

Daten- und
Methodenbericht
April 2023

Elke Middendorff | Marten Wallis

10.-21. Sozialerhebung 1982-2016

Daten- und Methodenbericht zum gepoolten Datensatz
der zwölf Studierendenbefragungen

Dieses Werk steht unter der Creative Commons Namensnennung – Nicht kommerziell – Weitergabe unter gleichen Bedingungen 3.0 Deutschland Lizenz (CC-BY-NC-SA)

<https://creativecommons.org/licenses/by-nc-sa/3.0/de/>



Projektleitung

Dr. Elke Middendorff
Telefon +49 (0)511 450670-432
E-Mail: middendorff@dzhw.eu

Marten Wallis
Telefon +49 (0)511 450670-434
E-Mail: wallis@dzhw.eu

Projektmitarbeiterin

Cagla Belgin Varol

Impressum

Herausgeber

Deutsches Zentrum für Hochschul- und
Wissenschaftsforschung GmbH (DZHW)
Lange Laube 12 | 30159 Hannover | www.dzhw.eu
Postfach 2920 | 30029 Hannover
Tel.: +49 511 450670-0 | Fax: +49 511 450670-960

Geschäftsführung:

Prof. Dr. Monika Jungbauer-Gans

Vorsitzender des Aufsichtsrats:

Ministerialdirigent Peter Greisler

Registergericht:

Amtsgericht Hannover | B 210251
Umsatzsteuer-Identifikationsnummer:
DE291239300

April 2023

Inhaltsverzeichnis

Inhaltsverzeichnis	I
Tabellen- und Abbildungsverzeichnis	I
1 Einleitung	2
2 Datennutzungshinweise	5
3 Harmonisierung	7
4 Schema der Variablennamen	10
4.1 Variablennamenschemata der einzubeziehenden Datensätze.....	11
4.2 Variablennamenschemata des gepoolten Datensatzes	13
4.3 Das Suffix.....	14
5 Variablen im Datensatz	18
5.1 Sortierung der Variablen im Datensatz.....	18
5.2 Übersicht zur sozialerhebungsbezogenen Variablen-Präsenz	18
5.3 Übersicht zur themenbezogenen Variablen-Präsenz.....	20
6 Literatur	23

Abbildungsverzeichnis

Abb. 1: Zeitreihenpotential: Anteil an Variablen im Datensatz für 1 bis 12 Erhebungszeitpunkte	20
Abb. 2: Themenpräsenz im kumulierten Datensatz: Anzahl Variablen je Themenfeld	21
Abb. 3: Themenspezifisches Zeitreihenpotential des Datensatzes.....	22

Tabellenverzeichnis

Tab. 1: Standards und Zugangswege der SUFs zur 10. – 21. Sozialerhebung.....	11
Tab. 2: Variablennamenschemata der Einzel-SUFs zur 10. – 21. Sozialerhebung	12
Tab. 3: Teilelemente und Zusammensetzung des Variablenstammes	13
Tab. 4: Themengebiete in den Variablennamen des gepoolten Datensatzes – 10.-21. Sozialerhebung (1982-2016).....	14
Tab. 5: Zusammenfassende Übersicht der verwendeten Suffix Kürzel	15
Tab. 6: Systematik fehlender Werte in den Quell-Datensätzen 10. - 21. Sozialerhebung.....	16
Tab. 7: Systematik fehlender Werte im gepoolten Datensatz der 10.-21. Sozialerhebung.....	17
Tab. 8: Übersicht zur sozialerhebungsbezogenen Variablen-Präsenz	19

1 Einleitung

Die Sozialerhebung des Deutschen Studentenwerks (DSW) ist eine seit 1951 bestehende Untersuchungsreihe zur wirtschaftlichen und sozialen Lage der Studierenden in Deutschland.¹ Sie wird seit 1982² (10. Sozialerhebung) im Auftrag des bzw. seit der 21. Sozialerhebung in Kooperation mit dem Deutschen Studentenwerk (DSW) durch das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung GmbH (DZHW)³ durchgeführt. Das Bundesministerium für Bildung und Forschung (BMBF) fördert die Studie seit der 6. Sozialerhebung (1967/1968). Die Sozialerhebung dient – in Ergänzung zur amtlichen Hochschulstatistik – unter anderem dem nationalen und internationalen Bildungsmonitoring. Darüber hinaus liefert sie wichtiges Steuerungswissen für hochschul- und sozialpolitische Fragen sowie belastbare und umfassende Daten für die Forschung.

Im Rahmen der Tätigkeit des vom BMBF geförderten Forschungsdatenzentrums für Hochschul- und Wissenschaftsforschung am DZHW (FDZ-DZHW) werden die Daten der Erhebungen dieser Reihe, die in der Verantwortung der HIS bzw. DZHW GmbH lagen, nachträglich zum Zweck der Nachnutzung aufbereitet und dokumentiert.⁴ Bisher stehen die 10. bis 21. Sozialerhebung als einzelne Scientific Use Files (SUF) für die wissenschaftliche Sekundärnutzung zur Verfügung. Neben dem Datensatz der einzelnen Erhebung wird jeweils auch Dokumentationsmaterial zum Datensatz und zur Durchführung der Studien bereitgestellt.⁵

¹ Die Berichte zur Sozialerhebung stehen auf der Website des Nachfolgeprojektes „Die Studierendenbefragung in Deutschland“, mit dem die drei großen Studierendenbefragungen (Sozialerhebung, Studentendensurvey Konstanz, beeinträchtigt studieren) zusammengeführt wurden, zur Verfügung (https://www.dzhw.eu/forschung/projekt?pr_id=650).

² Die 1. (1951) und 2. Sozialerhebung (1953) wurden vom Studentenwerk Frankfurt am Main im Auftrag des Verbands Deutscher Studentenwerke durchgeführt. Das Studentenwerk Frankfurt am Main führte auch die 3. (1956) bis 9. Sozialerhebung (1979) durch, die vom Deutschen Studentenwerk (DSW) beauftragt wurden. Einen detaillierten Überblick über Akteure, Methoden, Themen und projektbezogene Publikationen der Untersuchungsreihe von ihren Anfängen bis zur 21. Sozialerhebung bietet ein Working Paper von Middendorff (2022).

³ Das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung (DZHW, <http://www.dzhw.eu>) entstand im August 2013 durch eine Ausgründung aus der HIS Hochschul-Informationssystem GmbH. Im nachfolgenden Text wird stets der Begriff DZHW verwendet, auch wenn die Studie vor der Ausgründung durchgeführt wurde. Auf allen Originaldokumenten der 10. bis 19. Sozialerhebung (Fragebogen, Flyer etc.) sowie in den dazugehörigen Berichten ist entsprechend die HIS GmbH (HIS) bzw. für die 20. Sozialerhebung das HIS Institut für Hochschulforschung (HIS-HF) als Akteur gekennzeichnet.

⁴ Da zu den Erhebungszeitpunkten der Daten keine Datennachnutzung vorgesehen war, sind einige Informationen zu den Erhebungen nicht mit dem Fokus einer späteren Datennachnutzung dokumentiert worden und deshalb teilweise nicht mehr rekonstruierbar. Dies ist an entsprechenden Stellen im Text angemerkt.

⁵ Informationen zu verfügbaren Datensätzen und Dokumentationen können im Metadatensuchsystem des FDZ-DZHW (<https://metadata.fdz.dzhw.eu>) recherchiert und heruntergeladen werden.

Die zwölf Einzeldatensätze der 10. bis 21. Sozialerhebung wurden in einen gemeinsamen Datensatz gepoolt und stehen jetzt als Scientific Use File (ssypool-SUF) sowie als Campus Use File (ssypool-CUF) zum Download bereit. 2021 und 2022 waren bereits ein gepoolte Datensätze veröffentlicht worden, der die vergleichbaren bzw. harmonisierten Daten der 17.-21. bzw. der 13.-21. Sozialerhebung umfassten (doi: 10.21249/DZHW:ssypool:1.0.1 bzw. 10.21249/DZHW:ssypool:2.0.1).⁶ Diese Datensätze stehen weiterhin zur Verfügung.

Der Gesamtdatensatz der 10.-21. Sozialerhebung umfasst zum Teil auch Daten der 17. bis 21. Sozialerhebung, die im gepoolten Datensatz der 17.-21. Sozialerhebung nicht enthalten sind. Das betrifft Variablen, die (erst) durch die Erweiterung um die 10. bis 16. Sozialerhebung das Kriterium erfüllen, dass eine Variable mindestens dreimal vergleichbar bzw. harmonisierbar erhoben worden sein muss.

Die Daten der Studierenden mit ausländischer Staatsangehörigkeit (Bildungsausländer*innen) sind auch im gepoolten Datensatz der 10.-21. Sozialerhebung nicht enthalten.⁷

Die Dokumentationsmaterialien zum Datenpaket bestehen aus dem Datensatzreport, den Fragebogen und Variablen-Fragebogen der zwölf einbezogenen Sozialerhebungen sowie der Dokumentation zur Variablen-Harmonisierung inklusive einer Übersicht der enthaltenen Variablen. im Metadaten-system des FDZ-DZHW (<https://metadata.fdz.dzhw.eu>) sind alle Materialien zur Ansicht und zum Download frei zugänglich.

Die zentralen Informationen zur Nutzung des Datensatzes folgen im 2. Kapitel. In Kapitel 3 werden die bei der Harmonisierung der Variablen angewandten Prinzipien beschrieben; Kapitel 4 informiert über das Konzept der Variablennamen und Kapitel 5 enthält Informationen über die Reihenfolge und die sozialerhebungsbezogene Präsenz der Variablen im Datensatz. Im Unterschied zu den Scientific Use Files der Einzelstudien enthält der gepoolte Datensatz, der vor allem der sozialerhebungsübergreifenden Analyse dient, keine Gewichte.

Eine (weitere) Anonymisierung des gepoolten Datensatzes war nicht notwendig, weil (1) aufgrund des Querschnittsdesigns der Untersuchungsreihe mit der Datenzusammenfügung keine zusätzlichen Informationen je Fall generiert, sondern lediglich Variablen um Fälle ergänzt werden (Datensatz im long format) und (2) seine Datengrundlage die bereits faktisch anonymisierten SUF der einbezogenen Sozialerhebungen sind. Die Beschreibung der je Sozialerhebung vorgenommenen Anonymisierung sind den Daten- und Methodenberichten zu den Einzel-SUF zu entnehmen. In einigen wenigen Fällen werden anonymisierte Variablen einzelner SUF für die sozialerhebungsübergreifende Harmonisierung verwendet. Die Erstellung der Einzel-SUF erfolgte zu verschiedenen Zeitpunkten nach verschiedenen Standards (s. Tabelle 1), was auch zur Folge hatte, dass abweichende Anonymisierungskonzepte angewandt worden waren. Ein Beispiel hierfür ist das Alter der Befragten. Für die Einzel-

⁶ Berücksichtigt werden hier jeweils nur die Stichproben deutscher und bildungsinländischer Studierender. Die im Rahmen der fünf Sozialerhebungen ebenfalls erhobenen Daten der Bildungsausländer*innen (17.-21. Sozialerhebung) sind im Datenpaket nicht enthalten. Diese Daten wurden mit einem zielgruppenspezifischen Fragebogen zwei- (17.-19. Sozialerhebung) bzw. mehrsprachig (20. und 21. Sozialerhebung) generiert und vom Primärforschungsprojekt getrennt abgelegt und analysiert (vgl. Middendorff 2022, S. 9 ff.).

⁷ Eine Unschärfe besteht hierbei in Bezug auf die Daten der 13. Sozialerhebung. Aus den Unterlagen des Primärforschungsprojektes ist nicht ersichtlich, ob auch (bildungs)ausländische Studierende in die Erhebung einbezogen worden waren. Aufgrund fehlender Variablen, beispielsweise Staatsangehörigkeit oder spezifische Fragen für ausländische Studierende, kann auch anhand der Daten diese Frage nicht aufgeklärt werden. Die 10. bis 12. Sozialerhebung war auf Studierende mit deutscher Staatsangehörigkeit in (den elf alten Ländern) der Bundesrepublik Deutschland beschränkt. Bildungsinländer*innen sind nicht identifizierbar.

SUF waren unterschiedliche Altersgruppen gebildet worden. Für die Harmonisierung musste zu diesem Zweck auf die anonymisierte, nicht aggregierte Altersvariable zurückgegriffen werden, um eine einheitliche Gruppierung umzusetzen. Welche Variablen darüber hinaus mittels Rückgriff auf Variablen, die im Rahmen der Einzel-SUF nicht zum Download zur Verfügung standen, harmonisiert wurden, kann anhand der Dokumentation zur Variablen-Harmonisierung nachvollzogen werden.

2 Datennutzungshinweise

[Voraussetzungen der Datennutzung] Die Daten der gepoolten Sozialerhebung werden durch das FDZ des DZHW entsprechend der europäischen Datenschutzgrundverordnung (EU-DSGVO) anonymisiert bereitgestellt und ausschließlich zur wissenschaftlichen Nutzung freigegeben.⁸ Das FDZ bietet sowohl einen Scientific Use File (SUF) für die wissenschaftliche Sekundärnutzung als auch einen Campus Use File (CUF) für Studium und Lehre an.

Voraussetzungen für die Nutzung des SUF sind die Anstellung der Datennutzer*innen an einer wissenschaftlichen Einrichtung und der Abschluss eines Datennutzungsvertrags mit dem FDZ. Studierende oder Promovierende ohne eine Anstellung an einer wissenschaftlichen Einrichtung müssen gemeinsam mit einer betreuenden Person, die an einer wissenschaftlichen Einrichtung angestellt ist, einen Datennutzungsvertrag abschließen. Im Zuge des Vertragsabschlusses wird durch das FDZ das Vorliegen eines wissenschaftlichen Nutzungsinteresses geprüft. Das Formular für den Datennutzungsantrag kann von der Website des FDZ heruntergeladen werden.

[Datenzugang] Das SUF der gepoolten Sozialerhebung wird via Download angeboten.

Download: Die Daten werden verschlüsselt auf der Website des FDZ zum Download bereitgestellt. Datennutzer(innen) können die Daten auf ihrem lokalen Computer speichern, falls gewünscht selbst mit Daten aus externen Quellen verknüpfen und die Daten mit eigener Software analysieren.

[Datenprodukte] Über den Digital Object Identifier (DOI) 10.21249/DZHW:ssypool:3.1.0 ist eine Website mit zentralen Informationen zur Studie, weiteren Dokumentationsmaterialien sowie einer Übersicht der zur Verfügung stehenden Datenprodukte zur Studie erreichbar.

[Gebühren der Datenbereitstellung] Das SUF wird derzeit (Stand: Juni 2019) kostenfrei zur Verfügung gestellt. Änderungen bzw. die aktuelle Gebührenordnung können auf der Website des FDZ (<https://fdz.dzhw.eu>) eingesehen werden.

[Pflichten der Datennutzer*innen] Die Datennutzer*innen sind verpflichtet, folgende Regeln⁹ einzuhalten:

- Wissenschaftliche Nutzung: Die Daten dürfen ausschließlich für wissenschaftliche Zwecke verwendet werden. Eine kommerzielle Nutzung ist untersagt.

⁸ Das Datenschutzkonzept des FDZ ist angelehnt an den Portfolio-Ansatz von Lane, Heus und Mulcahy (2008, S. 6 ff.), an dem sich bereits das Leibniz-Institut für Bildungsverläufe (LifBi) (Koberg, 2016, S. 699 ff.) und das FDZ der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufsforschung (Hochfellner, Müller, Schmucker und Roß, 2012, S. 9 f.) orientieren. Das FDZ des DZHW hat diesen Ansatz an die Anforderungen der eigenen Datenbestände angepasst und nutzt vier Kategorien von Maßnahmen zur Sicherstellung des Datenschutzes, die in unterschiedlicher Weise kombiniert werden können: Rechtlich-institutionelle Maßnahmen, informationelle Maßnahmen, technische Maßnahmen und statistische Maßnahmen.

⁹ Der Datennutzungsvertrag regelt die Nutzungsbedingungen im Detail.

- De-Anonymisierungsverbot: Jeder Versuch der Re-Identifikation von Analyseeinheiten (z. B. Personen, Haushalten, Institutionen) ist verboten.
- Gebot zur Mitteilung von Sicherheitslücken: Falls Datennutzer(innen) Kenntnis von Sicherheitslücken hinsichtlich Datenschutz bzw. Datensicherheit erlangen, müssen diese dem FDZ unverzüglich angezeigt werden.
- Keine Weitergabe der Daten: SUF dürfen nur durch die Person(en) genutzt werden, die den Datennutzungsvertrag abgeschlossen hat/haben.
- Lösungsgebot: Download-SUF sind nach Ablauf der vereinbarten Nutzungsdauer (in der Regel 1,5 Jahre) von jeglichen Rechnern, Servern und Datenträgern zu löschen. Ebenso müssen alle Sicherungskopien, modifizierten Datensätze (z. B. Arbeits-, Auszugs- oder Hilfsdateien) sowie Ausdrücke vernichtet werden.
- Bereitstellung/Meldung von Publikationen: Jede Art von Publikation, die aus der Arbeit mit Daten des FDZ hervorgeht, ist dem FDZ unmittelbar nach Veröffentlichung anzuzeigen und – unabhängig vom Veröffentlichungsformat – als elektronische Version zur Verfügung zu stellen.
- Zitationspflicht: Die verwendeten Daten müssen in Veröffentlichungen, anderen Arbeiten (z. B. Abschlussarbeiten) und Vorträgen gemäß den Vorgaben des FDZ zitiert werden. Die Zitation kann aus dem MDM direkt übernommen werden.

3 Harmonisierung

Die Themen der Sozialerhebung fokussieren traditionell auf die wirtschaftliche und soziale Lage der Studierenden. Anhand eines Kernkatalogs von Fragen richtete sich das Erkenntnisinteresse sowohl auf die Momentaufnahme bestimmter Parameter als auch auf ihre Entwicklung über die Zeit. Darüber hinaus waren jeweils aktuelle Themen ein- oder mehrmalig Gegenstand der Erhebung.

[Vergleichbarkeit der Messungen] Trotz der angestrebten Zeitreihen, die eine Input-Harmonisierung der Instrumente voraussetzen, weisen auch die Fragen des Kernkatalogs aus unterschiedlichen Gründen im Vergleich der neun Sozialerhebungen vielfältige Abweichungen auf. Es können folgende Gruppen von Gründen für als Messunterschiede einzustufende Abweichungen differenziert werden:

- Sachbezogene Gründe:
 - Veränderungen von zu erhebenden Sachverhalten, wie z. B. Änderungen bei der Art bzw. Bezeichnung von (Hochschul-)Abschlüssen (z. B. DDR-Abschlüsse, akademische Abschlüsse nach der Studienstrukturreform), bei Bildungs- und Förderinstitutionen bzw. ihren Bezeichnungen, bei gesetzlichen Regelungen (z. B. Studienkredite, Studiengebühren, Wehrdienst), Herkunftsstaaten(gruppen) bzw. Zielstaaten(gruppen) für Auslandsmobilität
 - Veränderungen des Erkenntnisinteresses, wie z. B. Bedarf an detaillierteren Informationen zu bestimmten Untergruppen, z. B. zu Studierenden mit Kind, Studierende mit gesundheitlicher Beeinträchtigung, Studierende mit besonderen Bedingungen des Hochschulzugangs
 - Währungsumstellung durch Einführung des Euro am 1. Januar 2002.
- Methodische Gründe:
 - Anpassung als Reaktion auf das Antwortverhalten der Befragten, die z. B. das Kindergeld, das ihre Eltern erhalten, zunehmend häufiger als eigenständige Einnahmenquelle nannten oder die beim Zeitbudget angaben, dass es keine „typische“ Semesterwoche gäbe bzw. dass gerade die letzte Woche „untypisch“ gewesen wäre
 - Streichung von Items aus Item-Batterien, um die Länge des Fragebogens zu kürzen
 - Änderungen in Zusammenhang mit dem Switch der Erhebungsmethode von einem Paper-Pencil- (13. - 20. Sozialerhebung) zu einem Online-Survey (21. Sozialerhebung). Bei dem Online-Survey erfolgte z. B. eine automatische Filterung statt der Filterung mit entsprechenden Fragen, viele Frageformulierungen waren aufgrund vorheriger Antworten individualisiert, es wurde die Möglichkeit zur Randomisierung der Abfolge von Fragen oder Items genutzt etc.
 - Aufgrund des Querschnittsdesigns war eine Übereinstimmung der Variablen- und Wertetabellen identischer Fragen für die Primärforschungsprojekte nicht zwingend. Diese Abweichungen wurden auch im Prozess der Erstellung der einzelnen SUF nicht harmonisiert. Die SUF-Erstellung erfolgte zudem zu verschiedenen Zeitpunkten nach unterschiedlichen Standards (vgl. Tabelle 1 und Tabelle 2), so dass auch in diesem Prozess keine Einheitlichkeit hergestellt worden war.
- Sonstige Gründe:

- nicht intendierte, zufällige Abweichungen, wie z. B. geänderte Reihenfolge der Kategorien (männlich/weiblich vs. weiblich/männlich), auf Flüchtigkeit, individuellem Sprachgebrauch oder -verständnis beruhende Detailunterschiede in der Formulierung von Fragen, Items oder Antwortvorgaben.
- gesellschaftliche Veränderungen, z. B. zunehmende Akzeptanz bzw. Erwartungshaltung gegenüber der Erhebung von Kindergeld für die Befragten als eigenständige Kategorie, von eingetragenen Lebenspartnerschaften oder einer dritten Option bei der Geschlechtsangabe.

All diese Abweichungen müssen vor dem Poolen der Daten erkannt und in ihrer Systematik identifiziert werden, um für die ex post-Harmonisierung begründete Entscheidungen über (1) die Aufnahme einer Variable in den Datensatz und – im Falle ihrer Aufnahme – (2) die erforderliche bzw. mögliche Harmonisierung zu treffen.

Mit Hilfe des gepoolten SUF sollen sozialerhebungsübergreifende Analysen erleichtert werden. An diesem Ziel sind die drei nachfolgend dargestellten Hauptregeln für die Harmonisierung ausgerichtet.

[1. Vollständigkeit der Messungen] Ein zentrales Kriterium für die Aufnahme einer Variablen ist die Häufigkeit ihrer Repräsentanz in den Quell-Studien. Nur Variablen, die in mindestens drei der neun Sozialerhebungen in vergleichbarer bzw. harmonisierbarer Weise enthalten sind, wurden in den Gesamtdatensatz aufgenommen.

Beim vorliegenden kumulierten Datensatz der 10. bis 21. Sozialerhebung wird von diesem Dreifach-Prinzip beim Thema Studierende mit Kind abgewichen: Im Rahmen der 13. und der 18. Sozialerhebung erhielten diese einen zielgruppenspezifischen Zusatzbogen. Zentrale Variablen zur Thematik liegen somit nur zweimal vor und sollen dennoch für einen 15 (Nachwende-)Jahre umspannenden Zeitvergleich (1991 und 2006) zugänglich gemacht werden.

[2. Referenzmessung] Der Maßstab dafür, ob eine identische oder abweichende Messung vorliegt, wurde für den vorliegenden Datensatz historisch definiert, beschränkt sich jedoch auf die einbezogenen Datensätze. Als „Urmessung“ gilt demzufolge theoretisch die 10. Sozialerhebung, weil sie die älteste der neun zu poolenden Erhebungen ist bzw. die Sozialerhebung, die eine Frage zum ersten Mal enthält. Weil der erste gepoolte Datensatz die Daten der 17. bis 21. Sozialerhebung umfasste und demzufolge für dessen Erstellung die 17. Sozialerhebung als älteste galt, gilt die Version der 10. Sozialerhebung nur dann als „älteste“ Messung, wenn eine Variable kein Bestandteil der 10. bis 21. Sozialerhebung war bzw. wenn diese Festlegung für eine optimale Harmonisierung über alle zwölf Erhebungszeitpunkte erforderlich ist. Die nachfolgenden Sozialerhebungen werden an der Formulierung einer Frage, der dazugehörigen Ausfüllanweisung, den Items sowie dem Antwortmodell der als Referenzmessung festgelegten bzw. der späteren Ersterhebung gemessen. Ausnahmen von dieser Regel sind aus triftigem Grund möglich. Ein Beispiel hierfür ist der angestrebte Studienabschluss. Aufgrund der 2003 noch im Umsetzungsprozess befindlichen Studienstrukturreform sind Organisation und Bezeichnung der neuen Abschlüsse Bachelor und Master im Rahmen der 17. Sozialerhebung noch von Vorläufigkeit gekennzeichnet, so dass hier auf später etablierte Abschlüsse bzw. ihre Bezeichnung zurückgegriffen wird (s. Dokumentation zur Harmonisierung).

Ein Sonderfall ist die Währung, mit der Einnahmen, Ausgaben, Elterneinkommen etc. erfasst wurden. Bis einschließlich 16. Sozialerhebung wurden alle Geldbeträge in DM erhoben. Für den kumulierten Datensatz wurden diese Werte gemäß des seit der Euro-Einführung unveränderlichen Umrechnungskurses in Euro umgerechnet (1 DM = 0,51129 €). Wenn dennoch in DM gerechnet werden soll, können die DM-Beträge zurück berechnet werden (1 Euro = 1,95583 DM). Diese Vorgehensweise ermöglicht den direkten Vergleich der Beträge im Zeitverlauf. Welchen Zeitwert die jeweiligen

Beträge haben, welche realen Verluste bzw. Zuwächse über die Zeit den Beträgen zu entnehmen ist, kann nur mithilfe externer Daten ermittelt werden (z. B. Preisindex, Inflationsrate).

[3. Vollständigkeit nicht-referenzielle Messungen] Sollte eine spätere Variante häufiger als die Erstmessung und mindestens dreimal eingesetzt worden sein, so wird diese übernommen und die Erstmessung nach Möglichkeit angepasst.

Ausführliche Informationen zu den vorgenommenen Harmonisierungen stellt eine entsprechende Dokumentation zum Datenpaket zur Verfügung.¹⁰ Auf Anfrage kann die zur Harmonisierung verwendete Stata-Syntax zur Verfügung gestellt werden.

Wie oben bereits erwähnt, kommt der 21. Sozialerhebung eine Sonderstellung innerhalb der einbezogenen Datensätze zu, weil sie ausschließlich als Online-Survey durchgeführt wurde. Dieser Methodenswitch wurde – neben der Ausweitung des Fragekatalogs – auch für einen zielgruppenspezifischen Zuschnitt vieler Standardthemen genutzt, wie z.B. Zeitbudget, studentische Erwerbstätigkeit, monatliche Ausgaben. Die Formulierung der Fragen, Ausfüllanweisungen und Antwortoptionen weichen zum Teil stark ab von den methodischen Standards der Vorläuferbefragungen. Mit Blick auf das maximal mögliche Auswertungspotential des gepoolten Datensatzes für den inhaltlichen Kern der Untersuchungsreihe, wurde jeweils eine harmonisierte Variante für die Variablen der 21. Sozialerhebung entwickelt und aufgenommen. Weil es jedoch möglicherweise akzeptable, methodisch vertretbare Alternativen zur gewählten Harmonisierungs-Variante gibt bzw. diese sich u. U. nicht für jede Forschungsfrage eignet, werden neben den harmonisierten Variablen auch die jeweiligen Original-Variablen der 21. Sozialerhebung in den Datensatz aufgenommen. So können Nutzer*innen bei Bedarf eigene Adaptionen vornehmen. Welche Variablen das in Einzelnen betrifft, kann der Dokumentation zur Harmonisierung entnommen werden.

¹⁰ Middendorff, E. & Wallis, M. (2023). 10.-21. Sozialerhebung. Dokumentation der Variablen-Harmonisierung für den gepoolten Datensatz der 10. bis 21. Sozialerhebung (1982-2016). DZHW: Hannover

4 Schema der Variablennamen

Das FDZ-DZHW hat einen Standard zur Variablenbenennung entwickelt, der in den hier aufbereiteten SUF und CUF angewendet wird.¹¹ Es besteht aus einer Präfix-Stamm-Suffix-Systematik: Der Variablenname enthält in Präfix und Suffix zentrale Metadaten, die für die strukturierte Verarbeitung von Variablen nötig sind. Der Stamm enthält zwei hierarchisch zusammenhängende Differenzierungen: Kennzeichnung des Themas sowie eine numerische Ordnung innerhalb des Themas.

Die systematische Vergabe von Stamm und Präfix sind unerlässlich, da sie Metadaten enthalten, die für die weitere (Meta)Datenaufbereitung notwendig sind. Nach der Evaluation der bisherigen Erfahrungen wurde die „thematische Freigabe“ des Stamms als bestes Mittel der Ressourcenverminderung bei gleichzeitig möglichst hohem Informationsgehalt identifiziert.

Im Folgenden wird das Variablennamenschema dargestellt, das für den gepoolten Datensatz der 10. bis 21. Sozialerhebung entwickelt wurde und das sich am sogenannten Goldstandard des FDZ des DZHW entwickelten einheitlichen Variablennamenschema orientiert (vgl. ebenda). Die Zusammenfügung von Datensätzen setzt voraus, dass identische Variablen und/oder identische Fälle als solche eindeutig identifizierbar sind. Für den gepoolten Datensatz aus neun Sozialerhebungen (10. - 21. Sozialerhebung) muss deshalb ein einheitliches Variablennamenschema entwickelt werden, um dieser Anforderung zu entsprechen. Darüber hinaus gibt es kohortenbezogene Variablenspezifika, wie z. B. zusätzliche Items einer Itematterie, Modifikationen in der Formulierung der Frage und/oder Antwort(en). Diese Besonderheiten sollen im Variablennamen systematisch kenntlich gemacht werden, damit Nutzer*innen sowohl die thematische Zugehörigkeit als auch die Besonderheit einer Variablen erkennen können.

¹¹ Vgl. Daniel, Weber (2018). Einheitliches Variablennamenschema für das FDZ des DZHW. Gold- und Silberstandard. Version 3.0. Projektbericht. DZHW: Hannover

4.1 Variablennamenschemata der einzubeziehenden Datensätze

Für alle zwölf Quelldatensätze wurden im FDZ SUFs erstellt. Die SUF-Erstellung erfolgte zu unterschiedlichen Zeitpunkten und unter Anwendung unterschiedlicher Standards sowohl in Bezug auf die Zugangswege (Tabelle 1) als auch auf das Variablennamenschema (Tabelle 2).

Tab. 1: Standards und Zugangswege der SUFs zur 10.–21. Sozialerhebung

Sozialerhebung Nr.	Erhebungsjahr	Jahr der SUF-Erstellung	angewandter SUF-Standard	Zugangswege
10	1982			SUF: Download
11	1985			SUF: Download
13-21	1991-2016	2022		CUF: Download SUF: Download
12	1988			SUF: Download
13	1991			
14	1994		Bronze	SUF: Download
15	1997			
16	2000	2021		
17-21	2003-2016			CUF: Download SUF: Download
17	2003			
18	2006	2019		SUF: Download
19	2009			
20	2012	2017	Gold	CUF: Download SUF: Download, Remote-Desktop, On-Site
21	2016	2018	Silber	

Tab. 2: Variablennamenschemata der Einzel-SUFs zur 10.–21. Sozialerhebung

Sozialerhebung Nr.	Beispiele (Variablen-Name: Variablen-Label)	SUF-Variablenschema
10-16	stu02_g: angestrebter Abschluss par04: allgemeinbildender Abschluss der Mutter baf01: Im SoSe 2003 nach dem BAföG gefördert. fin02b: Ausgaben für Ernährung	FDZ-Präfix-Stamm-Suffix-Schema
17-21	stu02_h: angestrebter Abschluss par04_h: allgemeinbildender Abschluss der Mutter baf01_h: Im SoSe 2003 nach dem BAföG gefördert. fin02b_h/ fin02_v21: Ausgaben für Ernährung	
17	absartagg: angestrebter Abschluss bilmut: allgemeinbildender Abschluss der Mutter baf: Im SoSe 2003 nach dem BAföG gefördert. ausern: Ausgaben für Ernährung	projektseitig mnemotechnisch ¹²
18	absartagg: angestrebter Abschluss bilmut: allgemeinbildender Abschluss der Mutter baf: Im SoSe 2006 nach dem BAföG gefördert. ausern: Ausgaben für Ernährung	
19	stu03_g1: angestrebter Abschluss par04: allgemeinbildender Abschluss der Mutter baf01: Im SoSe 2009 nach dem BAföG gefördert. fin03b: Ausgaben für Ernährung	FDZ-Präfix-Stamm-Suffix-Schema
20	stu02_g1: angestrebter Abschluss par04: allgemeinbildender Abschluss der Mutter baf01: Im SoSe 2012 nach dem BAföG gefördert. fin03b: Ausgaben für Ernährung	
21	sabsan_g1: angestrebter Abschluss deltschum: allgemeinbildender Abschluss der Mutter fbafja: Im SoSe 2016 nach dem BAföG gefördert. fausgernehh: Ausgaben für Ernährung, Einzel-Haushalt	projektseitiges Präfix-Stamm-Suffix-Schema

¹² Mnemotechnischen Variablennamen werden auch als „sprechende“ Variablennamen bezeichnet. Mit ihnen vermittelt sich meist bereits der zentrale Inhalt der Variable, z. B. absart, geschl, eink, kind.

4.2 Variablennamenschemata des gepoolten Datensatzes

Die Harmonisierung der Variablennamen ist zwingend erforderlich, wenn die Daten gleicher Items aus unterschiedlichen Sozialerhebungen gemerged werden sollen.

Für das Variablennamenschema des Datensatzes, der die sozialerhebungsübergreifend vergleichbaren Daten der 10.-21. Sozialerhebung enthält, wird der im FDZ-DZHW entwickelte Goldstandard verwendet.¹³ Dieser Standard sieht vor, dass sich Variablennamen aus Präfix, Stamm und Suffix zusammensetzen. Das Präfix der Variable enthält bei Längsschnittbefragungen mit mehr als einem Befragungszeitpunkt (Panel) die Wellenkennung anhand eines Buchstabens. Da es sich bei den Sozialerhebungen um Querschnittsbefragungen handelt, entfällt das Präfix. Im Stamm geht der Themenbereich, dem die Variable zugeordnet ist, aus einem dreistelligen englischen Buchstabenkürzel hervor (Tabelle 3).

Die Unterscheidung, mit welcher der zwölf Sozialerhebungen die jeweiligen Daten einer Variable erhoben wurden, erfolgt über eine entsprechende Identifikatorvariable (ssynr) mit den Werten 10 bis 21.

Tab. 3: Teilelemente und Zusammensetzung des Variablenstammes

Teilelement	Beschreibung
Themen-differenzierung*	Mit einem (englischen) Kürzel aus drei Buchstaben wird die Variable einem inhaltlichen Themengebiet zugeordnet. Der gepoolte Datensatz enthält nur Themengebiete, die Bestandteil aller zwölf einzubeziehenden Sozialerhebungen (10.-21. Sozialerhebung) sind.
Nummerierung*	Innerhalb der definierten Themenbereiche werden die Variablen auf minimal zwei, maximal drei Stellen durchnummeriert.
Indizierung	Mit Hilfe eines Buchstabens am Ende des Stamms können verschiedene Variablen, die zur gleichen Frage gehören und dadurch die gleiche Themendifferenzierung und Nummerierung aufweisen (z. B. bei Itembatterien, Mehrfachnennungen oder Fragen, in denen geschlossene und offene Fragen kombiniert werden), gekennzeichnet werden (z. B. 01a, 01b, 01c, ...). Falls eine Frage den Umfang von 26 Einzelvariablen (a-z) überschreitet, wird die Itembezeichnung ab dem 27. Item mit zwei Buchstaben fortgesetzt (aa, ab, ac, ...).

* muss zwingend vergeben werden

Tabelle 4 stellt die verschiedenen Themenbereiche des gepoolten Datensatzes der 10.-21. Sozialerhebung sowie das zugehörige Kürzel für den Stamm des Variablennamens dar. Das Kürzel leitet sich jeweils von der englischen Bezeichnung für das Themenfeld ab. Entsprechend der Zielstellung dieses Datenpools, sozialerhebungsübergreifende Analysen zu erleichtern, enthält der Datensatz ausschließlich Themenbereiche, die Gegenstand von mindestens drei der neun einbezogenen Sozialerhebungen war und auf vergleichbare Weise erhoben wurden.

¹³ Vgl. Daniel, Weber (2017). Einheitliches Variablennamenschema für das FDZ des DZHW. Gold- und Silberstandard. Version 3.0. Projektbericht. DZHW: Hannover. S. 8 ff.

Tab. 4: Themengebiete in den Variablennamen des gepoolten Datensatzes – 10.-21. Sozialerhebung (1982-2016)

Nr.	Themengebiets-Kürzel (= Stamm)	Themengebiet (englisch)	Themengebiet (deutsch)
1	dem	socio-demographic characteristics	sozio-demographische Merkmale
2	par	characteristics of parents	Merkmale der Eltern
3	stu	characteristics of study	Merkmale des Studiums
4	ped	prior education and entry into HE	Vorbildung und Hochschulzugang
5	fin	financing (of living during studies)	Finanzierung (des Lebensunterhalts während des Studiums)
6	baf	BAföG (German Federal Grant on Training and Education Promotion)	BAföG (Bundesausbildungsförderungsgesetz)
7	tim	time usage (studies/job)	Zeitaufwand (Studium/ Erwerbstätigkeit)
8	job	job during studies	Erwerbstätigkeit während des Studiums
9	abr	studying abroad	studienbezogener Auslandsaufenthalt
10	lan	language skills	Sprachkenntnisse
11	liv	living (accommodation)	Wohnsituation
12	adv	demand for advice and information	Beratungs- und Informationsbedarf
13	nut	mensa and nutrition	Mensa und Ernährung
14	way	way and mode of transportation to university	Weg zur Hochschule und Verkehrsmittelwahl
15	kid	specific topics related to students with child	spezifische Themen für Studierende mit Kind
16	med	computer usage and new media	Computernutzung und neue Medien

4.3 Das Suffix

Das anhand eines Unterstriches vom Stamm abgetrennte Suffix enthält verschiedene Zusatzinformationen, wie in Tabelle 5 beschrieben:¹⁴

¹⁴ Vgl. Daniel, Weber (2017). Einheitliches Variablennamenschema für das FDZ des DZHW. Gold- und Silberstandard. Version 3.0. Projektbericht. DZHW: Hannover. S. 10 f.

Tab. 5: Zusammenfassende Übersicht der verwendeten Suffix Kürzel¹⁵

Suffix-Bedeutung (Kürzel)	Beschreibung
generiert (g#)	Generierte Variablen werden mit dem Kürzel g# (g1 bzw. bei weiteren Derivaten g2, g3 etc.) markiert. Unter den Typus der generierten Variablen fallen alle Variablen, die aus einer oder mehreren Variablen des Ursprungsdatensatzes erzeugt wurden (Recodierungen ¹⁶ , Indizes, vercodete Variablen, Aggregationen). Präfix und Stamm der generierten Variablen entsprechen jeweils der Ausgangsvariablen. Wird eine abgeleitete Variable aus mehreren Ausgangsvariablen gebildet, so wird der Stamm neu vergeben.
versioniert (v#)	<p>Bei Langzeitstudien wie der Sozialerhebung können einzelne Fragen im Laufe der Zeit abgewandelt werden (neue Fragen-, Item- oder Antwortkategorienformulierung). Dabei gilt, dass mit jeglicher Formulierungsänderung (Satzstellung, Verwendung von Synonymen oder Abkürzungen) immer eine neue Version der Variable einhergeht, die entsprechend gekennzeichnet wird. Referenz für eine Änderung ist die Fassung der Frage in der jeweils ältesten der einzubeziehenden Sozialerhebungen. Weil es jedoch bereits einen gepoolten Datensatz gibt (17.-21. Sozialerhebung), dient diese mit der 17. Sozialerhebung beginnende Zeitreihe auf für die jetzt neu einbezogenen älteren Sozialerhebungen (13.-16.) als Referenz. Für erstmals aufgenommene Variablen ist die jeweils älteste Erfassung der Maßstab für Abweichungen bei den Fragen, Ausfüllhinweisen, Antwortmodellen.</p> <p>Diese Variablenversionen erhalten denselben Variablenstamm wie die Ursprungsvariable, ergänzt um ein sozialerhebungsbezogenes Versionskürzel v# im Suffix (z. B. _v18 für eine Änderung, die im Rahmen der 18. Sozialerhebung vorgenommen wurde, _v19 für die Änderung im Rahmen der 19. Sozialerhebung usw.). Die ursprüngliche Version (hier: die Version, die als erstes in einer der neun einzubeziehenden Sozialerhebungen eingesetzt wurde) in der Variable erhält keine Kennzeichnung. Sollte nach einer zwischenzeitlichen Änderung wieder zur Ursprungsvariante zurückgekehrt worden sein, erfolgt ebenfalls keine Kennung.</p> <p>Als über die Zeit „unveränderte Variable“ werden dementsprechend nur Items mit exakt gleicher Formulierung aller Bestandteile einer Variable (Frage, Ausfüllanweisung, Antwortmodell, inhaltlicher und/oder zeitlicher Bezug) in den gepoolten Sozialerhebungen angesehen.¹⁷ Die Darstellung und die Anordnung der Frage im Erhebungsinstrument spielen keine Rolle. Ebenso ist nicht notwendig, dass die zugehörige Frage auch den gleichen Personengruppen wiederholt gestellt wurde.</p>
harmonisiert (h)	Zusätzlich zu den unterschiedlichen Versionen einer Variablen können auch Variablen erzeugt werden, welche orientiert am „kleinsten gemeinsamen Nenner“ der unterschiedlichen Variablenversionen harmonisiert werden. Diese Variablen erhalten ebenfalls den Namen der Ausgangsvariablen, ergänzt um das Suffix _h. Infolge der Erweiterung des gepoolten Datensatzes können Variablennamen, die im Gesamtdatensatz der 17.-21. Sozialerhebung noch kein Suffix enthielten, nunmehr mit Suffix versehen sein, weil die hinzugefügten Daten der 13. bis 16. Abweichungen der Fragestellung aufweisen, die harmonisierungsrelevant sind.

¹⁵ Die optionalen Suffix-Kürzel p für Variablen, die im Zuge des Datenaufbereitungsprozesses plausibilisiert wurden bzw. das Suffix-Kürzel für Varianten des Zugangsweges (c, d, r, o, a) entfallen, weil bereits aufbereitete SUF gepoolt werden und nur ein Zugangsweg (Download) zur Verfügung gestellt wird.

¹⁶ Darauf hinzuweisen ist, dass Variablen, bei denen lediglich im Rahmen der Plausibilisierung Editionen vorgenommen wurden, nicht als generierte Variablen anzusehen sind.

¹⁷ Auf Zeitvergleiche angelegte Langzeituntersuchungsreihen sollten gewährleisten, dass Änderungen an Fragetexten und/oder Antwortkategorien von Variablen nur dann erfolgen, wenn dies inhaltlich und/oder methodisch zwingend geboten ist. (Vermieden werden sollte vor allem das Einfügen oder Löschen von Füllwörtern und das Ersetzen von Wörtern durch Synonyme.)

4.4 Systematik fehlender Werte

Die Einzel-SUF - als Quell-Datensätze für den kumulierten Datensatz - weisen verschiedene Missing-Systematiken auf (vgl. Tabelle 6), die für den gepoolten Datensatz übernommen und lediglich mit einer einheitlichen Kodierung versehen wurden (vgl. Tabelle 7). Eine Harmonisierung der Missings ist nur begrenzt möglich, weil nicht alle Missings gleichermaßen existent sind bzw. vergeben worden waren. So weist die 21. Sozialerhebung onlinesurveyspezifische Missings auf („vorheriger Befragungsabbruch“, „nicht gesehen“). Das Missing „nicht gesehen“ wurde hier seitens des Primärforschungsprojektes auch vergeben, wenn Befragte aufgrund ihres vorherigen Antwortverhaltens über Fragen hinweggefiltert wurden. Bei der 17. bis 20. Sozialerhebung enthalten gefilterte Fragen das entsprechende Missing „filterbedingt fehlend“; bei der 21. Sozialerhebung hingegen werden diese Missings unter „nicht gesehen“ subsumiert¹⁸.

Tab. 6: Systematik fehlender Werte in den Quell-Datensätzen 10. - 21. Sozialerhebung

10. bis 16., 19., 20. Sozialerhebung		17. und 18. Sozialerhebung		21. Sozialerhebung	
Code	Wertelabel	Code	Wertelabel	Code	Wertelabel
-999	weiß nicht			-12	weiß nicht (selbstberichtet)
-998	keine Angabe	-2	keine Angabe	-9990	nicht beantwortet
-997	keine Angabe (Antwortkategorie)			-13	keine Angabe (selbstberichtet)
-996	Interviewabbruch			-9993	vorheriger Befragungsabbruch
-994	verweigert				
-989	filterbedingt fehlend				
-988	trifft nicht zu	-1	trifft nicht zu	-9981	Filter-Plausi
-987	designbedingt fehlend (Fragebogensplit)			-11	tnz (selbstberichtet)
-985	designbedingt fehlend (Kohorte)b			-9991	nicht gesehen (Filter, Split, Einblendbedingung)
-969	unbekannter fehlender Wert			-9981	Filter-Plausi
-968	unplausibler Wert				
-967	anonymisiert			-967	anonymisiert
-966	nicht bestimmbar			-9996	Zusatzvariable nicht bestimmbar
-965	ungültige Mehrfachnennung				

Da der gepoolte Datensatz auf der Grundlage der zwölf Einzel-SUF erstellt wird und in dem SUF für die 21. Sozialerhebung diese pauschalisierende Kodierung des Primärforschungsprojektes unverän-

¹⁸ Beispiel hierfür ist die Variable „Anzahl Kinder“ (dem06_h): Für die 21. Sozialerhebung werden fast 53.000 „nicht gesehen“-Fälle ausgewiesen, von denen mehr als 52.000 zuvor die Frage, ob sie Kinder haben, verneint hatten.

dert blieb, wurde auf eine ex post Plausibilisierung des Missings verzichtet, das verschiedene Gründe für fehlende Werte zusammenfasst.

Die Quell-Datensätze weisen verschiedene Missing-Systematiken auf (vgl. Tabelle 6), die für den gepoolten Datensatz übernommen und lediglich mit einer einheitlichen Kodierung versehen wurden (vgl. Tabelle 7). Eine Harmonisierung der Missings ist nur begrenzt möglich, weil nicht alle Missings gleichermaßen existent sind bzw. vergeben worden waren. So weist die 21. Sozialerhebung online-survey-spezifische Missings auf („vorheriger Befragungsabbruch“, „nicht gesehen“). Das Missing „nicht gesehen“ wurde hier seitens des Primärforschungsprojektes auch vergeben, wenn Befragte aufgrund ihres vorherigen Antwortverhaltens über Fragen hinweggefiltert wurden. Bei der 17. – 20. Sozialerhebung enthalten gefilterte Fragen das entsprechende Missing „filterbedingt fehlend“; bei der 21. Sozialerhebung hingegen werden diese Missings unter „nicht gesehen“ subsummiert. Da der gepoolte Datensatz auf der Grundlage der neun Einzel-SUF erstellt wird und in dem SUF für die 21. Sozialerhebung diese pauschalisierende Kodierung des Primärforschungsprojektes unverändert blieb, wurde auf eine ex post Plausibilisierung des verschiedene Gründe für fehlende Werte zusammenfassenden Missings verzichtet.

Tab. 7: Systematik fehlender Werte im gepoolten Datensatz der 10.-21. Sozialerhebung

Wertebereich	Code	Wertelabel
(selbstberichteter) Nonresponse	-999	weiß nicht (Antwortkategorie)
	-998	keine Angabe
	-997	keine Angabe (Antwortkategorie)
Nicht zutreffend / filterbedingt fehlend	-989	filterbedingt fehlend
	-988	trifft nicht zu
	-987	designbedingt fehlend (Split, Filter, Einblendbedingung) ^a
	-986	ssy-spezifisch fehlend ^b
	-985	kohortenspezifisch fehlend ^c
Editierter fehlender Wert	-969	unbekannter fehlender Wert ^d
	-968	unplausibler Werte
	-967	anonymisiert
	-966	nicht bestimmbar ^f
	-965	ungültige Mehrfachnennung

a Dieses Missing kommt nur in der 21. Sozialerhebung vor. Sie wurde als einzige online durchgeführt und hat Splits und Einblendbedingungen eingesetzt. Zu den Einblendbedingungen gehören auch Filterführungen, in deren Ergebnis Folgefragen nicht angezeigt wurden, d. h. von den Befragten „nicht gesehen“ wurden – ebenso wie bei Splits.

b Dieser Wert wird vergeben, wenn eine Variable ein Missing aufweist, weil es eine sozialerhebungsspezifische Variante dieser Variablen gibt. Diese Varianten sind in einer eigenen Variable im Datensatz enthalten und mit einem entsprechenden Suffix im Variablenamen gekennzeichnet.

c Dieses Missing bedeutet, dass die Variable bzw. der Wert in einer oder mehreren Sozialerhebungen nicht erhoben wurde/nicht enthalten ist.

d Dieser Missing wird vergeben, wenn keinerlei Ursache rekonstruiert werden kann.

e Angaben, die aufgrund unterschiedlicher Faktoren in der Codierphase als nicht plausibel eingestuft werden, erhalten diesen Wert. Eine exakte Rekonstruktion ist ggf. nicht mehr möglich.

f Diese Kategorie wird vergeben, wenn eine eindeutige Codierung nicht möglich ist, z. B. offene Angabe, die nicht vercodet werden konnte, da sie nicht lesbar ist.

5 Variablen im Datensatz

5.1 Sortierung der Variablen im Datensatz

Da die Reihenfolge der Fragen des Fragekatalogs im Vergleich der einbezogenen Sozialerhebungen unterschiedlich ist, sind die Variablen im Datensatz thematisch gruppiert (s. Kapitel 5.2).

Generierte Variablen werden, wenn möglich, unterhalb der Ausgangsvariablen positioniert. Wurde eine neue Variable aus verschiedenen Ausgangsvariablen generiert, wird sie hinter jene Variable(ngruppe) einsortiert, welche ihr thematisch am ehesten entspricht. Falls eine thematische Zuordnung nicht möglich ist, wird die Variable an das Ende des Datensatzes gestellt.

5.2 Übersicht zur sozialerhebungsbezogenen Variablen-Präsenz

Das Scientific Use File der gepoolten Daten enthält 909 Variablen, darunter drei Systemvariablen (Ident-Nr., Nr. und Jahr der Sozialerhebung). Die einbezogenen zwölf Sozialerhebungen sind in den 906 inhaltlichen Variablen unterschiedlich stark repräsentiert (vgl. Tabelle 8). Tendenziell waren die Fragekataloge jüngerer Sozialerhebungen länger und detaillierter als die älterer, so dass der Anteil an allen Variablen im kumulierten Datensatz mit der Kürze der zeitlichen Distanz einer Erhebung steigt. Der gepoolte Datensatz enthält neben harmonisierten Variablen ($n = 229$) eine Vielzahl an sozialerhebungsspezifischen Variablen. Letztere sind nur mit Vorbehalt bzw. ggf. nach weiteren Anpassungen mit den Messungen anderer Erhebungszeitpunkte vergleichbar. Aufgrund ihrer Ähnlichkeit wurden sie dennoch in den Gesamtdatensatz aufgenommen, um den Datennutzer*innen die Möglichkeit einzuräumen, je nach Forschungsfrage eigene Adaptionen vorzunehmen bzw. zu entscheiden, ob die hiermit gekennzeichneten inhaltlichen und/ oder methodischen Abweichungen für ihre Analyseziele relevant sind.

Der große Anteil an Variablen mit dem Suffix `_v21` (26 %, vgl. Tabelle 8) hängt damit zusammen, dass die erstmals als Online-Survey erhobene 21. Sozialerhebung die Zielgruppenspezifika des Fragekatalogs in den Vordergrund stellen konnte und gestellt hat. So wurden viele Kernfragen an Untergruppen angepasst, was die Vergleichbarkeit und das Harmonisierungspotential stark einschränkt. Das betrifft vor allem Fragen zu den monatlichen Ausgaben, die differenziert für vier Haushaltstypen gestellt wurden, sowie Fragen zum Zeitbudget, bei dem sieben Fallgruppen unterschieden wurden (s. auch Dokumentation zur Variablen-Harmonisierung von Middendorff/Wallis 2023).

Ein Indikator für das Potential des Datensatzes, Zeitreihenanalysen durchzuführen, ist die Repräsentanz der einbezogenen Kohorten in den Variablen. Ein realistisches Bild zeigt sich, wenn die Variablen, die spezifisch für die 21. Sozialerhebung sind `_v21`, unberücksichtigt bleiben (vgl. Abb. 1): Etwas mehr als 5 % der 672 Variablen ohne Suffix `_v21` enthalten Werte für alle zwölf Erhebungszeiträume. Ein Fünftel der Variablen (20 %) ermöglicht einen Vergleich von maximal acht Erhebungssemestern, mehr als ein Drittel (34 %) der Variablen liegen für fünf Erhebungen in identischer bzw. harmonisierter Varianten vor. 43 % der Variablen sind viermal vergleichbar erhoben worden und 60 % aller Vari-

ablen im gepoolten Datensatz umfassen drei Erhebungen. Ausschließlich auf eine einzelne Sozialerhebung beziehen sich 30 % der Variablen (ohne Variablen mit dem Suffix _v21). In Abhängigkeit von der Forschungsfrage sind dennoch die Einbeziehung weiterer Erhebungszeitpunkte möglich, wenn versionierte Variablen als vergleichbar mit harmonisierten oder unverändert gelassenen Variablen eingestuft und für die Analyse verwendet werden können.

Tab. 8: Übersicht zur sozialerhebungsbezogenen Variablen-Präsenz

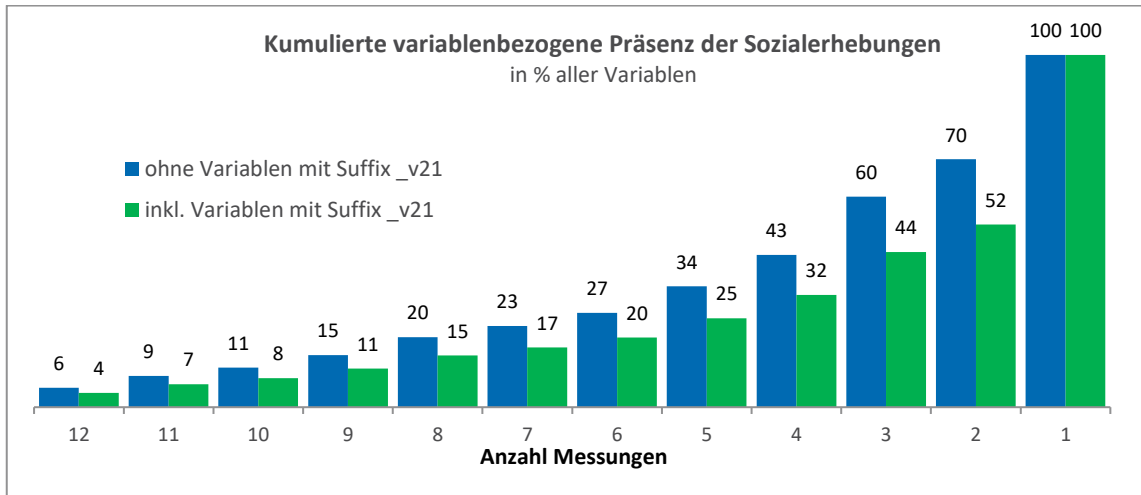
Variablen je Sozialerhebung	10	11	12	13	14	15	16	17	18	19	20	21	ges.
	Anzahl Variablen	147	171	206	238	260	246	273	256	298	292	266	378
Anteil an allen Variablen im Datensatz in %	16,2	18,9	22,7	26,3	28,7	27,2	30,1	28,3	32,9	32,2	29,4	41,7	100

Variablen je Suffix _h = harmonisiert _# = sozialerhebungs-spezifisch	Suffix												ges.	
	_h	_v10	_v11	_v12	_v13	_v14	_v15	_v16	_v17	_v18	_v19	_v20		_v21
Anzahl Variablen	229	27	16	19	16	41	10	27	0	32	20	32	236	705
Anteil an allen Variablen im Datensatz in %	25,3	3,0	1,8	2,1	1,8	4,5	1,1	3,0	0,0	3,5	2,2	3,5	26,0	77,8

Variablenbezogene Präsenz = Zeitreihenpotential	Mehrfachmessungen: Anzahl Messungen												ges.
	1	2	3	4	5	6	7	8	9	10	11	12	
Anzahl Variablen	436	71	110	60	50	25	21	34	24	16	22	37	906
Anteil an allen Variablen im Datensatz in %	48,1	7,8	12,1	6,6	5,5	2,8	2,3	3,8	2,6	1,8	2,4	4,1	100

ohne Suffix _21: Variablenbezogene Präsenz = Zeitreihenpotential	Mehrfachmessungen ohne Variablen mit Suffix _21: Anzahl Messungen												ges.
	1	2	3	4	5	6	7	8	9	10	11	12	
Anzahl Variablen	198	71	110	60	50	25	21	34	24	16	22	37	668
Anteil an übrigen Variablen im Datensatz in %	29,6	10,6	16,5	9,0	7,5	3,7	3,1	5,1	3,6	2,4	3,3	5,5	100

Abb. 1: Zeitreihenpotential: Anteil an Variablen im Datensatz für 1 bis 12 Erhebungszeitpunkte



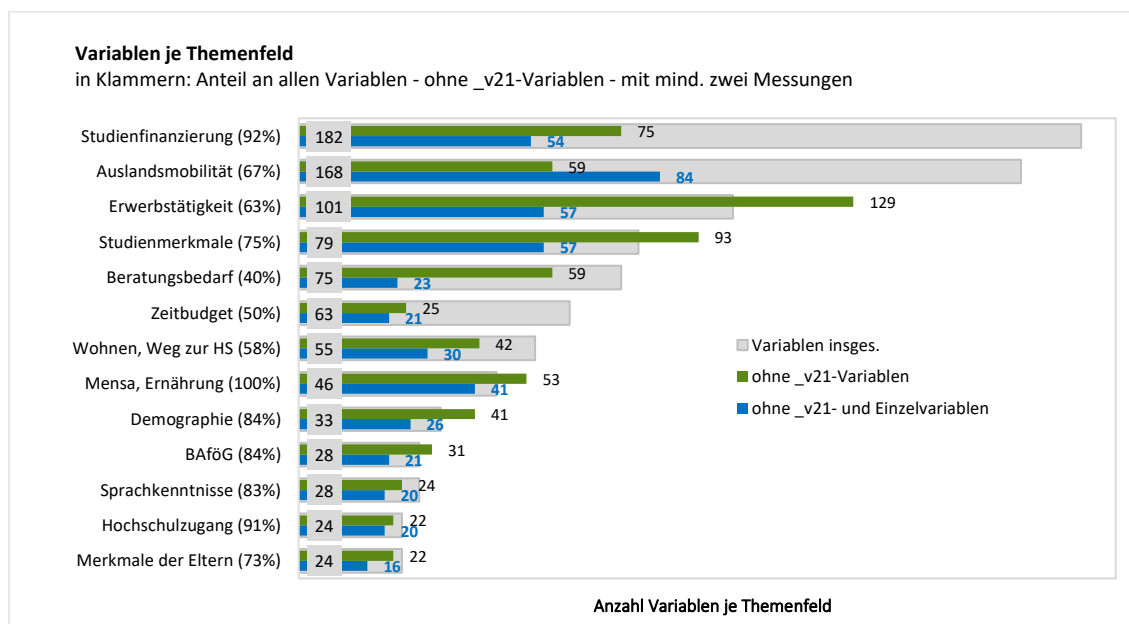
Eine Übersichtstabelle über alle Variablen des gepoolten Datensatzes der 10.-21. Sozialerhebung einschließlich der Kennzeichnung, aus welchen Erhebungen sie jeweils zur Verfügung stehen, ist Bestandteil der Dokumentation zur Variablen-Harmonisierung (Middendorff/Wallis, 2023).

5.3 Übersicht zur themenbezogenen Variablen-Präsenz

Wenngleich die soziale und wirtschaftliche Lage der Studierenden unverändert der inhaltliche Kern der Sozialerhebung blieb, sind – wie oben bereits beschrieben – über die Jahrzehnte unterschiedliche Schwerpunktsetzungen, einmalig aufscheinende oder periodisch aufgenommene Themen zu beobachten (vgl. Middendorff 2022). Selbst ein gleichbleibendes bzw. wiederholt integriertes Thema wurde häufig unterschiedlich ausführlich bzw. methodisch abweichend erfasst. Eine ex post-Harmonisierung im Rahmen der Erstellung des kumulierten Datensatzes stößt deshalb schnell an Grenzen, innerhalb derer Anpassungen inhaltlich und methodisch machbar bzw. vertretbar sind. Darüber hinaus gibt es zumeist mehrere Möglichkeiten, die Variablen einander anzupassen. Die zu bevorzugende Variante der Angleichung ist in vielen Fällen stark von der jeweiligen Forschungsfrage abhängig. Die Vielfalt potentieller Forschungsfragen und ihrer spezifischen Datenbedarfe kann von den Kurator*innen weder vorhergesehen noch berücksichtigt werden. Um das Analysepotential des kumulierten Datensatzes nicht a priori einzuschränken, wurden deshalb sozialerhebungsspezifische Varianten auch in Form von „Einzelvariablen“, die so nur in einer Sozialerhebung erhoben wurden, in den Datensatz aufgenommen. Das ermöglicht den Nutzer*innen, selbst zu entscheiden, ob bzw. wie Variablen in der Zeitreihe für ihre Zwecke vergleichbar sind bzw. vergleichbar gemacht werden können.

Abb. 2 veranschaulicht sowohl die inhaltliche Verteilung der Variablen im gepoolten Datensatz als auch die Abweichung zwischen der Anzahl der Variablen je Themenfeld und der Anzahl an themenfeldspezifischen Variablen, die für mindestens zwei Messzeitpunkte in identischer oder harmonisierter Form vorliegen. Zu den am häufigsten präsenten Themen gehören demnach einerseits *Studienfinanzierung*, *Auslandsmobilität* und *Erwerbstätigkeit* mit 182, 168 bzw. 101 Variablen im Datensatz.

Abb. 2: Themenpräsenz im kumulierten Datensatz: Anzahl Variablen je Themenfeld

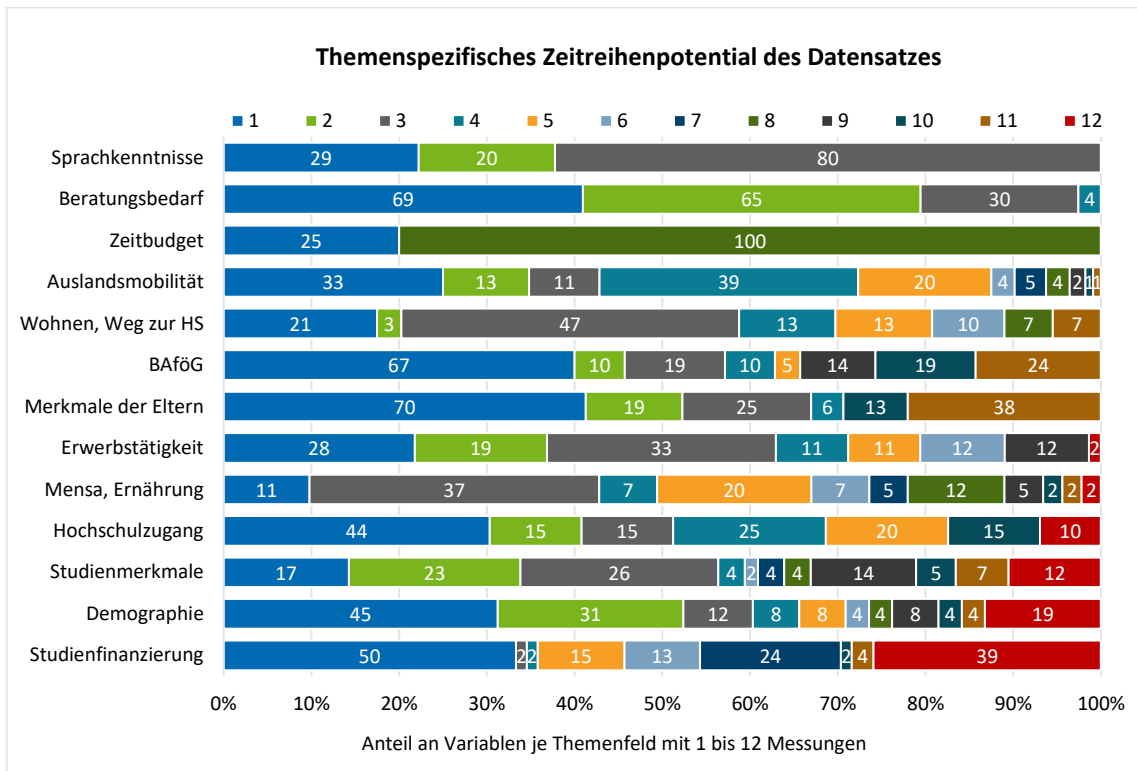


Die Erhebung der Themen unterlag im hier betrachteten, fast dreieinhalb Jahrzehnte umfassenden Zeitraum (1982-2016) mehr oder weniger starken methodischen Veränderungen, denn der Anteil an Variablen, die Bestandteil nur einer Sozialerhebung waren – insbesondere nur der 21. Sozialerhebung – ist unterschiedlich groß. Innerhalb der Variablen, die vergleich- oder harmonisierbar erhoben wurden und deshalb in den kumulierten Datensatz aufgenommen werden konnten, weisen die Themenfelder unterschiedliche Anteile an Wiederholungsmessungen auf. Besonders hoch ist dieser Anteil für Themenfelder wie *Mensa* (100 %), *Studienfinanzierung* (92 %), *Hochschulzugang* (91 %) sowie demografische *Merkmale* (84 %) (vgl. Abb.2). Gegenbeispiele sind die Themenfelder *Zeitbudget* (50 %) und *Beratungs- und Informationsbedarf* (40 %).

In Bezug auf das zeitreihenbezogenen Analysepotential haben die Variablen der Themenfelder *Studienfinanzierung*, *Demographie*, *Studienmerkmale* und *Hochschulzugang* einen relativ hohen Anteil an Variablen, die für alle zwölf integrierten Sozialerhebungen vorliegen (zwischen 39 % und 10 %; vgl. Abb.3). Auf der anderen Seite stehen Themen, die nicht regelmäßig bzw. unterschiedlich ausführlich, z. B. mittels Zusatzbogen, erhoben wurden und deshalb vergleichsweise wenig Wiederholungsmessungen aufweisen. Zu ihnen gehören die Themenfelder *Beratungsbedarf*, *Studium mit Kind* (*Demografie*), aber auch *Merkmale des Studiums* und *Sprachkenntnisse*.

Diese Analyse ist nur ein erster Überblick über das themenspezifische Zeitreihenpotential des gepoolten Datensatzes der 10.-21. Sozialerhebung. Welchen Wert er für welche Fragestellung tatsächlich hat, kann nur anhand eines detaillierten Blicks in die Daten gewonnen werden, denn selbstverständlich ist es nicht unerheblich, für welche Zeitpunkte, in welchen zeitlichen Abständen etc. inhaltlich relevante und vergleichbare Variablen enthalten sind. Unter Umständen sind weitere Harmonisierungen möglich, weil methodische Unterschiede für die Forschungsfrage vernachlässigbar sind.

Abb. 3: Themenspezifisches Zeitreihenpotential des Datensatzes



6 Literatur

- Becker, K., Baillet, F. & Weber, A. (2019). 21. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 2016. Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy21\\$-2.0.0/attachments/ssy21_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy21$-2.0.0/attachments/ssy21_MethodReport_de.pdf)
- Daniel, A. & Weber, A. (2017). Einheitliches Variablennamenschema für das FDZ des DZHW. Gold- und Silberstandard. Version 3.0. Projektbericht. Hannover: FDZ-DZHW.
- Daniel, A., Sarcletti, A. & Vietgen, S. (2017). 20. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 2012. Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy20\\$-1.0.0/attachments/ssy20_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy20$-1.0.0/attachments/ssy20_MethodReport_de.pdf)
- Ebel, T. & Meyermann, A. (2015). Hinweise zur Anonymisierung von quantitativen Daten (Forschungsdaten Bildung informiert Nr. 3). Verbund Forschungsdaten Bildung.
- Hoffstätter, U. & Sarcletti, A. (2017). 19. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 2009. Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy19\\$-1.0.0/attachments/ssy19_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy19$-1.0.0/attachments/ssy19_MethodReport_de.pdf)
- Koberg, T. (2016). Disclosing the National Educational Panel Study. In H.-P. Blossfeld, J. v. Maurice, M. Bayer & J. Skopek (Hrsg.), *Methodological Issues of Longitudinal Surveys. The example of the National Educational Panel Study* (S. 691–708). Wiesbaden: Springer VS. doi:10.1007/978-3-658-11994-2
- Middendorff, E. (2022). Die Sozialerhebungen des Deutschen Studentenwerks 1951–2016. Ein historischer Überblick über Akteure, Wellen, Methoden, Themen und projektbezogene Publikationen (Working Paper). Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW). <https://t1p.de/se-history>
- Middendorff, E. & Hoffstätter, U. (2020). 17. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 2003. Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy17\\$-2.0.0/attachments/ssy17_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy17$-2.0.0/attachments/ssy17_MethodReport_de.pdf)
- Middendorff, E. & Hoffstätter, U. (2019). 18. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 2006. Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy18\\$-1.0.0/attachments/ssy18_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/studies/stu-ssy18$-1.0.0/attachments/ssy18_MethodReport_de.pdf)
- Middendorff, E. & Wallis, M. (2023): 10.-21. Sozialerhebung. Dokumentation der Variablen-Harmonisierung und sozialerhebungsbezogene Übersicht aller Variablen für den gepoolten Datensatz der 10.-21. Sozialerhebung (1982-2016). Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssypool1021\\$/attachments/Doku_Harmonisierung_Var_Uebersicht.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssypool1021$/attachments/Doku_Harmonisierung_Var_Uebersicht.pdf)

- Middendorff, E. & Wallis, M. (2022a). 10. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 1982. Hannover: DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy10\\$-1.0.0/attachments/ssy10_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy10$-1.0.0/attachments/ssy10_MethodReport_de.pdf)
- Middendorff, E. & Wallis, M. (2022b). 11. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 1985. Hannover: DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy11\\$-1.0.0/attachments/ssy11_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy11$-1.0.0/attachments/ssy11_MethodReport_de.pdf)
- Middendorff, E., & Wallis, M. (2022c). 12. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 1988. Hannover: DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy12\\$-1.0.0/attachments/ssy12_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy12$-1.0.0/attachments/ssy12_MethodReport_de.pdf)
- Middendorff, E., & Wallis, M. (2022d). 13. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 1991. Hannover: DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy13\\$-1.0.0/attachments/ssy13_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy13$-1.0.0/attachments/ssy13_MethodReport_de.pdf)
- Middendorff, E. & Wallis, M. (2022e). 13.-21. Sozialerhebung. Daten- und Methodenbericht zum gepoolten Datensatz der neun Studierendenbefragung 1991-2016. Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW). [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssypool\\$-2.0.1/attachments/ssypoolv2.0.0_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssypool$-2.0.1/attachments/ssypoolv2.0.0_MethodReport_de.pdf)
- Middendorff, E., & Wallis, M. (2021a). 14. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 1994. Hannover: DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy14\\$-1.0.0/attachments/ssy14_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy14$-1.0.0/attachments/ssy14_MethodReport_de.pdf)
- Middendorff, E., & Wallis, M. (2021b). 15. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 1997. Hannover: FDZ-DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy15\\$-1.0.0/attachments/ssy15_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy15$-1.0.0/attachments/ssy15_MethodReport_de.pdf)
- Middendorff, E. & Wallis, M. (2021c). 16. Sozialerhebung. Daten- und Methodenbericht zur Studierendenbefragung 2000. Hannover: DZHW. [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy16\\$-1.0.0/attachments/ssy16_MethodReport_de_bf.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssy16$-1.0.0/attachments/ssy16_MethodReport_de_bf.pdf)
- Middendorff, E. & Wallis, M. (2021d). 17.-21. Sozialerhebung. Daten- und Methodenbericht zum gepoolten Datensatz der fünf Studierendenbefragung 2003-2016. Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW). [https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssypool\\$-1.0.1/attachments/ssypool_MethodReport_de.pdf](https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-ssypool$-1.0.1/attachments/ssypool_MethodReport_de.pdf)
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(2), 581–592.