfehler produzieren. Als Resultat von Korrekturen oder dem Löschen von Werten und Fällen kann es zu Repräsentationsfehlern durch *Non-Response* kommen.

2.2 Forschungsdatenzyklus

Die Einordnung dieser Beispiele von möglichen Fehlerquellen in das TSE-Schema macht deutlich, dass Fehler zu unterschiedlichen Zeitpunkten eines Umfrageprozesses auftreten. Für die Darstellung dieser Prozesse hat sich in den letzten Jahren der so-

genannte „Forschungsdatenzyklus“⁸ durchgesetzt. Forschungsdatenzyklus-Modelle bilden in der Regel komplette Forschungsprozesse ab. Je nach Datentyp, Perspektive und Zielgruppe kann man die einzelnen, zusammenhängenden und aufeinander aufbauenden Phasen eines solchen Forschungsdatenzyklus, in denen Informationen produziert oder manipuliert werden, sehr detailliert beschreiben. Obwohl der Begriff „Zyklus“ eine zirkuläre Darstellung suggeriert, können die Prozesse je nach Abhängigkeiten auch linear angeordnet sein. Ein Überblick zu Forschungsdatenzyklusmodellen und Darstellungsformen bietet z. B. Wissik und Durco 2015⁹. Da diese Handreichung auf verschiedene Datentypen mit ihren zum Teil studienspezifischen Workflows eingeht, wird an dieser Stelle kein umfangreicher Forschungsdatenzyklus skizziert. Alle hier beschriebenen Datentypen haben jedoch eine Gemeinsamkeit: An einer Stelle ihrer Entwicklung und Verarbeitung durchlaufen sie einen Plausibilisierungsprozess.



Abbildung 2: Plausibilisierungsrelevante Etappen eines Forschungsdatenzyklus

Um den Fokus auf diesen Plausibilisierungsprozess zu setzen, beschränkt sich der Forschungsdatenzyklus in [Abbildung 2](#) auf die drei linear angeordneten Phasen in der Plausibilisierung stattfindet: die Studienplanungsphase, die Datenerhebungsphase und die Datenaufbereitungsphase. Diese drei Phasen werden in den folgenden Kapiteln aufgegriffen, um die Einordnung in den Prozess zu verdeutlichen.

⁸Synonym werden die Begriffe „(Forschungs-) Datenlebenszyklus“ oder „(Research-) Data Lifecycle“ genutzt.

⁹Wissik, Tanja und Matej Ďurčo. 2015. „Research Data Workflows: From Research Data Lifecycle Models to Institutional Solutions“. In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*, 94–107. Linköping University Electronic Press, Linköping Universität.

3 Der Plausibilisierungsprozess

Percy Scheller, Maximilian Trommer

Im Folgenden sollen einige zentrale Überlegungen und Konzepte des Plausibilisierungsprozesses ausgeführt werden, die für das Verständnis der in den folgenden Kapiteln beschriebenen Use Cases hilfreich sind. Dort führen wir die Schritte zur Plausibilisierung von Wildcodes, Filterfehlern, verknüpften Daten und Episodendaten aus. Es können jedoch keine eindeutigen Handlungsempfehlungen gegeben werden, welche Schritte in welcher Reihenfolge vorzunehmen sind. Nach unserem Erfahrungsaustausch laufen die Prüfroutinen je nach Datenkollektion/Survey sehr unterschiedlich ab. Das heißt, die verschiedenen Teams haben jeweils eigene, von den Variablen im Datensatz abhängige Workflows. Prinzipiell wird meist mit einem ersten Sichten „oben“ angefangen.

3.1 Wo können Fehler in Datensätzen auftreten?

Die folgenden Ausführungen beziehen sich immer auf quantitative, standardisierte, „rechteckige“ Datensätze, bestehend aus Variablen und ihren dazugehörigen Ausprägungen. Jeder Fehler in sozialwissenschaftlichen Datensätzen betrifft zum einen Fälle und zum anderen die enthaltenen Variablen. Die Ausprägungen werden zwar meist als einzelne Fälle in einzelnen Variablen betrachtet, je nach Quelle des Fehlers gibt es aber in der Regel typische Muster von Fehlerhäufungen, die in 3.1.1 bis 3.1.4 unterschieden werden: So können Fehler von der Art der Frage und der Antwortmöglichkeiten begünstigt werden und dementsprechend variablenspezifisch sein, sich aber auch über viele Variablen eines Falls hinweg häufen. Im schlechtesten Fall sind ganze Erhebungen, Befragungswellen oder Fragebögen betroffen. Sowohl die Identifikation von Fehlern als auch der richtige Umgang damit sind von der Art dieser Fehler abhängig.

3.1.1 Variablenspezifische Fehler

Im Fokus stehen meist Fehlerhäufungen über mehrere Fälle innerhalb einzelner Variablen. Beispiele hierfür sind Formulierungsfehler in Befragungsinstrumenten, Änderungen gesetzlicher Definitionen (mit Bezug zum Inhalt, wie beispielsweise ein späteres mögliches Ende der Erwerbstätigkeit durch ein geändertes Renteneintrittsalter), Programmierfehler in der Fragefilterung oder bei der Codierung sowie nicht-disjunkte oder -erschöpfende Kategorien. Die Arten der Fehler sind sehr unterschiedlich, bei der typischen Betrachtung von Häufigkeitsverteilungen oder Kreuztabellen fallen viele Fehler schnell auf. Auch für sich genommen plausible Einzelwerte können als extreme Verteilung über die Fälle hinweg (bis hin zu nicht vorhandenen Ausprägungen im Fragebogen) leicht aufgedeckt werden. Darüber hinaus können auch unplausible Durchschnittswerte auf Fehler in der Variable hindeuten (beispielsweise, wenn der Alters-

durchschnitt von Studierende unter 18 liegen würde). Oft wird eine Anpassung vorgenommen, weil es nicht sinnvoll wäre, eine Korrektur den Datennutzer:innen zu überlassen. Ansonsten eignen sich insbesondere Flag-Variablen (siehe Informationskasten 2 in Kapitel 3.3.2) zur Markierung von variablenspezifischen Fehlern.

3.1.2 Fallspezifische Fehler

Quellen fallspezifischer Fehler sind beispielsweise mangelnde Sprachkenntnisse von Befragten, Missverständnisse, ungeeignete Endgeräte oder willkürliches Ausfüllen von Fragebögen. Auch wenn die Fehler zunächst einzeln auffallen, ist die Identifikation auf der einen Seite recht leicht, da mit jeder Prüfung einer Variablen der gleiche Fall hervorsteicht. Auf der anderen Seite ist sie schwierig, da in der elektronischen Datenverarbeitung meist Variablen und nicht Fälle betrachtet werden. Fallspezifische Fehler werden häufig als fehlender Wert behandelt, das heißt auf „Missing“ gesetzt. Dominieren die falschen und fehlenden Werte den gesamten Fall, kommt auch ein Löschen des Falls in Betracht.

3.1.3 Erhebungsübergreifende Fehler

Eine erhebungsübergreifende Fehlerhäufung kann durch die Auswahl- und Erhebungsmethodik verursacht werden. Die vollständige Fehlerhaftigkeit ganzer Datensätze oder großer Fallgruppen liegt vor, wenn z. B. die Sprache nicht zur Stichprobe passt. Auch kann der Erhebungszeitpunkt oder der Ort aus verschiedenen Gründen nicht der Stichprobendefinition entsprechen (bzw. die Stichprobendefinition kann selbst mangelhaft sein). Druckfehler in Papierfragebögen und Programmierfehler im Zugang oder der Darstellung auf digitalen Geräten sind ebenfalls denkbar. Aufgrund der umfangreichen Fehler in vielen Fällen und Variablen erübrigen sich aufwändige Prüfmechanismen, da die Probleme offensichtlich sind. Bei solch umfangreichen Fehlern ist die Markierung und Korrektur einzelner Variablen und Fälle nicht mehr sinnvoll; in der Regel ist hierzu eine eigene Dokumentation nötig. Methodische Schwachpunkte lassen sich zwar unter Umständen schwer korrigieren, aber oft gut durch die Datennutzer:innen bewerten.

3.1.4 Interviewer:innenbezogene Fehler

Wenn die Erhebung durch Interviewer:innen durchgeführt wird, sind auch systematische Fehler in allen von einer Interviewerin oder einem Interviewer erhobenen Fällen denkbar. Interviewer:inneneffekte zeigen sich beispielsweise in abweichenden Durchschnittswerten; Fälschungen zeigen sich durch untypische Reihen, Muster oder identische Antworten über mehrere Fälle.¹⁰

¹⁰Gefälschte Fälle werden meist aus dem Datensatz entfernt, bekannte Interviewer:inneneffekte schriftlich dokumentiert. Auch der Einsatz von Flag-Variablen zur Markierung kritischer Fälle ist

3.2 Auf welchen Wegen lassen sich fehlerhafte Werte identifizieren?

Bei der Identifikation von Unplausibilitäten in quantitativen Daten können zwei Wege unterschieden werden: **Die Identifikation erfolgt induktiv**, wenn aus den Daten heraus unbekannte Fehler identifiziert werden. Dies entspricht im Wesentlichen einer kritischen ersten Sichtung beziehungsweise der Exploration der Daten ohne spezielles Ziel und wird typischerweise bei neuen Erhebungen angewendet, um neuartige Fehler aufzuspüren. Beispielsweise könnte auffallen, dass eine Variable versehentlich gar keine Werte enthält.

Als zweiter Weg wird eine gezielte **deduktive Identifikation** von Unplausibilitäten verfolgt – die Suche nach aus anderen Datenaufbereitungen bekannten Fehlern, beziehungsweise die systematische und umfassende Prüfung von in vorheriger Induktion erkannten Fehlern in den Daten. Ein Beispiel ist die Suche nach Wildcodes, nach fehlenden Werten (Missings), nach Fehlern aus der Verknüpfung verschiedener Datenquellen und nach Fehlern, wie sie typischerweise bei der Erhebung von Episodendaten entstehen (siehe im Detail die folgenden Abschnitte zu diesen Use Cases).

Informationskasten 1: Deduktives vs Induktives Prüfen

In der Forschungspraxis werden zumeist die Wertebereiche aller Variablen mit angenommenen möglichen Werten abgeglichen. Anhand von Kontextinformationen lassen sich die angenommenen möglichen Wertebereiche in der Regel schon dann festlegen, wenn die konkreten Fehler noch nicht bekannt sind. Beispielsweise könnte die Regel aufgestellt werden, dass das angegebene Alter einen positiven Wert annehmen muss oder dass es das Alter des bisher ältesten Menschen der Welt nicht (wesentlich) überschreiten darf. Dieses Vorgehen entspricht einer deduktiven Prüfung. Die Regeln werden aus dem Kontext abgeleitet.

Abgesehen vom ersten Sichten ist das induktive Prüfen seltener. Es erfolgt zum Beispiel aufgrund zufällig aufgefallener Ungereimtheiten in der Datenverarbeitung und oft aus fallweiser Betrachtung. Induktive Identifikationen von Unplausibilitäten können zur Anpassung von Regeln der deduktiven Prüfung führen.^a

^aIn den DZHW-Studienberechtigtenbefragungen (https://www.dzhw.eu/forschung/projekt?pr_id=465) fiel z. B. auf, dass Schulabgänger:innen mit Fachhochschulreife nach einer Ausbildung an einer Universität studieren und somit einer früher aufgestellten Regel (Universitätsstudium nur mit Abitur/Immaturenprüfung) widersprechen. Aufgrund dieser induktiv festgestellten Abweichung musste die Regel für die deduktive, regelbasierte Prüfung angepasst werden.

Fehler lassen sich relativ einfach identifizieren, wenn sie gehäuft in mehreren Fällen einer Variablen auftreten. Die gängigste Vorgehensweise zur Identifikation von Fehlern ist dementsprechend, sich alle Ausprägungen einer (als Häufigkeitsverteilung) oder die Kombination zweier Variablen (als Kreuztabelle) anzusehen. Fehlerhafte Wer-

denkbar (siehe Informationskasten 2 zu Flag-Variablen in Kapitel 3.3.2).

te in einzelnen Variablen sind unserer Erfahrung nach am häufigsten. Zu unterscheiden sind hierbei (wie auch generell) die Prüfung von rein numerischen Variablen und alphanumerischen Variablen (Zeichenfolgen, sog. Strings).

3.2.1 Numerische Variablen

Die meisten Variablen in quantitativen sozialwissenschaftlichen Datensätzen sind numerisch codiert. Auch wenn das Vorgehen sehr ähnlich ist, sollte bei der Identifikation technisch unterschieden werden zwischen der Prüfung von Werten während der Erhebung durch den:die Interviewer:in oder den programmierten Fragebogen sowie der Prüfung von Fehlern nach der Erhebung durch Datenedition.

Die Identifikation von fehlerhaften Werten bei numerischen Variablen geschieht in beiden Fällen stets durch einen Abgleich mit einem Vergleichswert, und zwar entweder

- mit bestehenden Werten aus anderen Quellen, Erhebungen, Wellen, Fragen (mehrfache Messung),
- oder mit theoretisch zu erwartenden Werten, basierend auf Erkenntnissen bestehender Forschung und Alltagserfahrung.

Der Abgleich zwischen zu prüfendem Wert und Vergleichswert lässt sich dabei über bool'sche Operatoren (Gleichungen mit z. B. (un)gleich, größer, kleiner oder (nicht) Teilmenge von ...) prüfen. Auch eine unplausible Kombination von zwei Werten unterschiedlicher Variablen eines Falls wird anschließend noch einmal mit einem Vergleichswert abgeglichen. Im Vorfeld sind hierzu aber weitere mathematische Operationen denkbar (z. B. Summen, Differenzen, Verhältnisse, Produkte, Rang- und Reihenfolgen). So könnte z. B. das Bruttoeinkommen aus zwölf Monaten aufaddiert werden und anschließend mit dem abgefragten Jahreseinkommen oder auch mit dem maximal erwarteten Jahreseinkommen abgeglichen werden.

Diese Tests können in **online- und computer-unterstützten Befragungen** bereits in die Befragung eingebaut werden; auch eine Prüfung durch eine betreuende Person während der Erhebung ist denkbar.

Darüber hinaus können die Werte **aber auch per Hand oder Skript** im Nachhinein geprüft werden. Bei großen Datenmengen sind skriptbasierte Prüfroutinen effizienter. Eine sinnvolle Prüfroutine kann zum Beispiel sein, sich die Anzahl an unplausiblen Werten eines Falles innerhalb von Itemskalen oder insgesamt ausgeben zu lassen. Darüber hinaus lassen sich gelegentlich Werte selbst dann als fehlerhaft identifizieren, wenn die einzelnen Werte nicht unplausibel erscheinen. Dies gelingt durch die Ausgabe von Antwortmustern (z. B. alle Items mit identischer Ausprägung).

Aus der Prüfung resultiert die Erkenntnis, dass

- der Wert nicht anzuzweifeln ist, da nichts gegen diesen Wert spricht,
- der Wert unwahrscheinlich ist (beispielsweise das Alter eines Studierenden von 12 Jahren) oder
- der Wert sicher falsch ist (beispielsweise das Alter eines > Studierenden von 181 Jahren).

Während die Prüfbedingungen sich also mathematisch scharf formulieren lassen, sind die Ergebnisse keinesfalls immer so eindeutig und beruhen fast immer auf nicht zu beweisenden Annahmen.

3.2.2 Strings

Werte als Zeichenfolgen (Strings) entstehen in der Regel aus offenen Angaben in Befragungen oder Textanalysen (z. B. *screen scraping* als Auslesen von Daten aus Webseiten oder Dokumenten). Offene Angaben werden dabei insbesondere angewendet,

- wenn Antwortkategorien nicht vollständig bekannt sind,
- die Menge an Antwortkategorien zu groß zur Darstellung ist oder
- zur Ergänzung ansonsten geschlossener Angaben („Sonstige“).

Diese werden zur quantitativen Auswertung in der Regel codiert, das heißt unter weniger Kategorien numerischer Art subsumiert. Wenn die Codierung eindeutig ist, aber die Inkonsistenz im Abgleich mit anderen Daten entsteht, unterscheiden sich offene Angaben nicht von numerischen Daten (siehe Abschnitt 3.2.1 „Numerische Variablen“).

Umgekehrt können in seltenen Fällen ergänzende Angaben aus anderen Variablen die Zuweisung zu numerischen Werten ermöglichen (zum Beispiel, wenn Angaben zu einer aktuellen Erwerbstätigkeit und zusätzlich ein Erwerbstätigkeitstableau (siehe Kapitel 7 zu Episodendaten) erhoben wird).

Wenn die Zuweisung zu numerischen Werten nicht möglich oder nicht eindeutig ist, sollte kein Wert zugewiesen werden (Es kann für manche Auswertungen sinnvoll sein, hier einen speziellen „fehlenden“ Wert - Missingcode - beziehungsweise einen eigenen Wert zu codieren). Denkbar ist in manchen Fällen aber auch ein Zuweisen zu neu geschaffenen, eindeutigen Überkategorien (z. B. „Kontinent“ statt „Staat“, „Jahreszeit“ statt „Monat“).

Aufgrund der Individualität der enthaltenen Informationen sind Strings besonders häufig fallspezifisch und müssen zur Anonymisierung gelöscht oder in allgemeine Kategorien einsortiert werden. Folglich können Datennutzer:innen die Angaben nicht im Original einsehen. Eine behutsame Plausibilisierung und Dokumentation sind dementsprechend besonders wichtig.

3.3 Umgang mit fehlerhaften Werten

Entscheidungen zum Umgang mit fehlerhaften Werten sollten nicht willkürlich erfolgen. Grundsätzlich ist zu empfehlen, Entscheidungen für Plausibilitätsprüfungen und -anpassungen in einem Team zu treffen, um wenigstens eine Intersubjektivität zu gewährleisten. Im Idealfall könnte sich ein solches Team aus verschiedenen Expert:innen zusammensetzen: Einem solchen Team könnten beispielsweise neben erfahrenen Forscher:innen, also den inhaltlichen Expert:innen für das Forschungsfeld, auch die direkt an der Umfrage beteiligten Personen (bspw. Interviewer:innen, Umfrageinstitut, Befragte) angehören.

Nach der Identifikation eines (wahrscheinlichen) Fehlers bestehen folgende Möglichkeiten:

- Keine Änderung vornehmen
 - ohne Markierung (akzeptieren)
 - mit Markierung (über Flag-Variablen oder Umcodierung in Flag-Werte)
- Änderung vornehmen (korrigieren)
 - Ersetzen durch fehlenden Wert (Missing)
 - Ersetzen durch neuen Wert

Diese Maßnahmen sollen hier kurz allgemein erläutert werden. In den vier betrachteten Use Cases der Folgekapitel (4 „Filterfehler“, 5 „Wildcodes“, 6 „Inkonsistenzen bei der Verknüpfung mehrerer Datensätze/-quellen“ und 7 „Inkonsistenzen bei der Verknüpfung von längsschnittlichen Episodendaten“) wird der Umgang dann detaillierter dargestellt.

3.3.1 Keine Änderung vornehmen, ohne Markierung (Flag)

Keine Änderung vorzunehmen ist die Vorgehensweise, die am wenigsten Ressourcen seitens der Datenproduzent:innen benötigt. Außerdem wird der Datensatz schlank gehalten, weil keine Flag-Variablen oder korrigierten Variablen ergänzt werden. Bei diesem Vorgehen können keine Fehler durch fehlerhafte Datenänderungen/Datenaufbereitung entstehen, sondern nur durch das Unterlassen dieser. Denn es bleibt das Risiko bestehen, dass Datennutzer:innen die Daten fehlinterpretieren, wenn keine Änderungen vorgenommen werden, und nicht über Plausibilitätsprobleme informiert werden. Das Wissen der Datenproduzent:innen aus der Plausibilitätsprüfung wird nicht an die Datennutzer:innen weitergegeben, die Plausibilisierungsschritte müssten von den Datennutzer:innen vorgenommen werden und verursachen somit unter Umständen redundantes Arbeiten, das aus gesamtwissenschaftlicher Sicht die Aufwände erhöht.

3.3.2 Keine Änderung vornehmen, mit Markierung (Flag)

Die Markierung kann durch eine Anmerkung im Codebuch, eigene Ausprägungen innerhalb von Variablen oder durch eigene Variablen zur Markierung, sogenannte Flag-Variablen (siehe Informationskasten 2 in Kapitel 3.3.2), geschehen. Wenn sich Fehlermuster wiederholen, kann manchmal auf eine ganze Reihe fehlerhafter Fälle (oder Variablen) mit einer einzigen übergreifenden Flag-Variable hingewiesen werden. Flag-Variablen sind nur dann von Nutzen, wenn die Interpretation durch eine gute Dokumentation unterstützt wird (sinnvolle Benennungen als Label, schriftliche Erläuterungen und Empfehlungen in der Dokumentation). Die Erstellung von verständlichen Markierungen kann sehr aufwändig sein. Zu zahlreich eingesetzt können Flag-Variablen den Datensatz aufblähen und die Dokumentation unübersichtlich machen. Ein Vorteil ist jedoch, dass den Datennutzer:innen bereits Hinweise auf unplausible Werte – manchmal in Kombination mit Interpretationsmöglichkeiten – gegeben werden können und sie weniger leicht übersehen werden. Die Nutzer:innen können im Idealfall auf diese Weise aufgeklärt werden, aber selbständig entscheiden, wie sie mit der Situation umgehen. Darüber hinaus können in Einzelfällen die unplausiblen Werte selbst von Erkenntnisinteresse insbesondere für methodische Forschungsfragen sein.¹¹

Informationskasten 2: Flag-Variable

Flag-Variablen sind in der einfachsten Form dichotome Variablen, die den Wert „wahr“ oder „unwahr“ annehmen. Im Fall der Plausibilisierung **markieren** sie, ob die Werte einer korrespondierenden Variable unplausibel sind (0 = unwahr, 1 = wahr).

Für jede Variable muss entschieden werden, ob eine eigene Flag-Variable angelegt werden soll. Würde für jede Variable eine Flag-Variable angelegt, kann ein Datensatz mit vielen Variablen unübersichtlich werden.

3.3.3 Änderung auf unwidersprüchliche „valide“ Werte

Durch die Anpassung auf den (wahrscheinlich) „validen“ Wert wird entweder die naive Interpretation eines unplausiblen Wertes oder die vielfache, aufwändige Plausibilisierung durch die Datennutzer:innen vermieden. In der Regel haben die Datenproduzent:innen durch ihr Prozesswissen und allumfassende Datensätze (die beispielsweise auch offene Angaben und Metadaten enthalten) mehr Informationen zur Hand, um

¹¹Für die Beantwortung methodischer Fragestellungen sind Rohdaten meist nützlicher. Dies gilt aber manchmal auch für inhaltliche Forschungsfragen. Ein Beispiel zur Verdeutlichung: In einigen Fällen fehlt das Geschlecht des:der Befragten; dieses soll anhand der offenen Berufsangabe auf „weiblich“ gesetzt werden, wenn der Beruf eine weibliche Bezeichnung trägt; es bleibt Missing, falls der Beruf kein Geschlecht andeutet. Dies ist ein vertretbares Vorgehen, wenn für eine Auswertung nur Frauen herangezogen werden sollen. Das gesamte Sample wird aber in Hinblick auf die Geschlechterverteilung verzerrt. Auch Analysen, bei denen das Geschlecht und der Beruf der Befragten in einen direkten Zusammenhang gebracht werden, produzieren zwangsläufig verzerrte Ergebnisse.

auf Fehlerquellen und korrekte Werte schließen zu können. Trotzdem ist eine solche Anpassung von Werten mit Problemen verbunden:

In der Regel ist davon auszugehen, dass bei Widersprüchen der wahre Wert einer Messung nicht bekannt ist. Dementsprechend sind Änderungen (beispielsweise die Interpretation der Angabe von 81 Jahren für Studienanfänger:innen als Zahlendreher) immer mit Unsicherheit behaftet und stellen eine eigene Fehlerquelle dar.

Einige Widersprüche scheinen zwar so offensichtlich, beispielsweise die Verwechslung von Brutto- und Nettogehalt, dass die Anpassung auf einen besseren Wert nahe liegt. Die korrekte, also sinnvolle Interpretation eines unplausiblen Wertes ist aber abhängig von den Forschungsfragen beziehungsweise der Auswertungsstrategie der Nutzer:innen, die zum Zeitpunkt der Plausibilisierung meist noch nicht bekannt sind.

Dementsprechend sollte mit Korrekturen sehr sparsam umgegangen werden. Eine gängige und empfehlenswerte Möglichkeit ist es, verschiedene – korrigierte und unbereinigte – Versionen von einzelnen Variablen oder ganzen Datensätzen zu veröffentlichen.

3.3.4 Änderung auf fehlende Werte (Missings)

Die Anpassung von widersprüchlichen Werten durch Recodierung in fehlende Werte (Missings) ist für sozialwissenschaftliche Datensätze eine gängige Methode. Sie vermeidet die Fehlinterpretation durch Dateneditor:innen oder Datennutzer:innen und ist generell ressourcensparend. Bei einem großen Anteil widersprüchlicher Werte kann dieses Verfahren allerdings zu geringen Fallzahlen bei Auswertungen führen. Darüber hinaus können fehlende Werte die Auswertungen verzerren, wenn die Fehler nicht zufällig verteilt sind.¹²

Alternativ zu den Änderungen auf plausible oder fehlende Werte kommt auch eine Imputation (im Anschluss an eine Änderung auf fehlende Werte) in Betracht. Diese wird hier nicht behandelt, da fachgerechte Imputationen aufwändig sind und nicht pauschal für jeden fehlenden Wert durch die Datenproduzent:innen durchgeführt werden können. Auch um die Auswirkungen auf die eigenen Analysen einschätzen zu können, sollten sie in der Regel den Nutzer:innen überlassen werden.

3.4 Entscheidungskriterien zum Umgang mit unplausiblen Werten

Die Entscheidung für den richtigen Umgang mit unplausiblen Werten ist immer abhängig vom jeweiligen Datensatz und der Art der Fehler. Prinzipiell ist dabei aber auch die Akzeptanz des ursprünglichen Wertes eine Entscheidung „für“ etwas und nicht neutral.

¹²Allison, Paul. 2002. *Missing Data*. Thousand Oaks: SAGE Publications. <https://doi.org/10.4135/9781412985079>, sowie Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.

Auch wenn hier keine generelle Empfehlung für Maßnahmen gegeben werden kann, sollen Entscheidungskriterien genannt werden.

- **Forschungsfrage/Auswertungsstrategie:** Ob eine Änderung der Daten vorgenommen werden sollte, ist für Datennutzer:innen nicht zuletzt abhängig von der Forschungsfrage und Auswertungsstrategie. Bei Subsample-Analysen können Fehler gravierende Anteile ausmachen, die sonst im statistischen Rauschen untergehen, oder sogar subsample-spezifisch sein (z. B. bei sprachlichen Verständnisproblemen). Ein anderes Beispiel sind Wildcodes, deren inhaltliche Information für eine Forschungsfrage relevant sein kann, während entsprechende Fälle bei anderen Forschungsfragen besser standardisiert recodiert werden. Da Datenproduzent:innen die Forschungsfragen der Nutzer:innen in der Regel nicht bekannt sind, sollte – wie bereits erwähnt – sehr sparsam mit Änderungen umgegangen werden. Originalwerte sollten nach Möglichkeit verfügbar bleiben und alle Änderungen gut dokumentiert werden.
- **Relevanz des Fehlers:** Auch die potenzielle Relevanz eines Fehlers kann eine Rolle spielen: Sind Werte zwar im Einzelnen nicht zwangsläufig fehlerhaft, hätten aber als Fehler große Auswirkungen auf die Auswertungen, ist ein Hinweis umso wichtiger.
- **Häufigkeit des Fehlers:** Ob eine Kennzeichnung durch Flag-Werte und Variablen sinnvoll ist, hängt von der relativen Häufigkeit eines Fehlers (Fallzahlen) und der Menge an Kennzeichnungen insgesamt ab. Flag-Werte und -Variablen können den Datensatz schnell aufblähen: Wird für jeden potenziellen Fehler eine eigene Variable erstellt, so hat sich die Zahl der Variablen schnell verdoppelt.
- **Wahrscheinlichkeit des Fehlers:** Ob eine Änderung auf neue Werte vorgenommen werden sollte, ist von der Wahrscheinlichkeit des Fehlers abhängig. Ist der korrekte Wert aus anderen Daten oder Prozesswissen über systematische Fehler bekannt – handelt es sich also *sehr* wahrscheinlich um einen Fehler – spricht wenig gegen eine Korrektur. Dabei ist das Risiko, durch die Anpassung neue Fehler zu produzieren, in die Abwägung einzubeziehen. Dieses Risiko muss nicht nur gegenüber dem ursprünglichen, fehlerhaften Wert kleiner sein, sondern auch gegenüber der Implementierung fehlender Werte oder einer Imputation. Die Messlatte ist dementsprechend hoch. Auch ein bloßer Vorschlag neuer Werte in einer neuen Variable kann uninformierte Datennutzer:innen schnell zur Übernahme verleiten und ist mit Bedacht abzuwägen. Je wahrscheinlicher ein Fehler ist, desto eher sollte auf den Fehler hingewiesen und mögliche Korrekturen vorgeschlagen werden. Vollkommen unplausible Werte sollten nicht ohne Hinweis Teil von (geteilten) Datensätzen sein. Zumindest sollte auf die unbearbeitete Form der Daten hingewiesen werden.
- **Fallzahlen:** Die Fallzahlen spielen bei der Änderung auf neue Werte eine untergeordnete Rolle. Zwar sind Anpassungen umso folgenreicher, je zahlreicher sie sind- allerdings ist die Qualität der Auswertungen umso besser, je mehr Werte korrekt sind. Es gibt also weder ein Mindestmaß an Fällen noch ein Maximum, um Korrekturen vorzunehmen.

Grundsätzlich muss bei der Datenaufbereitung abgewogen werden, ob überhaupt eine Plausibilisierung der Daten stattfinden soll. Die Verlockung und Gefahr, im Rahmen des Plausibilisierungsprozesses möglichst saubere Daten zu produzieren und dafür zu starke Annahmen zu treffen (die letztlich zu einer Verfälschung der Daten führen können), ist groß. Zudem nimmt es den Nutzer:innen potenziell die Möglichkeit, das Vorhandensein beziehungsweise den Umfang von Plausibilitätsproblemen selbst zu bewerten und den Umgang damit mit eigenen Annahmen zu versehen.

Die meisten anpassenden Schritte der Datenverarbeitung sind (zumindest logisch) unumkehrbar und beeinflussen somit direkt die Auswertung. Umgekehrt ist der Sinn der Anpassung immer auch von den Forschungsfragen und der Auswertungsstrategie abhängig. Da in der Datenerhebung selten alle später bearbeiteten Forschungsfragen bekannt sind, ist es sinnvoll, die Anpassung den Datennutzer:innen zu überlassen.

Dagegen sprechen aber:

- die mangelnde Effizienz durch mehrfache, parallele Bearbeitung,
- dass Nutzer:innen meistens Prozesswissen fehlt und ihnen Daten oft nur in bearbeiteter und anonymisierter Form vorliegen und
- eine verminderte Auswertungsobjektivität durch mangelnde Standardisierung.

Insofern kann eine Anpassung durch Datenproduzent:innen und -editor:innen durchaus sinnvoll sein, bleibt aber stets eine Abwägung im Hinblick auf die Forschungsfragen, Auswertungsstrategie und die wahren Werte.

3.5 Dokumentation des Umgangs mit unplausiblen Werten

Da **Nachvollziehbarkeit oder Replizierbarkeit** zentraler Bestandteil aller wissenschaftlichen Schritte ist, sollte transparent, übersichtlich, begründet und verständlich sein, welche Anpassungen vorgenommen wurden. Jede Änderung sollte zumindest auf Nachfrage reproduzierbar und somit im Ideal auch umkehrbar sein. Dementsprechend wird bereits vorweg empfohlen, alle Vorgänge zur Plausibilisierung während der Datenaufbereitung möglichst ausführlich und verständlich zu dokumentieren, um den Nutzer:innen Hinweise zu geben, welche Korrekturen an den Daten vorgenommen oder nicht vorgenommen wurden.

Hiermit verbunden ist die grundsätzliche Empfehlung, bei einer etwaigen **Plausibilisierung regelbasiert** vorzugehen (siehe Regelbeispiele in Informationskasten 1 „Deduktives vs. induktives Prüfen“ in Kapitel 3.2 und in den Folgekapiteln zu den Use Cases) und lediglich im Ausnahmefall und in sehr geringem Umfang Einzelfälle zu plausibilisieren (etwa, weil sie nach der regelbasierten Plausibilisierung übrig geblieben und durch das Raster gefallen sind). Regelbasierte Vorgänge zur Plausibilisierung sind wesentlich einfacher zu dokumentieren und für die Nutzer:innen nachvollziehbar.

Die korrekte Identifikation und der richtige Umgang mit fehlerhaften beziehungsweise unplausiblen Werten ist immer abhängig vom Wesen der verwendeten Daten. Allgemein sollte eine Überprüfung und Korrektur fehlerhafter Werte im Forschungsprozess jedoch **bereits während oder direkt nach der Datenerfassung und -codierung** geschehen. Von einer Korrektur fehlerhafter Werte auf Papierfragebögen vor einer Erfassung ist abzuraten, da dies wesentlich schwieriger zu standardisieren und nachzuvollziehen ist. Des Weiteren sollte eine Korrektur unbedingt vor eventueller Gewichtung und Imputation geschehen.

4 Filterfehler

Kristina Krell, Tanja Kunz, Senta-Melissa Pflüger

4.1 Beschreibung

Filterfragen werden eingesetzt, um eine oder mehrere für die Befragten nicht zutreffende oder nicht (sinnvoll) zu beantwortende Fragen zu überspringen. So können beispielsweise Fragen zur Partnerschaft nur von Personen sinnvoll beantwortet werden, die aktuell in einer Partnerschaft leben. Mittels Filterinstruktionen werden Interviewer:innen bzw. Befragte (im Falle einer selbst-administrierten Befragung) auf bestimmte Fragen oder Abschnitte im Fragebogen verwiesen bzw. weitergeleitet. Filterinstruktionen setzen sich aus einer Filterbedingung („wenn“-Komponente) und einer Filterführung („dann“-Komponente) zusammen (z. B. „Wenn Antwort X ausgewählt, dann springe zu Frage Y“).

Handelt es sich, wie im folgenden Beispiel aus einer Befragung des Sozio-ökonomischen Panels (SOEP)¹³, um eine nicht-computergestützte Datenerfassung, werden Filterinstruktionen durch verbale Aufforderungen bzw. grafische Hinweise (z. B. Pfeile) umgesetzt (Abbildung 3). Bei einer computergestützten Datenerfassung hingegen werden Filterbedingung und -führung programmiert und automatisiert ausgeführt, wodurch auch deutlich komplexere Filterführungen möglich sind.

192. Haben Sie derzeit eine feste Partnerschaft?
 Ja..... Nein ➔ Frage 197!
 ↓

193. Wohnt Ihr Partner / Ihre Partnerin hier im Haushalt?
 Ja..... Nein ⚙

Abbildung 3: Beispiel für eine Filterfrage (Quelle: SOEP Personenfragebogen 2018)

Im Beispiel sollen nur diejenigen Befragten, die derzeit eine feste Partnerschaft haben und auf die Filtereingangsfrage 192 mit „Ja“ geantwortet haben, die darauffolgenden Fragen zu den Partner:innen beantworten. Haben die Befragten mit „Nein“ geantwortet, sind die Interviewer:innen – bzw. in selbst-administrierten Erhebungen die Befragten – aufgefordert, die Folgefragen zu überspringen und direkt mit Frage 197 fortzufahren.

¹³https://www.diw.de/documents/publikationen/73/diw_01.c.622046.de/diw_ssp0608.pdf.

Es können zwei Arten von Filterfehlern unterschieden werden, die sowohl in nicht-computergestützten als auch computergestützten Erhebungen auftreten können.¹⁴

- a. Zutreffende Fragen werden von Interviewer:innen oder Befragten fälschlicherweise übersprungen und von den Befragten nicht beantwortet (*error of omission*). Die Folge sind fehlende Angaben (Datenverlust; *item non-response*).
- b. Nicht zutreffende Fragen werden von Interviewer:innen oder Befragten fälschlicherweise nicht übersprungen und von den Befragten beantwortet (*error of commission*). Die Folge sind unnötige oder widersprüchliche Angaben (Dateninkonsistenz).

Im obigen Beispiel liegt ein *error of omission* vor, wenn Interviewer:innen oder Befragte fälschlicherweise von Frage 192 direkt zu Frage 197 übergehen, d. h. die Fragen 193 bis 196 überspringen, obwohl sie beantwortet werden sollten. Wenn Befragte angegeben haben, derzeit in einer festen Partnerschaft zu leben und die Frage 192 mit „Ja“ beantwortet haben, fehlen entsprechend die Antworten auf die darauffolgenden Fragen zu dem bzw. der Partner:in.

Im Falle eines *error of commission* fahren Befragte trotz der Antwort „Nein“ auf Frage 192 mit Frage 193 fort. Die Angaben auf die Folgefragen sind in diesem Fall widersprüchlich, da eine Angabe über eine:n Partner:in getroffen wird, der/die nicht existiert.

4.2 Ursachen

4.2.1 In der Studienplanungsphase

Filterfehler können bereits bei der Konstruktion des Fragebogens auftreten, womit die Ursache im Erhebungsinstrument selbst begründet liegt. Wenn bereits in der Fragebogenvorlage die Definition der Filterbedingung und/oder Filterführung fehlerhaft ist, führt dies fast immer zu fehlenden oder unnötigen bzw. widersprüchlichen Angaben – sowohl in nicht-computergestützten als auch in computergestützten Erhebungen. Werden in der Fragebogenvorlage korrekt definierte Filterbedingungen und/oder Filterführungen fehlerhaft in den Papierfragebogen einer nicht-computergestützten Erhebung übernommen oder im Rahmen einer computergestützten Erhebung falsch programmiert, führt dies ebenfalls zu Filterfehlern.

Fehlerhafte Filterinstruktionen in der Programmierung eines computergestützten Fragebogens führen zu systematischen Filterfehlern, wenn Folgefragen aufgrund eines Programmierfehlers nicht angezeigt werden (*errors of omission*). Das bedeutet, dass

¹⁴Redline, Cleo D. und Don A. Dillman. 2002. „The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-Administered Surveys“. In *Survey Nonresponse*. Hrsg. Robert M. Groves, Don A. Dillman, John L. Eltinge und Roderick J. A. Little. New York: Wiley, 179–93.

alle Befragten bzw. Interviewer:innen, die eine fehlerhafte Filterinstruktion im Fragebogen vorfinden, gleichermaßen von dem Problem betroffen sind. Somit ist auch die davon betroffene Befragtengruppe nicht selektiv. Ein Beispiel hierfür ist die Filtereingangsfrage „Rauchen Sie gegenwärtig?“ („Ja“/„Nein“). Fehlt eine geeignete Filterinstruktion, wird die Folgefrage (z. B. „Wie viele Zigaretten rauchen Sie pro Tag?“) auch denjenigen Befragten gestellt, die zuvor mit „Nein“ geantwortet hatten. Dies führt zu a) unnötigen Angaben, wenn die Befragten eine „0“ als Antwort auf die Folgefrage angeben oder b) widersprüchlichen Angaben, wenn die Befragten eine beliebige Zahl größer als 0 angeben.

4.2.2 In der Datenerhebungsphase

Filterfehler können auch während der Datenerhebung auf eine unbeabsichtigte Missachtung von Filterinstruktionen durch die Befragten bzw. Interviewer:innen zurückzuführen sein. Bei einer nicht-computergestützten Datenerfassung liegt dies meist an einer für die Befragten oder Interviewer:innen unzureichenden oder unverständlichen Kennzeichnung der Filterbedingung und/oder Filterführung im Fragebogen (z. B. aufgrund von uneindeutigen verbalen Anweisungen oder schlecht erkennbaren Pfeilen von einer Antwortoption zur Folgefrage).¹⁵ Werden Papierfragebögen von Interviewer:innen administriert, können derartige Filterfehler in mangelhaften bzw. fehlenden Interviewer:innen-Schulungen begründet liegen. Filterfehler infolge von Unachtsamkeit und Verständnisschwierigkeiten seitens der Befragten oder Interviewer:innen können zum einen dazu führen, dass die Filtereingangsfrage korrekt beantwortet wird und die Angaben auf die Folgefragen fehlerhaft sind. Zum anderen ist ebenso möglich, dass die Filtereingangsfrage falsch beantwortet wird und die Angaben auf die Folgefragen wiederum korrekt sind. Oftmals ist nicht zweifelsfrei feststellbar, an welcher Stelle der Fehler passiert, der die unnötigen bzw. widersprüchlichen Angaben zur Folge hat. Relevant ist dies vor allem beim Umgang mit Filterfehlern, da vorab Annahmen getroffen werden müssen, an welcher Stelle ein Filterfehler fußt (d. h. Filtereingangsfrage oder Folgefrage) bzw. welche Angaben korrigiert oder ersetzt werden müssen. Bei einer computergestützten Datenerfassung können Filterfehler infolge von Unachtsamkeit und Verständnisschwierigkeiten seitens der Befragten oder Interviewer:innen durch die korrekte Programmierung der Filterinstruktionen weitestgehend ausgeschlossen werden.

Eine weitere Ursache für Filterfehler kann die beabsichtigte Missachtung von Filterinstruktionen durch die Befragten bzw. Interviewer:innen sein. Daraus resultierende Filterfehler können sowohl im Rahmen einer nicht-computergestützten als auch computergestützten Datenerfassung auftreten. Meist sind sie die Folge einer absichtli-

¹⁵Siehe. bspw. Olson, Kristen, Megumi Watanabe, und Jolene D. Smyth. 2018. „A Comparison of Full and Quasi Filters for Autobiographical Questions“. *Field Methods* 30 (4): 371–85. <https://doi.org/10.1177/1525822X18795868>; Redline, Cleo D., Don A. Dillman, Aref N. Dajani und Mary A. Scaggs. 2003. „Improving Navigational Performance in U.S. Census 2000 by Altering the Visually Administered Languages of Branching Instructions“. *Journal of Official Statistics* 19 (4): 303–419.

chen Falschbeantwortung der Filtereingsfrage durch die Befragten bzw. Interviewer:innen, um detaillierte Folgefragen zu überspringen und den Fragebogen abzukürzen (*motivated underreporting*).¹⁶

Ungeachtet dessen, ob der Filterfehler Folge einer unbeabsichtigten oder beabsichtigten Missachtung von Filterinstruktionen ist, hängt das Auftreten eines Filterfehlers in der Datenerhebungsphase vom individuellen Verhalten der Befragten bzw. der Interviewer:innen ab. Deshalb handelt es sich um unsystematische Filterfehler. Die Zusammensetzung der Gruppe der Befragten bzw. Interviewer:innen, die von dieser Art von Filterfehlern betroffen ist, kann also unter Umständen selektiv sein. So hat eine Studie gezeigt, dass Interviewer:innen, die Interviews face-to-face durchführen (d. h. keine Überwachung, Bezahlung pro abgeschlossenem Interview), im Vergleich zu Interviewer:innen, die Interviews am Telefon durchführen (d. h. direkte Überwachung mittels Bildschirm- und Audioaufnahmen, Bezahlung pro Stunde), im Laufe der Feldphase zunehmend dazu neigen, Interviews durch die absichtliche Missachtung von Filterinstruktionen abzukürzen.¹⁷

4.3 Identifikation

Es empfiehlt sich, bereits während der Fragebogenkonstruktion, aber spätestens im Rahmen der Datenaufbereitung festzulegen, welche Kombinationen von Merkmalsausprägungen zweier oder mehrerer Variablen theoretisch Filterfehler darstellen. So lassen sich mittels Kreuztabellierung der entsprechenden Variablen oder anderen Prüfroutinen im Datenaufbereitungsprozess Filterfehler aufdecken.

Das folgende Beispiel basiert auf Daten der TwinLife-Studie¹⁸. In der ersten Datenerhebung der Studie wurden einige Fragen zum Rauchverhalten gestellt. Diese Fragen wurden bei einem Teil des Samples mittels eines selbst-administrierten Papierfragebogens¹⁹ erhoben, sodass keine computergestützte Filterführung stattfand.

¹⁶Daikeler, Jessica, Ruben L. Bach, Henning Silber und Stephanie Eckman. 2022. „Motivated Misreporting in Smartphone Surveys“. *Social Science Computer Review* 40 (1): 95–107. <https://doi.org/10.1177/0894439319900936>; Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau und Stanley Presser. 2014. „Assessing the Mechanisms of Misreporting to Filter Questions in Surveys“. *Public Opinion Quarterly* 78 (3): 721–33. <https://doi.org/10.1093/poq/nfu030>; Kosyakova, Yuliya, Jan Skopek und Stephanie Eckman. 2015. „Do Interviewers Manipulate Responses to Filter Questions? Evidence from a Multilevel Approach“. *International Journal of Public Opinion Research* 27 (3): 417–31. <https://doi.org/10.1093/ijpor/edu027>.

¹⁷Kosyakova, Yuliya, Jan Skopek und Stephanie Eckman, „Do Interviewers Manipulate Responses“.

¹⁸<https://www.twin-life.de/>.

¹⁹Diewald, Martin, Christian Kandler, Rainer Riemann, Frank M. Spinath, Anastasia Andreas, Tina Baier, Annika Bartling, et al. 2022. „TwinLife. CATI Survey of wave 1, Scientific use file ZA6701_person_wid2“. <https://doi.org/10.4232/1.13987>.

Zunächst wurden die Teilnehmer:innen gefragt, ob sie rauchen (Variable hbe0100, Abbildung 4):

5. Rauchen Sie?

Ja, ich bin starker Raucher	<input type="checkbox"/>	1
Ja, ich bin mäßiger Raucher	<input type="checkbox"/>	2
Ja, ich bin Wenig-Raucher	<input type="checkbox"/>	3
Ja, ich bin Gelegenheits-Raucher	<input type="checkbox"/>	4
Nein, ich bin Ex-Raucher (rauche nicht mehr, habe aber früher geraucht)	<input type="checkbox"/>	5
Nein, ich bin Nie-Raucher (rauche jetzt nicht und habe früher nicht geraucht)	<input type="checkbox"/>	6 → Weiter mit Frage 8

Abbildung 4: Filtereingangsfrage zum Rauchverhalten mit Ausgangsfilterinstruktion im selbst-administrierten Papierfragebogen der ersten Erhebungswelle der TwinLife-Studie.

Hatten die Befragten bei Frage 5 die Kategorie 6 („Nein, ich bin Nie-Raucher“) ausgewählt, wurden sie mittels einer Ausgangsfilterinstruktion („Weiter mit Frage 8“) aufgefordert, direkt mit Frage 8 fortzufahren. Alle übrigen Befragten sollten mit Frage 6 fortfahren. Anschließend sollten laut Eingangsfilterinstruktion in Frage 7 („Wenn Sie jetzt immer noch zumindest gelegentlich rauchen“) lediglich diejenigen Befragten, die Frage 5 mit „Ja“ beantwortet hatten (Kategorie 1-4), angeben, ob sie lieber allein oder in Gesellschaft rauchen (Variable hbe0102, Abbildung 5):

Wenn Sie jetzt immer noch zumindest gelegentlich rauchen:

7. Rauchen Sie eher alleine oder in Gesellschaft von Bekannten oder Freunden?

Alleine	<input type="checkbox"/>	1
In Gesellschaft	<input type="checkbox"/>	2
Teils/teils	<input type="checkbox"/>	3

Abbildung 5: Folgefrage zum Rauchverhalten mit Eingangsfilterinstruktion im selbst-administrierten Papierfragebogen der ersten Erhebungswelle der TwinLife-Studie.

Eine Kreuztabellierung der beiden Variablen in der vom Erhebungsinstitut erfassten Rohdatenversion (Tabelle 1²⁰) zeigt, dass ein nicht unerheblicher Teil der Befragten die Folgefrage fälschlicherweise dennoch beantworteten (Ziffern in Fettdruck und kursiv). Darunter waren sogar vier Personen, die in der Filtereingangsfrage angegeben hatten, noch nie im Leben Raucher:in gewesen zu sein.

Ein solches Prüfungsverfahren kann auch automatisiert und um weitere Dimensionen erweitert werden, um Filterfehler im gesamten Befragungsprogramm bzw. für alle relevanten Variablen effizient zu prüfen und aufzudecken.

²⁰Quelle: Erste Erhebungswelle der TwinLife-Studie: Diewald, Martin et al. 2022. TwinLife. GESIS, Cologne. ZA6701 Data file Version 6.1.0, <https://doi.org/10.4232/1.13987>.

Tabelle 1: Kreuztabellierung der Filtereingangsfrage und Folgefrage zum Rauchverhalten in der TwinLife-Studie.

hbe0100: Rauchverhalten	hbe0102: Rauchverhalten – in Gesellschaft oder allein		
	1: (eher) alleine	2: In Gesellschaft	3: Teils/teils
1: Ja, ich bin starker Raucher	30	10	81
2: Ja, ich bin mäßiger Raucher	64	66	391
3: Ja, ich bin Wenig-Raucher	14	63	83
4: Ja, ich bin Gelegenheits-Raucher	13	181	67
5: Nein, ich bin Ex-Raucher (rauche nicht mehr, habe aber früher geraucht)	5	49	25
6: Nein, ich bin Nie-Raucher (rauche jetzt nicht und habe früher nicht geraucht)	0	4	0

Im Datenaufbereitungsprozess der TwinLife-Studie wird dabei folgendermaßen vorgegangen: Die Filterbedingungen aller Fragen werden aus der Fragebogenvorlage in eine Datenbank importiert und automatisch in Stata-Sprache übersetzt. Durch Anwendung der Filterbedingungen auf die Rohdaten können alle Befragungspersonen identifiziert werden, die auf Grundlage ihrer Antwort auf die Filtereingangsfrage auch die Folgefrage beantwortet haben müssten. Im Anschluss kann überprüft werden, ob für diese Personen nur gültige Werte bzw. fehlende Werte für alle anderen Personen vorhanden sind. Etwaig identifizierte Filterfehler können schließlich bearbeitet werden (siehe Abschnitt 4.4 in diesem Kapitel). Dieses Verfahren kann sehr aufwendig sein und bietet sich vor allem an, wenn viele Variablen von mehrdimensionalen Filterbedingungen betroffen sind.

Ein weiteres Beispiel wird anhand der Absolvent:innenstudie des Deutschen Zentrums für Hochschul- und Wissenschaftsforschung (DZHW)²¹ skizziert. Zunächst wird eine Kreuztabelle der Filtervariable (Filtereingangsfrage) und der zu filternden Variable (Folgefrage) ausgegeben. Gibt es Fälle, in denen die eigentlich gefilterte Variable gültige Werte aufweist, wird versucht, die Ursache des Filterfehlers zu ermitteln. In den Online-Befragungen handelt es sich meist um zunächst versehentlich ausgewählte Antwortkategorien der Filtereingangsfrage, die zu falschen Angaben in der Filtervariable führen. Da die gefilterte Folgefrage auf dem nächsten Screen dann häufig nicht sinnvoll erscheint, entscheiden sich viele Befragte, zur Filtereingangsfrage auf dem vorherigen Screen zurückzugehen (sofern ein Zurück-Button vorhanden ist) und den filternden Wert zu korrigieren, indem bei der Filtereingangsfrage eine andere Antwortkategorie ausgewählt wird. Bereits gemachte Angaben in der Folgefrage bleiben jedoch aus technischen Gründen im Datensatz bestehen, was die aufgezeigte Widersprüchlichkeit in den Angaben begründet. Etwas komplizierter gestaltet sich die Identifikation bei Papierfragebögen: Hier handelt es sich bei der Beantwortung der Filtereingangsfrage und dem „Springen“ zur Folgefrage um zwei getrennte Prozesse. Es ist selten möglich

²¹https://www.dzhw.eu/forschung/projekt?pr_id=467 (abgerufen am 30.03.2023).

zu unterscheiden, ob die Filtereingangsfrage falsch beantwortet oder zur falschen Folgefrage gesprungen wurde. Je nachdem, wie sich die Ursachen und die Möglichkeiten der Identifikation der Filterfehler darstellen, unterscheiden sich die Optionen im Umgang damit.

4.4 Umgang

Grundsätzlich ist die Entscheidung zu treffen, ob eine Überprüfung von Filterfehlern im Vorfeld der Datenbereitstellung erfolgen soll. Ist keine Überprüfung vorgesehen, obliegt die Prüfung der Daten auf etwaige Filterfehler den Datennutzer:innen.

Wird im Rahmen des Datenaufbereitungsprozesses eine Überprüfung von Filterfehlern vorgenommen, sind grundsätzlich zwei Vorgehensweisen möglich (vgl. für einen Überblick Kapitel 3 „Der Plausibilisierungsprozess“ sowie Abbildung 6).

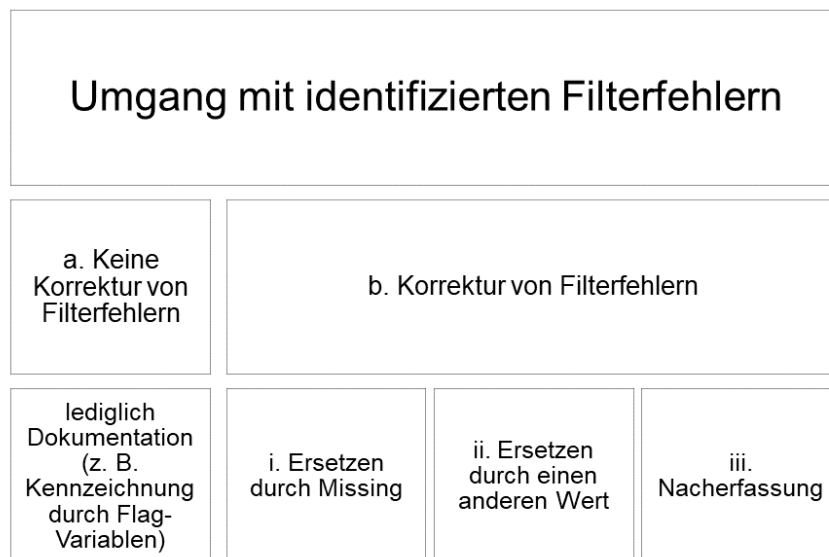


Abbildung 6: Umgang mit identifizierten Filterfehlern (eigene Darstellung).

- a. Identifizierte Filterfehler werden nicht korrigiert, sondern lediglich dokumentiert. Fehlende oder widersprüchliche Angaben infolge von Filterfehlern werden beispielsweise mittels zusätzlich generierter Variablen, sogenannter Flag-Variablen²², für die Nachverfolgung im Datensatz gekennzeichnet, um ggf. später von den Datennutzer:innen korrigiert zu werden (z. B. durch Imputation). Dies kann folgendermaßen aussehen: Ein Skript erzeugt eine Flag-Variable und listet alle Beobachtungen auf, die nicht fehlende Werte für Folgefragen bzw. -variablen aufweisen, obwohl der Wert der Filtervariablen angibt, dass diese Folgefrage

²²Für eine Erläuterung des Begriffs siehe Informationskasten 2 in Kapitel 3.3.2.

nicht hätte gestellt werden dürfen (und umgekehrt). Die Datennutzer:innen können nun selbst entscheiden, ob sie die von Befragten fälschlicherweise gemachten Angaben für ihre Analysen verwenden oder löschen bzw. fehlende Angaben ersetzen. Flag-Variablen werden für alle von Filterfehlern betroffenen Variablen angelegt und zeigen das Vorliegen eines potenziellen Filterfehlers an (z. B. 0/1 codiert). Sie vergrößern jedoch auch die Komplexität des bereitgestellten Datensatzes. Dieser Problematik kann begegnet werden, indem die Flag-Variablen nicht in den Hauptdatensatz integriert, sondern als zusätzlicher Datensatz bereitgestellt werden.

b. Identifizierte Filterfehler werden korrigiert; das Vorgehen für die Datennutzer:innen dokumentiert. Hier können verschiedene Methoden zur Anwendung kommen.

- (i) Ersetzen der fehlerhaften Werte durch einen explizit definierten Missing-Code (fehlender Wert): Fehlende Angaben auf fälschlicherweise nicht beantwortete Folgefragen (*errors of omission*) werden durch einen speziellen Missing-Code gekennzeichnet. Unnötige oder widersprüchliche Angaben auf fälschlicherweise beantwortete Folgefragen (*errors of commission*) werden gelöscht und durch einen Missing-Code (i. d. R. der in der Studie gültige Code für Überfilterung/„trifft nicht zu“) ersetzt.

Dies ist beispielsweise das von der TwinLife-Studie, und auch teilweise von der DZHW-Absolvent:innenstudie verfolgte Vorgehen. Bei letzterer werden die nachträglich durch die korrigierte Antwort in der Filtereingangsfrage gefilterten, aber aus oben beschriebenen Gründen (siehe Abschnitt 4.2 in diesem Kapitel) noch vorhandenen Werte in den Folgefragen gelöscht. Wenn sich die Ursache oder die korrekten Werte nicht ermitteln lassen, werden sowohl die Angaben der Filtereingangsfrage als auch die Angaben der Folgefragen in der Regel auf Missing gesetzt. Als weiteres Beispiel kann die mixed-mode Befragung des European Values Study (EVS)²³ 2017 angeführt werden. Filterfehler, die in Papierfragebögen entstehen, werden auf Missing gesetzt, wenn diese Fragen im Online-Fragebogen den Befragten nicht hätten gestellt werden dürfen.²⁴

Das Ersetzen von fehlerhaften Angaben durch Missing-Codes basiert grundsätzlich auf der Annahme, dass die Filtereingangsfrage korrekt beantwortet wurde und die Angaben auf die Folgefragen fehlerhaft sind. Insbesondere bei nicht-computergestützten Datenerhebungen können vermeintlich unnötige oder widersprüchliche Angaben jedoch auch darauf zurückzuführen sein, dass die Filtereingangsfrage falsch, die Folgefragen aber richtig beantwortet wurden. In diesem Fall führt das beschriebene Vorgehen dazu, dass gültige Angaben der Befragten gelöscht werden, die Falschangabe in der Filtereingangsfrage jedoch erhalten bleiben. Eine Möglichkeit, mit dieser Problematik umzugehen, ist, einen zusätzlichen Datensatz mit unbe-

²³<https://europeanvaluesstudy.eu/>.

²⁴EVS, GESIS. 2020. „EVS 2017 – Variable Report – Integrated Datasets (ZA7500, ZA7502).“ GESIS-Variable Reports 2020/16. S. 36. GESIS Data Archive, Cologne.

reinigten Daten zu veröffentlichen. Dies ermöglicht den Datennutzer:innen, auf die unbereinigten Daten zurückzugreifen und eigene Anpassungen oder Korrekturen vorzunehmen, insbesondere wenn sehr viele Werte im Zuge der Datenbereinigung gelöscht wurden.

- (ii) Ersetzen durch einen anderen Wert:
 - Fehlende oder widersprüchliche Angaben aufgrund von Filterfehlern werden basierend auf vorab festgelegten Regeln, Expert:innenwissen, Angaben aus anderen Variablen etc. korrigiert. Da es sich häufig um Einzelfallentscheidungen handelt, kann diese Methode sehr aufwendig und ressourcenintensiv sein. Daher ist dieses Vorgehen vor allem dann denkbar, wenn die Anzahl der betroffenen Fälle eher gering ist und die Variable eine Information erfasst, die gut durch Expert:innenwissen oder Angaben aus anderen Variablen (z. B. Längsschnittdaten) ermittelbar ist.²⁵ Vereinzelt wurde so im Rahmen der DZHW-Absolvent:innenstudie verfahren, beispielsweise wenn Filtervariablen aufgrund ihrer zentralen Rolle im Panel mehrfach erhoben wurden.
 - Fehlende Angaben werden auf Basis statistischer Schätzverfahren (z. B. Regressionsmodelle) imputiert.
- (iii) Nacherfassung: Die Befragten können erneut kontaktiert werden, um fehlende oder widersprüchliche Angaben zu klären. Sind allerdings viele Fälle von einem Filterfehler betroffen, ist eine Nacherfassung mit einem hohen zusätzlichen Aufwand und Mehrkosten verbunden. Dieser Punkt relativiert sich wiederum, wenn im Rahmen von Längsschnittbefragungen die Befragten ohnehin erneut kontaktiert werden.

Je nach Art und Ausmaß der Filterfehler können auch mehrere unterschiedliche Verfahren innerhalb einer Studie zum Einsatz kommen. Wann welches Verfahren im Umgang mit einem Filterfehler am sinnvollsten ist, ist abhängig von der Relevanz der betroffenen Variable und dem Ausmaß des Fehlers (z. B. wie viele Fälle im Datensatz von einem Filterfehler betroffen sind). In jedem Falle sollten entdeckte Filterfehler und der Umgang damit ausreichend dokumentiert werden. Unabhängig von der gewählten Vorgehensweise ist die Bereitstellung der Aufbereitungssyntax wünschenswert.

4.5 Dokumentation

Vorab ist festzuhalten, dass der prinzipielle Umgang mit Filterfehlern in einer Studie an einer übergeordneten Stelle (z. B. auf der Dokumentationswebseite oder im Codebook der Studie) dokumentiert werden sollte – vor allem dann, wenn bestimmte Filterfehler für die Datennutzer:innen selbst nicht als solche erkennbar sind. Dies ist beispielsweise der Fall, wenn Filterbedingungen und/oder Filterführungen in der Fragebogenvor-

²⁵Beim Einsatz von Längsschnittdaten ist selbstverständlich zu berücksichtigen, dass die verwendeten Variablen zeitkonstant sein oder sich über die Zeit gleichmäßig entwickeln müssen (z. B. Alter o. ä.).

lage und somit auch in der Fragebogendokumentation korrekt waren, jedoch bei der Fragebogenprogrammierung falsch umgesetzt wurden.

Eine Dokumentation der Filterfehler kann auf verschiedene Arten erfolgen:

- a. Man beschränkt sich auf eine allgemeine Beschreibung zum Umgang mit Filterfehlern im Codebook oder auf einer Dokumentationswebseite, ohne dass auf die Filterfehler selbst im Einzelnen eingegangen wird (z. B. wenn alle *errors of omission* durch einen speziellen Missing-Code gekennzeichnet sind und alle *errors of commission* gelöscht und durch ein „trifft nicht zu“-Missing ersetzt wurden). Diese Art der Dokumentation kann auch dann erfolgen, wenn keinerlei Filterfehler ermittelt oder behandelt werden. Auf diese Weise wird der Umgang mit Filterfehlern beispielsweise in der TwinLife-Studie dokumentiert.²⁶
- b. Falls Filterfehler ermittelt, aber keinerlei Korrekturen an den Daten selbst vorgenommen werden, bieten sich Flag-Variablen zur Dokumentation an. Flag-Variablen geben an, welche Fälle im Datensatz von Filterfehlern betroffen sind, sodass die Datennutzer:innen diese Fälle identifizieren und selbst entscheiden können, wie sie mit diesen umgehen.
- c. Die Dokumentation von Filterfehlern und deren Korrektur kann auf Variablenebene erfolgen, etwa als Hinweis in der Dokumentation zu einzelnen Variablen in den Variablen-Hinweisen (vgl. Befehl `-note-` in Stata) oder im Codebook. So werden beispielsweise planmäßige Filterabweichungen der nationalen EVS-Fragebögen vom internationalen Master-Fragebogen dokumentiert (z.B. zusätzliche Filterbedingungen oder nicht umgesetzte Filterinstruktionen).²⁷

4.6 Vermeidung

Idealerweise werden bereits in der Studienplanungsphase sogenannte systematische Filterfehler identifiziert und behoben, die auf eine fehlerhafte Fragebogenvorlage oder im Falle einer computergestützten Erhebung auf eine falsche Programmierung zurückzuführen sind. Dazu sollten bereits bei der Fragebogenkonstruktion alle Filterbedingungen und -führungen möglichst umfassend von mehreren Personen überprüft werden, um eine korrekte Fragebogenvorlage sicherzustellen. Dafür hat sich das Anfertigen von Filterführungsdiagrammen bewährt, die auch später für die Dokumentation bereitgestellt werden sollten.²⁸ Falls mit einem Erhebungsinstitut zusammengearbei-

²⁶Siehe <https://www.twin-life.de/documentation/check-routines/data-adjustment> (abgerufen am 30.03.2023).

²⁷EVS, GESIS. 2020. „EVS 2017 - Variable Report – Integrated Datasets (ZA7500, ZA7502).“ GESIS-Variable Reports 2020/16. S. 36. GESIS Data Archive, Cologne.

²⁸Siehe zum Beispiel das im Rechercheportal des FDZ-DZHW verfügbare Filterführungsdiagramm zum Fragebogen des DZHW-Studienberechtigtenpanels 2008 – dritte Welle: https://metadata.fdz.dzhw.eu/public/files/instruments/ins-gsl2008-ins3§-1.0.0/attachments/gsl2008_W3_QuestionFlow_de.pdf (abgerufen am 30.03.2023).

tet wird, ist es sinnvoll, dabei auch auf dessen Unterstützung und Erfahrung zurückzugreifen.

Bei einer computergestützten Datenerfassung helfen eine sorgfältige Programmierung und umfassende Testung der Filterinstruktionen, Fehler in der Programmierung bereits in der Testphase zu identifizieren und zu korrigieren (z. B. in einem Pretest mit realen Befragten oder durch die automatisierte Generierung eines Testdatensatzes).

Unsystematische Filterfehler infolge der (un)beabsichtigten Missachtung der Filterinstruktionen durch die Befragten oder Interviewer:innen können bei nicht-computergestützten Papierfragebögen durch eine Optimierung der grafischen Darstellung der Filterinstruktionen vermieden werden.²⁹ Filterinstruktionen sollten als solche klar erkennbar (z. B. mit einem Filterpfeil markiert) und vom restlichen Fragetext eindeutig zu unterscheiden sein (z. B. durch eine Filterspalte, Fettdruck oder eine farbliche Hervorhebung).

Werden Papierfragebögen von Interviewer:innen administriert, sollten die Interviewer:innen im Umgang mit Filterfragen und den damit verbundenen Problemen umfassend geschult werden. Werden Papierfragebögen von den Befragten im selbst-administrierten Modus ausgefüllt, sollte die Funktionsweise von Filterinstruktionen zu Beginn des Fragebogens den Befragten anschaulich erklärt werden. Sind komplizierte Filterinstruktionen für die Forschungsfrage unumgänglich, sollte auf einen anderen, idealerweise computergestützten Erhebungsmodus ausgewichen werden.

Zusammenfassend kann gesagt werden, dass das Auftreten von Filterfehlern durch eine sorgfältige Fragenbogenkonstruktion und eine ausführliche Testung des Instruments minimiert werden kann. Kommt es dennoch zu Filterfehlern, sind die Vor- und Nachteile der verschiedenen Verfahren des Umgangs mit Filterfehlern sorgfältig gegeneinander abzuwägen sowie das Auftreten und die Behandlung der Filterfehler in jedem Fall nachvollziehbar und umfassend zu dokumentieren.

²⁹Redline, Cleo D. und Don A. Dillman, „The Influence of Alternative Visual Designs“.

5 Wildcodes

Pablo Christmann, Andreas Franken

5.1 Beschreibung

Als „Wildcodes“ bezeichnet man Variablenwerte, die sich außerhalb des zulässigen Wertebereiches von Antwortmöglichkeiten zu einer Frage befinden. Im Zuge der Codierung von Umfragedaten werden die in einer Befragung gegebenen Antworten eindeutigen Kategorien mit zugehörigen numerischen Werten zugewiesen. Bei geschlossenen Fragen erfolgt diese Zuweisung zu den vordefinierten Antwortmöglichkeiten in der Regel bereits während der Fragebogenerstellung, wohingegen die Codierung der Antworten auf offene Fragen im Zuge der Datenaufbereitung erfolgt. Werden Befragte beispielsweise gefragt, ob sie Linkshänder:innen (1) oder Rechtshänder:innen (2) sind, sind als Antwortmöglichkeiten zwei gültige Werte und gegebenenfalls noch ein weiterer Missingcode für „keine Angabe“ (z. B. -2) vorgesehen. Alle anderen Werte wie zum Beispiel „.“, „.“, „.“, „9“, „-5“ oder „K“ würden hier als „wilde“ Werte bezeichnet werden, da sie außerhalb des vordefinierten und gültigen Wertebereiches liegen.

Wildcodes lassen sich exemplarisch mit Hilfe des integrierten Datensatzes der European Values Study (EVS) 2017³⁰ veranschaulichen. Die EVS ist ein länderübergreifendes Umfrageprogramm, welches die Einstellungen und Werte der Bevölkerung zu Themen wie Umwelt, nationale Identität, Familie, Politik, Gesellschaft, Religion, Moral und Arbeit erfasst. In der fünften Querschnittserhebung in 2017/18 beteiligten sich insgesamt 37 Länder. In Deutschland wurde die EVS in drei verschiedenen Umfragemodi erhoben: 1. im computerunterstützten persönlichen Interview-Modus („Computer Assisted Personal Interviewing“ CAPI), 2. im Mixed-Modus (Web, postalisch) und 3. im Mixed-Modus (Web, postalisch) mit zusätzlichem Matrix-Design („Split Questionnaire Design“), d. h. jede:r Befragte hat nur einen Teil des Fragebogens beantwortet. Aufgrund der unterschiedlichen Erhebungsmodi und der Einbettung in ein international harmonisiertes Studienprogramm mit genauen Zielvorgaben wird die deutsche EVS Erhebung in diesem Kapitel wiederholt zu Illustrationszwecken herangezogen.

Abbildung 7 zeigt nun eine Fragebatterie zu Geschlechterrollen (Q25), wie sie im CAPI-Modus und im Mixed-Modus des EVS im Fragebogen abgefragt wurde. Die ordinale Antwortskala hat vier Werte (1 „Stimme voll und ganz zu“, 2 „Stimme zu“, 3 „Stimme nicht zu“ und 4 „Stimme überhaupt nicht zu“).

³⁰EVS. 2019. „European Values Study 2017: Integrated Dataset (EVS 2017). ZA7500 Data file Version 2.0.0“. GESIS Data Archive. <https://doi.org/10.4232/1.13314>.

26. Bitte geben Sie zu jeder der folgenden Aussagen zur Rolle von Mann und Frau an, wie sehr Sie dieser zustimmen oder nicht zustimmen.

	Stimme voll und ganz zu	Stimme zu	Stimme nicht zu	Stimme überhaupt nicht zu	Weiß nicht
Wenn eine Mutter einer bezahlten Arbeit nachgeht, dann leiden ihre Kinder darunter.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Ein Beruf ist gut, aber was die meisten Frauen wirklich wollen, ist ein Heim und Kinder.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Alles in allem leidet das Familienleben darunter, wenn die Frau in Vollzeit berufstätig ist.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Es ist die Aufgabe des Mannes, Geld zu verdienen, die Frau ist für Haushalt und Familie zuständig.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
In politischen Führungspositionen sind Männer allgemein besser als Frauen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Eine Hochschulausbildung ist für Jungen wichtiger als für Mädchen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Insgesamt sind Männer bessere Führungskräfte in der Wirtschaft als Frauen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Zu meinen Lebenszielen gehört es, meine Eltern mit Stolz zu erfüllen.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Abbildung 7: Frage zu Geschlechterrollen ohne Mittelkategorie im Mixed-Modus-Fragebogen (Web/postalisch) der European Values Study 2017 in Deutschland

Durch eine abweichende Programmiervorlage für die EVS-Umfrage im Mixed-Modus Matrix-Design in Deutschland wurde jedoch eine zusätzliche Mittelkategorie („Weder noch“) in die Antwortskala eingefügt, sodass die Antwortskala von 1 „Stimme voll und ganz zu“ bis 5 „Stimme überhaupt nicht zu“ reicht. Diese Abweichung wurde erst bei der Tabellierung der einzelnen Variablen nach der Datenerhebung identifiziert, da nun ein zusätzlicher Wert 5 in den Daten existiert (vgl. Tabelle 2 für die Variable „child suffers with working mother“).³¹

³¹Quelle: Eigene Berechnung (EVS. 2020. „European Values Study 2017: Integrated Dataset (EVS 2017) - Matrix Design Data. ZA7502 Data file Version 1.0.0“. GESIS Data Archive. <https://doi.org/10.4232/1.13092>)

Tabelle 2: Häufigkeitstabelle einer Frage zu Geschlechterrollen mit zusätzlicher Mittelkategorie in der Mixed-Mode Matrix Umfrage (Web/postalisch) der European Values Study 2017 in Deutschland (Variable v72_DE)

Wert	Wertelabel	absolute Häufigkeit	relative Häufigkeit
-10	multiple answers Mail	1	0,03
-9	no follow-up	1623	49,57
-2	no answer	33	1,01
-1	don't know	32	0,98
1	agree strongly	82	2,50
2	agree	333	10,17
3	neither agree nor disagree (additional category)	305	9,32
4	disagree	577	17,62
5	disagree strongly	288	8,80
Gesamt		3274	100

Die Variablenausprägung der Fragebatterie im Mixed-Mode-Matrix Design entspricht so aber nicht mehr den Vorgaben des internationalen Umfrageprogramms. Zwar könnten die Antworten „Stimme nicht zu“ und „Stimme überhaupt nicht zu“ inhaltlich richtig auf die Werte 3 und 4 recodiert werden, die Mittelkategorie („Weder noch“) befindet sich jedoch außerhalb des vorab definierten Wertebereichs für diese Frage und wird daher als Wildcode behandelt.

Wildcodes sind nicht zu verwechseln mit extremen Werten oder Ausreißern. Obwohl Ausreißer unwahrscheinlich oder unplausibel erscheinen, können sie sich im Gegensatz zu Wildcodes noch in einem möglichen Wertebereich befinden (z. B. zwölf leibliche Kinder). Die Unterscheidung zwischen Wildcodes und Ausreißern ist nicht immer einfach, da in manchen Fällen der gültige Wertebereich nicht klar definiert werden kann. So scheint es zum Beispiel aus heutiger Sicht ausgeschlossen, dass Menschen 140 Jahre alt werden können. Ein maximales Alter von 125 Jahren zu erreichen, scheint aber theoretisch möglich, wenn auch sehr unwahrscheinlich.³² Im Gegensatz hierzu kann die Angabe, wie viele Stunden jemand am Tag mit einer bestimmten Tätigkeit verbringt, die maximale Anzahl von 24 Stunden nicht übersteigen. Hiermit hat die Verteilung eine „natürliche“ Begrenzung. Eine Angabe von 35 Stunden wäre ein Wert, welcher außerhalb der möglichen Verteilung liegt und somit ein Wildcode.

³²Beer, Joop de, Anastasios Bardoutsos und Fanny Janssen. 2017. „Maximum Human Lifespan May Increase to 125 Years“. *Nature* 546 (7660): E16–17. <https://doi.org/10.1038/nature22792>.

5.2 Ursachen

Die Ursachen von Wildcodes sind vielfältig und können sowohl während der Studienplanungsphase, der Datenerhebungsphase, als auch während der Datenaufbereitungsphase entstehen. Auslöser sind unter anderem sich unterscheidende Programmiervorlagen (siehe EVS-Beispiel), Fehler bei der Dateneingabe (z. B. durch das falsche Übertragen von Codes aus standardisierten Listen wie Studienfachlisten oder Berufslisten in Papierfragebögen), oder auch durch fehlerhafte Skripte zum Codieren der Daten. Zudem können unvollständige Auswahlmöglichkeiten bei geschlossenen Fragen, wie beispielsweise eine vergessene „Sonstige“-Kategorie, zu Wildcodes führen. In Papierfragebögen verleitet eine nicht vorhandene Antwortmöglichkeit dazu, sie händisch neben den Antwortkategorien anzumerken.

Im Folgenden werden die unterschiedlichen Ursachen für Wildcodes anhand des Auftretens in unterschiedlichen Phasen des Forschungszyklus erläutert.

5.2.1 In der Studienplanungsphase

- a. Fehler in der Fragebogen- und/oder Programmiervorlage: voneinander abweichende Fragebogen- oder Programmiervorlagen als Ursache für Wildcodes können in jedem Befragungsmodus auftreten und sind häufig ein Resultat von Unachtsamkeit oder mangelnder Abstimmung im Entwicklungsprozess. Programmiervorlagen sind i. d. R. Textdokumente, die auf Fragebogenvorlagen beruhen, aber noch mit weiteren wesentlichen Informationen für die Umsetzung angereichert sind, wie z. B. numerische Codes für Antwortkategorien oder Missingwerte, Filterbedingen zur Steuerung der Umfrage sowie mode-spezifische Hinweise, beispielsweise für die Darstellung von Fragen. Gründe für Wildcodes in Fragebogen und Programmiervorlagen können z. B. Zeitdruck, nicht vorhandene Qualitätskontrollen, Probleme im Workflow bei der Zusammenarbeit von mehreren Personen oder Organisationen, aber auch Änderungen in letzter Minute sein.
- b. Fehler in der Fragebogenprogrammierung: Eine mangelhafte Umsetzung der Programmiervorlage in einer Befragungssoftware (ggf. durch ein Feldinstitut) ist ebenfalls eine häufige Ursache für Wildcodes. Die Umsetzung ist insbesondere dann problematisch, wenn für die Befragung andere Werte und Wertelabels verwendet werden (müssen) als für den finalen Datensatz.

5.2.2 In der Datenerhebungsphase

- a. Hinzufügen weiterer Antwortmöglichkeiten bei geschlossenen Fragen: Insbesondere bei einer nicht-computergestützten Datenerfassung können Interviewer:innen oder Befragte zusätzliche Antwortmöglichkeiten im Papierfragebogen notieren, die im Fragebogen nicht standardmäßig vorgesehen sind. Das passiert

vornehmlich dann, wenn Interviewer:innen nicht hinreichend geschult werden und/oder die im Fragebogen vorgesehenen Antwortmöglichkeiten nicht erschöpfend sind. Beispielweise könnte das Auslassen einer Partei in den Antwortmöglichkeiten bei der „Sonntagsfrage“ vermehrt dazu führen, dass Anmerkungen in nicht-computergestützten Instrumenten gemacht werden. Auch Listenangaben von Befragten sind fehleranfällig, da hier leicht mögliche Angaben übersehen werden können (z. B. können bei Herkunftslandabfragen mögliche Länder vergessen werden).

- b. Antworten auf offene Fragen: Offene Fragen in Form eines einfachen Textfeldes werden zur Abfrage numerischer Informationen, wie beispielsweise Häufigkeiten oder Zeitangaben, als auch für freie Textangaben, wie beispielsweise Sprach- oder Berufsangaben verwendet. Anstatt aus einer vorgegebenen Liste von Antwortmöglichkeiten auszuwählen, erfolgt die Dateneingabe bei offenen numerischen Fragen „händisch“ (d. h. handschriftliche Angaben in Papierfragebögen sowie Tastatur- oder Touchscreen-Eingaben in computergestützten Erhebungen). Ein Beispiel für eine offene narrative Frage wäre „Wie viele Stunden pro Woche nutzen Sie das Internet durchschnittlich, entweder für berufliche oder private Zwecke?“. Neben Bereichsangaben (z. B. „4-6“, „10 bis 15“) und Schätzwerten (z. B. „ca. 5“, „ungefähr 20“), welche in erster Linie eine Herausforderung für die Vercodung offener Angaben darstellen, zählen auch Wildcode-Angaben (z. B. „350“, „990“) zu den unerwünschten Antworten bei offenen narrativen Fragen. So können im genannten Beispiel alle Werte als ungültig ausgeschlossen werden, die die maximale Stundenanzahl pro Woche übersteigen. Wildcodes bei offenen numerischen Fragen können entweder unbeabsichtigt (z. B. durch versehentliches Vertippen auf der Tastatur oder dem Touchscreen) oder beabsichtigt (z. B. durch Falschangaben aufgrund mangelnder Motivation der oder des Befragten) entstehen.

5.2.3 In der Datenaufbereitungsphase

- a. Abweichungen bei der Datenübertragung: Die Angaben eines Papierfragebogens werden entweder händisch abgetippt oder mittels Software automatisiert eingelesen. Wildcodes können bei der nicht-computergestützten Datenerfassung infolge einer falschen Datenübertragung oder durch Schreibfehler in ein elektronisches Datenformat entstehen. Der richtige Wert ist im Nachhinein häufig nicht mehr zu ermitteln.³³
- b. Fehler bei der Datenrecodierung: Bei der Aufbereitung der Rohdaten kann es geschehen, dass Werte und Labels inkorrekt umcodiert werden. Häufig werden Variablen und Werte umcodiert, damit sie z. B. den Vorgaben von international harmonisierten (Umfrage)-Programmen entsprechen oder damit sie konsistent

³³Rutsch, Martin. 1986. „Datenerfassung und Datenmanagement“. In *Statistik 1. Mit Daten umgehen*. Hrsg. Martin Rutsch. Basel: Birkhäuser. 125–66.

mit Daten aus vergangenen (Wiederholungs-)Befragungen sind. Dieser Aufbereitungsschritt kommt in der Praxis häufig vor, wenn zum Beispiel neue Variablen aus bereits bestehenden Variablen generiert werden, oder Daten aus verschiedenen Umfragemodi in einem einzigen Datensatz harmonisiert werden. So weichen z. B. Datensätze aus Webumfragen von Datensätzen aus Papierumfragen i. d. R. voneinander ab, was beispielsweise den Umgang mit fehlenden Werten oder Mehrfachnennungen angeht. Zudem werden für die Durchführung von persönlich-mündlichen Interviews, aber auch für Telefon- oder Webinterviews oder auch zur automatisierten Erfassung von Papierfragebögen verschiedene Softwareprogramme eingesetzt, die oft eigene Konventionen in Bezug auf das hinterlegte Werte- und Variablenschema haben. So entsprechen die erhobenen Rohdaten meistens nicht dem finalen Werte- und Variablenschema und müssen entsprechend recodiert werden. Dieser Prozess kann problematisch werden, da oft sehr viele einzelne Werte verschiedener Variablen umcodiert werden müssen und sich hier z. B. Tippfehler einschleichen können. Erschwerend kommen unter Umständen unterschiedliche länderspezifische Konventionen für den Umgang mit Fehlerwerten (z. B. der Einsatz von System-Missings wie „.“, negative Missings wie „-1“ oder positive Missings wie „8“ oder „9“) hinzu.

- c. Vergabe von zusätzlichen Codes: Ein Beispiel für die Übertragung von Informationen in eine international vergleichbare Variable ist die Codierung der Antworten auf offene Fragen in ein internationales Codierungsschema, wie z. B. die ILO-ISCO-Variable („International Standard Classification of Occupations“³⁴) für Berufsklassifikationen. Nicht für jede länderspezifische Berufsbezeichnung sind Werte im Schema vorgesehen, was Verocoder:innen dazu verleitet, mehr oder weniger zur schematischen Struktur passende, nicht vorgesehene Werte zu vergeben. Diese Werte werden mit Labels versehen, welche wiederum in den Ländern freihändig ins Englische übersetzt werden. An der Stelle, an der die Daten zusammengeführt werden (z. B. in einem Datenarchiv) werden Datenaufbereiter:innen unter Umständen mit identisch gelabelten, unterschiedlichen Werten konfrontiert, deren genaue Bedeutung nur mit der entsprechenden länderspezifischen Expertise verstanden werden kann. Ähnliche Fälle kann es zum Beispiel bei der Erstellung von internationalen Bildungsvariablen wie ISCED („International Standard Classification of Education“)³⁵ geben, bei der Abfrage des Geschlechts mit einer „diversen“ Kategorie oder aber auch beim Familienstand, der in verschiedenen Ländern unterschiedlich erfasst wird. Bei ISO-Länderlisten kann es zum Beispiel ebenfalls vorkommen, dass Befragte Geburtsländer nennen, die so heute nicht mehr bestehen, beispielsweise Jugoslawien oder die Tschechoslowakei.

³⁴International Labour Office. 2012. *International Standard Classification of Occupations 2008 (ISCO-08): Structure, group definitions and correspondence tables*. https://www.ilo.org/wcmsp5/groups/public/@dgreports/@dcomm/@publ/documents/publication/wcms_172572.pdf.

³⁵OECD, Eurostat und UNESCO Institute for Statistics. 2015. *ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*. OECD Publishing. <https://doi.org/10.1787/9789264228368-en>.

5.3 Identifikation

Allgemein ist bei der Identifikation von Wildcodes anzumerken, dass eine möglichst frühe Identifikation im Forschungsprozess den Umgang mit fehlerhaften Werten später meist vereinfacht, da dann unter Umständen noch die Möglichkeit besteht, direkt während der Befragung einzugreifen. Zusätzliche Informationen, die früh nacherfasst werden, können später hilfreich für die Bereinigung der Wildcodes sein. Die am Anfang beschriebene Differenzierung zwischen Wildcodes, extremen Werten und Ausreißern gilt es auch bei der Identifikation zu beachten.

5.3.1 In der Studienplanungsphase

- a. Manueller Abgleich (*eyeballing*, idealerweise nach dem Vier-Augen-Prinzip) des Erhebungsinstrumentes (Papierfragebogen, CAPI- oder Web-Programmierung) mit der Fragebogen-/Programmiervorlage, mit besonderem Augenmerk auf die Antwortkategorien.
- b. Durchführung eines Pretests: Im Zuge einer Testbefragung können Probleme mit einzelnen Fragen und Antwortkategorien frühzeitig erkannt werden. Der dadurch gewonnene Testdatensatz kann zudem für eine systematische Identifizierung von Wildcodes genutzt werden (vgl. Abschnitt 5.3.2 „In der Datenaufbereitungsphase“).
- c. Softwareseitige Generierung eines Testdatensatzes: In den meisten Fällen bietet Umfragesoftware die Möglichkeit sich einen Testdatensatz generieren zu lassen, der für die systematische Auswertung genutzt werden kann (vgl. Abschnitt 5.3.2 „In der Datenaufbereitungsphase“). Zudem können mit diesen Testdatensätzen bereits Skripte zur Datenaufbereitung (insb. Vergabe der Wertelabel) geschrieben werden. Dieser Bearbeitungsschritt, der später meist sowieso notwendig wird, beinhaltet üblicherweise einen intensiveren Blick auf die Werte.

5.3.2 In der Datenaufbereitungsphase

- a. Tabellierung aller Verteilungen (inklusive Missings) und manueller Abgleich (*eyeballing*) mit dem Codebuch: Es empfiehlt sich, sich sowohl die numerischen Werte als auch die Wertelabels anzeigen zu lassen.³⁶ So ist der EVS-Fehler erst bei dem systematischen Abgleich aller Häufigkeitsverteilungen mit dem Codebook aufgefallen. In manchen Fällen eignet sich auch ein Test auf extreme Werte, um Wildcodes aufzuspüren.

³⁶Jensen, Uwe. 2012. *Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten*. GESIS Technical Reports 2012 (07). Köln: GESIS - Leibniz-Institut für Sozialwissenschaften. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-320650>.

- b. Automatisierter Abgleich von vorab definierten Wertebereichen für verschiedene numerische Variablentypen und Auflistung von Abweichungen³⁷: So lässt sich z. B. der zulässige Wertebereich bei Geburtsjahren aus dem Stichprobenrahmen mit den Antworten aus der Umfrage abgleichen. Oftmals lässt sich auch eine Aussage darüber treffen, ob eine Variable nur einen positiven oder auch einen negativen Wertebereich hat. So wäre z. B. bei einer Variable mit der Anzahl der Kinder oder dem Einkommen nur ein positiver Wertebereich zulässig.
- c. Für String-Variablen kann die tatsächliche Zeichenlänge mit einer vorab definierten Zeichenlänge automatisiert abgeglichen und Abweichungen aufgelistet werden.³⁸ So dürften beispielsweise Länderkürzel nach ISO 3166 (ALPHA-2) nur eine zweistellige Zeichenlänge aufweisen („DE“, „UK“ oder „FR“ usw.).
- d. Für alle Antwortmöglichkeiten Wertelabels vergeben: Es empfiehlt sich, alle Werte mit Labels zu versehen, da Wildcodes so besser aufgespürt werden können. Bei einer anschließenden Tabellierung der Variablen fallen Werte ohne Labels so einfacher ins Auge.³⁹

5.4 Umgang

Es ist die grundsätzliche Entscheidung zu treffen, ob eine Überprüfung von Wildcodes im Vorfeld der Datenbereitstellung erfolgen soll. Wird im Zuge des Datenaufbereitungsprozesses eine Überprüfung vorgenommen, gibt es mehrere Möglichkeiten des Umgangs mit Wildcodes⁴⁰, welche auch miteinander kombiniert werden können.

- a. Wildcodes berichtigen und die richtigen Werte zuweisen: Obwohl oftmals nicht zweifelsfrei klar ist, was der „richtige“ Wert ist, lassen sich durch Informationen aus anderen Fragen, Nachbefragungen der Personen⁴¹ oder aus vorherigen Befragungen in bestimmten Situationen Fehler identifizieren und berichtigen. Bei Papierfragebögen besteht zudem die Möglichkeit, die ausgefüllten Originalfragebogen zu prüfen.⁴² Idealerweise werden die entstandenen Wildcodes direkt

³⁷Schwarz, Hannah. 2018. „Data consistency“. In *Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research* GESIS Papers (22). Hrsg. Sebastian Netscher und Christina Eder. Köln: GESIS – Leibniz-Institut für Sozialwissenschaften. 25-33. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-59492-3>.

³⁸Schwarz, Hannah, „Data consistency“.

³⁹Für einen automatisierten Test in R siehe: <https://rdrr.io/rforge/memisc/man/wild.codes.html> (abgerufen am 30.03.2023).

⁴⁰Lück, Detlev und Uta Landrock. 2014. „Datenaufbereitung und Datenbereinigung in der quantitativen Sozialforschung“ In *Handbuch Methoden Der Empirischen Sozialforschung*, Hrsg. Nina Baur und Jörg Blasius. Wiesbaden: Springer VS. 457-471.

⁴¹Hillmert, Steffen. 2003. „Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutschland: die jüngste Teilstudie der Lebensverlaufsstudie als Scientific Use File.“ In *ZA-Information / Zentralarchiv für Empirische Sozialforschung* 53 (173) (2003). 167-177, <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-198919>.

⁴²Porst, Rolf. 2000. *Praxis Der Umfrageforschung*. Wiesbaden: VS Verlag für Sozialwissenschaften. <https://doi.org/10.1007/978-3-663-11135-1>.

nicht für die Aufbereitung eines Datensatzes, da die Antwortskalen nicht identisch sind und dies Implikationen für das Antwortverhalten der Befragten und die Analyse der Daten haben kann.

- f. Eine weitere Möglichkeit besteht darin, Wildcodes in den Daten beizubehalten. Dies ist eine Option, wenn die Antwortmöglichkeiten einer Frage nicht ausreichend bestimmt wurden, z. B. Antwortmöglichkeiten vergessen wurden und anschließend, wenn möglich, im Erhebungsprozess separat dokumentiert wurden. Dieses Vorgehen ist unter Umständen auch bei Mehrfachnennungen möglich, wenn eigentlich Einfachnennungen vorgesehen waren. So wurde zum Beispiel im Fragebogen des SOEP aus dem Jahr 2019 die Parteineigung erfragt – eigentlich angedacht als Einfachnennung (siehe Abbildung 9⁴⁶).

169. Welcher Partei neigen Sie zu?

SPD

CDU

CSU

FDP

Bündnis 90 / Die Grünen

Die Linke →

AfD

NPD / Republikaner / Die Rechte...

Andere

und
zwar:

Abbildung 9: Frage zur Parteineneigung (SOEP-Core – 2019: Personenfragebogen, Stichproben A-L3, M1-M2 + N-P)

Einige Befragte haben hier jedoch Parteienkombinationen angegeben; es ist also mehr als eine Partei angekreuzt worden. Üblicherweise würden man diese Fälle zu „unplausiblen Werten“ umcodieren. In diesem Fall wurden die Angaben jedoch als „Sondercodes“ aufgenommen, wie beispielsweise eine 11 für die Neigung zur „SPD und FDP“ (vgl. Tabelle 3). Da diese Werte bei Analysen klar erkenntlich als Wildcodes auftauchen, wird den Datennutzer:innen hier die Verantwortung überlassen, wie sie mit diesen Werten umgehen. Es besteht jedoch die Gefahr, dass sich uninformierte Datennutzer:innen nicht über die Besonderheit dieser Codierung im Klaren sind. In diesem Beispiel hatten lediglich Personen, welche den Papierfragebogen erhalten haben, die Möglichkeit, sich über die Anweisung einer Einfachnennung hinwegzusetzen. Daher wurde auch ausschließlich dieser sehr selektiven Personengruppe ein Wert außerhalb des gültigen Wertebereiches zugeordnet. In den Daten wurde dies wie in Tabelle 3⁴⁷ dargestellt abgelegt.

⁴⁶Kantar Public, „SOEP-Core - 2019: Personenfragebogen, Stichproben A-L3, M1-M2 + N-P“.

⁴⁷Quelle: Eigene Berechnung SOEPCore v36 (Datensatz bjp). Sozio-oekonomisches Panel (SOEP).

Tabelle 3: Häufigkeitstabelle Parteilichkeitszugehörigkeit für das Jahr 2019 (Variable bjp_169_01)

Wert	Wertelabel	absolute Häufigkeit	relative Häufigkeit
-5	In Fragebogenversion nicht enthalten	3857	12,90
-4	Unzulässige Mehrfachantwort	1	0
-2	trifft nicht zu	14210	47,51
-1	keine Angabe	425	1,42
1	SPD	2137	7,15
2	CDU	3157	10,56
3	CSU	650	2,17
4	FDP	674	2,25
5	Bündnis90/Grüne	2750	9,19
6	Die Linke	915	3,06
7	NPD/Republikaner/Die Rechte	29	0,1
8	Andere	265	0,89
9	SPD/Bündnis'90/Grüne	18	0,06
10	SPD/CDU	2	0,01
11	SPD/FDP	1	0
13	CDU/CSU	15	0,05
14	CDU/FDP	15	0,05
15	CDU/Bündnis'90/Grüne	13	0,04
16	Bündnis'90/Grüne/Die Linke	12	0,04
17	SPD/Die Linke	4	0,01
22	CSU/FDP	4	0,01
23	Grüne/FDP	2	0,01
25	FDP/NPD	1	0
27	AFD	743	2,48
30	CDU+/oder CSU/AfD	7	0,02
31	SPD/AFD	1	0
Gesamt		29908	100

Alternativ kann auch eine zusätzliche Flagvariable⁴⁸ den Hinweis liefern, dass sich der Wert außerhalb des gültigen Wertebereiches befindet.

2021. „Socio-Economic Panel, data from 1984-2019, (SOEP-Core, v36, EU-Edition)“. SOEP Socio-Economic Panel Study. <https://doi.org/10.5684/SOEP.CORE.V36EU>.

⁴⁸Flagvariablen sind binär kodierte Variablen, die anzeigen, ob eine Bedingung zutrifft, oder nicht.

5.5 Dokumentation

Prinzipiell gilt, dass alle Datenerhebungs- und Datenverarbeitungsprozesse möglichst standardisiert sein sollten. Dies vereinfacht die Nachvollziehbarkeit und ist insbesondere bei längeren Zeithorizonten sowie wechselnden Mitarbeiter:innen einer Studie hilfreich. Alle am Datensatz getätigten Änderungen sowie alle identifizierten, aber nicht angepassten Widersprüche sollten übersichtlich, begründet und verständlich dokumentiert werden.

Werden im Zuge des Datenaufbereitungsprozesses die Daten auf das Vorhandensein von Wildcodes überprüft, gilt es als Best Practice, in der Dokumentation auf das Vorhandensein oder der Veränderung von möglichen Wildcodes hinzuweisen und den Umgang mit Wildcodes zu erläutern. Auch ist ein einheitliches Vorgehen im Umgang mit Wildcodes für zukünftige Datennutzer:innen leichter zu verstehen und nachzuvollziehen.

Im Beispiel der EVS wurde im Variablenreport auf das Problem der fehlerhaften Antwortskala im Mixed-Modus Matrix-Design hingewiesen und erläutert, dass alle Angaben der Frage für Deutschland für den entsprechenden Modus auf Missing gesetzt wurden, um die Vergleichbarkeit der Variablen zwischen den teilnehmenden Ländern der EVS zu gewährleisten (vgl. Abbildung 10⁴⁹). Um die Daten jedoch für Forschungszwecke zu sichern, wurden nachträglich ebenfalls neue, länderspezifische Variablen in dem Datensatz angelegt. In diesen zusätzlichen Variablen wurden dann die ursprünglichen Antworten mit den Wildcodes („Weder noch“) erhalten (vgl. Abbildung 11⁵⁰).

⁴⁹Quelle: Brislinger, Evelyn und Ivet Solanes (Hrsg). 2019. „EVS 2017 - Variable Report Integrated Dataset: Documentation of the second pre-release 2019/7/15. Related to the Integrated Dataset ZA7500 Version 2.0.0“. *Variablen Report 2019/12*. Cologne: GESIS. <https://doi.org/10.4232/1.13314>.

⁵⁰Quelle: Brislinger und Solanes, „EVS 2017 - Variable Report.“

European Values Study 2017 - Second pre-release.

GESIS Study No. 7500 (v2.0.0), <http://dx.doi.org/doi:10.4232/1.13314>

Variable, Label

Question Text (English Language)

v72 - child suffers with working mother (Q25A)

Q25

<SHOWCARD 25>

<READ OUT AND CODE ONE ANSWER PER LINE>

For each of the following statements I read out, can you tell me how strongly you agree or disagree with each. Do you strongly agree, agree, disagree, or strongly disagree?

in CAWI,MAIL:

For each of the following statements, please indicate how strongly you agree or disagree with each.

How much do you agree or disagree with this statement?

Q25.A When a mother works for pay, the children suffer

- 10 multiple answers Mail
- 9 no follow-up
- 8 follow-up non respons
- 7 matrix not applied
- 6 na (survey break-off)
- 5 other missing
- 4 item not included
- 3 not applicable
- 2 no answer
- 1 dont know
- 1 agree strongly
- 2 agree
- 3 disagree
- 4 disagree strongly

Germany CAWI & Mail matrix-design: Due to an error in the field questionnaire used for the CAWI/Mail matrix-design survey a middle answer category ('neither agree nor disagree') was added in Q25. The question is not further comparable with the Master questionnaire. Concerned cases are coded -4 'item not included'. For MODIFIED VARIABLES, see v72_DE - v79_DE.

Abbildung 10: Dokumentation einer Frage zu Geschlechterrollen ohne Mittelkategorie (Items auf Missing gesetzt)

European Values Study 2017 - Second pre-release.

GESIS Study No. 7500 (v2.0.0), <http://dx.doi.org/doi:10.4232/1.13314>

Variable, Label

Question Text (English Language)

v72_DE - DE MODIFIED (additional middle cat): child suffers with working mother (Q25A)

Germany CAWI & Mail matrix-design survey: MODIFIED VARIABLE (additional middle category)

Q25_DE

<SHOWCARD 25>

<READ OUT AND CODE AN ANSWER PER LINE>

For each of the following statements I read out, can you tell me how strongly you agree or disagree with each. Do you strongly agree, agree, disagree, or strongly disagree?

in CAWI,MAIL:

For each of the following statements, please indicate how strongly you agree or disagree with each.

How much do you agree or disagree with this statement?

Q25_DE.A When a mother works for pay, the children suffer

-10 multiple answers (Mail)

-9 no follow-up

-8 follow-up non respons

-7 matrix not applied

-6 na (survey break-off)

-5 other missing

-4 not asked in survey

-3 not applicable

-2 no answer

-1 dont know

1 agree strongly

2 agree

3 neither agree nor disagree (additional category)

4 disagree

5 disagree strongly

Due to an error in the field questionnaire used for the CAWI/Mail matrix-design survey a middle answer category ('neither agree nor disagree') was added in Q25. The question is not further comparable with the Master questionnaire. Concerned cases are coded -4 'item not included' in the master variables v72 - v79.

Abbildung 11: Dokumentation einer Frage zu Geschlechterrollen mit zusätzlicher Mittelkategorie (Daten durch neue Variable gesichert)

5.6 Vermeidung von Wildcodes

Der beste Umgang mit Wildcodes ist die Vermeidung ihrer. Hierfür gibt es unterschiedliche Vorgehensweisen:

- a. Den Programmier:innen anstelle des Originalfragebogens Programmiervorlagen mit eindeutig definierten Variablennamen, Wertebereichen und Wertelabels bereitstellen: Bei einer computergestützten Datenerfassung lassen sich die Antwortmöglichkeiten auf dieser Grundlage so programmieren, dass keine Wildcodes entstehen können.
- b. Sorgfältiges Pretesting, um systematische Wildcodes schon vor Beginn der eigentlichen Erhebung zu identifizieren und zu beseitigen. Das Entwickeln von potenziellen Befragtenprofilen mit spezifischen Eigenschaften (z. B. mittelalt, hochgebildet, weiblich) kann ggf. bei der Fragebogentestung helfen, fehlende Antwortmöglichkeiten bereits vor der Erhebung auszumachen.
- c. Zulässige Wertebereiche in der Umfragesoftware definieren: Bei elektronischen Erhebungsverfahren können die validen Antwortmöglichkeiten von geschlossenen Fragen sowie in einem gewissen Umfang auch offenen Fragen so programmiert werden, dass keine Wildcodes entstehen bzw. eingegeben werden können. Interviewer:innen oder Befragte werden so bereits bei der Antworteingabe auf mögliche Fehler aufmerksam gemacht.
- d. Für fehlende Werte negative numerische Werte vorsehen und für alle anderen Werte immer positive Werte vergeben: In einigen Umfragen werden fehlende Werte immer noch im positiven Bereich vergeben – oft ein Relikt der Notwendigkeit zur Datensparsamkeit vergangener Zeiten. So werden z. B. fehlende Werte mit 8 „Weiß nicht“ oder 9 „Keine Angabe“ für Variablen mit einem Zeichen vergeben. Für Variablen mit zwei Zeichen werden fehlende Werte mit 88 und 99 vergeben, für Variablen mit drei Zeichen 888 und 999 und so weiter. Die Vielzahl der somit möglichen fehlende Werte erhöht nicht nur die Wahrscheinlichkeit für versehentlich falsch codierte fehlende Werte (Wildcodes), sie erschwert auch die Definition des zulässigen Wertebereichs für eine automatisierte Fehlerprüfung. Zudem erhöht sich durch die Vergabe positiver Werte für fehlende Werte die Gefahr, dass Datennutzer:innen fehlende Werte als substantielle Antwortmöglichkeiten auswerten – negative Werte fallen hier üblicherweise eher ins Auge.
- e. Bei computergestützten Befragungen können aufwendige – und für Wildcodes anfällige – Codierungen vereinfacht werden, indem direkt auf eine Datenbank mit möglichen Antwortmöglichkeiten zugegriffen wird. Ein Beispiel hierfür ist die Abfrage des Berufs und der dazugehörigen ISCO-Kategorie. Hierbei lassen sich Wildcodes (aber auch falsch codierte Antworten) reduzieren, indem nach der Eingabe eines Stichwortes alle passenden Antwortkategorien aufgelistet werden. Im Vergleich zu einer offenen Angabe, die möglicherweise uneindeutig ist, oder einer unhandlichen, langen Liste von Antwortkategorien, wird Interviewer:innen oder Befragten durch die Suchfunktion eine schnelle und komfortable Möglichkeit gegeben, die passende Kategorie zu finden, die direkt mit einem gül-

tigen Code verknüpft ist. Hierbei sollte immer eine Kategorie für mögliche sonstige Werte berücksichtigt werden.

Insgesamt können trotz sorgfältiger Vorbereitung in der Studienplanungsphase und durch Prüfroutinen bei der Datenbereinigung Wildcodes vermutlich nie gänzlich ausgeschlossen werden. Umso wichtiger ist es deshalb, diese Fehler zu identifizieren und transparent für die Nutzer:innen zu dokumentieren.

6 Inkonsistenzen bei der Verknüpfung mehrerer Datensätze/-quellen

Fabio Franzese, Kristina Krell, Senta-Melissa Pflüger

6.1 Beschreibung

Die Verknüpfung der Informationen aus mehreren Datensätzen oder -quellen kann das Analysepotenzial häufig immens erweitern. Unter Plausibilisierungsperspektive betrachten wir die Verknüpfung von

1. Befragungen der gleichen Personen zu verschiedenen Zeitpunkten (Panelbefragungen).
2. Befragungen der gleichen Personen in inhaltlich unterschiedlichen Befragungen oder Messinstrumenten (Haushalts- vs. Personenfragebögen, zeitgleich verlaufende Befragungen, etc.)
3. Befragungen von unterschiedlichen Personen zu verschiedenen oder zum gleichen Zeitpunkt (z. B. Proxyinterviews⁵¹, Erfassung von Selbst- und Fremdeinschätzung zur Persönlichkeit einer Befragungsperson, etc.).⁵²

Wird dieselbe Information zu einer Person in einer Querschnittsbefragung in zwei verschiedenen Messinstrumenten erhoben und werden diese Datenquellen miteinander verknüpft, können Ungereimtheiten bzw. Differenzen zwischen den Informationen zutage treten. Dies kann z. B. der Fall sein, wenn dieselbe Information in mehreren Fragebögen abgefragt oder von verschiedenen Personen gegeben wurde. Auch bei Panelbefragungen, also Befragungen, bei denen in verschiedenen Durchläufen der Studie immer wieder dieselben Befragungspersonen befragt werden, werden die Daten nach jeder Welle verknüpft. Gibt es von einer Person mehrere Interviews mit teilweise gleichen inhaltlichen (und objektiv messbaren) Variablen, kann es über die Zeit zu (unplausiblen) Inkonsistenzen kommen – insbesondere bei in der Regel über die Zeit konstanten Merkmalen, wie beispielsweise dem Geburtsland. Außerdem ist es möglich, dass Informationen zwar nicht zeitlich konstant sind, sich jedoch konsistent über die Zeit entwickeln müssen. Beispielsweise kann ein Schulabschluss in einer Folgebefragung nicht niedriger sein als in der vorherigen Befragung einer Person.

⁵¹Proxyinterviews sind Interviews, bei denen Dritte über die eigentlichen Befragungspersonen befragt werden.

⁵²Eine weitere Möglichkeit der Plausibilisierung bietet die Verknüpfung von Befragungsdaten mit administrativen – also externen – Daten, wenn sie auf Individualebene vorhanden und kompatibel sind. Administrative Daten können hilfreich bei der Identifikation von Inkonsistenzen (Validierung auf Individualebene oder im Aggregat), beim Umgang (z. B. für die Entwicklung eines Imputationsmodells oder das Training eines Machine Learning Algorithmus) und bei der Dokumentation (für Aussagen über Measurement Errors bestimmter Variablen im Aggregat) von Inkonsistenzen sein. Die in diesem Kapitel beteiligten und beschriebenen Studien haben leider keine Möglichkeit, compatible administrative Daten auf Individualebene an die Befragungsdaten anzuschließen. Aus diesem Grund wird diese Möglichkeit der Plausibilisierung hier nicht weiter vertieft.

Dieses Kapitel behandelt diese Art von Inkonsistenzen und die Möglichkeiten, damit umzugehen. Dabei wird das Augenmerk auf Inkonsistenzen in den (objektiv messbaren) Informationen zu einer bestimmten Untersuchungseinheit (z. B. Person oder Haushalt) gelegt. Unterschiede in der Abfrage bestimmter Informationen in international vergleichenden Studien, die wiederum zu Inkonsistenzen und verminderter Vergleichbarkeit führen, sollen hier nicht im Fokus stehen.

Generell sind Inkonsistenzen dieser Art in Befragungsdaten nicht erwünscht und können Datenbereitstellende vor operationale Probleme stellen, etwa wenn Informationen aus früheren Befragungswellen zur Vorbereitung des Preloads für zukünftige Befragungswellen verwendet werden.⁵³ Auch für Datennutzende sind solche Unplausibilitäten problematisch, da sie meist noch weniger Informationen als die Datenbereitstellenden haben, um zu entscheiden, welcher Information getraut werden kann bzw. wie mit unplausiblen Fällen umgegangen werden soll.

Die Harmonisierung inkonsistenter Angaben in verschiedenen Datenquellen kann je nach Studie mehr oder weniger sinnvoll und praktikabel sein. Für die Entscheidung für oder gegen eine Plausibilisierung spielen u. a. die in der Datenaufbereitungsphase vorhandenen Ressourcen und die Komplexität der Studie (z. B. die Anzahl und Zuverlässigkeit verschiedener Quellen) eine Rolle.

Voraussetzung für die Identifikation und Handhabung von Inkonsistenzen zwischen mehreren Datensätzen ist, dass tatsächlich die Informationen von ein und derselben Untersuchungseinheit verglichen werden. Dies kann etwa mittels eines sogenannten *identity checks* überprüft und sichergestellt werden. Die Voraussetzung dafür ist, dass jede Untersuchungseinheit eine eindeutige Identifikationsnummer (z. B. Personen-ID) besitzt. Weiterhin werden personenidentifizierende Informationen – etwa das genaue Geburtsdatum, das Geschlecht⁵⁴ und der Vorname – aus allen Datenquellen über die Personen-ID zusammengespielt und abgeglichen. Hierzu können auch personenidentifizierende Informationen aus externen Datenquellen (z. B. Daten aus dem Stichprobenziehungsverfahren oder der amtlichen Statistik) herangezogen werden, falls solche verfügbar sind. Auch weitere zeitkonstante Merkmale von Personen, wie etwa die Körpergröße bei erwachsenen Befragungspersonen, können bei *identity checks* hilfreich sein. Inkonsistenzen müssen in der Regel einzelfallgeprüft werden. Sie können auf Verwechslungen bzw. eine Fehlzuzuweisung der Personen-ID in der Befragung hinweisen, aber ihre Ursache auch in Korrekturen z. B. des Geburtsdatums haben, wenn dieses zuvor falsch erfasst wurde. Wenn die Ursache von Inkonsistenzen bei den identifizierenden Angaben unklar ist, kann im Einzelfall durch das Erhebungsinstitut noch

⁵³Preloading bezeichnet das Einspielen und Einblenden dieser früheren Information in das Befragungsinstrument der neuen Erhebung.

⁵⁴Hierbei ist festzuhalten, dass das Geschlecht über die Zeit hinweg variabel sein kann, wenn – wie in den meisten Studien – nicht explizit nach dem biologischen bzw. genetischen Geschlecht gefragt wird. *Identity checks*, die lediglich auf dieses Merkmal anschlagen, sollten also besonders gründlich geprüft werden, denn möglicherweise liegt keine Verwechslung der Befragungsperson vor. Wenn möglich, kann eine erneute Kontaktaufnahme des Erhebungsinstituts mit der Befragungsperson für Aufklärung sorgen.

einmal bei der betroffenen Person nachgefragt werden. Hilfreich für die *identity checks* ist es, wenn die Interviewer:innen in einem Kommentarfeld offene Angaben zur vorgenommenen Korrektur machen können. Im Folgenden wird davon ausgegangen, dass sichergestellt ist, dass es sich in allen Datenquellen tatsächlich um dieselbe Untersuchungseinheit handelt.

6.2 Ursachen

Es gibt mehrere mögliche Gründe für die Entstehung inkonsistenter Informationen über dieselbe befragte Person in verschiedenen Datenquellen, sei es im Quer- oder im Längsschnitt.

6.2.1 In der Studienplanungsphase

Fehler in der Programmiervorlage oder dem Erhebungsinstrument

Inkonsistente Angaben über mehrere Datenquellen hinweg können im Erhebungsinstrument begründet sein. Beispielsweise kann eine Frage in einer spezifischen Welle einer Panelbefragung im Erhebungsinstrument inkorrekt oder einfach anders als in den Vorwellen konzipiert worden sein (Fehler in der Programmiervorlage) oder die Programmiervorlage falsch umgesetzt sein. In diesem Fall können die Antworten derselben Person auf dieselbe Frage zwischen den Befragungswellen abweichen.

In der TwinLife-Studie⁵⁵ wurde beispielsweise versehentlich über die ersten beiden Wellen hinweg bei der Frage nach der derzeit besuchten Schulform die Antwortoption „Orientierungsschule“ im Fragebogen vorgegeben. Diese Schulform existierte zum Zeitpunkt der ersten TwinLife-Befragung 2014 in Deutschland nicht mehr, dennoch wurde diese Antwort in wenigen Fällen gegeben. In der dritten Face-to-Face-Erhebung wurde „Orientierungsschule“ nicht mehr als Antwortoption angeboten. Entsprechend ergaben sich zweierlei Arten von Unplausibilität: einerseits die Angabe dieser Kategorie in Einzelfällen in den ersten beiden Wellen, obwohl die Schulform nicht existierte und andererseits abweichende Angaben in den Folgewellen, weil die Kategorie nicht mehr verfügbar war.

6.2.2 In der Datenerhebungsphase

Eingabefehler

Im Fall von Befragungen durch Interviewende, z. B. CAPI (*computer-assisted personal interviewing*) oder CATI (*computer-assisted telephone interviewing*), kann ein Teil der Inkonsistenzen zudem auf menschliche Fehler zurückgeführt werden, z. B. durch unkorrekte Bedienung der Software. Es ist zum Beispiel möglich, dass Interviewende

⁵⁵<https://www.twin-life.de/>

während der Erhebung aufgrund eines Missverständnisses eine falsche Information in den Fragebogen eintragen (etwa, weil die Befragungsperson keine klare Antwort gegeben hat oder diese falsch interpretiert wurde) oder beim Eintragen der Antworten Tippfehler entstehen (zum Beispiel ein Zahlendreher in der ID oder dem Geburtsjahr). Denkbar ist auch, dass Interviewende kategoriale Antwortmöglichkeiten unvollständig oder fehlerhaft vorlesen, was dazu führen kann, dass sich die Befragungspersonen in die dem wahren Wert am nächsten liegende Kategorie einordnen.

Bei Papierfragebögen kann eine undeutliche Handschrift zu widersprüchlichen Informationen in Umfragedaten führen.

Erinnerungsfehler

Ein weiterer Faktor, der unabhängig vom Befragungsmodus zu Unstimmigkeiten in den Informationen zwischen zwei oder mehreren Datenquellen führen kann, sind Erinnerungslücken der Befragungsperson selbst. So kann die Frage nach dem Jahr des Schul- oder Studienabschlusses zu zwei unterschiedlichen Zeitpunkten zu unterschiedlichen Antworten führen, insbesondere wenn das Ereignis schon sehr lange zurückliegt (siehe auch Abschnitt 7.6 im Kapitel zu längsschnittlichen Episodendaten).

Mangel- oder fehlerhaftes Befragungsinstrument

Bei der mehrfachen Abfrage derselben Information kann es zu unterschiedlichen Angaben kommen, wenn die Frage missverständlich formuliert oder nicht gut erläutert ist und somit auf verschiedene Weisen interpretiert werden kann. So können Angaben der Anzahl der (Enkel-)Kinder in verschiedenen Befragungen abweichen, wenn etwa in einem Fall auch Stief-, Pflege-, adoptierte und bereits verstorbene (Enkel-) Kinder mitgezählt, im anderen Fall jedoch nur lebende, leibliche (Enkel-)Kinder angegeben werden. Ein weiteres Beispiel ist die Abfrage des Erwerbsstatus. Unter Umständen könnte eine Person, die Rente bezieht und nie oder nur kurze Zeit erwerbstätig war, hier abweichend „in Rente“ oder „Hausfrau/-mann“ angeben, wenn eine Erläuterung fehlt.

Verständnisprobleme bei den Befragten

Zudem kann es zu Verständnisproblemen und somit inkonsistenten Informationen führen, wenn die Befragungsperson abgelenkt oder akustischen Störungen ausgesetzt ist, z. B. bei Anwesenheit anderer Haushaltsmitglieder. Auch Sprachbarrieren zwischen Befragungsinstrument oder -personal und Befragungsperson können zu Inkonsistenzen führen.

Falsche oder abweichende Proxyinformationen

Außerdem können Unstimmigkeiten bei bestimmten Merkmalen einer Befragungsperson entstehen, wenn die Fragen in unterschiedlichen Datenquellen (auch innerhalb einer Studie und innerhalb einer Erhebungswelle) von verschiedenen Personen beantwortet werden. Beim Survey of Health, Ageing and Retirement in Europe (SHARE)⁵⁶

⁵⁶<https://share-eric.eu/>

muss beispielsweise vor Beginn des Face-to-Face-Interviews ein kurzer Fragebogen zur Haushaltszusammensetzung von einer Person aus dem Haushalt beantwortet werden.⁵⁷ Dies muss nicht unbedingt die Befragungsperson selbst sein, der Fragebogen kann von einer beliebigen Person aus dem Haushalt ausgefüllt werden. Falls sich diese Person nicht korrekt an alle Informationen (z. B. Geburtsjahre) der anderen Haushaltsmitglieder erinnert, können diese Daten von den Informationen aus der Hauptbefragung, die die Befragungsperson selbst angibt, abweichen. Ein weiteres Beispiel sind Proxyinterviews, also Interviews, die aus bestimmten Gründen (beispielsweise aufgrund von physischen oder kognitiven Einschränkungen der Befragungsperson) nicht mit der Befragungsperson selbst, sondern mit einem oder einer Stellvertretenden der Befragungsperson durchgeführt werden. Ein Spezialfall eines Proxyinterviews sind Befragungen zu dem Tod der Befragungsperson wie sie in SHARE durchgeführt werden.⁵⁸ Da nicht in allen Fällen nahe Angehörige ermittelt oder zur Teilnahme an der Befragung motiviert werden können, werden diese sogenannten „End-of-life-Interviews“ teilweise auch von Freund:innen, Bekannten oder Nachbar:innen beantwortet, die möglicherweise nicht alles über die verstorbene Befragungsperson wissen und daher nur ungenaue oder geschätzte Angaben machen können. Die Informationen aus diesem Interview können von Informationen aus anderen Datenquellen, etwa früheren Befragungswellen abweichen, wenn der oder die Stellvertretende (versehentlich) falsche Angaben über die Befragungsperson macht. Auch aufgrund des Erhebungsdesigns kann es zu der mehrfachen und mitunter redundanten Erhebung von Informationen kommen. In der TwinLife-Studie beispielsweise werden insbesondere zu den jüngeren Zwillingen viele Angaben von den Eltern gemacht. Ältere Zwillinge hingegen geben die Auskünfte selbst. Für die Informationen zu Körpergröße und -gewicht werden für die Kinder im Alter von 10 bis 13 Jahren sowohl Selbst- als auch Fremdberichte erhoben. Diese stimmen nicht in allen Fällen überein.⁵⁹

⁵⁷Stuck, Stephanie, Sabrina Zuber, Fabio Franzese, Stefan Gruber, Tim Birkenbach, Senta-Melissa Pflüger, Josefine Atzedorf, Carolina Brändle, Magdalena Gerum. 2022. „SHARE Release Guide 8.0.0“. In: *Munich Center for the Economics of Aging*. https://share-eric.eu/fileadmin/user_upload/Release_Guides/SHARE_release_guide_8-0-0.pdf (abgerufen am 20.03.2023).

⁵⁸Börsch-Supan, Axel, Martina Brandt, Christian Hunkler, Thorsten Kneip, Julie Korbmacher, Frederic Malter, Barbara Schaan, Stephanie Stuck, Sabrina Zuber, on behalf of the SHARE Central Coordination Team, Data Resource Profile. 2013. „The Survey of Health, Ageing and Retirement in Europe (SHARE).“ *International Journal of Epidemiology*, Volume 42, Issue 4, August 2013, 992-1001. <https://doi.org/10.1093/ije/dyt088>

⁵⁹Christoph H. Klatzka, Lena Paulus, Franziska Lenau, Elisabeth Hahn. 2021. „Documentation TwinLife Data: Height, Weight, and BMI. F2F2. v.2.0.0.“ TwinLife Technical Report Series 06. <https://pub.uni-bielefeld.de/record/2951444> (abgerufen am 30.03.2023).

6.2.3 In der Datenaufbereitungsphase

Unterschiede oder Fehler in den Verarbeitungsschemata

Bei Variablen, die nach der Erhebung anhand (offen) abgefragter Merkmale codiert oder generiert werden, kann es passieren, dass durch unterschiedliche (oder auch fehlerhafte) Codierungsschemata dieselbe Information in zwei Datenquellen zu unterschiedlichen (und ggf. inkonsistenten) Ergebnissen führt. Beispiele hierfür sind die ISCED-Klassifikation, bei welcher Ausbildungsabschlüsse in ein ordinalskaliertes Schema überführt werden, oder die ISCO-Klassifikation für berufliche Tätigkeiten. Insbesondere die ISCO-Klassifizierung kann in vielen Fällen nicht automatisiert werden, sondern muss nach festgelegten Regeln händisch durchgeführt werden. Bei der TwinLife-Studie, bei der zwischen der zweiten und dritten Befragung ein Wechsel des Erhebungsinstituts stattfand, kam es aufgrund unterschiedlicher Codierungsregeln der Erhebungsinstitute zu Abweichungen in der ISCO-Klassifikation bezüglich bestimmter Berufsgruppen, obwohl der Codierung identische Angaben (offene Angabe des Berufs-, Schul- und Ausbildungsabschlusses) zugrunde lagen.

Vermeintliche Unplausibilität

Eine weitere mögliche Ursache für Inkonsistenzen in bestimmten Merkmalen einer Befragungsperson in unterschiedlichen Datenquellen ist die tatsächliche Veränderung der Information. Es ist durchaus möglich, dass Informationen, die theoretisch über die Zeit hinweg konstant sein oder sich konsistent entwickeln sollten, sich über die Zeit verändern. Beispiele hierfür sind ein weiterer Bildungsabschluss im hohen Alter oder eine angepasste Geburtsurkunde mit geändertem Geburtsdatum. Außerdem ist es möglich, dass sich subjektive Einstellungen einer Person (beispielsweise die Lebenszufriedenheit oder politische Einstellung) auch in kurzer Zeit stark verändern und somit unterschiedliche Informationen in zwei ungefähr zeitgleich erhobenen Umfragen auftreten.

Identifikation

Inkonsistenzen in Informationen aus unterschiedlichen Datenquellen können in der Datenaufbereitungsphase grundsätzlich durch das (automatisierte) Zusammenführen der einzelnen Datensätze und das Abgleichen mehrfach erhobener Informationen aufgedeckt werden. Hierfür muss ein eindeutiger Identifikator der Befragungsperson in allen Datenquellen enthalten sein. Werden beispielsweise einheitliche Informationen in bestimmten Variablen (z. B. ID, Geschlecht, Geburtsjahr) über mehrere Datenquellen für den Preload der nächsten Befragungswelle benötigt, bietet es sich an, alle verfügbaren Datenquellen anhand dieser Variablen zusammenzuführen, um etwaige Inkonsistenzen zwischen den Informationen zu identifizieren.

In SHARE und TwinLife werden nach jeder Datenerhebung alle vorhandenen Datensätze innerhalb einer Befragungswelle und über Befragungswellen hinweg anhand der ID,

des Geschlechts und des Geburtsjahres der Befragten kombiniert, um mögliche Inkonsistenzen in diesen für den Preload relevanten Variablen aufzudecken und gegebenenfalls zu korrigieren.

Insgesamt ist es sinnvoll, die Variablen zu ermitteln, die zwischen den verschiedenen Datenquellen identisch sein oder sich in eine logische Richtung entwickeln sollten und dort gezielt nach Inkonsistenzen zu suchen. Hierfür sollten die zu prüfenden Variablen in den einzelnen Datensätzen unterschiedlich benannt (z. B. `birthyear_1`, `birthyear_2`, ...) und anschließend über die identische (und als korrekt angenommene) ID-Variable zusammengeführt werden, sodass sie nebeneinander bestehen und direkt miteinander verglichen werden können. Typische Variablen für einen solchen Check wären beispielsweise:

- Geburtsjahr/Alter,
- Geschlecht⁶⁰,
- höchster Schul-/Bildungsabschluss, oder
- Körpergröße/-gewicht.

Diese sind in der Regel objektiv messbar und über die Zeit konstant oder variabel in einem einigermaßen festgelegten Rahmen. Dieser Rahmen muss für die Konsistenzprüfung definiert werden. In der TwinLife-Studie etwa wurden für die Variablen Körpergröße und Körpergewicht vorab Kriterien festgelegt, in welche Richtung und wie stark sich diese Indikatoren innerhalb eines Jahres entwickeln sollten (Erwachsene sollten nicht mehr als 5 cm wachsen oder schrumpfen; Mädchen unter 14 Jahren und Jungen unter 16 Jahren sollten wachsen, etc.). Die Einhaltung dieser Kriterien wird nach jeder Datenerhebung geprüft. Fälle, auf die diese Kriterien nicht zutreffen, die obersten 2,5 % der Größendifferenzen sowie Gewichtsunterschieden von mehr als 20 kg werden im Einzelfall geprüft.⁶¹

Falls in den Rohdaten vorhanden, können auch offene Angaben aus mehreren Befragungswellen oder -quellen, beispielsweise die Vornamen der Befragten (siehe auch Abschnitt Umgang) oder offene Berufsangaben, zur Fehleridentifikation herangezogen werden.

⁶⁰Wie bereits oben erwähnt ist das Geschlecht ein möglicherweise problematisches Merkmal, das besondere Beachtung und Prüfung verdient, da es sich in seltenen Fällen plausibel ändern kann. Dennoch ist Geschlecht in den meisten Studien ein zentrales Merkmal für Preloads und sollte daher nicht grundsätzlich bei Plausibilitätsprüfungen außen vor gelassen werden.

⁶¹Klatzka, Christoph H., Lena Paulus, Franziska Lenau und Elisabeth Hahn. 2021. „*Documentation TwinLife Data: Height, Weight, and BMI. F2F2. v.2.0.0.*“ TwinLife Technical Report Series 06 <https://pub.uni-bielefeld.de/record/2951444> (abgerufen am 30.03.2023).

6.3 Umgang

Grundsätzlich gibt es zwei Optionen wie Datenbereitstellende in der Vorbereitung auf die Datenveröffentlichung mit Inkonsistenzen in Umfragedaten umgehen können (siehe auch Kapitel 3 „Der Plausibilisierungsprozess“): Bearbeitung und Nicht-Bearbeitung.

Einerseits können die Daten ohne jegliche Bearbeitung den Datennutzenden zur Verfügung gestellt werden. Die Entscheidung, wie mit inkonsistenten Informationen umgegangen wird, wird dann den Datennutzenden überlassen. Der Vorteil für die Datenbereitstellenden ist, dass sie sich auf die Dokumentation von Inkonsistenzen (etwa durch Flag-Variablen⁶²) beschränken können. Der Vorteil für Datennutzende wiederum ist, dass die Entscheidung des Umgangs in Abstimmung auf die jeweilige Fragestellung vorgenommen werden kann (allerdings einhergehend mit höherem Aufwand bei der Analysevorbereitung).

Andererseits können die Datenaufbereitenden die Daten vor der Veröffentlichung auf Inkonsistenzen überprüfen und ggf. plausibilisieren. Für eine Plausibilisierung durch die Datenaufbereitenden spricht, dass ihnen oftmals mehr Informationen über die Befragten zur Verfügung stehen als den Datennutzenden (z. B. durch den Kontakt zu den Umfrageinstituten, Notizen der Interviewenden oder Informationen, die aus datenschutzrechtlichen Gründen nicht veröffentlicht werden) und sie somit auch besser informierte Entscheidungen bei der Plausibilisierung treffen können.

Nutzungsfreundliche, jedoch aufwendige Möglichkeiten sind die Veröffentlichung der unplausibilisierten Daten in Kombination mit plausibilisierten Daten oder auch die Erstellung von Flag-Variablen, die Personen oder Variablen markieren, bei denen Plausibilitäts- bzw. Konsistenzprobleme festgestellt, jedoch nicht korrigiert wurden. Diese zeigen den Datennutzenden problematische Fälle an, überlassen ihnen jedoch die Entscheidung über den Umgang damit. Ein Beispiel für die erste Methode ist die Vorgehensweise beim Sozio-oekonomischen Panel (SOEP)⁶³. Hier werden fehlerhafte Geburtsdaten in den Pfad-Datensätzen⁶⁴ korrigiert abgelegt. Die Befragungsdaten (z. B. aus dem Fragebogen für Jugendliche) bleiben jedoch unberührt. Eine Korrektur kann von den Datennutzenden auf Grundlage der Pfad-Datensätze dann selbst durchgeführt werden. Weiterhin wird beim SOEP zwischen unbearbeiteten Datensätzen, die direkt vom Erhebungsinstitut kommen (§pl-Datensätze), und aufbereiteten Datensätzen, die neben Generierungen auch Plausibilisierungen nach diversen Kriterien enthalten (§pgen-Datensätze), unterschieden.⁶⁵ Die zweite Methode, das Bereitstellen sogenannter Flag-Variablen, wird beispielsweise bei der German Longitudinal Election

⁶²Für eine Erläuterung des Begriffs siehe Informationskasten 2 in Kapitel 3.3.2.

⁶³https://www.diw.de/de/diw_01.c.412809.de/sozio-oekonomisches_panel_soep.html

⁶⁴Diese enthalten zu jeder jemals am SOEP beteiligten Befragungsperson und jedem Befragungshaus den Befragungsstatus in jeder Welle sowie eine Reihe unveränderlicher und auch veränderlicher Informationen.

⁶⁵Für mehr Informationen zur Datensatzstruktur der SOEP-Studie siehe <http://companion.soep.de/Data%20Structure%20of%20SOEPcore/Data%20Sets.html> (abgerufen am 30.03.2023).

Study (GLES)⁶⁶ angewandt. Für das Sample B des Panels wird eine Flag-Variable erstellt, die die Anzahl der Unstimmigkeiten bei den Verifizierungsvariablen (Geschlecht, Geburtsjahr, -monat und -ort (erster Buchstabe)) angibt.⁶⁷

Die radikalste Möglichkeit, mit inkonsistenten Informationen umzugehen, ist die Löschung der betroffenen Fälle. Dies führt jedoch zu einer kleineren und möglicherweise selektiven Stichprobe. Wie bereits oben thematisiert, können bei der Betrachtung der inkonsistenten Daten auch Fälle aufgedeckt werden, bei denen es sich nicht um Inkonsistenzen an sich handelt, sondern um eine Verwechslung der Befragungsperson. Wurde eine Person befragt, die nicht in die Stichprobe gehört, sollten die Daten nicht weiterverwendet werden. Ist eine Korrektur der Zuordnung von Befragungsdaten zu einer Personen-ID zweifelsfrei möglich, kann diese Korrektur durchgeführt werden. Weiterhin kann lediglich der als inkonsistent identifizierte Wert gelöscht und durch einen Missing-Code ersetzt werden. In der TwinLife-Studie wurde in der ersten Welle eine Familie der Kohorte 1 (mit dem Geburtsjahr 2009/2010 der Zwillinge) befragt. In der zweiten Welle wurde diese Familie als der Kohorte 4 (Geburtsjahr 1990-1993 der Zwillinge) zugehörig erfasst. Diese Inkonsistenz konnte nicht aufgeklärt werden, da die Familie im weiteren Verlauf nicht mehr kontaktierbar war. Ihre Befragungsdaten wurden folglich gelöscht.

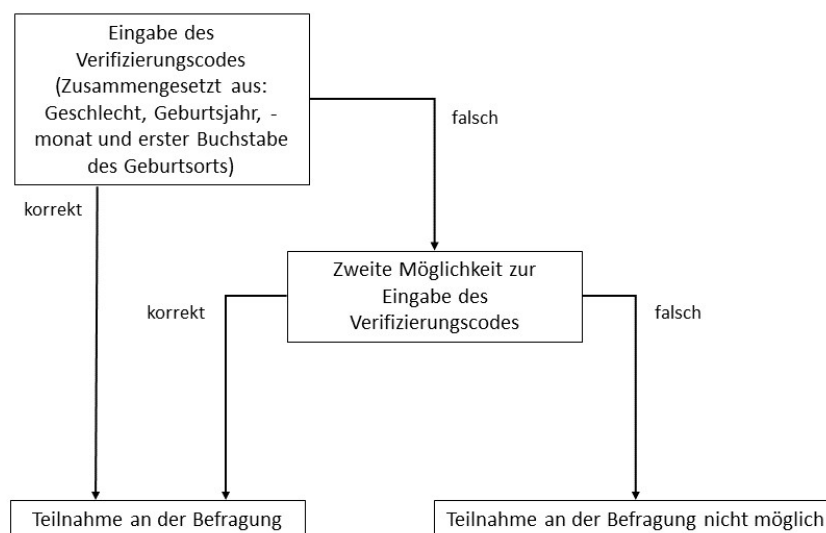


Abbildung 12: Verifizierungsprozess der Befragten im GLES Panel, Sample A.

Das Vorgehen bei Sample A des GLES-Panels (siehe Abbildung 12⁶⁸) ist ein Beispiel für

⁶⁶<https://gles.eu/>

⁶⁷Siehe dazu den Wave Report des Samples B des GLES Panels (ZA6838_wr_w14_sB.pdf), verfügbar unter: GLES (2021): GLES Panel 2016-2021, Wellen 1-15. GESIS Datenarchiv, Köln. ZA6838 Datenfile Version 5.0.0, <https://doi.org/10.4232/1.13783>.

⁶⁸Eigene Darstellung in Anlehnung an den Wave Report des Samples A des GLES Panels (ZA6838_wr_w15_sA.pdf), verfügbar unter: GLES (2021): GLES Panel 2016-2021, Wellen 1-15. GESIS Datenarchiv, Köln. ZA6838 Datenfile Version 5.0.0, <https://doi.org/10.4232/1.13783>.

das konsequente Ausblenden von Befragten, die sich nicht korrekt verifizieren können. Hier werden zu Beginn der Befragung Verifizierungsfragen⁶⁹ (Geschlecht⁷⁰, Geburtsjahr, -monat und erster Buchstabe des Geburtsorts) gestellt, die sicherstellen sollen, dass immer dieselbe Person am Panel teilnimmt. Falls Inkonsistenzen auftreten, die nicht aufgehoben oder geklärt werden können, werden diese Befragten ausgeblendet und können nicht weiter an der Befragung teilnehmen.⁷¹

Fällt die Entscheidung für eine Plausibilisierung aus, so sollte eine Korrekturstrategie festgelegt werden. Ein mögliches Vorgehen sind klar definierte Regeln wie etwa, bei Inkonsistenzen die Information aus der aktuellsten Datenquelle oder die am häufigsten genannte Variante zu verwenden. Auf diese Weise geht die TwinLife-Studie bei der Harmonisierung der Geschlechts- und Geburtsdatumsangaben zwischen den Erhebungswellen vor. Bei geringfügigen Differenzen zwischen Selbst- und Proxybericht von Körpergröße und -gewicht wird der Mittelwert zwischen den beiden Angaben gebildet (größere Differenzen werden im Einzelfall geprüft). Ein großer Vorteil solcher (einfacheren) Regeln ist, dass sie im Datenaufbereitungsprozess durch einen Algorithmus automatisiert durchgeführt werden können. Zudem sind sie für die Nutzer:innen einfach nachvollziehbar. Ein Nachteil besteht jedoch darin, dass insbesondere mit steigender Anzahl an Datenquellen, möglicherweise die Regeln den verschiedenen auftretenden Mustern nicht gerecht werden bzw. ein einfacher Algorithmus für eine zuverlässige Entscheidung nicht ausreicht. Je komplexer die Datenstruktur ist, desto eher besteht die Gefahr, dass der Algorithmus nicht in jedem Fall greift oder Werte fälschlicherweise umcodiert werden.

Falls nach regelbasierter Überprüfung oder Korrektur weiterhin Inkonsistenzen auftreten, bleibt meist nur die Einzelfallprüfung. Häufig ist es jedoch schwer abzuwägen, welche Quelle die wahre Information enthält. Jede Inkonsistenz einzeln zu betrachten ist sehr zeitaufwendig und birgt die Gefahr von uneinheitlichen und subjektiven Entscheidungen, die zudem schwierig zu dokumentieren sind. Es ist möglich, dass Einzelfallkorrekturen mehrfach an unabhängigen Stellen durchgeführt werden und die Fälle, in denen die Art der Korrekturen voneinander abweichen, noch einmal gesondert betrachtet werden. In Fällen, in denen unmöglich festgestellt werden kann, welche Angabe die korrekte ist, kann ggf. ein (speziell benannter und dokumentierter) Missing-Code vergeben werden.

⁶⁹Die genaue Formulierung und Darstellung der Verifizierungsfragen sowie der Frage nach dem Grund für einen evtl. falschen Verifizierungscode lässt sich in der Fragebogendokumentation zum GLES Panel nachlesen (ZA6838_fb.pdf). Diese Datei sowie weitere Dokumente zum Fragebogen, wie z. B. Bildschirmansichten, sind verfügbar unter: GLES (2021): GLES Panel 2016-2021, Wellen 1-15. GESIS Datenarchiv, Köln. ZA6838 Datenfile Version 5.0.0, <https://doi.org/10.4232/1.13783>.

⁷⁰Aufgrund der inzwischen bestehenden Möglichkeiten einer dritten Geschlechtsangabe „divers“ sowie einer Änderung der Geschlechtszugehörigkeit, wird in Zukunft (ab Welle 15) beim GLES Panel auf die Verwendung der Angabe des Geschlechts als Verifizierungsmerkmal verzichtet.

⁷¹Siehe dazu den Wave Report des Samples A des GLES Panels (ZA6838_wr_w15_sA.pdf), verfügbar unter: GLES (2021): GLES Panel 2016-2021, Wellen 1-15. GESIS Datenarchiv, Köln. ZA6838 Datenfile Version 5.0.0, <https://doi.org/10.4232/1.13783>.

Bei der Harmonisierung abweichender Informationen sollte berücksichtigt werden, wie vertrauenswürdig die einzelnen Datenquellen sind und nach welchen Kriterien entschieden wird, welche der Informationen als wahr angesehen wird. Die Vertrauenswürdigkeit der Informationen kann beispielsweise davon abhängen, ob die Befragungsperson auch die Auskunftsperson war oder ob es sich um ein Interview eines oder einer Stellvertretenden handelt (siehe Abschnitt *Ursachen*). Auch wie die Daten erhoben wurden, ist hier relevant. So kann z. B. das Geschlecht in einer Onlinebefragung (bewusst) falsch angegeben werden. Ob eine Selbstauskunft oder amtliche Daten, etwa dem Melderegister, vertrauenswürdiger sind, ist unter Umständen im Einzelfall zu entscheiden.

Für eine Harmonisierung der Umfragedaten sollten grundsätzlich alle zur Verfügung stehenden Informationen verwendet werden. Hierzu können auch die Kontakte zum Umfrageinstitut genutzt werden, um beispielsweise direkt bei den Interviewenden oder im Einzelfall sogar bei den Befragungspersonen nachzufragen.

6.4 Dokumentation

Die Dokumentation inkonsistenter Werte kann auf zwei unterschiedliche Arten erfolgen. Auf der einen Seite kann innerhalb der Daten dokumentiert werden und auf der anderen Seite durch die Bereitstellung zusätzlicher Dokumente. Zudem können beide Arten der Dokumentation miteinander verbunden werden.

Die Dokumentation direkt im Datensatz erfolgt meistens über die Vergabe von Missing-Codes oder das Erzeugen von Flag-Variablen. So wird beispielsweise beim Sozio-oekonomischen Panel (SOEP) der Missing-Code „-3 unplausibler Wert“ in längsschnittlichen Datensätzen vergeben, wenn bei der Generierung der längsschnittlichen Daten auffällt, dass ein Wert inkonsistent ist. In den querschnittlichen Datensätzen bleibt der Originalwert jedoch weiterhin bestehen. Eine alternative Vorgehensweise wäre hier, anstelle der Vergabe des Missing-Codes eine Flag-Variable zu generieren, welche die Datennutzenden darauf hinweist, dass die Variable bei bestimmten Beobachtungen inkonsistente Werte aufweist. In dem bereits oben beschriebenen Beispiel der Variablen zu Körpergröße und -gewicht in der TwinLife-Studie wird beides bereitgestellt, speziell definierte Missing-Codes und Flag-Variablen. Erstere ersetzen den eigentlich angegebenen Wert und erläutern den Nutzenden, was der Grund für das Löschen der Angabe ist (z. B. bei einer extremen Abweichung zwischen Selbst- sowie Proxy- bzw. Fremdbbericht, wobei der wahre Wert nicht ermittelt werden kann). Letztere zeigen den Nutzenden an, dass Korrekturen an der Variable vorgenommen wurden (z. B. Korrektur eindeutiger Tippfehler) oder dass Werte auffällig sind, jedoch keine Korrekturen vorgenommen wurden (z. B. extremes aber mögliches oder auch negatives Körperwachstum innerhalb eines Jahres).⁷²

⁷²Christoph H. Klatzka, Lena Paulus, Franziska Lenau, Elisabeth Hahn. 2021. „Documentation TwinLife Data: Height, Weight, and BMI. F2F2. v.2.0.0.“ TwinLife Technical Report Series 06. <https://pub.uni-bielefeld.de/record/2951444> (abgerufen am 30.03.2023).

Alternativ hierzu kann auch in einem Codebook (oder einer anderen Art von Variablen-dokumentation) darauf hingewiesen werden, dass eine Variable Werte aufweist, welche nicht konsistent sind, und Regeln bereitgestellt werden, mithilfe derer mögliche Fälle identifiziert werden können.⁷³

Insgesamt gilt, dass eine Dokumentation (etwa mittels eines technischen Reports) der durchgeführten Konsistenzprüfungen sowie der Korrekturmaßnahmen für Datennutzende hilfreich und wünschenswert ist.

6.5 Vermeidung

Da Inkonsistenzen bei der Verknüpfung mehrerer Datenquellen sowohl Datenbereitstellende als auch Datennutzende vor Probleme stellen, sollten vor allem in der Studienplanungsphase vor der Datenerhebung Maßnahmen zur Vermeidung solcher Inkonsistenzen ergriffen werden.

6.5.1 In der Studienplanungsphase

In der Entwicklungsphase des Fragebogens können Entscheidungen getroffen werden, die die Wahrscheinlichkeit für inkonsistente Informationen reduzieren. Sogenannte Identifier (z. B. eindeutig identifizierende Personen-IDs), die grundsätzlich in jeder Studie angelegt werden sollten, sind die Voraussetzung dafür, im Forschungsdatenmanagement herausstechende Fälle jederzeit identifizieren zu können. Dies ist oft eine laufende Nummer oder in Spezialfällen eine ID, die bestimmte Informationen enthalten kann (s. u. ein Beispiel für sogenanntes *classroom sampling*).

Um Fragen und Antwortkategorien, die zu Verständnisproblemen und somit zu abweichenden Angaben führen, zu vermeiden, sollten diese eindeutig und verständlich formuliert sein. Falls nicht auf bereits getestete Formulierungen zurückgegriffen werden kann, sollten die Fragen und Antwortkategorien vor dem Einsatz im Fragebogen getestet werden.⁷⁴ Um Verständnisprobleme zu vermeiden, ist es je nach Art des Fragebogens auch möglich, den Befragten bei Bedarf zusätzliche Erläuterungen zur Verfügung zu stellen, etwa durch optionale Einblendungen in Onlinebefragungen (Hilfeschaltfläche) oder die Möglichkeit, dass Interviewende bei Unsicherheit der Befragten die Frage weiter spezifizieren oder Beispiele nennen.⁷⁵

⁷³Siehe z. B. die Dokumentation zu Inkonsistenzen im Erhebungsinstrument bei den Variablen eca0200 (Ausbildungs- oder Studienabschluss) und eca0700 (Gewünschter Schulabschluss) im Codebook einer F2F-Erhebung in der TwinLife-Studie (verfügbar unter <https://www.twin-life.de/documentation/downloads>, abgerufen am 30.03.2023).

⁷⁴Porst, Rolf. 2013. *Fragebogen. Ein Arbeitsbuch*. Springer-Verlag.

⁷⁵Laferrère, Anne und Frederic Malter. 2015. „SHARE questionnaire encyclopaedia (or“question-by-question manual“ or „Q-by-Q“). In *SHARE Wave 5: Innovations & Methodology*. Hrsg. Axel Börsch-Supan und Frederic Malter. 48-50. https://share-eric.eu/fileadmin/user_upload/Methodology_Volumes/Method_vol5_31March2015.pdf (abgerufen am 23.03.23).

Sollen Umfragedaten mit einer bestehenden Datenquelle verknüpft werden, kann es hilfreich sein, sich bei der Erhebung der Daten an der bestehenden Datenquelle zu orientieren. Um Inkonsistenzen über verschiedene Befragungen hinweg zu vermeiden, ist es von Vorteil, die Formulierung der Fragen und die Antwortkategorien zwischen den Befragungen nicht (stark) zu verändern. Eine konsistente Erhebung erleichtert auch den Datennutzenden den Umgang mit den Daten. Dennoch kann es gerechtfertigt sein, Inkonsistenzen in Kauf zu nehmen und die Fragen zu verändern, um die Qualität der erhobenen Daten zu verbessern – insbesondere, wenn die Frage zuvor falsch oder missverständlich formuliert war. Auch hier gilt, dass Kosten und Nutzen von Änderungen klar sein und gegeneinander abgewogen werden müssen.

Bei Panelbefragungen ist außerdem zu erwägen, einige (über die Zeit konstante) Informationen aus vorherigen Befragungen als Preload einfließen zu lassen. Das heißt, die Information wird von einer früheren in die aktuelle Befragung weitergegeben, wo sie für die Filterführung verwendet werden kann. Dies hat den Vorteil, dass Befragte bestimmte Fragesequenzen nicht in jeder Befragung wiederholen müssen und das Interview somit kürzer wird. Ohne erneute Eingabe der Information kann es nicht zu widersprüchlichen Angaben zwischen den Befragungen kommen. Allerdings können dann Fehler in der Information der vorherigen Befragung oder Änderungen in als über die Zeit konstant angenommenen Merkmalen, wie ein weiterer Bildungsabschluss im hohen Alter, nicht abgebildet werden. Weiterhin ist denkbar, Informationen, die in früheren Befragungswellen schon abgefragt wurden, in der aktuellen Befragung als Preload einzusetzen und sie nur noch bestätigen zu lassen und nicht wieder komplett neu zu erfassen. Dieses Vorgehen kann (unbeabsichtigte) Inkonsistenzen vermeiden und trotzdem können Änderungen noch eingetragen werden. Unter Umständen verlängert sich jedoch die Interviewdauer im Vergleich zu der Verwendung von Preloads ohne Bestätigung durch die Befragten. Außerdem kann dieses Vorgehen Anreize zu einer „Bestätigungsstrategie“ durch die Befragten oder die Interviewenden führen, die Preloads pauschal bestätigen – unabhängig davon, ob eine Änderung stattgefunden hat oder nicht – um die Befragungsdauer zu verkürzen.⁷⁶

Bei Querschnittserhebungen bzw. Panelwellen mit mehreren Datenquellen oder Fragebögen ist eine pragmatische Lösung, auf die mehrfache Erhebung ein und derselben Information (z. B. Alter oder Geburtsdatum) zu verzichten, um inkonsistente Angaben zu vermeiden. Die Instrumente können (abgesehen von PAPI-Fragebögen) in der Regel so programmiert werden, dass die identifizierenden Informationen von einem Instrument in das nächste transferiert werden. So kann beispielsweise bei Proxyinterviews der Vorname der Person, über die Auskunft gegeben wird, eingeblendet und die erhobenen Daten im Anschluss automatisch an die entsprechende Person angespielt werden. Ein Nachteil dieser Vorgehensweise ist jedoch, dass etwaige Verwechslungen nicht mehr so einfach aufzudecken sind. Gibt es in einem Haushalt laut Befragung zur Haushaltszusammensetzung beispielsweise zwei Personen, ID 01 geboren im Jahr

⁷⁶Eggs, Johannes, Annette Jäckle. 2015. „Dependent Interviewing and Sub-Optimal Responding.“ *Survey Research Methods*, 9(1), 15-29. <https://doi.org/10.18148/srm/2015.v9i1.5860>.

1970 und ID 02 geboren im Jahr 1975, und würde das Geburtsjahr gemeinsam mit der ID in das Interviewinstrument transferiert werden, würde es in den Daten nicht sofort auffallen, wenn Person 01 versehentlich das Interview auf ID 02 durchgeführt hätte. Dies kann auch passieren, wenn der Name der eigentlich zu befragenden Person einblendet wird und liegt meist in Fehlern durch Interviewer:innen wie dem Ignorieren der Einblendung begründet.

Vor Beginn der Datenerhebungsphase der Studie sind bei Face-to-face-Befragungen gute und genaue Schulungen der Interviewer:innen unabdingbar. Je besser die Interviewer:innen trainiert werden, desto geringer fallen vermutlich die Inkonsistenzen aufgrund von Eingabefehlern aus. Bei der Ausbildung der Interviewer:innen kann zudem darauf geachtet werden, dass diese ihre Arbeit ausführlich dokumentieren (von der Kontaktaufnahme bis zum Verlauf des Interviews). Die Dokumentation kann bei der Datenverarbeitung bzw. Plausibilisierung eine wichtige Quelle für Informationen sein, die bei Inkonsistenzen hinzugezogen werden kann.

Postalische bzw. schriftliche Erhebungen, wie sie zum Beispiel für sogenanntes *classroom sampling* durchaus noch eingesetzt werden, stellen einen Spezialfall dar. In der Erhebungsphase empfiehlt sich die Verwendung einer prüfziffergeschützten ID (insbesondere falls die Fragebögen händisch erfasst werden) und eine sehr genaue Erfassung des Rücklaufs. In den Studienberechtigtenbefragungen des DZHW⁷⁷ wurden z. B. alle über die Schulen ausgegebenen Fragebögen mit einer ID versehen, die neben dem Bundesland, der Schulart, einer Schul- und Schüler:innenummer auch eine Prüfziffer nach dem „Modulo 10“-Verfahren enthielt.⁷⁸ Der Rücklauf, die postalisch zurückgeschickten Fragebögen, wurde über einen sogenannten Paginierstempel mit einer laufenden Nummer versehen und diese mit der ID und dem Datum tagesaktuell in einer Tabelle erfasst. Dies ermöglichte das sichere gleichzeitige Handling in Datensatz und Papierform sowie in der zweiten Welle die Zuordnung der Erste-Welle-Fragebögen.

Grundsätzlich ist eine absolute Vermeidung von Inkonsistenzen in Befragungsdaten nicht möglich. Umfragedaten sind fehleranfällig und die Wahrscheinlichkeit für Unstimmigkeiten steigt, je mehr Datenquellen miteinander kombiniert werden. Gleichzeitig sind multiple Datenquellen für Informationen über die Befragten für die Forschung äußerst relevant. Auch ist festzuhalten, dass Inkonsistenzen nicht zwangsläufig Fehler in den Daten identifizieren.

⁷⁷https://www.dzhw.eu/forschung/projekt?pr_id=465

⁷⁸In einer Gleichung mit den übrigen Zahlen müssen Prüfziffern den logischen Wert „Wahr“ ergeben.

7 Inkonsistenzen in längsschnittlichen Episodendaten

Christina von Rotz, Maarten Koomen, Maximilian Trommer, Percy Scheller

7.1 Beschreibung

7.1.1 Was sind Episoden?

Episoden (Spells) sind zeitliche Verläufe von Tätigkeiten oder Zuständen, die in verschiedenen Formen erhoben werden können und typischerweise in einem Episoden-File (im Längsschnittdatenformat) publiziert werden. Episoden sind meist definiert durch einen Anfangs- und einen Endzeitpunkt (z. B. Jahre, Monate, Wochen, Tage, Stunden, Minuten etc.), seltener auch durch einen Zeitpunkt und eine Länge oder durch verschiedene, diskrete Zeiträume.⁷⁹ Episoden werden bestimmt durch ein oder mehrere Merkmale eines Zustandes oder einer Tätigkeit. Dies können in Abhängigkeit des Forschungsinteresses und des Zeithorizonts sowohl sehr langfristige Merkmale, wie zum Beispiel Erwerbstätigkeit, Arbeitslosigkeit, Ehedauer, Vereinszugehörigkeit oder Wohnorte sein, als auch sehr kurzfristige wie zum Beispiel Essen, Joggen, Einkaufen oder Telefonieren. Dabei können einer Episode auch mehrere Charakteristiken zugeordnet werden, beispielsweise ein bestimmter Job (1.) wird befristet (2.) mit einem bestimmten Gehalt (3.) in einer bestimmten Firma (4.) ausgeübt; gegessen (1.) wird vegetarisch (2.) zuhause (3.).

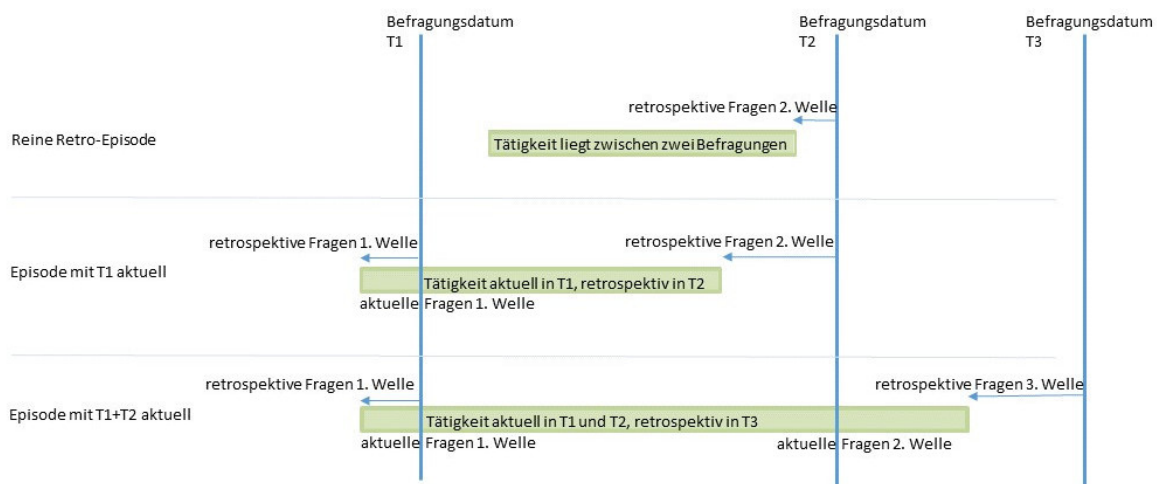


Abbildung 13: Episoden zu unterschiedlichen Erhebungszeitpunkten

⁷⁹Eine dreimonatige Episode mit Arbeitslosigkeit kann beispielsweise definiert sein als „Juni - August: arbeitslos“, „Ab Juni drei Monate arbeitslos“ oder „Juni: arbeitslos, Juli: arbeitslos, August: arbeitslos“.

Wie Abbildung 13 zeigt, werden bei längsschnittlichen Episoden meist nicht nur aktuelle Tätigkeiten zu einem bestimmten Erhebungszeitpunkt erfasst, sondern auch retrospektive Phasen dieser Tätigkeiten.⁸⁰ So können sich Episoden zu den Erhebungszeitpunkten unterschiedlich positionieren.

Messungen sich überschneidender Zeiträume sind für Episodendaten in standardisierten Panelbefragungen üblich (vgl. Abbildung 13) und bei lang andauernden Episoden können sich auch Mehrfachmessungen zu der gleichen Tätigkeitsepisode ergeben (vgl. Abbildung 13, 3. Zeile).⁸¹ Beides kann Widersprüche zwischen den Messwerten generieren. Ferner kann sich die Kombination verschiedener Episoden zu einem bestimmten Zeitpunkt inhaltlich auch völlig ausschließen (z. B. arbeitslos und 100 % arbeitstätig; essen und schlafen) ohne dass sich die Episodendaten, zum Beispiel in den Mehrfachmessungen, widersprechen.

Auch die mehrfache Erhebung von Episoden in einer Erhebung(swelle) kann widersprüchlich sein. Abbildung 2 und 3 zeigen ein fiktives Beispiel der Erhebung von Erwerbsepisoden, angelehnt an das „DZHW-Absolventenpanel“.⁸² Um unterschiedliche Episoden, wie Ausbildungen oder Erwerbstätigkeiten und Unterbrechungen leicht zugänglich abbilden zu können, werden diese zunächst anhand eines Kalenders erfasst:



Abbildung 14: Erhebung von Erwerbsepisoden in Form eines Kalenders, DZHW-Absolventenstudien

Danach werden zu den wichtigsten Erwerbsepisoden weitere Daten in Tabellenform erhoben. Die Episodendaten aus dem Kalender können dabei nicht übernommen werden, weil es sich teilweise um eine Paper-Pencil-Befragung handelt oder die verwendete Software keine Übernahme der Episoden ermöglicht.

⁸⁰Dies unterscheidet die retrospektive Erhebung von Episoden von Erhebungen zu einem Panelbefragungszeitpunkt. Hier würde nur der Zustand zu den Befragungszeitpunkten gemessen.

⁸¹Meist werden Inkonsistenzen auch erst bei Widersprüchen durch Mehrfachmessung offensichtlich.

⁸²Brandt, Gesche, Kolja Briedis, Gunther Dahm, Thorsten Euler, Gregor Fabian, Saskia Klüver, Torssten Rehn, Maximilian Trommer und Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW). 2022. „DZHW Graduate Panel 2009/DZHW-Absolventenpanel 2009“. German Centre for Higher Education Research and Science Studies (DZHW). <https://doi.org/10.21249/DZHW:GRA2009:2.0.0>.

Zeitraum (Monat/Jahr)	Art des Arbeits- verhält- nisses	Arbeitszeit (vertraglich vereinbart)	Berufliche Stellung	Arbeitsort	Firma/ Betrieb
von: 01 / 20 10 bis: 03 / 20 10 läuft noch <input type="checkbox"/>	1 Schlüssel siehe unten	<input checked="" type="checkbox"/> Vollzeit mit <input type="checkbox"/> Teilzeit mit ohne fest <input type="checkbox"/> vereinbarte Arbeitszeit mit ca. 39 Std./Woche	3 Schlüssel siehe unten	Bundesland bzw. Land (bei Ausland) Ort: _____ (erste 3 Ziffern der PLZ) falls PLZ nicht bekannt, bitte Ort angeben:	A Schlüssel siehe unten
von: 05 / 20 10 bis: _____ / 20 _____ läuft noch <input checked="" type="checkbox"/>	1 Schlüssel siehe unten	<input type="checkbox"/> Vollzeit mit <input checked="" type="checkbox"/> Teilzeit mit ohne fest <input type="checkbox"/> vereinbarte Arbeitszeit mit ca. 25 Std./Woche	2 Schlüssel siehe unten	Bundesland bzw. Land (bei Ausland) Ort: _____ (erste 3 Ziffern der PLZ) falls PLZ nicht bekannt, bitte Ort angeben:	B Schlüssel siehe unten
von: _____ / 20 _____ bis: _____ / 20 _____ läuft noch <input type="checkbox"/>	<input type="checkbox"/> Schlüssel siehe unten	<input type="checkbox"/> Vollzeit mit <input type="checkbox"/> Teilzeit mit ohne fest <input type="checkbox"/> vereinbarte Arbeitszeit mit ca. _____ Std./Woche	<input type="checkbox"/> Schlüssel siehe unten	Bundesland bzw. Land (bei Ausland) Ort: _____ (erste 3 Ziffern der PLZ) falls PLZ nicht bekannt, bitte Ort angeben:	<input type="checkbox"/> Schlüssel siehe unten

Abbildung 15: Detaillierte Erhebung von Erwerbsepisoden, DZHW-Absolventenstudien

Es fallen zwei Inkonsistenzen auf: Zum einen hört die gleiche Stelle A in Abbildung 14 erst im April auf, in Abbildung 15 dagegen bereits im März. Zum anderen wird die Stelle B für sechs Monate unterbrochen (Abbildung 14), danach aber durchgängig angegeben.⁸³ Spätestens bei der Auswertung muss eine Entscheidung getroffen werden, welche Angabe als die wahrscheinlichere angenommen wird, was in diesem Beispiel keinesfalls offensichtlich ist.

Darüber hinaus werden Konsistenzprobleme evident, wenn eine Episode in zwei oder mehr Messzeitpunkten erfasst wird. Es können Widersprüche zwischen den Angaben in mehreren Befragungen auftreten. Diese Widersprüche unterscheiden sich abhängig davon, ob ein *dependent* oder *independent interviewing* gewählt wurde und ob die Befragungsrichtung reaktiv oder proaktiv ist.

7.1.2 Wie werden Episodendaten erhoben?

Es gibt zwei Möglichkeiten Episodendaten zu erheben. In einem *dependent interviewing* werden die Episodendaten aus früheren Befragungswellen in Form von Preloads in die aktuelle Befragung eingespeist.⁸⁴ In der Regel sind dies computerbasierte Erhe-

⁸³In beiden Fällen ist durchaus möglich, dass unterschiedliche Fragestellungen den Unterschied erklären und keine echte Inkonsistenz vorliegt: So könnte einmal die vertragliche und einmal die faktische Situation abgefragt worden sein; je nach Art der Unterbrechung ist ein fortgeführter Arbeitsvertrag nicht unplausibel. Allerdings führen Interpretationsspielräume der Befragten auch immer wieder zu Interpretationsschwierigkeiten der Auswertung.

⁸⁴Beispielsweise verwendet die TREE-Studie *dependent interviewing*: Rudin, Melania und Christoph Müller. 2013. „Kann es sein, dass ich das Beginndatum falsch erfasst habe?‘ Reactive Dependent Interviewing bei der Erhebung von Ausbildungs- und Erwerbsverläufen junger Erwachsener“. *Methoden, Daten, Analysen* 7 (3), S. 433-463, <http://doi.org/10.12758/mda.2013.020>. University of Bern. 2016. „TREE: Transitions from Education to Employment, Cohort 1 – 2000-2014. Version 1.0.0“ Distributed by FORS, Lausanne. <https://doi.org/10.23662/FORS-DS-816-7>.

bungsmodi wie CAPI (*computer assisted personal interviewing*), CATI (*computer assisted telephone interviewing*) oder CAWI (*computer assisted web interviewing*).

Im Unterschied zu einem *dependent interviewing* werden in einem *independent interviewing* keine in früheren Wellen erfassten Episodendaten (Preloads) eingespeist. Dadurch ist kein Abgleich mit bereits früher erfassten Episoden direkt in der Befragung möglich. Diese Form der Datenerfassung wird typischerweise bei Paper-Pencil-Fragebögen (wie den initialen Befragungswellen des DZHW-Studienberechtigtenpanels) verwendet.

In einem *dependent interviewing* kann die Befragungsrichtung proaktiv oder reaktiv erfolgen. In einem *proactive dependent interviewing* wird prototypisch gefragt: „Sie haben uns das letzte Mal gesagt, dass...“. Offene Episoden (zum Zeitpunkt der letzten Befragung war eine Episode „aktuell“) werden im Interview als Ausgangspunkt genommen, um die weitere Entwicklung dieser Episoden zu verfolgen. Dabei wird die eingespeiste Episode zuerst bestätigt und danach wird gefragt, ob diese Episode immer noch aktuell ist.⁸⁵ Die Befragung bezieht sich dabei auf den Zeitraum zwischen dem letzten Interview bis in die Gegenwart.

The screenshot shows a software interface for a CATI interview. On the left, there is a vertical menu with categories: Ausbildung, Schule, Ausbildungsbetrieb, Lehrvertragsbetrieb, Praktikum, Erwerbstätigkeit, and Andere. The main area displays a timeline from 2017 to 2019. A specific episode is highlighted in purple, labeled '[7262] Malermeister GmbH'. Below the timeline, a text box contains the following text:

Sie haben bei der letzten Befragung im Februar 2019 angegeben, dass Sie folgende Ausbildung machen:
 3-4-jährige Berufslehre mit Lehrbetrieb
 Malerin EFZ.

Wie ist es heute: Sind Sie aktuell noch in dieser Ausbildung, ohne grösseren Unterbruch?
Oder haben Sie diese Ausbildung beendet oder abgebrochen?
 (Epi-Nr: [7262])

Grösserer Unterbruch = Phase ohne Ausbildungsinstitution über mehr als 1 Monat (weder Schule noch Betrieb)

immer noch aktuell, ohne grösseren Unterbruch dazwischen
 nicht mehr aktuell oder grösserer Unterbruch
 Fehler: Eingblendete Angabe von früher falsch (ACHTUNG: Episode wird wenn möglich rückwirkend korrigiert)
 Widerspruch: Habe zum Zeitpunkt der letzten Befragung nie eine solche oder ähnliche Ausbildung gemacht (ACHTUNG: Episode wird GELÖSCHT)
 Weiss nicht
 Angabe nicht möglich
 Angabe explizit verweigert

At the bottom, there are two buttons: 'Zurück' and 'Weiter'.

Abbildung 16: Proactive dependent interviewing, TREE-Studie 2. Kohorte (CATI-Version)

Abbildung 16 zeigt ein Beispiel aus der TREE-Studie (Transitionen von der Erstausbil-

⁸⁵Jäckle, Annette. 2006. „Dependent Interviewing: A Framework and Application to Current Research“. ISER Working Paper 2006–32. <https://www.econstor.eu/bitstream/10419/92099/1/2006-32.pdf> (abgerufen am 30.03.2023).

derung ins Erwerbsleben)⁸⁶ für ein *proactive dependent interviewing*. Dabei werden die Daten aus der letzten Befragung eingeblendet (Preload) und darauf aufbauend die aktuelle Situation der befragten Person erfasst. Bei falschen Daten können die Angaben entweder korrigiert oder gelöscht werden.

Beim *reactive dependent interviewing* werden Episoden zuerst unabhängig von den eingespeisten Daten erfragt und danach mit den vorgeladenen Episoden verglichen. Die neu erfassten Episoden können nun vollständig, teilweise oder überhaupt nicht mit den bereits bestehenden Episoden abgeglichen werden. Diese Abgleiche (*matches*) basieren auf den jeweiligen Episodendefinitionen, wobei Widersprüche bei Matches direkt im Interview behandelt werden können. Meist wird *reactive dependent interviewing* verwendet, wenn die Befragung von der Gegenwart in die Vergangenheit erfolgt.⁸⁷ Beispielsweise wird ein aktueller Erwerb erfasst und, wenn das Beginn-Datum des Erwerbs vor der letzten Befragung liegt, mit den vorhandenen Daten (Preload) aus früheren Befragungen abgeglichen.

Tabelle 4 zeigt, angelehnt an die TREE-Studie Kohorte 1, wie eine solche Bewertung anhand der Kriterien Erwerbstätigkeit, Beginn-Datum und Betrieb erfolgen kann.

Tabelle 4: Beispiel Matching-Kriterien bei Erwerbsepisoden

	Job	Beginn	Betrieb
Fall 1: alle drei Kriterien matchen	+	+	+
Fall 2: Job gleich, Beginn-Datum gleich, Betrieb unterschiedlich	+	+	-
Fall 3: Job unterschiedlich, Beginn-Datum gleich, Betrieb gleich	-	+	+
Fall 4: Job gleich, Beginn-Datum unterschiedlich, Betrieb gleich	+	-	+
Fall 5: Job untersch., Beginn-Datum gleich, Betrieb untersch.	-	+	-
Fall 6: Job gleich, Beginn-Datum untersch., Betrieb untersch.	+	-	-
Fall 7: Job untersch., Beginn-Datum untersch., Betrieb gleich	-	-	+
Fall 8: alle drei Kriterien unterschiedlich	-	-	-

Die Differenzen können im Interview nun gezielt angesprochen und geklärt werden.

7.1.3 Arten von widersprüchlichen und fehlenden Angaben in Episodendaten

In Episodendaten können sehr unterschiedliche Widersprüche auftreten und auch fehlende Angaben können eine Plausibilisierung lohnenswert machen. Je nach Fragestellung können aber auch gerade diese ge- beziehungsweise unterbrochenen Episoden im Analysefokus stehen:

⁸⁶University of Bern. 2021. „TREE. Transitions from Education to Employment, Cohort 2 (TREE2), Panel waves 0-2 (2016-2018)“. Distributed by FORS, Lausanne. <https://doi.org/10.23662/FORS-DS-1255-1>.

⁸⁷Jäckle „Dependent Interviewing“, 5.

- *vollständiger Widerspruch zu vorgeladenen Episoden oder zu Episodenangaben in der gleichen Befragung*: beispielsweise wird negiert, überhaupt eine Ausbildung gemacht zu haben.
- *fehlende Episoden aus früheren Befragungen*: zum Beispiel wird eine Ausbildung angegeben, die vor der letzten Befragung beginnt, aber in der letzten Befragung nicht erwähnt wurde.
- *Teilwiderspruch zu vorgeladenen Episoden, oder zu früheren oder in der gleichen Befragung erfassten Episoden in independent interviewing (Teilübereinstimmung)*: zum Beispiel wird eine andere Ausbildung angegeben, als die bereits in früheren Befragungen erfasste.
- *doppelte Episoden*: beispielsweise wird eine Tätigkeit einmal als Ausbildung und einmal als Praktikum angegeben
- *fehlende Episoden, die eine Lücke im Verlauf erzeugen*: generiert Phasen ohne Angaben
- *fehlende Episoden, die keine Lücken erzeugen*: zum Beispiel werden nicht alle parallelen Erwerbstätigkeiten angegeben, oder bestehende Episoden überstreichen Lücken. Solche Übergangseffekte werden in der Literatur auch als *seam effects* bezeichnet.⁸⁸
- *Episoden, die in sich unplausibel sind*: zum Beispiel mehr als 12 Monate im Jahr, mehr als 24 h am Tag.

7.2 Ursachen von Widersprüchen und Lücken innerhalb Episoden

7.2.1 In der Studienplanungsphase

In der Studienplanungsphase können durch das Survey Design und die Verwendung von *dependent* und *independent interviewing* Widersprüche verursacht werden. Der Unterschied zwischen *dependent* und *independent interviewing* ist die Bereitstellung von Stimuli für die Befragten. Die Einführung solcher Stimuli kann ihrerseits zu Widersprüchen führen, weil (1) die Befragten in einer vorherigen Panelwelle einen Irrtum begangen haben und/oder (2) die Befragten in der aktuellen Panelwelle ein Ereignis fehlerhaft erinnern.⁸⁹

Definition der Eigenschaften von Episoden

Auch die Episodendefinition im Survey Design bestimmt, welches Ausmaß an Variation in den Eigenschaften bzw. Merkmalen mit einer Veränderung im biografischen Verlauf des oder der Befragten verbunden ist. Beispielsweise führt ein Filialwechsel

⁸⁸Rips, Lance J., Frederick G. Conrad und Scott S. Fricker. 2003. „Straightening the Seam Effect in Panel Surveys“. *Public Opinion Quarterly* 67 (4): 522–54. <https://doi.org/10.1086/378962>.

⁸⁹Mathiowetz, Nancy A. und Katherine A. McGonagle. 2000. „An Assessment of the Current State of Dependent Interviewing in Household Surveys“. *Journal of Official Statistics* 16 (4): 401–18. Lynn, Peter, Annette Jäckle, Stephen P. Jenkins und Emanuele Sala. 2006. „The Effects of Dependent Interviewing on Responses to Questions on Income Sources“. *Journal of Official Statistics* 22 (3): 357–84.

im Einzelhandel zu zwei Erwerbsepisoden, wenn der Arbeitgeber auf der Betriebsebene definiert wird. Ist er dagegen auf der Unternehmensebene definiert, wird nur eine Episode erstellt.

Survey Design als Modell eines Ausschnittes der Wirklichkeit

Bei der Konstruktion eines Fragebogens auf der Basis von zeitlichen Verläufen wird immer ein Wirklichkeitsausschnitt, den man untersuchen möchte, als Modell formuliert. Dabei muss die Realität in ihrer Komplexität reduziert und strukturiert werden. Da kein Modell absolut passgenau auf die Situation aller Befragten erstellt werden kann oder die Fragen missverstanden werden können, sind Inkonsistenzen und Widersprüche nicht zu vermeiden. In dieser Design-Phase kann es sich lohnen, mögliche Widersprüche und Inkonsistenzen zu berücksichtigen und entsprechende Lösungen zu implementieren. Zum Beispiel gibt es in der zweiten TREE Kohorte eine Schule, die eine Ausbildung sowohl mit einem gymnasialen Abschluss als auch einem Berufslehrausbildung anbietet. Im Survey Design war diese Option nicht genügend modelliert, so dass Befragte nur entweder eine Berufsausbildung oder ein Gymnasium auswählen konnten. In den meisten Fällen führte dies bei jeder Befragung zu Widersprüchen, weil das Survey Design die reale Situation nicht genau genug abgebildet hat. Durch die Erweiterung der Ausbildungsliste um die Kategorie „Berufsausbildung und Gymnasium“ und die Recodierung der bestehenden Fälle, konnten diese Widersprüche eliminiert werden.

Wandel gesellschaftlicher Strukturen

Auch Veränderungen gesellschaftlicher Strukturen können zu Widersprüchen zwischen Erhebungswellen führen. Zum Beispiel erfolgten mit dem Bologna-Prozess zur Vereinheitlichung des europäischen Hochschulraumes in einigen Ländern strukturelle und inhaltliche Veränderungen am Bildungssystem. Ein Kernteil des Bologna-Prozess war die Harmonisierung der akademischen Ausbildung durch ein zweistufiges System mit zwei separaten Abschlüssen (Bachelor und Master Diplom). In manchen europäischen Ländern bestand vor dem Bologna-Prozess die Hochschulausbildung aus nur einer Phase, die zu einem einzigen Abschluss führte. Eine derartige strukturelle Veränderung im Studienaufbau kann zu Inkonsistenzen in Befragungen führen, insbesondere, wenn die frühere und die heutige Realität im Survey Design nicht genügend berücksichtigt werden.

Ein anderes Beispiel geben Lynn et al.⁹⁰: In vielen Wohlfahrtssystemen gibt es eine Vielzahl von staatlichen Leistungen, die die Befragten erhalten könnten. Die meisten dieser Leistungen ändern sich recht häufig im Namen oder in den Anspruchsvoraussetzungen. Dies kann also zu (vermeintlichen) Inkonsistenzen führen, wenn ein und dieselbe Episode von einer Definitions- oder Strukturveränderung betroffen ist.

⁹⁰Lynn et al. „Effects of Dependent Interviewing“.

Semantische Veränderungen

Veränderungen können auch weniger formale Entwicklungen in unserem alltäglichen Sprachgebrauch widerspiegeln – wie zum Beispiel wechselnde oder neu geschaffene Berufsbezeichnungen. Solche externen strukturellen Veränderungen oder Entwicklungen stellen eine Herausforderung sowohl für die Befragten als auch für das Fragebogendesign dar. Die Fragebögen müssen von Welle zu Welle auf dem neuesten Stand gehalten werden, und die Befragten, die sich solchen Veränderungen nicht vollständig bewusst sind, erkennen ihre spezielle Situation möglicherweise nicht mehr, wenn ihnen der neueste Stand vorgelegt wird.

Ein Beispiel aus der schweizerischen Liste der Lehrberufe (berufliche Grundbildung):

Tabelle 5: Schweizerischen Liste der Lehrberufe: der Beruf „Polybauer:in“ bis 1.1.2017

	Lehrberuf bis 1.1.2017	ISCED Ausbildungsfeld		Fachrichtung
51908	Polybauer:in EFZ	0732 Baugewerbe, Hoch- und Tiefbau	51908	Abdichten
			51909	Dachdecken
			51910	Fassadenbau
			51911	Gerüstbau
			51912	Sonnenschutz-Systeme

Der Lehrberuf Polybauer EFZ (eidgenössisches Fähigkeitszeugnis) wurde aufgehoben und durch folgende neue Lehrberufe ersetzt (ehemals Fachrichtungen):

Tabelle 6: Schweizerischen Liste der Lehrberufe: der Beruf „Polybauer:in“ ab 1.1.2017

	Lehrberuf ab 1.1.2017	ISCED Ausbildungsfeld		Fachrichtung
51914	Abdichter:in EFZ	0732 Baugewerbe, Hoch- und Tiefbau		–
51915	Dachdecker:in EFZ			–
51916	Fassadenbauer:in EFZ			–
51917	Gerüstbauer:in EFZ			–
51918	Storenmonteur:in EFZ			–

Inkonsistenzen können dann beispielsweise entstehen, wenn eine befragte Person eine Berufslehre als Polybauer EFZ vor 2017 begonnen hat, diese Ausbildung aber aufgrund von aktualisierten Listen ab 2017 nicht mehr existiert.

Benutzung von Nomenklaturen und Klassifikationssystemen

Da Episodendaten eine bestimmte Tätigkeit oder einen Zustand beschreiben (z. B. Beschäftigung, Ausbildung, Krankheit), ist es oft notwendig, offizielle Nomenklaturen, Klassifizierungssysteme oder projektspezifische Codelisten zu verwenden, um den Status zu ermitteln und ihn in Bezug auf Spells desselben Typs zu klassifizieren. Sowohl bei *dependent* als auch bei *independent interviewing* ist die Verwendung solcher Codesysteme (z. B. ISCO88 zu ISCO08)⁹¹ gleichzeitig eine Hauptquelle für Widersprüche.

Das Potenzial für Widersprüche ist bei Codesystemen, die einen relativ niedrigen Aggregationsgrad haben, am größten. Zum Beispiel können einige amtliche Berufscodelisten mehrere Einträge für Berufe haben, die sehr ähnlich klingende Beschreibungen haben. Wenn Codes, die solchen Berufen entsprechen, aufgezeichnet und über Erhebungswellen hinweg verglichen werden, kann eine kleine Änderung in der Formulierung durch die Befragten oder eine geringfügige Verschiebung in der Interpretation durch eine:n Interviewer:in leicht zu einer fälschlichen Änderung führen, die keine echte biografische Veränderung widerspiegelt.⁹²

7.2.2 In der Datenerhebungsphase

Die kognitive Psychologie unterscheidet sieben Komponenten des Antwortprozesses, *encoding, storage, comprehension, retrieval, integration, mapping* und *editing*.⁹³ Sozialforscher:innen haben keinen Einfluss auf die Kodierung (*encoding*) und Speicherung (*storage*) von Informationen; sie können in der Datenerhebungsphase den Abrufprozess (*comprehension -> retrieval -> integration -> mapping -> editing*), eine autobiografische Frage zu beantworten, aber durch geeignete Stimuli erleichtern. Stimuli können den Gedächtnisabruf erleichtern, ohne den Befragten Antwortvorgaben aufzudrängen. Die Beantwortung einer autobiographischen Frage, also a) die Erinnerung und b) die Datierung eines autobiographischen Ereignisses, sind dabei zwei unabhängige Prozesse.⁹⁴ Neben den richtigen Stimuli hängt das Abrufen einer Information von weiteren Faktoren ab (siehe 7.6 Vermeidung). Es hat sich gezeigt, dass eine Verlängerung des Zeitintervalls zwischen einem bestimmten Ereignis und der Messung zu einer Verzerrung bei der Erinnerung an das genaue Datum dieses Ereignisses führen kann – be-

⁹¹ISCO = International Standard Classification of Occupations.

⁹²Collins, C. 1975. „Comparison of Month-to-Month Changes in Industry and Occupation Codes with Respondent's Report of Changes“. In U.S. Bureau of the Census, *CPS Job Mobility Study. Response Research Staff Report*, 75-5. Mathiowetz u. McGonagle, „Assessment“.

⁹³Tourangeau, R., L. Rips und K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press: Cambridge.

⁹⁴Sudman, Seymour, Norman M. Bradburn und Norbert Schwarz. 1996. *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass. Conway, Martin A. 1996. „Autobiographical knowledge of autobiographical memory“. In *Autobiographical knowledge and autobiographical memories*. Hrsg. D.C. Rubin. 67–93. Cambridge, MA: Cambridge University Press.

kannt als *forward telescoping*.⁹⁵ Widersprüche in den Episodendaten können demnach immer sowohl auf Messfehler in der aktuellen Panelwelle als auch auf Messfehler in früheren Wellen hindeuten.

7.3 Identifikation

Schon die Identifikation von Inkonsistenzen bei den Episoden kann schnell eine große Herausforderung im Datenmanagement werden. Die Identifikation erfolgt über den Vergleich der Informationen zu den Episoden, zum Beispiel wenn sie in mehreren Wellen für die gleichen Episoden erhoben worden sind, oder über die Prüfung von mit den Episoden verbundenen Rahmenbedingungen und Gesetzmäßigkeiten. Beispiele für letzteres sind, dass eine Professur normalerweise ein Hochschulstudium und besondere akademische Leistungen voraussetzt oder, dass für das Leisten eines Militärdienstes in der Regel eine bestimmte Dauer gesetzlich festgelegt ist. Abweichungen wären, wenn es keine (gesetzlichen) Ausnahmeregelungen gibt, unplausibel und sollten im Einzelfall geprüft werden. Da Einzelfallprüfungen schnell exorbitant umfangreich werden könnten, sollten gegebenenfalls zuerst häufige Ausnahmeregelungen (wie vielleicht eine unterschiedliche Dauer des Militärdienstes für Männer und Frauen) in die Prüfkriterien aufgenommen werden.

Bei der Identifikation von inkonsistenten Informationen, die in mehreren Wellen für die gleichen Episoden erhoben worden sind, ist zwischen den beiden oben beschriebenen Arten der Erhebung von Episodendaten zu unterscheiden: Bei einem *independent interviewing* gibt es die Möglichkeit, die bereits früher erhobenen Episoden mit den neu erhobenen Daten zu vergleichen. Dabei werden idealerweise alle Merkmale verglichen und auf Differenzen hin untersucht. Bei einem *dependent interviewing* fließen die früher erhobenen Episoden als Preload in die aktuelle Befragung ein. Dabei können Widersprüche bereits bei der Befragung selbst auftreten. Um das Interview überhaupt fortsetzen zu können, sollte den Befragten die Möglichkeit gegeben werden, die Widersprüche zu klären und Episoden zu ändern, was meist zu einer Steigerung der Datenqualität führt. In einem solchen Widerspruchsmodul können verschiedene Änderungen zur Unterstützung des nachfolgenden Plausibilisierungsprozesses entsprechend markiert werden. Trotzdem wird es auch bei einem *dependent interviewing* oft notwendig sein, abhängig vom Survey Design weitere Datenprüfungen zu etablieren, um beispielsweise Doppelnennungen einer Tätigkeit oder Gleichzeitigkeit von widersprüchlichen Tätigkeiten (z. B. eine Ausbildung an einem Gymnasium zum Abitur und gleichzeitig ein Universitätsstudium) aufzudecken. Diese Tests sind bei größeren Datenmengen praktisch nur skriptbasiert zu leisten.

Bei der Identifikation von Widersprüchen spielt die Definition der Episodenmerkmale eine große Rolle: je feinteiliger die Definition, desto mehr Widersprüche treten auf.

⁹⁵Neter, John und Joseph Waksberg. 1964. „A Study of Response Errors in Expenditures Data from Household Interviews“. *Journal of the American Statistical Association* 59 (305): 18–55. <https://doi.org/10.1080/01621459.1964.10480699>.

Beispielsweise können für die gleiche Erwerbstätigkeit zu verschiedenen Messzeitpunkten unterschiedliche Bezeichnungen wie:

- Zollbeamter / Zollbeamtin,
- Zollangestellter / Zollangestellte,
- Zollinspektor / Zollinspektorin,
- Zollassistent / Zollassistentin,
- Zollexperte / Zollexpertin,
- Zollfachmann / Zollfachfrau

genannt werden⁹⁶ und zu Widersprüchen führen. Werden dagegen aggregierte Listen (Klassifikationen) wie ISCO verwendet:

- 3351 Fachkräfte im Zolldienst und Grenzschutz,

werden alle diese Bezeichnungen in einem ISCO-Code zusammengeführt und scheinbare Widersprüche können von bedeutsamen Widersprüchen unterschieden werden.

Für die Datenaufbereitung ist es hilfreich, wenn das Survey Design den Befragten ermöglicht, schon während eines Interviews Widersprüche direkt zu klären: Liegt ein Fehler im Startdatum vor oder gibt es ein Problem mit einer bestimmten Klassifizierung oder Typisierung einer bestimmten Episode?

Beispielsweise wird in der TREE-Studie bei der Aufnahme von Ausbildungen zwischen einer 2-jährigen (EBA) und 3-4-jährigen Berufsausbildung (EFZ) unterschieden. Wenn die angegebene Dauer mit der Ausbildung nicht korrespondiert, wird folgender Screen angezeigt, um die Situation zu klären:

Q: check_educ

Ausbildung

Schule

Ausbildungsbetrieb

Lehrvertragsbetrieb

Praktikum

Erwerbstätigkeit

Andere

[-9] 2-jährige Alterslehre (EBA) mit Lehrbetrieb

J J A S O N D 2017 F M A M J J A S O N D 2018 F M A M J J A S O N D 2019 F M

Sie haben angegeben, dass es sich um eine berufliche Grundbildung mit EBA handelt. Diese dauert in der Regel 2 Jahre. Könnte es vielleicht eine andere Art von Ausbildung sein, oder wurde vielleicht die Dauer nicht korrekt erfasst?

Korrekturmöglichkeiten:

> Die Dauer kann direkt auf dem nächsten Bildschirm korrigiert werden, falls nötig.

> Falls der Typ der Ausbildung falsch ist >> ZURÜCKBLÄTTERN und am Anfang der Episode richtig aufnehmen!

Stimmt so nicht, Angaben müssen korrigiert werden

Stimmt so

Zurück Weiter

Abbildung 17: Plausibilisierung von Ausbildungskategorien, TREE-Studie 2. Kohorte (CATI-Version)

⁹⁶Angaben gemäß: Berufsdatenbank des Bundesamtes für Statistik Schweiz.

Um Widersprüche zu identifizieren, können auch Daten mit überschneidenden oder redundanten Informationen erhoben werden. Zum Beispiel können redundante Informationen, wie der Zeitpunkt des Ausbildungsbeginns, die Dauer der Ausbildung und das erwartete Jahr des Abschlusses, die Datenaufbereitung unterstützen. Mittels weiterer erhobener Merkmale kann das Datenmanagement idealerweise die fehlerhaften Angaben identifizieren und den Fall entsprechend beurteilen.

Neben direkt von Befragten erhobene Informationen, können in computergestützten Surveys auch automatisch erfasste Informationen zu den Befragungsabläufen von Wert sein. Beispielsweise können spezifische Routing-Informationen erfasst werden, um anzuzeigen, ob Befragte bestimmten Fragebogenteile vollständig ausgefüllt haben oder wo genau ein Abbruch des Interviews stattgefunden hat. Bei Mehrfacherfassung (in Schleifen) von aktuellen und retrospektiven Episoden können diese Informationen die Beurteilung der Abbrüche unterstützen, zum Beispiel in der Beurteilung, ob die Episoden vollständig oder unvollständig erfasst wurden. Unvollständig erfasste Episoden können zu einer vorgängigen oder einer nächsten Befragungswelle zu Inkonsistenzen führen.

7.4 Umgang

Im *independent interviewing* ist im Gegensatz zum *dependent interviewing* kein Abgleich mit bereits früher erfassten Episoden möglich. Die Plausibilisierung besteht primär aus einem Abgleich, indem die Episoden aus früheren Befragungen mit den Episoden der letzten Befragung verglichen und zusammengeführt werden, wenn Episoden als identisch identifiziert werden. Aber auch über das *dependent interviewing* erhobene Daten werden nicht frei von Widersprüchen sein und müssen ggf. noch aufbereitet werden. Obwohl es unbestreitbar wichtig ist, dass die Preloads für die Befragten erkennbar sind,⁹⁷ ist es das Ziel der meisten Panelerhebungen, Längsschnittdaten zu produzieren, die nicht nur eine hohe Reliabilität, sondern auch eine hohe Validität aufweisen. *Dependent interviewing* kann helfen, die Konsistenz von Längsschnittdaten über die Zeit bzw. die Erhebungswellen zu steigern. Als Nebenwirkung kann die Validität der individuellen Angabe sinken, weil manche Befragten dazu tendieren, die Preload-Episoden zu akzeptieren, auch wenn diese nicht (ganz) korrekt sind (*acquiescence bias*)⁹⁸. Um die Balance zwischen Reliabilität und Validität zu erreichen, muss jedes Projekt, das Episodendaten bearbeitet, ein Verfahren für den Umgang mit widersprüchlichen Informationen entwickeln. Dabei sind, wie bei anderen Widersprü-

⁹⁷Um den Erinnerungsprozess zu erleichtern, weisen Trahms et al. darauf hin, dass Preloads für die Befragten als Erinnerungshilfe erkennbar sein sollen. Siehe: Trahms, Annette, Britta Matthes und Michael Ruland. 2016. „Collecting Life-Course Data in a Panel Design: Why and How We Use Proactive Dependent Interviewing“. In *Methodological Issues of Longitudinal Surveys*. Hrsg. Hans-Peter Blossfeld, Jutta von Maurice, Michael Bayer und Jan Skopek. Wiesbaden: Springer Fachmedien Wiesbaden. 349–66. https://doi.org/10.1007/978-3-658-11994-2_20.

⁹⁸Eggs, Johannes und Annette Jäckle. 2015. „Dependent Interviewing and Sub-Optimal Responding“. *Survey Research Methods* 9 (February): 15-29. <https://doi.org/10.18148/SRM/2015.V9I1.5860>.

chen auch, technisch zwei Schritte zu unterscheiden: erstens die Identifikation der Widersprüche (*inspection*) und zweitens das Anpassen (*alteration*). Generell sollte versucht werden, möglichst alle Widersprüche zu identifizieren und anschließend eine Entscheidung des Umgangs getroffen werden. Dies kann insbesondere bei Episodendaten einen erheblichen Teil der Projektzeit in Anspruch nehmen.

In der Regel werden in der Datenaufbereitungsphase zunächst die Existenz und Konsistenz einzelner Variablen geprüft, anschließend die Kombination von Variablen. Auch bei Episodendaten sollten also zunächst das Vorhandensein, die Ausprägungen von Start- und Endpunkten (Format, Vergleich zum Beobachtungszeitraum) sowie die Ausprägungen der Inhalte geprüft werden; danach die Kombination der Ausprägungen innerhalb von Episoden (z. B. Start vor Ende, Ende vor Beobachtungszeitraum), die Kombination verschiedener Episoden (z. B. Zweitstudium vor Erststudium, Essen während Schlaf, Lücken) und anschließend Widersprüche, die durch Mehrfacherfassung von gleichen Episoden über verschiedene Messpunkte entstehen. Als Ergebnis der Prüfung werden so unmögliche (und damit ungültige), aber auch unrealistische (also unwahrscheinliche) Variablenwerte identifiziert.

Der Umfang, in dem Widersprüche behandelt werden müssen, wird zu einem Teil vom Erhebungsdesign und zum anderen Teil von der Offenheit der Produzent:innen abhängen, die Datennutzer:innen mit Widersprüchen in veröffentlichten Scientific-Use-Files selbst umgehen zu lassen. Wenn Formen des *dependent interviewing* verwendet werden, ist eine entsprechende Bearbeitung von widersprüchlichen Informationen bereits beim Design der Erhebungen wichtig. Da Preload-Episoden als Stimuli oder Vergleich verwendet werden, ist es sinnvoll, bestehende Widersprüche vor jeder Folgewelle zu klären. Bei der Datenpublikation selbst besteht etwas mehr Freiheit, die Bearbeitung auf ein Minimum zu beschränken und zum Beispiel Flag-Variablen in ein Scientific-Use-File aufzunehmen. Abhängig von der Größe der Stichprobe, der Anzahl der Messungen und der Art und Spezifität der gesammelten Daten können solche Dateien leicht unhandlich werden. Dies kann die Fähigkeiten durchschnittlicher Sekundärnutzer:innen zum Datenmanagement übersteigen. Auch Zusatzinformationen, wie Paradata oder Textangaben, stehen den Datennutzer:innen nicht zur Verfügung. Daher ist ein gewisser Aufwand bei der Datenedition kaum zu vermeiden. Da die Datenaufbereitung an sich eine große Fehlerquelle sein kann, besteht die Kunst bei jeder Längsschnittstudie darin, ein effektives Gleichgewicht der potenziellen Editionsfehler und dem Umgang mit Widersprüchen aus der Befragung zu finden.

7.5 Dokumentation

Prinzipiell sollten alle Datenerhebungs- und Datenverarbeitungsprozesse innerhalb einer Studie so weit wie möglich standardisiert sein, da nur so gültige und zuverlässige Vergleiche vorgenommen werden können. Diese Standardisierung ist insbesondere über verschiedene Kohorten, Erhebungswellen und Mitarbeiter:innen zu verfolgen. Episodendaten werden häufig in Paneldesigns erhoben. Eine Standardisierung ist dabei

aufgrund der langen Zeithorizonte, mit wechselnden Mitarbeiter:innen, sich ändernden thematischen Schwerpunkten und häufig auch sich ändernden Methoden eine besondere Herausforderung.

Darüber hinaus sollten alle am Datensatz getätigten Änderungen, im besten Fall auch alle identifizierten, aber nicht angepassten Widersprüche übersichtlich, begründet und verständlich nachvollziehbar sein. In der Regel macht es Sinn, wichtige Punkte mit dem Datensatz extern zu veröffentlichen, alle anderen intern zu archivieren und auf Nachfrage zugänglich zu machen. Interne Dokumentation beinhaltet vor allem das „Wer hat was, wozu geändert und warum?“. Dabei ist eine skriptbasierte Bearbeitung der Daten sinnvoll, das „Wer“ und „Warum“ kann dort eingearbeitet werden. Darüber hinaus kann aber bereits intern auch eine Ergänzung durch Entscheidungsprotokolle, Working Paper, Codebooks oder andere Formate sinnvoll sein und der Übersichtlichkeit dienen. Eine externe Dokumentation kann umfassend oder zusammenfassend (mit einem Fokus auf die wichtigsten Punkte) erfolgen.

Da sich bei der Plausibilisierung von Episodendaten viele Kleinigkeiten anhäufen und Datennutzer:innen schnell überfordert sein können, ist es mitunter sinnvoll, einen komplett plausibilisierten Datensatz und einen Originaldatensatz zu veröffentlichen. Im besten Fall werden die Prinzipien der gemachten Änderungen zusätzlich schriftlich veröffentlicht. Da die Dokumentation zu episodischen Widersprüchen schnell ausufern kann, ist eine genaue Abwägung der veröffentlichten Inhalte besonders wichtig.

7.6 Vermeidung

Bereits in der Studienplanungsphase können Schritte zur Vermeidung von Widersprüchen vorgenommen werden. Ein wichtiger Teil des Befragungsdesigns bei der Erhebung von biografischen Episoden ist sicherzustellen, dass keine Lücken oder Missings die Analysen behindern. Die Überprüfung auf solche Lücken oder fehlende Informationen zu einer Episode (vgl. Abbildung 18) ist ein zentrales Designmerkmal, das dazu beiträgt, Widersprüche zu entschärfen, die auf eine unvollständige Erfassung von Episoden zurückzuführen sind. Es muss definiert werden, wie explizit die Umfrage nach Informationen zu den Episoden fragt und was eine Lücke in diesen Informationen darstellt.

Q: xholecheck

Ausbildung

Schule

Ausbildungsbetrieb

Lehrvertragsbetrieb

Praktikum

Erwerbstätigkeit

Andere

[-1] 3-4-jährige Berufslehre mit Lehrbetrieb

[-1] Gewerblich-Industrielle Berufsfachschule

[-1] Ausbildungsbetrieb A

[-1] Vertragsbetrieb A

A S O N D 2017 F M A M J J A S O N D 2018 F M A

Zwischen dem 01.11.2017 und dem 27.02.2018 haben wir eine Phase, wo Sie mir keine Tätigkeit angegeben haben. Was haben Sie in dieser Zeit gemacht?

INT: VORLESEN!

Eine Tätigkeit ausgeübt, über die wir bereits gesprochen haben (Daten anpassen)

oder eine Tätigkeit ausgeübt, über die wir noch nicht gesprochen haben (neue Episode erfassen)

Weiss nicht

Angabe nicht möglich

Angabe explizit verweigert

Zurück Weiter

Abbildung 18: Lückencheck der TREE-Studie 2. Kohorte (CATI-Version)

Dependent interviewing ist in fast allen Fällen vorzuziehen, als den Befragten überhaupt keine Stimuli anzubieten.⁹⁹ Es wird angenommen, dass *dependent interviewing* nützlich ist, um einige besondere Messfehler zu umgehen. *Dependent interviewing* kann hilfreich sein, um falsche Veränderungsraten (*rates of change*) zu unterdrücken, die durch leichte Variationen bei offenen Antworten entstehen. Darüber hinaus zeigen Lynn et al. in einer Studie mit experimentellem Design, dass *dependent interviewing* bei Problemen wie Underreporting auf Fragen zu Einkommensquellen helfen kann.¹⁰⁰ Schließlich kann es helfen, die Fehldatierung von Episodengrenzen zu reduzieren und damit den *seam effect* bei wiederholten Messungen zu verringern (siehe 7.1.3 „Arten von widersprüchlichen und fehlenden Angaben in Episodendaten“).

Ein längsschnittliches Fragebogendesign,¹⁰¹ *proactive* oder *reactive dependent interviewing*, die Wahl der richtigen Stimuli bzw. Matching-Kriterien und der sensible Umgang mit Unplausibilitäten sind wirksame Instrumente zur Reduzierung von Messfehlern und zur Reduktion von Widersprüchen in aufeinanderfolgenden Panelwellen.

⁹⁹Mathiowetz u. McGonagle, „Assessment“; Lynn et al., „Effects of Dependent Interviewing“.

¹⁰⁰Lynn et al., „Effects of Dependent Interviewing“.

¹⁰¹Eine andere wichtige Voraussetzung dafür, im Forschungsdatenmanagement herausstechende Fälle jederzeit identifizieren zu können, ist die Anlage von Identifier (z. B. eindeutig identifizierende Personen-IDs) (siehe Kapitel 6 „Inkonsistenzen bei der Verknüpfung mehrerer Datensätze/-quellen“, Abschnitt 6.1 „Beschreibung“).

7.6.1 Dependent Interview Design

Der Umgang mit Unplausibilitäten beginnt bereits mit dem Survey Design. Entscheidend ist, dass verschiedene Arten von Fragen unterschiedliche Lösungen erfordern können. Wenn das Hauptanliegen eine Reduzierung der falschen Übergangsraten oder der Fehldatierung von historischen Ereignissen ist, kann *proactive dependent interviewing* das richtige Werkzeug sein. Allerdings kann die proaktive Befragung auch zu einem gewissen *acquiescence bias* – oder *cognitive satisficing* – führen, bei dem die Befragten eher den gegebenen Stimuli zustimmen, als Energie für die Erinnerung an ergänzende oder widersprechende Informationen aufzuwenden.¹⁰² Es ist daher möglich, dass *proactive dependent interviewing* zu einer Verzerrung oder Untererfassung von Übergängen zwischen Panelwellen führen kann. In Fällen, in denen ein solches *cognitive satisficing* als nicht hinzunehmendes Risiko angesehen wird, kann im Fragebogendesign stattdessen eine Form des *reactive dependent interviewing* verwendet werden. Zum Beispiel hat die *Canadian Longitudinal Survey of Labour and Income Dynamics* (SLID) ein *reactive dependent design* in Bezug auf die aktuelle Lohnhöhe der Befragten implementiert. Es wurde befürchtet, dass ein *proactive dependent design* die realen Variationen der Lohnrate unterschätzen würde. Das *reactive dependent design* der SLID veranlasste eine Überprüfung mit den Befragten, wenn die neu gemeldete Lohnhöhe eine Diskrepanz zu der zuvor erfassten Lohnhöhe von mehr als zehn Prozent aufwies. Nur in diesen Fällen wurden die Befragten an die zuvor erfasste Lohnhöhe erinnert und gebeten, die Diskrepanz zu überprüfen und zu begründen.¹⁰³

Stimuli sollten den Gedächtnisabruf erleichtern, ohne den Befragten Antwortvorgaben aufzudrängen, das heißt das Erhebungsdesign sollte darauf abzielen, die Gedächtnisleistung der Befragten zu entlasten und ihnen gleichzeitig angemessene Möglichkeiten bieten, frei zu antworten. Daneben hängt das Abrufen von Erinnerungen davon ab, den Befragten für die Bedeutung eines Ereignisses entsprechend ausreichend Antwortzeiten zu gewähren und im Survey Design einzuplanen. Die berichteten Zeitintervalle sollten sorgfältig gewählt werden (z.B. kein Rückblick über fünf Jahre in Wochenintervallen, siehe *forward telescoping* in 7.2 zu Ursachen). Auch sollte durch die Studienvorbereitung und Erhebungsdurchführung allgemein für eine hohe Kooperationsbereitschaft seitens der Befragten gesorgt werden.

7.6.2 Widerspruchsmodul

Außerdem sollte die Umfrage so konzipiert sein, dass sie mit Widersprüchen umgehen kann, die während der Befragung auftreten können. Widersprechen Befragte in

¹⁰²Eggs u. Jäckle, „Dependent Interviewing“.

¹⁰³Hale, Alison und Sylvie Michaud. 1995. „Dependent Interviewing: Impact on Recall and on Labour Market Transitions“. In *Income Research Paper Series, Catalogue No. 95-06, Statistics Canada*. <https://www150.statcan.gc.ca/n1/en/pub/75f0002m/75f0002m1995006-eng.pdf?st=bdcfbhK0> (abgerufen am 30.03.2023).

einem *proactiv dependent interviewing* den vorgeladenen Episoden, muss ein Widerspruchsmodule in der Befragungssoftware eingebaut werden, damit Widersprüche behandelt und das Interview danach weitergeführt werden kann. Bei einer computergestützten Umfrage sollten festgelegte Prozeduren zur Behandlung von Widersprüchen einprogrammiert werden. Die Entscheidung, was mit widersprüchlichen Angaben über die Wellen hinweg passiert, erfolgt im Plausibilisierungsprozess in der Datenaufbereitungsphase.

7.6.3 Gestaltung der Stimuli

In Abschnitt 7.3 Identifikation wurde erwähnt, dass die Erfassung redundanter Informationen den Umgang mit Widersprüchen bei der Datenaufbereitung unterstützen kann, um den Preis, dass zusätzliches Potenzial für widersprüchliche Aussagen eröffnet wird. Generell ist es ratsam, Redundanzen bei Informationen, die für die Definition von Episoden zentral sind, zu reduzieren. Dies kann auf mehrere Arten geschehen. Erstens müssen bei *proactive dependent interviewing* die Preload-Episoden als Stimulus nicht vollständig beschrieben werden. Anstatt beispielsweise das genaue Anfangsdatum einer zuvor erfassten Beschäftigung zum Beispiel als „Datenmanager:in“ explizit anzugeben, kann ein vager Stimulus gegeben werden, zum Beispiel „während des letzten Interviews erwähnten Sie, dass Sie als Datenmanager:in für die Firma X arbeiten“. Zweitens ist es wichtig, sicherzustellen, dass Redundanzen in Codelisten nicht zu falschen Übergängen führen. Beispielsweise ist es nicht ungewöhnlich, dass Listen mit Stellenbeschreibungen verschiedene Einträge enthalten, die auf semantischer Ebene relativ ähnlich sind, aber mit unterschiedlichen Codes versehen sind.

Dieses Problem kann sich noch verschärfen, wenn mehrsprachige Codelisten in Erhebungen verwendet werden, die in mehr als einer Sprache durchgeführt werden. Beispielsweise führt das Schweizer Bundesamt für Statistik ein Lexikon der Berufsbilder, in dem es zwei Einträge für Schornsteinfeger:innen gibt:

Tabelle 7: Mehrsprachige Codelisten

Code	Gruppe	Bezeichnung deutsch	Bez. französisch	Bez. italienisch
41104003	62104	Schornsteinfeger/ Schornsteinfegerin	Ramoneur/ Ramoneuse	Spazzacamino/ Spazzacamino
41104001	62104	Kaminfeger/ Kaminfegerin	Ramoneur/ Ramoneuse	Spazzacamino/ Spazzacamino

Die beiden Einträge sind völlig äquivalent, der einzige Unterschied ist, dass „Kaminfeger“ im Schweizerdeutsch umgangssprachlicher ist, während „Schornsteinfeger“ eher hochdeutsch ist. Bei der Gestaltung von Fragebögen und Umfragen kann es wichtig sein, sprachliche Varianten anzubieten, um entsprechende Vorschläge zur Autovervollständigung zu erzeugen. Dann können die Befragten – unabhängig von Sprache

oder Dialekt – ihre bevorzugten Ausdrücke leichter finden. Die französischen und italienischen Übersetzungen sind in diesem Beispiel vollständig redundant und spiegeln keine Unterschiede im Dialekt oder regionalen Ausdruck wider. Auch wenn sie für den Erhebungsprozess hilfreich sein können, ist es bei der Definition von Schreibweisen wichtig, solche Redundanzen zu umgehen, indem Einträge auf einer höheren Ebene aggregiert und gruppiert werden („Gruppe“ in diesem Beispiel). Dies hilft, Widersprüche und falsche Übergänge einzuschränken, die durch leichte Abweichungen in der Wortwahl der Befragten oder der Interpretation ihrer Angaben durch Interviewer:innen verursacht werden.

7.6.4 Grobe Matching-Kriterien

Eine Erhebung könnte zum Beispiel lediglich darauf abzielen, zu messen, ob und wann ein:e Befragte:r beschäftigt ist. Beschäftigungsepisoden dieser Art sind unempfindlich gegenüber Veränderungen in spezifischen Merkmalen des Beschäftigungsstatus, zum Beispiel der Stellenbeschreibung oder der Arbeitgeber:innen. So definierte Episoden werden aufgrund der begrenzten Informationen weniger zu Widersprüchen zwischen Erhebungswellen führen und eine Erhöhung der Spezifität der Spell-Definitionen führt potenziell zu mehr Widersprüchen zwischen den Messungen.

Wird das Matching der Episoden auf der ISCO-Klassifizierung durchgeführt (siehe Beispiel „Fachkräfte im Zolldienst und Grenzschutz“ in Abschnitt 7.3 „Identifikation“) statt auf dem viel differenzierteren Thesaurus der Berufsbezeichnungen, können Widersprüche in einem erheblichen Umfang vermieden werden.

7.6.5 Schulung der Interviewer:innen

Bei CATI/CAPI-Befragungen ist es außerdem entscheidend, die Interviewer:innen sorgfältig in der Episodenerfassung und im Umgang mit Widersprüchen während des Interviews zu schulen. Die Datenerhebung ist eine Interaktion zwischen Interviewer:innen und Befragten. Auch wenn diese Interaktion technisch durch ein solides Umfragedesign gesteuert werden kann, liegt es in erster Linie an den Interviewer:innen, dieses Design auf effektive Weise umzusetzen.¹⁰⁴

Bei allen Vermeidungsmaßnahmen, der Identifikation und dem Umgang in Bezug auf Episodendaten sollten tatsächlich vorhandene Widersprüche und Lücken in den Biografien der Befragten mit Offenheit und Unvoreingenommenheit gesehen werden. Auch wenn diese Analysen behindern können, sind sie unter Umständen gerade die interessantesten Daten.

¹⁰⁴Die Möglichkeit besteht, Befragte mit eher komplizierten biografischen Verläufen durch erfahrene Interviewer:innen befragen zu lassen, um den Interviewprozess zu erleichtern und Widersprüche zu vermeiden. Allerdings muss hier die Reduktion von Widersprüchen gegen die selektive Steigerung von Interview-Effekten abgewogen werden.

8 Resümee und Ausblick

Insa Bechert, Kristina Krell, Percy Scheller

Gute Datenqualität ist die Voraussetzung für gute Forschung.¹⁰⁵ Für Datenaufbereitende sind die Qualität der Daten und der Dokumentation die Kriterien, an denen sich die Qualität ihrer Arbeit messen lässt. Das Erreichen und Erhalten guter Datenqualität anhand vielschichtiger Kontrollen ist daher ein grundlegender Aspekt, der sich wie ein roter Faden durch alle Phasen des Forschungsdatenzyklus zieht. Diese Handreichung befasst sich ausgiebig mit einem essenziellen Teilbereich der Datenqualitätskontrolle: der Plausibilisierung.

Die einzelnen Kapitel dieser Handreichung zeigen deutlich wie wichtig Plausibilitätskontrollen für die Datenqualität sind. Sie zeigen allerdings auch wie vielschichtig Inkonsistenzen und ihre Ursachen sind und wie aufwendig der Prozess sein kann, ihrem Ursprung auf die Spur zu kommen. Als Handlungsempfehlungen lassen sich hier deshalb nur grobe Eckpunkte festhalten die auf den vier wesentlichen Maßnahmen **Identifikation**, **Umgang**, **Dokumentation** und **Vermeidung** beruhen:

- Für die **Identifikation** unplausibler Werte empfiehlt es sich automatisierte Prüfroutinen zu entwickeln.
- Beim **Umgang** mit unplausiblen Werten ist es wichtig sich bewusst zu machen, dass vollkommen plausible Daten zwar gut aussehen, aber leider eher selten der Realität entsprechen. Es liegt in der Natur der Menschen, hin und wieder unplausible (jedoch mitunter durchaus valide) Angaben zu machen. Ein zu starker Eingriff in die Daten ist daher nicht empfehlenswert. Welche Maßnahmen konkret angewendet werden ist jeweils studien- und fallabhängig, wobei Einzelfallprüfungen häufig zu zeitintensiv sind, um zweckmäßig zu sein.
- Jeder Eingriff in die Daten birgt das Risiko neue Fehler zu produzieren. Daher sollten Datenmanipulationen immer **dokumentiert** werden, damit Nutzende und andere Datenaufbereitende sie nachvollziehen können. Auch Inkonsistenzen, die nicht korrigiert werden können, sollten immer **dokumentiert** werden.
- Ob über die Dokumentation hinaus Flagvariablen integriert werden sollten, ist von der Anzahl der Fälle und der betroffenen Variablen sowie dem durch die Flagvariablen entstehenden Handlungsspielraum für die Nutzenden abhängig.

¹⁰⁵Katsanidou, Alexia, Laurence Horton und Uwe Jensen. 2016. „Data Policies, Data Management and the Quality of Academic Writing“. *International Studies Perspectives*. Online first. 1-18. <https://doi.org/10.1093/isp/ekv014>.

- Technische Fehler oder uneindeutige Fragestellungen in Fragebögen, die zu unplausiblen Werten in den Daten führen, können durch sorgfältige Kontrollen in der Studienplanungsphase teilweise **vermieden** werden.
- Sind Datenkontrollen in Hinblick auf Plausibilität, z. B. aufgrund fehlender Ressourcen nicht möglich, sollte auch das im Methodenbericht dokumentiert werden. Nutzenden wird dadurch bewusst, dass sie eigene Kontrollen durchführen und ggf. Maßnahmen ergreifen müssen.

Durch die Beschreibung der Ursachen von verschiedenen Arten von Unplausibilitäten, Techniken zur Identifikation, Vorschläge zum Umgang und zur Dokumentation sowie Maßnahmen zur Vermeidung, hat sich diese Handreichung nach unserem Kenntnisstand erstmals systematisch und ausführlich mit der Thematik befasst. Da zu diesem praxisnahen Thema kaum Fachliteratur existiert, bzw. unplausible Werte üblicherweise lediglich am Rande besprochen werden, gibt es auch keine allgemein gültigen Standards, wie am besten damit zu verfahren ist. Um diese Lücke zu adressieren, sind in dieser Handreichung eine Reihe von Fallbeispielen (Best Practices) aus verschiedenen Umfrageprogrammen zusammengetragen worden. Über die entscheidende Frage, wie mit unplausiblen Werten umgegangen werden soll, müssen Datenaufbereitende jedoch im Einzelfall entscheiden. Sie müssen individuell abwägen, wie viel analyserelevante Information die betroffenen Daten tragen und ob Abweichungen gelöscht oder beibehalten und dokumentiert werden, denn bisweilen lassen sich auch aus unvorhergesehenen Informationen, die Befragte liefern, noch Erkenntnisse destillieren.

Ein Ausblick für weitere praxisnahe Handreichungen dieser Art ergibt sich durch die Bereiche, die in dieser Publikation nicht beleuchtet werden konnten. Zum einen befasst sie sich ausschließlich mit nationalen Daten. Internationale Daten erfordern und ermöglichen Plausibilisierungsschritte über den Ländervergleich. Auf der anderen Seite gestaltet sich die Nachvollziehbarkeit von Unplausibilitäten über Ländergrenzen und verschiedene Kulturen hinweg häufig noch schwieriger.¹⁰⁶ Auch die Verknüpfung von Individualdaten mit Daten auf Länderebene oder Social Media Daten erfordern und ermöglichen Plausibilitätskontrollen, deren nähere Betrachtung lohnend erscheint.

Weitere Anschlussmöglichkeiten bietet das weite Spektrum der Datenqualitätskontrollen, von denen die Plausibilisierung ja nur einen, wenn auch wesentlichen, Baustein darstellt. Eine umfassende Darstellung möglicher Fehlerquellen in Umfragedaten bietet das *Total Survey Error* Schema¹⁰⁷. Mit Blick auf dieses Schema bietet es sich an, die entsprechenden Strategien (Identifikation, Umgang, Dokumentation, Vermeidung) auf

¹⁰⁶So lässt sich beispielsweise eine gedrehte Skala bei der persönlichen Einschätzung der sozialen Schicht durch Korrelationen mit dem Familieneinkommen leicht identifizieren. Der Zusammenhang solle in jedem Land stark sein. Wenn in einem Land das Vorzeichen dieser Korrelation von den anderen abweicht, ist die Wahrscheinlichkeit für eine gedrehte Skala hoch. Andererseits machen es die verschiedenen Sprachen schwer den genauen Wortlaut und die Skalierung der Fragestellungen zu überprüfen und Unterschiede im Steuerrecht erschweren es Angaben der Befragten zu ihrem Einkommen zu interpretieren.

¹⁰⁷Siehe dazu auch Kapitel 2 „Inhaltliche Einbettung in den Forschungsprozess“.

weitere mögliche Fehlerquellen im Rahmen von sozialwissenschaftlichen Umfragen zu betrachten. Für Nutzende und Datenaufbereitende interessant wären z. B. Strategien, Interviewfälschungen zu begegnen oder verzerrte Datenstichproben zu identifizieren.

Es ist eine erfreuliche Entwicklung, dass die Forschungsgemeinschaft und in der Konsequenz auch Drittmittelgeber zunehmend Wert auf die Möglichkeit der Nachnutzbarkeit von Daten legen, wofür sorgfältiger Datenaufbereitung und Datenqualitätskontrollen eine Voraussetzung sind.¹⁰⁸ Forschungsdatenmanagementpläne, die Strategien und Verfahren zum Schutz, zur Validierung und Dokumentation von Forschungsdaten und somit auch zu Datenqualitätskontrollen umfassen, sind mittlerweile ein fester Bestandteil erfolgreicher Drittmittelanträge.¹⁰⁹ Das Wissen um die verschiedenen Möglichkeiten von Plausibilitätskontrollen und mögliche Fallstricke bei der Behebung von Fehlern ist die Voraussetzung dafür, ausreichend Ressourcen für diese Arbeitsphasen in einem Projekt einzuplanen. Diese Handreichung stellt dieses Wissen über verschiedene Anwendungsfälle hinweg bereit.

¹⁰⁸IFDO (International Federation of Data Organization). <http://ifdo.org/data-archiving-information/#Policies> (abgerufen am 30.03.2023).

¹⁰⁹Rat für Sozial- und Wirtschaftsdaten. 2018. Forschungsdatenmanagement in den Sozial-, Verhaltens- und Wirtschaftswissenschaften: Orientierungshilfen für die Beantragung und Begutachtung datengenerierender und datennutzender Forschungsprojekte. *RatSWD Output 3 (5)*. <https://doi.org/10.17620/02671.7>.