Daten- und
Methodenbericht
Juni 2025

Verena Jahn | Heike Spangenberg | David Ohlendorf | Dennis Föste-Eggers | Johanna Niebuhr | Sandra Vietgen | Thorsten Euler

DZHW-Studienberechtigten-panel 2012

Daten- und Methodenbericht zur 3. Befragungswelle des Studienberechtigtenjahrgangs 2012

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

fdz.DZHW.
Forschungsdatenzentrum für Hochschul- und Wissenschaftsforschung

Autor*innen
Verena Jahn
Dr. Heike Spangenberg
Dr. David Ohlendorf
Dennis Föste-Eggers
Johanna Niebuhr
Sandra Vietgen
Dr. Thorsten Euler

Impressum

Herausgegeben von
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

Geschäftsführerin
Prof. Dr. Monika Jungbauer-Gans

Vorsitzender des Aufsichtsrats
Ministerialdirigent Peter Greisler

Registergericht
Amtsgericht Hannover | B 210251
Umsatzsteuer-Identifikationsnummer:
DE291239300

Juni 2025

Inhaltsverzeichnis

Tabellen-/Abbildungsverzeichnis

1 Einleitung

Die DZHW-Studienberechtigtenbefragungen sind eine Untersuchungsreihe zu den Bildungs-, Berufs- und Lebenswegen von Schulabsolvent*innen mit einer schulischen Hochschulzugangsberechtigung.[1] Sie werden durch das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung GmbH (DZHW) durchgeführt, vom Bundesministerium für Bildung und Forschung (BMBF) gefördert und dienen – in Ergänzung zur amtlichen Hochschulstatistik – dem nationalen Bildungsmonitoring sowie der Beantwortung von Fragestellungen der Hochschul- und Wissenschaftsforschung. Seit 1976 werden ausgewählte Schulabschlussjahrgänge befragt.

Die Daten einiger Studienberechtigtenjahrgänge werden zum Zweck der Nachnutzung aufbereitet und dokumentiert. Sie werden über verschiedene Zugangswege als Scientific Use Files (SUF) für die wissenschaftliche Sekundärnutzung und als Campus Use Files (CUF) für Lehr- und Übungszwecke zur Verfügung gestellt. Neben den Datensätzen der Erhebungen werden auch Dokumentationsmaterialien zu den Datensätzen und zur Durchführung der Studien bereitgestellt.[2]

Der vorliegende Daten- und Methodenbericht ist Teil der Dokumentation zur dritten Befragungswelle des Studienberechtigtenpanels 2012. Weitere Dokumentationsmaterialien zur Studie (Datensatzreports, Fragebögen etc.) können frei im Metadatensuchsystem des FDZ-DZHW (https://metadata.fdz.dzhw.eu) heruntergeladen werden.

Das Kapitel 2 stellt Inhalt und Anlage aller Studienberechtigtenbefragungen bis 2012 im Allgemeinen und des Studienberechtigtenpanels 2012 im Speziellen vor. Die weitere Gliederung des Berichts orientiert sich im Wesentlichen am Ablauf des Forschungsprozesses. In Kapitel 3 werden die eingesetzten Erhebungsinstrumente und in den Kapiteln 4 bis 7 der Erhebungsprozess beschrieben (Stichprobenziehung, Erhebungsablauf, Rücklauf, Datenaufbereitung). In den Kapiteln 8 und 9 folgt die Darstellung der vorgenommenen Gewichtung und Anonymisierung.

[1] Aktuelle Informationen zum DZHW-Studienberechtigtenpanel können über die Webseite des Projektes (http:/bildungswege.dzhw.eu) abgerufen werden.
[2] Informationen zu verfügbaren Datensätzen und Dokumentationen werden auf der Webseite des FDZ-DZHW (https://fdz.dzhw.eu) zur Verfügung gestellt.

2 Inhalt und Anlage der Studie

Das DZHW-Studienberechtigtenpanel 2012 ist Teil der DZHW-Befragungsreihe zu Studienberechtigten, in der anhand von standardisierten Mehrfachbefragungen sowohl Informationen zum Übergang von Studienberechtigten von der Schule in Studium und Berufsausbildung als auch die nachschulischen Bildungs- und Erwerbsverläufe erfasst werden. Das erste Studienberechtigtenpanel wurde 1976 durchgeführt, seitdem wurden 19 Studienberechtigtenjahrgänge (Kohorten) untersucht. Die Grundgesamtheit der Kohorten umfasst jeweils die (angehenden) Schulabsolvent*innen, die an allgemeinbildenden und berufsbildenden Schulen in Deutschland die allgemeine bzw. fachgebundene Hochschulreife, die volle Fachhochschulreife bzw. die fach- oder landesgebundene Fachhochschulreife erwerben.

In der Regel werden für jeden Studienberechtigtenjahrgang mehrere Befragungswellen zu unterschiedlichen Zeitpunkten vor und nach dem Erwerb der Hochschulzugangsberechtigung durchgeführt. Es handelt sich somit um ein Multikohorten-Panel-Design. Die Befragungen der Studienberechtigtenkohorten umfassen ein bis vier Wellen. Bis 1986 verfolgte das DZHW das Ziel, ausgewählte Studienberechtigtenjahrgänge kurz nach sowie 2 ½, 4 ½ und 12 ½ Jahre nach Schulabgang zu befragen. Ab 1990 wurde der Befragungszeitpunkt „2 ½ Jahre nach Schulabgang“ gestrichen. Seit 2005 wird eine zusätzliche Befragung ein halbes Jahr vor dem Erwerb der Hochschulzugangsberechtigung (HZB), also noch in der Schulabschlussklasse durchgeführt. Nach dieser ersten Befragungswelle folgt die zweite Welle einer Kohorte, die ein halbes Jahr nach dem Erwerb der HZB erfolgt. Die dritte Befragungswelle folgte zuletzt in der Regel etwa viereinhalb Jahre nach dem Schulabschluss. Die dritte Welle der Kohorte 2012 fand jedoch in Abweichung davon erstmals sechseinhalb Jahre nach Schulabgang statt. Über eine Spanne von 40 Jahren bis zur vorliegenden Befragung entstand so eine Zeitreihe von 19 Jahrgängen mit insgesamt 47 Befragungswellen (vgl. Abbildung 1).

Abbildung 1: Zeitreihe von befragten Abschlussjahrgängen des Studienberechtigtenpanels seit 1976 bis 2020

Die dritte Erhebung des Jahrgangs 2012 wurde als schriftlich-postalische Paper-and-Pencil-Befragung (PAPI) und wahlweise auch als Online-Befragung (CAWI) durchgeführt (alternatives Mixed-Mode-Design). Das Erhebungsinstrument enthält Fragen zum bisherigen Tätigkeitsverlauf, zu Studium, Berufsausbildung und beruflicher Fortbildung, zu Intentionen für nachfolgende Qualifizierungen, zum Übergang vom Bachelor- ins Masterstudium sowie in den Beruf, zu Berufszielen, zum Erwerbsverlauf und zur aktuellen Erwerbstätigkeit. Zudem wurden monetäre und nicht-monetäre Bildungserträge im Vergleich von Studien- und Berufsausbildungsabsolvent*innen vertiefend erhoben. Der wellenübergreifende, thematische Schwerpunkt der Studienberechtigtenbefragung 2012.3 ist Diversität mit besonderem Fokus auf dem Migrationshintergrund.

[Analysepotential] Mit Kohortenvergleichen können langfristige Trends der nach-schulischen Werdegänge in den Blick genommen werden. Zudem wird innerhalb einer Kohorte ein Teil der Fragen in den verschiedenen Befragungswellen wiederholt gestellt. Dies ermöglicht die Betrachtung intra-individueller Veränderungen zwischen den Wellen. Darüber hinaus werden in Abhängigkeit von aktuellen Entwicklungen und Forschungsinteressen in einzelnen Kohorten bestimmte Aspekte vertiefend oder ergänzend abgefragt. Des Weiteren wurden monatsgenaue kontinuierliche Verlaufsdaten zu den seit Schulabschluss ausgeübten Tätigkeiten erhoben, die sich für Ereignis- und Sequenzmusteranalysen eignen. Die Daten der Studienberechtigtenpanels werden in der Regel anhand der Merkmale Geschlecht, Bundesland, Art der Schule und der Hochschulreife gewichtet und an die Grundgesamtheit angeglichen. Zusätzlich werden paneltypische Ausfallprozesse in der Gewichtung der Daten berücksichtigt.

[Einordnung ins Forschungsfeld] Das Stichproben- und Erhebungsdesign sowie die damit verbundenen Analysemöglichkeiten unterscheiden das DZHW-Studienberechtigtenpanel von anderen in Deutschland durchgeführten Befragungen von Studienberechtigten. Keine andere Befragung ermöglicht bundesweite Analysen. Zusätzlich weisen andere Erhebungen in diesem Feld keine oder deutlich kürzere Zeitreihen auf. Andere Studienberechtigtenbefragungen sind beispielsweise die sächsische Abiturientenbefragung[3] (durchgeführt durch das Kompetenzzentrum für Bildungs- und Hochschulplanung an der TU Dresden), die TOSCA-Studie[4] (inzwischen durchgeführt vom Hector-Institut für Empirische Bildungsforschung an der Universität Tübingen) sowie das Berliner-Studienberechtigtenpanel Best Up[5] (durchgeführt vom Deutschen Institut für Wirtschaftsforschung (DIW) und dem Wissenschaftszentrum Berlin (WZB)).

[Spezifika des Studienberechtigtenpanels 2012] Neben den allgemeinen Charakteristika der Studienreihe weist die hier betrachtete Studienberechtigtenkohorte 2012 folgende Spezifika auf: Erstmalig erfolgte die dritte Befragungswelle erst sechseinhalb Jahre nach Schulabschluss, sodass zwischen der zweiten und dritten Erhebung ein sechs-jähriger Abstand liegt. Zudem erfolgte die Befragung ohne Beteiligung des Saarlands. In den Bundesländern Baden-Württemberg, Bremen, Berlin und Brandenburg wurden 2012 an allgemeinbildenden Gymnasien zwei Jahrgänge zum Abitur geführt, nach neun- bzw. achtjähriger Schulzeit (G9/G8). Thematisch wird ein Schwerpunkt auf Diversität mit einem besonderen Fokus auf Migrationshintergrund gelegt.

[3] vgl. dazu https://tu-dresden.de/zqa/forschung/Forschungsprojekte/saechsische-abiturientenstudie
[4] vgl. dazu https://uni-tuebingen.de/fakultaeten/wirtschafts-und-sozialwissenschaftliche-fakultaet/faecher/fachbereich-sozialwissenschaften/hector-institut-fuer-empirische-bildungsforschung/forschung/aktuelle-studien/tosca/
[5] vgl. dazu https://www.wzb.eu/de/forschung/dynamiken-sozialer-ungleichheiten/ausbildung-und-arbeitsmarkt/best-up

3 Erhebungsinstrumente

In der dritten Befragungswelle 2012 wurden als Erhebungsinstrumente ein standardisierter Papierfragebogen und ein Online-Fragebogen eingesetzt.[6] Kapitel 3.1 stellt die zentralen Inhalte der Erhebungsinstrumente vor. Kapitel 3.2 beschreibt die zur Prüfung und Verbesserung der Fragebögen durchgeführten Pretests.

3.1 Inhalte der Erhebungsinstrumente

[Charakteristika der Studienreihe] Im Fokus des Studienberechtigtenpanels 2012 steht, wie bei den übrigen Kohorten der Studienreihe, die Beschreibung und Erklärung von Bildungsentscheidungen. Da bei der Studienberechtigtenkohorte 2012 erstmalig ein neuer dritter Befragungszeitpunkt (sechseinhalb Jahre nach Schulabschluss) gewählt wurde, der später in der Bildungsbiografie liegt als bei vorherigen Kohorten, erfolgte zudem eine Schwerpunktsetzung auf Fragen zum bisherigen Qualifizierungsverlauf, zu geplanten Weiterqualifizierungen und insbesondere auf Fragen zur Erwerbstätigkeit. Dabei wurden in Vorbereitung der Student Life Cycle Panel-Studie einige Fragen identisch zur Absolventenbefragung 2017[7] erhoben.

Die Fragen der dritten Welle beziehen sich auf sechs Themenkomplexe: Fragen zu Persönlichkeit und Werdegang, Ausbildungs- und Studienverlauf, Einstellungen und Engagement, Erwerbstätigkeit und Erwerbsverlauf, Weiterbildung und berufliche Zukunftspläne sowie Personenangaben. Im Detail enthalten diese sechs Themenkomplexe folgende Inhalte:

Tabelle 1: Themenkomplexe des DZHW-Studienberechtigtenpanels 2012 (Welle 3)

Themenkomplex	Zentrale Inhalte
Persönlichkeit und Werdegang	Wichtigkeit von Lebensbereichen, Risikobereitschaft; Persönlichkeitseigenschaften; monatsgenaue Erfassung der Tätigkeiten seit Schulabschluss; idealistische und realistische Berufsaspirationen; Berufs- und Lebensziele; Lebenszufriedenheit
Ausbildungs- und Studienverlauf	Bisherige Qualifizierungsschritte (einschl. Zeitraum, Hauptstudienfächer bzw. Berufsbezeichnung, angestrebte Abschlussprüfung, Name und Ort der Hochschule bzw. des Aus-bildungsbetriebes, Abschussnote); geplante Qualifizierungen (Master, Promotion, Aufstiegsfortbildung, etc.); Realistische Berufsaspirationen; Selbsteinschätzung Kompetenzen

[6] Der Papierfragebogen kann von der Website des FDZ heruntergeladen werden.
[7] vgl. dazu https://www.dzhw.eu/pdf/22/ap2017.pdf

Einstellungen und Engagement	Interesse an Politik; Freizeitverhalten und ehrenamtliches Engagement; Einstellungen zu beruflichen Chancen von Menschen mit Migrationshintergrund
Erwerbstätigkeit und -verlauf	Erwerbstätigkeit nach Erwerb der Hochschulzugangsberechtigung; Suchverhalten und Schwierigkeiten bei der Stellensuche; tabellarische Erfassung aller seit Schulabschluss ausgeübten Erwerbstätigkeiten (inkl. Zeitraum, Art, Stundenumfang, beruflicher Stellung); zusätzlich zur aktuellen Erwerbstätigkeit: berufliche Zufriedenheit, Adäquanz, Einkommen, Anforderungsprofil
Weiterbildung und berufliche Zukunftspläne	Einstellungen zu Weiterbildung; absolvierte Weiterbildungen innerhalb der letzten 12 Monate
Personenangaben	Beziehungsstatus, beruflicher Abschluss Partner*in; Kinder; Religionszugehörigkeit; Geburtsland der Großeltern; aktueller Wohnort; Selbsteinschätzung des Gesundheitszustands

[Spezifika des Studienberechtigtenpanels 2012] Da das Studienberechtigtenpanel 2012 das Thema Diversität mit einem besonderen Fokus auf Migrationshintergrund als einen Hauptschwerpunkt betrachtet, werden sowohl Indikatoren zur Bestimmung des Migrationshintergrundes (Geburtsland der Studienberechtigten sowie von deren Eltern und Großeltern, im Elternhaus gesprochene Sprache, Staatsangehörigkeit der Studienberechtigten), als auch Einstellungen gegenüber der Situation von Migrant*innen in Deutschland erfragt. Ein weiteres Spezifikum des Studienberechtigtenjahrgangs 2012 ist der im Jahr 2012 doppelte Abiturjahrgang (G8/G9) an den allgemeinbildenden Gymnasien in Baden-Württemberg, Bremen, Berlin und Brandenburg. In der ersten Welle wurde deshalb die Gymnasialschuldauer (8 bzw. 9 Jahre) erhoben, sodass alle Analysen in der Differenzierung G8/G9 möglich sind. Die dritte Erhebungswelle der Studienberechtigtenkohorte 2012 fand einmalig erst sechseinhalb Jahre nach Schulabschluss statt und somit vergleichsweise spät in der Bildungsbiografie bzw. bereits am Übergang in die Erwerbstätigkeit. Dies wurde im Zusammenhang mit der geplanten Zusammenführung von DZHW-Studienberechtigten- und Absolventenuntersuchungen im Student Life Cycle Panel (SLC) zum Anlass genommen, im Studienberechtigtenpanel 2012 und der Absolventenkohorte 2017 zum Teil identische Instrumente einzusetzen.[8] Anders als bei zuvor durchgeführten dritten Erhebungswellen des Studienberechtigtenpanels wurde schließlich für alle Tätigkeiten seit Erwerb der Hochschulzugangsberechtigung zusätzlich erhoben, ob sie im Ausland ausgeübt wurden.

3.2 Pretests

[Ziel und Verfahren] Die Erhebungsinstrumente wurden im Vorfeld durch Pretests geprüft. Dabei sollte erstens für die bereits in vorherigen Kohorten und den DZHW-Absolventenbefragungen eingesetzten Fragen und Antwortvorgaben geprüft werden, ob sie von dem Studienberechtigtenjahrgang 2012 gleich perzipiert werden. Zweitens sollte für die neu eingesetzten Messinstrumente deren Verständlichkeit, Beantwortbarkeit, theoretische Aussagekraft, Reliabilität und Validität getestet werden.

[Probanden] Zur Überprüfung des Papierfragebogens wurde die erste Version der Erhebungsinstrumente zunächst im Rahmen von acht internen und anschließend von zwei externen Expertenbewertungen durch wissenschaftliche Mitarbeiter*innen und Experten aus dem Bereich der Hoch-

[8] Zur Dokumentation der Herkunft sekundär genutzter Instrumente siehe die Übersicht unter: https://metadata.fdz.dzhw.eu/public/files/instruments/ins-gsl2012-ins3$-3.0.0/attachments/gsl2012_W3_literature_secondary_used_instruments_de.pdf

schulforschung begutachtet. Darüber hinaus wurden ausgewählte (neu entwickelte oder angepasste) Instrumente sowie das Layout des Papierfragebogens im Rahmen von fünf kognitiven Pretests mit Proband*innen der Zielgruppe geprüft. Neben diesen Bewertungen fanden neun weitere Zielgruppenpretests statt. Diese wurden als sogenannter Pretest im Feld, also unter möglichst ähnlichen Bedingungen wie in den tatsächlichen Befragungen, durchgeführt.

[Durchführung] Die Phase des Zielgruppenpretests im Feld fand von Ende Oktober bis Anfang November 2017 statt. Dabei wurden an die ausgewählten Tester*innen Fragebögen mit einer Durchführungsanweisung postalisch versendet, mit der Bitte, Verständnisprobleme, Kritik oder Anmerkungen zu notieren. Im Anschluss an die Bearbeitung des Fragebogens sollten die Befragten einen Feedbackfragebogen zu verschiedenen Aspekten der Befragung ausfüllen. Dabei wurden Informationen zur Ausfülldauer, zu Inhalt und Länge des Fragebogens, zu Aufbau und Layout, zur Verständlichkeit der Fragen und Ausfüllanweisungen sowie zur Vollständigkeit der Antwortmöglichkeiten erhoben. Auf Grundlage der Ergebnisse der Pretests wurden die Formulierungen verschiedener Fragetexte präzisiert sowie die Reihenfolge von Items und Antwortkategorien in einzelnen Itembatterien und Mehrfachnennungen überarbeitet. Zudem wurde die Länge des Fragebogens verringert, indem Itembatterien gekürzt und Fragen gestrichen wurden. Der grundsätzliche Aufbau hingegen blieb unverändert.

4 Grundgesamtheit und Stichprobenverfahren

[Grundgesamtheit und Inferenzpopulation] Die Grundgesamtheit des Studienberechtigtenpanels 2012 umfasst alle Schulabsolvent*innen, die im Schuljahr 2011/2012 an allgemeinbildenden und berufsbildenden Schulen in Deutschland die allgemeine bzw. die fachgebundene Hochschulreife, die volle Fachhochschulreife bzw. die fach- oder landesgebundene Fachhochschulreife erworben haben.

[Stichprobenverfahren] Die erste Welle dieser Panel-Befragung fußt auf einer disproportional gezogenen geschichteten zufälligen Klumpenstichprobe von 894 Schulen bzw. Schulzweigen (Schneider & Franke, 2014, S. 1). Der Umfang der Stichprobe betrug 66.750 Personen. Insgesamt wurden somit rund 13 Prozent der Grundgesamtheit in die Untersuchung einbezogen. 34.465 Befragte beteiligten sich an der ersten, im Schulabschlussjahr durchgeführten Befragung. Für die ein Jahr später durchgeführte zweite Befragung lagen Adressangaben von 27.277 Studienberechtigten vor, von denen 11.686 erneut an der Befragung teilnahmen. Sie wurden im Dezember 2018 schließlich zur dritten Erhebung des Studienberechtigtenjahrgangs 2012 eingeladen.

5 Durchführung der Erhebung

[Kontaktaufnahme und Adresspflege] Für die dritte Befragung des Studienberechtigtenjahrgangs 2012 wurden alle Personen kontaktiert, die sich bereits an den beiden ersten Erhebungswellen beteiligt hatten. Die Postadressen der verbliebenen Teilnehmer*innen der zweiten Befragung des DZHW Studienberechtigtenpanels 2012 wurden vor Beginn der Feldphase zunächst einer Prüfung durch die Deutsche Post AG unterzogen, da zwischen den beiden Befragungen sechs Jahre lagen. Dadurch konnten bereits 257 notwendige Adressaktualisierungen durchgeführt werden. Im Anschluss daran wurde an 11.631 Teilnehmer*innen aus der zweiten Befragungswelle postalisch vom DZHW ein Schreiben inklusive der Befragungsunterlagen versendet, in dem um Teilnahme gebeten wurde. Den vom DZHW versendeten Unterlagen lagen bereits frankierte Briefumschläge sowie Kugelschreiber mit Touchpen-Funktion bei, sodass sowohl ein unkompliziertes Ausfüllen, als auch eine kostenfreie Rücksendung für die Befragten möglich war. Zudem war auf dem Papierfragebogen der Link zur Online-Befragung aufgedruckt, sodass alternativ zur Papierbefragung eine Online-Teilnahme möglich war.

Um die weiterhin teilnahmebereiten Personen für potentiell nachfolgende Untersuchungen des DZHW kontaktieren zu können, wurden im Fragebogen geänderte Kontaktdaten (Postanschrift und/oder E-Mail-Adresse) erfragt. Beim Eingang eines Fragebogens im DZHW wurde sowohl auf den Fragebogen als auch auf den Adressabschnitt des Fragebogens per Paginierstempel eine eindeutige Identifikationsnummer gestempelt.[9] Auch für die Online-Teilnehmer*innen liegen entsprechende Kontaktdaten vor.

[Erhebungsunterlagen] Die Erhebungsunterlagen der dritten Welle bestanden pro zu befragender Person aus einem Einladungsanschreiben (inkl. Datenschutzinformationen), einem Fragebogen, einer Studienfachliste, einem Kugelschreiber mit Touchpen-Funktion und einem an das DZHW adressierten portofreien Umschlag zur Rücksendung des ausgefüllten Fragebogens. Der Fragebogen enthielt auf dem Deckblatt einen Zugangscode für die Teilnahme an der alternativ angebotenen Onlinebefragung. Das alternative Mixed-Mode-Design mit Erstkontakt per Papierfragebogen wurde gewählt, da ausschließlich Postadressen aus den Vorwellen vorlagen, aber dennoch auch Befragte zur Teilnahme motiviert werden sollten, die eine Online-Teilnahme präferieren.

Im Februar 2019 wurde ein erstes Erinnerungsschreiben versandt, in dem auch nochmals auf eine mögliche Online-Teilnahme verwiesen wurde. Dem Schreiben lag ein Informationsflyer bei, der über die Studie und bereits veröffentlichte Ergebnisse aus vorangegangenen Wellen informierte. Ein zweites Erinnerungsschreiben, inklusive eines Fragebogenexemplars, Fächerliste und frankiertem Rückumschlag wurde im Monat darauf versandt. Die letzte Erinnerung erfolgte dann Anfang April, in Form einer Motiv-Postkarte. Dort wurde auf die letzte Teilnahmemöglichkeit hingewiesen.

[9] Zur Gewährleistung des Datenschutzes wurde der Adressabschnitt bei Eintreffen vom Fragebogen abgetrennt und nach der Erfassung getrennt von den Befragungsdaten auf einem geschützten Server gespeichert und verarbeitet.

[Feldphase] Der Erhebungszeitraum der dritten Befragungswelle erstreckte sich von Dezember 2018 bis Mitte des Jahres 2019. Aufgrund des langen Zeitraums, der zwischen der zweiten und dritten Befragung dieser Kohorte lag, mussten intensivere Adressrecherchen durchgeführt werden, was die Feldphase durch Wartezeiten bei Einwohnermeldeamts-Abfragen verlängerte. Die ersten Erinnerungsschreiben wurden Anfang Februar 2019 verschickt, wovon jedoch 1.763 Stück nicht direkt zustellbar waren. Für 1.706 dieser Adressen musste eine Einwohnermeldeamts-Abfrage gestartet werden, um so umgezogene Proband*innen dennoch zu erreichen. Anfang März 2019 wurden alle bis dahin erreichbaren Teilnehmer*innen ein zweites Mal postalisch mit einem weiteren Fragebogenexemplar erinnert. Die dritte und letzte Erinnerung erfolgte Anfang April 2019 in Form einer Postkarte mit dem Zugangscode für die Teilnahme am Online-Fragebogen.

[Rücklaufsteigernde Maßnahmen] Die rücklaufsteigernden Maßnahmen zielten auf die individuelle Motivation der Befragten ab. Jede*r Befragte erhielt als Prepaid-Incentive einen Touchpen-Kugelschreiber mit dem Erstanschreiben. Dem ersten Erinnerungsschreiben wurde ein Flyer mit Ergebnissen der Vorwellen beigelegt. Zudem wurde alternativ zum Papierfragebogen ein Onlinefragebogen angeboten. Außerdem wurde eine Webseite mit Informationen zur Studienberechtigtenbefragung erstellt (www.bildungswege.dzhw.eu). Als materieller individueller Anreiz wurden unter allen teilnehmenden Schulabsolvent*innen ein (Tablet-)PC im Wert von 1.000 €, ein Set Kopfhörer im Wert von 300 €, ein Bahngutschein im Wert von 200 € und 20 Buchgutscheine im Wert von jeweils 50 € verlost.

6 Rücklauf

[Rücklauf] Auf der Grundlage der 11.685 Teilnehmer*innen[10] der zweiten Welle, wurden für die dritte Erhebungswelle 11.659 Proband*innen angeschrieben. Von diesen sandten 6.120 einen auswertbaren Papierfragebogen zurück oder nahmen online an der Befragung teil, was einer Netto-Rücklaufquote von 52,5 Prozent für die dritte Befragung entspricht. Bereinigt um die trotz Adressrecherchen nicht zustellbaren Befragungsunterlagen (627) und verstorbene Proband*innen (11), erhöht sich die bereinigte Rücklaufquote auf 56 Prozent.

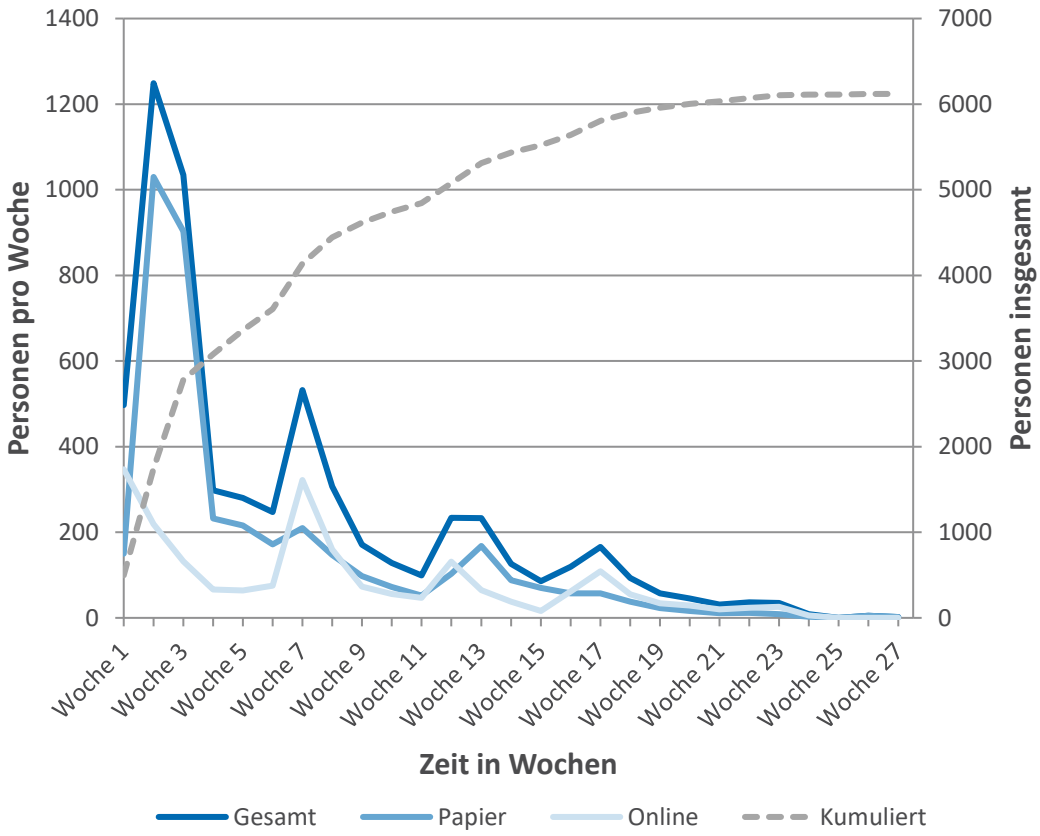
Tabelle 2: Brutto-, Nettostichproben und Rücklaufquoten des DZHW-Studienberechtigtenpanels 2012 (Welle 3)

Quotenart	Anzahl
Bruttostichprobe	11.685
Nettostichprobe	11.659
Nettorücklauf absolut	6.120
Rücklaufquote	52,5 %
Bereinigte Rücklaufquote	56,0 %

Wie Abbildung 2 zeigt, wurde die Hälfte der eingegangenen Fragebögen innerhalb der ersten vier Befragungswochen (20.12.2018 bis 16.01.2019) zurückgesandt. Aufgrund des durch zusätzliche Adressrecherchen teilweise verzögerten Versands von Einladungen wurde die Feldphase bis Juni verlängert.

[10] Darunter auch Personen ohne Angabe einer gültigen Adresse oder Personen, die nach einer Adressprüfung durch Premiumadress als verstorben gemeldet wurden.

Abbildung 2: Rücklauf des DZHW-Studienberechtigtenpanels 2012 im Zeitverlauf (Welle 3)



dargestellt sind nur die auswertbaren Fälle

[Panelausfälle] Das Studienberechtigtenpanel 2012 ist von paneltypischen Ausfallprozessen betroffen. Hier sind die grundsätzliche Verweigerung der Teilnahme an Folgebefragungen, veraltete oder falsche Adressangaben bei Kontaktierungsversuchen und das Versterben von Proband*innen zu nennen. Um diese Panelausfälle zu kompensieren, können die zur Verfügung gestellten Ausfallgewichte genutzt werden (Kapitel 8).

7 Datenaufbereitung

Im Folgenden werden die verschiedenen Schritte der Datenaufbereitung beschrieben. Die Aufbereitungsprozesse Gewichtung und Anonymisierung werden in den beiden folgenden Kapiteln 8 und 9 gesondert erläutert.

7.1 Datenübertragung

Zur weiteren Verarbeitung wurden die Angaben der Befragten aus den Papierfragebogen auf Basis eines Codeplans in computerlesbares Format (*.csv = comma-separated values*) übertragen. Zuvor wurden auf den Papierfragebögen numerische Codierungen für einen Teil der offenen Angaben vermerkt (offene Angaben zu Berufen wurden im Klartext erfasst, vgl. Kapitel 7.2). Zusätzlich wurden bereits manuelle Vorkorrekturen zur Erleichterung der Datenübertragung als auch erste Plausibilitätschecks der Daten vorgenommen (vgl. Kapitel 0). Für die online ausgefüllten Fragebögen erübrigt sich die Datenübertragung, da die Angaben hier bereits in digitaler Form vorliegen. Vorkorrekturen und erste Plausibilitätsprüfungen wurden bei den online ausgefüllten Fragebögen nach denselben Regeln wie auch bei den Papierfragebögen durchgeführt, allerdings über ein syntaxbasiertes Verfahren.

[Erstellung eines Codeplans] Auf Basis des Fragebogens der Papierbefragung wurde ein Codeplan erstellt. Dabei wurde vermerkt, welcher Frage bzw. Teilfrage eine Variable zugeordnet ist, welchen Namen diese Variable trägt und welche numerischen Codierungen für die standardisierten Antworten der Befragten verwendet werden. Um die Erfassungsreihenfolge festzulegen, wurden die Variablen zusätzlich nummeriert.

[Datenerfassung] Für die Datenübertragung wurden der Codeplan, weitere Anweisungen zur Datenerfassung sowie die vorbereiteten Papierfragebögen an einen externen Dienstleister übergeben. Die Erfassung der Angaben erfolgte dort manuell durch Schreibkräfte. Eine Ausnahme stellt die Erfassung von Frage 1.6, dem „Kalendarium“ zur Eintragung der monatlichen Tätigkeiten seit Erwerb der Hochschulzugangsberechtigung dar. Die Angaben wurden fallweise nach vorheriger Anleitung von studentischen Hilfskräften in hierfür erstellte, standardisierte Excel-Eingabemasken übertragen, die im Anschluss durch eine Syntax automatisiert eingelesen wurden.

7.2 Codierung offener Angaben

Vor der Datenübertragung erfolgte eine Codierung der (halb-)offenen Angaben. Hierzu wurden diesen anhand von Codierlisten numerische Codierungen zugeordnet. Bei diesen Listen handelt es sich um Klassifikationsschlüssel der amtlichen Statistik (Klassifikation der Berufe, Schlüsselverzeichnis der Studenten- und Prüfungsstatistik etc.) oder um bereits in vorherigen Wellen und Kohorten eingesetzte projekteigene Schlüssel (siehe Tabelle 3). Für einige Variablen, insbesondere zu halboffenen Angaben, wurden neue projekteigene Codierlisten entwickelt. Das Vorgehen variiert dabei leicht, je nach Variablen:

- Sämtliche Angaben zu Berufen (einschließlich Wunschberuf, Ausbildungsberuf, aktueller Beruf) wurden zunächst im Klartext erfasst und anschließend so weit wie möglich automatisiert über ein syntaxbasiertes Vorgehen codiert (Föste-Eggers 2021). Angaben, die nicht auf diese Weise codiert werden konnten, wurden anschließend durch geschulte studentische Hilfskräfte anhand einer Codieranleitung und Codierlisten manuell nachcodiert.
- Alle übrigen offenen Angaben (zu Hochschulen, Studienfächern, Wohnort, Arbeitsort und den Geburtsländern der Großeltern) wurden in den Papierfragebögen bereits vor der Erfassung codiert. Für die online ausgefüllten Fragebögen wurden die offenen Angaben zunächst in Excel-Listen exportiert, ebenfalls manuell codiert und anschließend wieder an den finalen Datensatz angefügt.
- Alle halboffenen Angaben wurden anhand projekteigener Codierlisten soweit möglich den vorhandenen geschlossenen Antwortkategorien zugeordnet oder mit erweiterten Codes versehen. Die Codierlisten wurden auf Basis der in den Daten des Studienberechtigtenjahrgangs 2012 vorkommenden Nennungen neu erstellt. Lediglich für die Frage 6.3 wurde für die korrekte Zuordnung auf die Klassifikation der Wirtschaftszweige des Statistischen Bundesamtes von 2008 zurückgegriffen.

In der folgenden Tabelle 3 sind die codierten Merkmale sowie die jeweils verwendeten Codierlisten dargestellt. Die Ausprägungen der einzelnen Variablen sind im Datensatzreport dokumentiert. Der Datensatz beinhaltet ausschließlich die codierten numerischen Variablen, die offenen Nennungen selbst sind aus Gründen des Datenschutzes nicht im finalen Datensatz enthalten.

Tabelle 3: Vercodete Merkmale und verwendete Codierlisten im Studienberechtigtenpanel 2012 (Welle 3)

Merkmal	Codierliste
Berufsbezeichnung (inkl. Ausbildungs-/Wunschberuf)	Destatis Klassifikation der Berufe 2010
Studienfächer	Destatis Schlüsselverzeichnis für die Studenten- und Prüfungsstatistik (WiSe 2018/19 und SoSe 19), ergänzt um projektspezifische Codes
Hochschulen	Hochschulen in Deutschland: Destatis Schlüsselverzeichnis für die Studenten- und Prüfungsstatistik (WiSe 2018/19 und SoSe 19), ergänzt um projektspezifische Codes Hochschulen im Ausland: Projekteigene Codierung nach Land der Hochschule
Geburtsland der Großeltern	Projekteigene Codierung
Wohn- /Arbeitsort	PLZ-Verzeichnis der Deutschen Post
Schwierigkeiten bei Übergang ins Masterstudium; Schwierigkeiten bei der Stellensuche; Art der Stellensuche; Religionszugehörigkeit	Zuordnung zu geschlossenen Kategorien bzw. neu erstellten Kategorien nach projekteigenen Codierlisten
Wirtschaftszweige	Zuordnung zu geschlossenen Kategorien nach Destatis Klassifikation der Wirtschaftszweige von 2008

7.3 Datenprüfung und Datenbereinigung

[Manuelle Vorkorrektur] Bereits vor der Übertragung der Daten wurden auf den Papierfragebögen eine manuelle, einzelfallbasierte Prüfung und gegebenenfalls eine Anpassung von Angaben der Befragten durchgeführt.[11] Dies sollte vor allem die Erfassbarkeit der Daten erleichtern. Dafür wurde in erster Linie die Form der bestehenden Angaben verändert. Beispielsweise wurden schwer lesbare Angaben oder Streichungen der Befragten verdeutlicht oder Zahlenangaben rechtsbündig in die dafür vorgesehenen Kästchen eingetragen. Fehlende oder ungültige Angaben wurden mit den entsprechenden Missingcodes versehen (siehe Abschnitt 0). Unzulässige Mehrfachnennungen bei Fragen mit Einfachauswahlen wurden gestrichen und mit dem Missingcode -965 „ungültige Mehrfachnennung“ versehen. Wurde bei Skalen zwischen zwei Kästchen gekreuzt, sodass die Antwort keiner Kategorie eindeutig zuzuordnen war, wurde die Angabe gestrichen und der Missingcode -966 „nicht bestimmbar“ vergeben. Zusätzlich wurde bereits eine erste Korrektur inkonsistenter Angaben innerhalb des Fragebogens durchgeführt. Dies umfasste zum Beispiel den Abgleich von Zeiträumen zu Qualifikations- und Erwerbsphasen zwischen einzelnen Fragekomplexen sowie die Überprüfung der Filterführung und Wertebereiche. Die entsprechenden Codieranweisungen wurden in einer Anleitung[12] dokumentiert. Die Durchführung der Codierung geschah fallweise durch zuvor geschulte studentische Hilfskräfte und wurde von mindestens einer weiteren Person geprüft.

[Softwaregestützte Korrektur] Im Anschluss an die Datenübertragung erfolgte eine umfassende Prüfung und Korrektur der Daten mit Hilfe einer DZHW-eigenen Software. Dabei sollten zum einen mögliche Fehler bei der vorherigen manuellen Vorkorrektur und Datenübertragung identifiziert werden und zum anderen zusätzliche Inkonsistenzen aufgedeckt werden, die sich durch einen Abgleich der Angaben im Fragebogen der dritten Welle mit den Angaben aus den Fragebögen der beiden Vorwellen ergeben.

Zu diesem Zweck wurden die erfassten Fragebogendaten, sowie relevante Informationen aus den beiden Vorwellen in eine Datenbank eingelesen. Anschließend wurden anhand formaler Regeln gültige Wertebereiche und Antwortkombinationen definiert und geprüft. Folgende Typen von Prüfungen wurden vorgenommen:

- Prüfung von Wertebereichen: Es wurde geprüft, ob die erfasste Ausprägung einer Variablen in dem für diese Variable definierten Wertebereich lag. Beispielsweise durften bei Variablen mit 5er-Skalen nur Werte von 1 bis 5 auftreten; im Zuge der Codierung der offenen Berufsangaben mussten vergebene Codes 5-stellig sein.
- Prüfung der Einhaltung der Filterführung: Auf Grundlage der definierten Filterführung des Fragebogens wurde zum einen geprüft, ob für die jeweilige befragte Person Angaben zu erwarten gewesen wären, die aber nicht vorhanden waren (Vollständigkeitsprüfung), und zum anderen, ob für die jeweilige Person, Angaben vorhanden waren, die nicht hätten erfolgen dürfen (Filterverstöße).[13]
- Prüfung von Merkmalskombinationen: Es wurde die Konsistenz der Angaben innerhalb des Fragebogens sowie wellenübergreifend überprüft.

Insgesamt wurden mehrere hundert Konsistenzregeln definiert und getestet. Bei fehlenden, fehlerhaften oder unplausiblen Werten wurde zunächst mit Hilfe des Papierfragebogens geprüft, ob der entsprechende Wert falsch (bzw. nicht) übertragen worden war. War dies nicht der Fall, wurden auf Basis für die einzelnen Fragen spezifisch festgelegter Regeln sowie anderer Angaben im Fragebogen ggf. Umsetzungen oder Korrekturen vorgenommen. Die entsprechenden Regeln bzw. Codieranwei-

[11] Die Zahl der vorgenommenen Korrekturen wurde nicht zentral, sondern nur auf den Papierfragebogen dokumentiert und ist daher nicht systematisch rekonstruierbar.
[12] Die Codieranleitung kann bei Bedarf durch das Projekt zur Verfügung gestellt werden.
[13] Die Filterführung kann im Filterführungsdiagramm nachvollzogen werden.

sungen bei Inkonsistenzen wurden wiederum in einer Anleitung[14] festgehalten. Fehlerkorrekturen wurden auf den Fragebögen dokumentiert[15] und jeweils von mindestens einer weiteren Person geprüft.

[Löschung von Fällen] Vorsätzlich von den Befragten falsch ausgefüllte Fragebögen der dritten Welle wurden ebenso aus dem Datensatz entfernt wie Fragebögen, die nicht, oder aber weitgehend unvollständig ausgefüllt waren. Als Regel galt hier, dass ein Fall gelöscht wurde, sofern eine Person die Befragung vor Erreichen der Frage 1.6 („Kalendarium“) abgebrochen hatte oder aber keine Angaben auf sämtlichen zentralen Fragekomplexen zu Qualifikations- und Erwerbsphasen gemacht hatte (Fragen 1.6 2.1 und 5.1). Weiterhin wurden Fälle aus dem Datensatz entfernt, die doppelt eingegangen sind, da Befragte sowohl den Papierfragebogen als auch den Onlinefragebogen ausgefüllt haben. Hier wurde in der Regel der später eingegangene Fragebogen gelöscht, es sei denn der erste Bogen war unvollständig ausgefüllt. Schließlich wurden alle Fälle gelöscht, die nach Ende der Feldphase eingegangen sind (siehe Kapitel 5).

Insgesamt wurden nach den genannten Kriterien 33 Fälle der dritten Befragungswelle gelöscht. Im Rahmen der Datenbereinigung zur Erstellung des Scientific Use File wurden zudem Befragte entfernt, die gemäß Angaben der zweiten Welle nicht die ausreichende Studienberechtigung erreicht hatten und somit nicht Teil der Grundgesamtheit sind. Abschließend enthält der SUF 6.079 analysierbare Fälle für die dritte Welle.

7.4 Generierung von Variablen

Neben den Variablen, die die codierten Antworten der Befragten enthalten, beinhaltet der Datensatz zum Studienberechtigtenjahrgang 2012 auch generierte Variablen. Dabei handelt es sich zum einen um Variablen mit numerischen Codierungen von ursprünglich offenen Nennungen (vgl. Kapitel 7.2). Zum anderen wurden Variablen des Datenschutzes wegen verändert (vgl. Kapitel 9) und im Forschungsfeld häufiger benötigte Variablen aus den Werten einer Quellvariablen generiert (z. B. Aggregation der Studienfächer zu Studienbereichen und Fächergruppen oder Ableitung von Hochschultyp aus den Hochschulvariablen). Eine Übersicht aller für das Studienberechtigtenpanel 2012 generierten Variablen sowie eine Dokumentation der einzelnen Variablen mit Angabe ihrer jeweiligen Ausprägungen findet sich im Datensatzreport sowie im Metadatensuchsystem[16].

7.5 Erstellung der Datensätze

[Zusammenführung der Wellen] Die Daten der dritten Befragungswelle wurden mit den Daten der beiden Vorwellen zusammengeführt. Die Zuordnung der Fälle erfolgte über die bei der ersten Erhebungswelle vergebene Identifikationsnummer der Befragten.

[Erstellung von Personen- und Episodendatensatz] Die so zusammengeführten Daten wurden in zwei getrennten Datensätzen abgelegt. Der Personendatensatz enthält den Großteil der Befragungsdaten sowie die zusätzlichen generierten Variablen. Pro befragter Person existiert eine Datenzeile (wide-Format). Die Reihenfolge der Variablen orientiert sich an der Reihenfolge der zugehörigen Fragen im Fragebogen.

Der Episodendatensatz enthält die Antworten aus dem Kalendarium der dritten Welle (Frage 1.6). Für jede befragte Person werden eine oder mehrere Episoden gespeichert. Dabei ist eine Episode

[14] Die Anleitung kann bei Bedarf vom Projekt zur Verfügung gestellt werden.
[15] Die Dokumentation der Fehlerkorrekturen erfolgte handschriftlich auf den Papierfragebögen und ist daher nicht systematisch rekonstruierbar.
[16] https://metadata.fdz.dzhw.eu/

definiert als ein Zeitraum, in dem eine bestimmte Tätigkeitsart (z. B. Berufsausbildung, Studium) ausgeübt wird bzw. ein konkreter Status (z. B. Arbeitslosigkeit) besteht. Für jede Episode einer Person existiert jeweils eine Datenzeile (long-Format), welche monatsgenau den Anfang- und Endzeitpunkt der Tätigkeit und die Episodennummer angibt. Die Episoden wurden fallweise sortiert, das heißt alle Episoden einer Person folgen direkt aufeinander. Das Zusammenführen von Personendatensatz und Episodendatensatz wird über die Identifikationsnummer der Person (Variable: pid) ermöglicht.

[Datenformat] Die Datensätze werden standardmäßig im Stata-Format bereitgestellt. Möglichkeiten zur Nutzung weiterer Formate bzw. Analysesoftware sind der Website zu entnehmen.

[Erstellung eines Rohdatensatzes] Die codierten und erfassten Befragungsdaten wurden in Rohform als Stata-Datensatz gespeichert. In diesen Rohdaten sind lediglich die manuellen Vorkorrekturen (siehe Kapitel 0) vorgenommen worden, jedoch keine weiteren softwaregestützten Korrekturen.

7.6 Vergabe von Variablennamen, Variablenlabels und Wertelabels

[Variablen- und Wertelabelvergabe] Für Variablen- und Wertelabels wurden Formulierungen des Fragebogens übernommen oder prägnante Kurzformen dieser Formulierungen gewählt. Dabei basieren die Variablenlabel in der Regel auf dem entsprechenden Fragetext. Grundlage für die Wertelabels sind je nach Fragetyp die Texte der Antwortoptionen bzw. eine Kombination der Texte von Frage und Antwortoption.

[Variablenbenennung im Personendatensatz] Die Benennung der Variablen erfolgt anhand von inhaltlichen Kriterien, orientiert sich also am Frageinhalt. Für Indikatoren, die in mehreren Befragungswellen verwendet werden, wurden die Namen der zugehörigen Variablen durch die Vergabe eines identischen Stammes harmonisiert. Bei Fragebatterien wurde ein einheitlicher Variablenstamm gewählt und die einzelnen Items mit einer Zahl versehen (z. B.*pereig01-pereig19*). Darüber hinaus sind Variablen, die im Rahmen der Anonymisierung (vgl. Kapitel 9) neu generiert wurden, mit einem anhand eines Unterstrichs abgetrennten g-Suffix gekennzeichnet (_g#).

Bestimmte Variablen sind aus Anonymisierungsgründen nicht über alle Zugangswege (Remote-Desktop-SUF, On-Site-SUF) einsehbar. In diesen Fällen wird bei den Personendaten im Variablennamen derjenige Zugangsweg angegeben, ab dem die Variable nutzbar[17] ist:

- _r: Variable ist im Remote-Desktop-SUF und im On-Site-SUF nutzbar.
- _o: Variable ist im On-Site-SUF nutzbar.
- _a: Variable ist über keinen Zugangsweg nutzbar. Sie wird aber dokumentiert, da es zugehörige Fragen im Fragebogen gibt.

[Variablenbenennung im Episodendatensatz] Die Variablen im Episodendatensatz sind die Identifikationsnummer der befragten Person (pid), die Identifikationsnummer der jeweiligen Episode (eid), die ausgeübte Tätigkeitsart (status) sowie Beginn und Ende des Episodenzeitraums, der über vier Variablen (Monat: begin_m und end_m; Jahr: begin_j und end_j) codiert wird.

[17] „Nutzbar“ heißt: die Variable enthält nicht ausschließlich Fälle mit dem Missing-Wert „anonymisiert“.

7.7 Codierung fehlender Werte

Zur Codierung fehlender Werte wurde für die dritte Welle des Studienberechtigtenpanels 2012 folgende Systematik genutzt:

Tabelle 4 Systematik für fehlende Werte im Studienberechtigtenpanel 2012 (Welle 3)

Code	Label	Erklärung
-999	weiß nicht	„Weiß nicht“ (wenn dieses explizit hingeschrieben wurde) oder durch den Befragten anders vermerkt wurde (etwa durch ein Fragezeichen).
-998	keine Angabe	Befragte Person hat keine Angabe gemacht, also nichts angekreuzt (PAPI) bzw. keine Eingabe getätigt (Online).
-996	Interviewabbruch	Fehlende Werte nach einem Interviewabbruch. Wurde nur bei den Online-Fragebogen verwendet.
-995	keine Teilnahme Panelwelle	Befragte Person hat an Panelwelle nicht teilgenommen. Wenn eine Person zum Beispiel nur an der ersten Welle teilgenommen hat, erhalten alle fehlenden Werte in den folgenden Wellen den Missingcode „-995“.
-989	filterbedingt fehlend	Fehlender Wert aufgrund von Filterführung des Fragebogens.
-988	trifft nicht zu	Dieser Code wird nicht bei Filterführung vergeben, sondern wenn - explizit eine Antwortoption "trifft nicht zu" vorgesehen ist - eine Antwortkategorie durch erfolgte Verwendung anderer Antwortkategorien bereits ausgeschlossen wurde.
-987	designbedingt fehlend	Ergibt sich durch unterschiedliche Befragungsmodi, insofern vereinzelte (Hilfs-)Variablen nur online oder nur im Papierfragebogen abgefragt wurden
-968	unplausibler Wert	Wird vergeben, sofern zwar wie vorgesehen eine Angabe gemacht wurde, diese aber außerhalb eines vordefinierten Wertebereichs liegt. Zum Beispiel: Bei Monatsangaben liegt der vordefinierte Wertebereich zwischen 1 und 12. Wird vom Befragten bspw. eine „14“ genannt, wird „-968“ vergeben.
-966	nicht bestimmbar	Z.B. offene Angabe, die nicht vercodet werden konnte, oder wenn bei Skalen zwischen zwei Kästchen angekreuzt wurde.
-965	ungültige Mehrfachnennung	Z. B. wenn auf einer Skala von 1 bis 5 die Werte 4 und 5 angekreuzt wurden.
-964	nicht valide	Nicht valide (z. B. unkenntlich gemacht oder gestrichen) sowie bei offenen Items ohne auswertbaren Inhalt (z. B. durchgestrichene Wörter, Striche)
-929	Datenverlust	Wurde nur in sehr seltenen Fällen vergeben, wenn Angaben aus den Vorwellen nicht zugespielt werden konnten

8 Gewichtung

Das Ziel quantitativer Erhebungen besteht für gewöhnlich in einer möglichst validen Schlussfolgerung von Stichprobendaten auf eine bestimmte Grundgesamtheit. Um potenzielle Verzerrungen durch Ausfallprozesse oder ein bestimmtes Stichprobendesign zu minimieren, ist in der sozialempirischen Forschungspraxis die Verwendung von *Gewichten* üblich. Für den Datensatz der dritten Welle des Sudienberechtigtenpanels 2012 wurden verschiedene Gewichte berechnet und zur Verfügung gestellt. Im Folgenden wird zunächst eine kurze allgemeine Einführung zur Gewichtung gegeben und dann die Gewichtungsprozedur im Detail beschrieben.

8.1 Vorgehen und Anwendungshinweise[18]

[Ursachen für die Verzerrungen der Stichprobe] Für die Verzerrung der Stichprobe sind zwei Prozesse maßgeblich:

- Designbedingte Verzerrung: Disproportionalitäten werden bewusst erzeugt, um in relevanten Subgruppen die Fallzahlen zu erhöhen. Im Fall des Studienberechtigtenpanels 2012 wurde eine disproportionale geschichtete Klumpenstichprobe gezogen (siehe Kapitel 4). Die dadurch erzeugten Verzerrungen gilt es über entsprechende Gewichtungen wieder auszugleichen.
- Verzerrung aufgrund von Nonresponse: Ausfallprozesse und die dadurch entstehenden fehlenden Werte stellen eines der fundamentalsten Probleme quantitativ empirischer Forschung dar (zusammenfassend: Allison, 2001). Nur wenn diese Ausfallprozesse komplett zufällig sind (Missing Completely at Random, MCAR), können sie ignoriert werden (Rubin, 1976). Üblicherweise ist dies jedoch nicht der Fall, weshalb die Ausfälle einer Modellierung bedürfen. Es kann grundsätzlich zwischen Item-Nonresponse – also fehlenden Werten auf einzelnen Fragen – und Unit-Nonresponse – also fehlenden Werten aufgrund der Nicht-Teilnahme an der gesamten Befragungswelle – unterschieden werden. Während zum Umgang mit Item-Nonresponse üblicherweise Imputationsverfahren Anwendung finden, lassen sich Verzerrungen durch Unit-Nonresponse über Gewichte minimieren. In Panel-Studien sind wiederum zwei Formen von Unit-Nonresponse relevant: Einerseits kommt es zu Ausfällen zwischen Brutto- und Nettostichprobe (z. B. aufgrund von Nichtteilnahmen, fehlender Erreichbarkeit, Verlust auf dem Postweg, etc.) und zweitens kommt es aufgrund von Panelmortalität zu Ausfällen zwischen einzelnen Befragungswellen.

[Konzeptuelles Vorgehen] Üblicherweise wird bei der Erstellung von Gewichten in mehreren Schritten vorgegangen (vgl. hierzu Daniel et al., 2017, S. 26). Zunächst werden designbedingte Proportionalitäten über sogenannte Designgewichte ausgeglichen, die sich direkt aus den Ziehungswahrscheinlichkeiten des jeweiligen Stichprobendesigns ergeben. Daran anschließend werden die Designgewichte mit Hilfe von Nonresponse-Gewichten im Quer- bzw. Längsschnitt adjustiert. Diese werden auf der Grundlage von Informationen über die Teilnehmer*innen und Nichtteilnehmer*innen auf Individualebene erzeugt. In einem letzten Schritt können diese nonresponse-

[18] Der Abschnitt orientiert sich in weiten Teilen an Daten- und Methodenberichten zu früheren Kohorten des Studienberechtigtenpanels, die durch das Forschungsdatenzentrum (FDZ) des DZHW zur Verfügung gestellt wurden (exemplarisch: Daniel et al., 2017, S. 26f.).

adjustierten Designgewichte schließlich anhand von Merkmalsverteilungen aus der Grundgesamtheit kalibriert werden (Kalibrierung).

Für die Berechnung der Gewichte der dritten Welle des Studienberechtigtenpanels 2012 war ein solches idealtypisches Vorgehen nicht durchführbar, da keine separaten Designgewichte für die erste Befragungswelle berichtet werden können. Die existierenden Gewichte der Vorwellen wurden über ein Zellgewichtungsverfahren erstellt und stellen somit kombinierte Design- und Ausfallgewichte dar, die bereits auf einzelne Merkmale[19] der Grundgesamtheit kalibriert wurden (Geschlecht, Schulart mit Art der Hochschulreife, Bundesland bei Erwerb der HZB). Aus diesem Grund wurde entschieden, für die dritte Welle ein separates Ausfallgewicht zu schätzen, das zunächst lediglich die Panelmortalität zwischen zweiter und dritter Welle abbildet. Das Gewicht wird anhand eines logistischen Regressionsmodells mit Prädiktoren aus der Vorwelle erstellt (propensity scores). Über eine Multiplikation des so erstellten Ausfallgewichts mit dem kombinierten Design- und Ausfallgewicht der Vorwellen (gewa bzw. gewb) erhält man schließlich ein Längsschnittgewicht der dritten Welle.

Tabelle 5: Bereitgestellte Gewichte zum DZHW-Studienberechtigtenpanel 2012 (Welle 3)

Variablenname	Beschreibung
gew3	Längsschnittgewicht 3-Wellen-Panel (getrimmt & normiert) auf Bundesebene
gew3ausf	Ausfallgewicht zwischen Welle 2 und 3 (getrimmt & normiert) auf Bundesebene
gew3land	Längsschnittgewicht 3-Wellen-Panel (getrimmt & normiert) auf Länderebene
gew3landausf	Ausfallgewicht zwischen Welle 2 und 3 (getrimmt & normiert) auf Länderebene

[Hinweise zur Anwendung der Gewichte] Bei dem erstellten Ausfallgewicht handelt es sich um probability weights, die sich in Stata spezifizieren lassen.[20] Bei dem Gewicht gew3 handelt es sich um ein Längsschnittgewicht für Auswertungen des Dreiwellenpanels. Das Gewicht gew3ausf modelliert nur den Ausfallprozess zwischen den Wellen zwei und drei und kann zur Multiplikation mit anderen Gewichten verwendet werden. Grundlegend ist zu beachten, dass Gewichte nur dann sinnvolle Korrekturgrößen darstellen, wenn das verwendete Analysemodell die zur Gewichtung herangezogenen Variablen enthält oder mit diesen in einem Zusammenhang steht. Aus diesem Grund müssen Gewichte immer mit Fokus auf die analysierte Fragestellung verwendet werden. Im Folgenden wird die Vorgehensweise bei der Erstellung der Gewichte für die dritte Welle näher dargestellt.

8.2 Modellierung der Ausfallgewichte

[Propensity Score Matching] Für die Welle 3 wurden die Ausfallgewichte mithilfe von Propensity Scores berechnet.[21] Dieses Verfahren macht sich die Tatsache zu Nutze, dass anders als bei Ausfallprozessen zwischen Brutto- und Nettostichprobe in der ersten Welle bei Ausfällen in nachfolgenden Wellen bereits eine Vielzahl an Informationen zu den Befragten aus den jeweiligen Vorwellen vorliegen. Diese Informationen werden als Kovariaten in einer logistischen Regression genutzt, die die Vorhersage der Teilnahmewahrscheinlichkeit in der nachfolgenden Welle zum Ziel hat. Die Wahrscheinlichkeit des Verbleibs $P_r(Y_i = 1)$ einer Person i im Panel lässt sich dann berechnen als

[19] Die Merkmale waren Schulart, Geschlecht, Bundesland, Erwerb der HZB nach 8 bzw. 9 Jahren sowie in der zweiten Welle zusätzlich die Studierneigung.
[20] Siehe hierzu die Stata-Hilfe (Befehl: help weights).
[21] Das Verfahren entspricht seiner Logik nach dem Propensity Score Matching, das auf Rosenbaum und Rubin (1983) zurückgeht (siehe auch Blumenstiel & Gummer, 2015).

$$Pr(Y_i = 1 \mid X_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} ,$$

wobei $X_i = (X_{i1}, X_{i2} \ldots X_{ip})$ die Liste der relevanten Prädiktoren aus der Vorwelle darstellt. Der Gewichtungsfaktor jeder Person ergibt sich dann aus der Inverse der individuellen Bleibewahrscheinlichkeit $P_r(Y_i = 1 \mid X_i)^{-1}$. Befragte mit hoher Bleibewahrscheinlichkeit erhalten einen niedrigen Gewichtungsfaktor und Befragte mit niedriger Bleibewahrscheinlichkeit erhalten einen hohen Gewichtungsfaktor.

[Spezifikation des Schätzmodells] Die Güte der errechneten Gewichte hängt maßgeblich von der Spezifikation des geschätzten logistischen Regressionsmodells und insbesondere von der Auswahl der relevanten Prädiktoren ab (vgl. Chen et al., 2015, S. 3; Little & Vartivarian, 2003). Um diesem Umstand Rechnung zu tragen, wurde ein exploratives, mehrstufiges Auswahlverfahren genutzt: In einem ersten Schritt wurden alle Variablen entfernt, die aufgrund der Filterführung des Fragebogens nicht für alle Befragten vorliegen. In einem zweiten Schritt wurden für alle übrig gebliebenen Variablen bivariate, logistische Regressionen auf die Bleibewahrscheinlichkeit in Welle 3 berechnet. Variablen die einen signifikanten Effekt ($p<0,05$) aufwiesen und ein Pseudo-R^2 von mehr als 0,01 wurden schließlich in das logistische Modell aufgenommen.[22] Variablen, die unter Kontrolle aller übrigen Kovariaten nicht zur Verbesserung des Schätzmodells beigetragen haben, wurden wieder entfernt. Da aus vergleichbaren Panel-Studien (z.B. dem SOEP) bekannt ist, dass zusätzlich zu inhaltlichen Befragungsdaten auch Informationen über die Feldarbeit wichtige Prädiktoren für die weitere Teilnahmewahrscheinlichkeit darstellen können (Schupp, 2004), wurden zusätzlich eine Variable zum Zeitpunkt des Fragebogeneingangs in der zweiten Welle aufgenommen.[23] Um für jede Person die Teilnahmewahrscheinlichkeit schätzen zu können, wurden fehlende Werte bei allen Variablen als zusätzliche Kategorie in das Modell integriert. Für metrische Variablen wurden die fehlenden Werte auf 0 gesetzt und jeweils eine zusätzliche dummy-Variable in das Modell aufgenommen, die ausgibt, ob ein Fall auf der entsprechende metrischen Variabel einen fehlenden Wert aufweist. Maßgebliches Kriterium für die Modellierung der Ausfallprozesse war, Modelle mit größtmöglicher Erklärungskraft bei gleichzeitiger Sparsamkeit zu finden, die die Varianzerhöhung so gering wie möglich halten.

[Trimmung und Normierung] Eine weitere Schwierigkeit des Propensity Score Matchings zur Berechnung von Ausfallgewichten besteht darin, dass Befragte eine sehr geringe geschätzte Teilnahmewahrscheinlichkeit aufweisen können und in der Folge besonders hohe Gewichte erhalten, was zu einer hohen Varianz bei gewichteten Schätzern von Populationsanteilen führen kann. Ein gängiges Vorgehen besteht in der Trimmung besonders großer Gewichte (Potter, 1990). Dem Verfahren liegt die Annahme zugrunde, dass Gewichte einer Betaverteilung folgen. Alle Gewichte, die über dem 99-Prozent-Quantil lagen, wurden auf diese Grenze trunkiert.

Im Anschluss an die Gewichtung wurden die Ausfallgewichte auf die Fallzahl der Stichprobe normiert, um in späteren Analysen keine Abweichungen in der Sampling-Größe zu erzeugen.

[Berechnung des finalen Längsschnittgewichts] Das finale Längsschnittgewicht gew3 berechnet sich schließlich aus der Multiplikation des (noch nicht getrimmten und nicht normierten) Ausfallgewichts mit dem kombinierten Designgewicht (gewb) aus der Vorwelle. Die Trimmung und Normierung des so entstandenen Längsschnittgewichts wurde in diesem Fall erst nach der Multiplikation durchgeführt.

---

[22] Verwendet wurde das von Tjur (2009) vorgeschlagene Pseudo-R^2 zur Beurteilung logistischer Regressionen (siehe hierzu auch: Cramer, 1999).

[23] Weitere Informationen zur Incentivierung und zum schulspezifischen Rücklauf wurden testweise in die Modelle aufgenommen. Im Vergleich zur Erklärungskraft wird dadurch allerdings nur die Varianz unnötig in die Höhe getrieben.

9 Anonymisierung

[Datenschutzrechtlicher Rahmen] Für personenbezogene Daten[24], die in freiwilligen Befragungen durch das DZHW erhoben werden, gelten die EU-Datenschutz-Grundverordnung (EU-DSGVO) und das Bundesdatenschutzgesetz in seiner Neufassung vom 30. Juni 2017.[25] Danach sind personenbezogene Daten für die Weitergabe zur wissenschaftlichen Sekundärnutzung (ohne Vorliegen einer Einverständniserklärung zur Sekundärnutzung der personenbezogenen Daten) in der Regel derart aufzubereiten, dass „die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden können“ (Art. 4 Abs. 5 DSGVO; s. auch Art. 89 DSGVO sowie Erwägungsgrund 26 DSGVO). Das heißt, für die Weitergabe von Daten aus wissenschaftlichen Forschungsprojekten an Dritte sind die Daten derart zu anonymisieren, dass kein Bezug zur Person mehr hergestellt werden kann.

[Datenzugang, Anonymisierungsgrad und Analysepotential] Das FDZ des DZHW stellt für das Studienberechtigtenpanel 2012 ein SUF für die wissenschaftliche Sekundärnutzung zur Verfügung. Die Anonymität der Befragten wird dabei über eine Kombination aus statistischen Maßnahmen und technischen Zugriffsbeschränkungen sichergestellt. Je stärker der Datenzugang technisch kontrolliert wird, desto geringer ist das Risiko einer De-Anonymisierung der Daten, desto weniger müssen die Daten mittels statistischer Maßnahmen um Informationen reduziert werden und desto größer bleibt ihr Analysepotential.

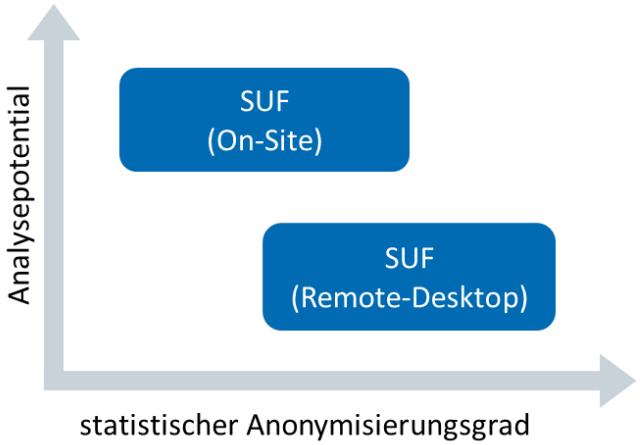
Das SUF wird über zwei verschiedene Zugangswege angeboten: Remote-Desktop und On-Site. Für jeden Zugangsweg wird eine andere SUF-Variante bereitgestellt, die unterschiedlich stark anonymisiert worden ist und entsprechend weniger oder mehr Informationen umfasst. Abbildung 3 gibt einen Überblick über den jeweiligen Grad der statistischen Anonymisierung und dem damit verbundenen Analysepotential. Im Folgenden werden die durchgeführten statistischen Anonymisierungsmaßnahmen in Abhängigkeit vom Zugangsweg erläutert.

---

[24] „Personenbezogene Daten (sind) alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden „betroffene Person“) beziehen; als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen identifiziert werden kann, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind“ (Art. 4 DSGVO, S. 1).

[25] Die DSGVO gilt grundsätzlich innerhalb der EU und somit ebenfalls für das DZHW. Das BDSG in seiner Neufassung vom 30. Juni 2017 (Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Datenschutzanpassungs- und Umsetzungsgesetz EU DSAnpUG-EU)) kommt teils zusätzlich zur Anwendung, da die DZHW GmbH juristisch als öffentliche Stelle des Bundes betrachtet wird (vgl. § 2 Abs. 3 BDSG). Der Bund hält die absolute Mehrheit der Anteile der DZHW GmbH und das Institut erfüllt Aufgaben der öffentlichen Verwaltung des Bundes im weitesten Sinn.

Abbildung 3: Datenzugangswege, statistischer Anonymisierungsgrad und Analysepotential der Daten des DZHW-Studienberechtigtenpanels 2012



[Statistische Anonymisierungsmaßnahmen] Im Rahmen der Anonymisierung sind zunächst alle Informationen, mit denen sich Personen oder Institutionen direkt identifizieren lassen, zu löschen. Diese sogenannten direkten Identifikatoren, wie Namen, Adressen oder E-Mail-Adressen, wurden im Rahmen des Studienberechtigtenpanels 2012 bereits während der Feldphase in einem separaten Datensatz erfasst und nur genutzt, um eine Kontaktaufnahme zu ermöglichen. Diese Angaben sind somit nicht in den verschiedenen SUF-Varianten enthalten. Des Weiteren wurde, um einen Rückbezug auf die Originaldaten zu verhindern, die Original-Identifikationsnummer aus dem Datensatz entfernt und durch eine neue, zufällig vergebene Identifikationsnummer ersetzt.

Anschließend wurden die Quasi-Identifikatoren bestimmt, also Informationen, die in Kombination oder durch die Anspielung externer Informationen geeignet sind, eine Person indirekt zu identifizieren. Für das Studienberechtigtenpanel 2012 wurden beispielsweise folgende Quasi-Identifikatoren identifiziert: regionale Informationen (Geburtsland, Hochschule oder Arbeitsort), Staatsangehörigkeit, Sprache im Elternhaus, Schulart, schulische Prüfungsfächer, Hochschule, Studienfach, Abschlussart, Berufsangaben. Um eine eindeutige Zuordnung der Studienberechtigtendaten zu unterbinden, wurden diese Schlüsselmerkmale – je nach Zugangsweg – aggregiert oder gelöscht (vgl. Tabelle 6: Maßnahmen der statistischen Anonymisierung der Daten des DZHW-Studienberechtigenpanels 2012 nach Zugangsweg). Beispielsweise ist das Merkmal Studienfach im SUF für die On-Site-Nutzung uneingeschränkt verfügbar. Im Remote-Desktop-SUF hingegen wird das Merkmal zu Studienbereichen aggregiert.

Darüber hinaus empfehlen Ebel und Meyermann offene Angaben zu löschen „selbst wenn die jeweiligen Fragestellungen an sich unproblematisch sind. Denn es besteht die Gefahr, dass Studienteilnehmer*innen bei eigentlich unbedenklichen Fragen mit offener Antwortmöglichkeit kritische Informationen preisgegeben haben, die zu einer Identifikation führen könnten“ (Ebel & Meyermann, 2015, S. 5). Die offenen Angaben wurden größtenteils bereits im Rahmen der Datenaufbereitung durch das Primärforschungsprojekt vercodet und werden in dieser Form in allen SUF-Varianten zur Verfügung gestellt. Teilweise wurden jedoch – in Abhängigkeit von der Sensibilität der enthaltenen Informationen und vom Zugangsweg – die vom Primärforschungsprojekt vorgenommenen Codierungen zusätzlich aggregiert. Nicht codierte offene Angaben wurden in allen SUF-Varianten gelöscht.

Zuletzt wurde geprüft, ob in den Daten sensible Informationen, z. B. zur Gesundheit, sexuellen Orientierung oder zu politischen Einstellungen, enthalten waren. Diese eignen sich zwar nicht notwendig zur Re-Identifikation von Individuen oder Institutionen, jedoch können die Informationen im Falle einer De-Anonymisierung nutzbringend sein (Koberg, 2016, S. 694) und sind daher besonders schützenswert (Art. 9 DSGVO, Erwägungsgrund 51 DSGVO). Die nachfolgende Tabelle 6 stellt in Kurzform die durchgeführten statistischen Anonymisierungsmaßnahmen je nach Zugangsweg dar. Variablen, die in den beiden SUF-Varianten aus Datenschutzgründen nicht verfügbar sind, enden mit dem Suffix „_a“.

Tabelle 6: Maßnahmen der statistischen Anonymisierung der Daten des DZHW-Studienberechtigenpanels 2012 nach Zugangsweg

Merkmal	On-Site-SUF	Remote-Desktop-SUF
Direkte Identifikatoren	Löschung	Löschung
Original-ID (Befragte und Schulen)	Löschung und Vergabe einer zufälligen ID	Löschung und Vergabe einer zufälligen ID
Schulart	Aggregation von geringbesetzten Schularten	Aggregation zu allgemeinbildenden und beruflichen Schulen
Prüfungsfächer/Schwerpunktfach (Schule, Berufsschule)	Freigabe	Aggregation zu 15 Fächergruppen
Voraussichtliches Studienfach	Freigabe	Freigabe
Studienfach	Freigabe	Aggregation zu Studienbereichen[a]
Hochschule	Information zur Hochschulart	Information zur Hochschulart[b]
Hochschulort/Aufenthaltsort im Dezember 2012	Aggregation zum Bundesland oder Bundeslandgruppen und Ausland	Aggregation zum Bundesland oder Bundeslandgruppen und Ausland
(Angestrebter) Studienabschluss	Aggregation von geringbesetzten Abschlussarten	Aggregation von geringbesetzten Abschlussarten
Beruf (Praktikumsberuf, (voraussichtlicher) Ausbildungsberuf, berufliche Tätigkeit, Berufe der Eltern)	Aggregation zu Berufsuntergruppen (KldB 4-Steller) [c]	Aggregation zu Berufsgruppen (KldB 3-Steller)[c]
Staatsangehörigkeit, Geburtsland, Geburtsland der Groß-/Eltern	Aggregation zu maximal 30 Kategorien	7 Staaten einzeln ausgewiesen; weitere Staaten zu Weltregionen[d]
Sprache im Elternhaus	Erste Sprache: 20 Sprachen einzeln ausgewiesen; andere Sprachen zu „Sonstiges“ Zweite Sprache: 3 Sprachen einzeln ausgewiesen; andere Sprachen zu „Sonstiges“ Dritte Sprache: nur „genannt“	Erste Sprache: 10 Sprachen einzeln ausgewiesen; andere Sprachen zu „Sonstiges“ Zweite Sprache: 3 Sprachen einzeln ausgewiesen; andere Sprachen zu „Sonstiges“ Dritte Sprache: nur „genannt“
Anzahl der Geschwister	Aggregation zu „1“, „2“ und „3 und mehr“	Löschung
Alter der Geschwister	Aggregation zu „älter“, „gleichalt“ oder „jünger“	Aggregation zu „älter“, „gleichalt“ oder „jünger“
Probleme bei Wahl des nachschulischen Werdegangs	Aggregation: Zusammenfassung der Gründe „Gesundheitsprobleme“ und „Sonstiges“	Aggregation: Zusammenfassung der Gründe „Gesundheitsprobleme“ und „Sonstiges“
Gründe, die gegen ein Studium gesprochen haben	Aggregation: Zusammenfassung der Gründe „Krankheit“ und „Sonstiges“	Aggregation: Zusammenfassung der Gründe „Krankheit“ und „Sonstiges“
Bedeutung für gewählten Werdegang: Gesundheit	Löschung	Löschung
Schwierigkeiten bei der Stellensuche: Offene Angabe	Aggregation: Aussehen, gesundheitl. Beeinträchtigung und Religion zu „Diskriminierungserfahrung“ zusammengefasst	Aggregation: Aussehen, gesundheitl. Beeinträchtigung und Religion zu „Diskriminierungserfahrung“ zusammengefasst
Arbeitsort (PLZ)/ Wohnort (PLZ)	PLZ-3-Steller	PLZ-2-Steller

Arbeitsort (Land)/ Wohnort (Land)	Aggregation	Aggregation
Kinder	Freigabe	Geburtsjahr: Aggregation: bis 1997; 1998- 2003; 2004-2009; 2010-2012; ab 2013 Geburtsmonat: Löschung
Religion	Löschung	Löschung
Tätigkeitsart	Aggregation	Aggregation

a Nach Schlüsselverzeichnis der Studenten- und Prüfungsstatistik WiSe 2011/12 (Welle 1 und 2) und WiSe 2018/19 (Welle 3) des Statistischen Bundesamts.
b Nur Unterscheidung von Fachhochschule vs. Universität (inklusive Pädagogische Hochschulen, Theologische Hochschulen, Kunst- und Musikhochschulen und Verwaltungsfachhochschulen).
c Nach Klassifikation der Berufe von 1992 (Welle 1 und 2) und 2010 (Welle 3) des Statistischen Bundesamts.
d Nach Klassifikation der Vereinten Nationen.

10 Literaturverzeichnis

Allison, P. D. (2001). *Missing data* (Sage university papers. Quantitative applications in the social sciences, 07-136). Thousand Oaks, Calif.: Sage Publications.

Blumenstiel, J. E. & Gummer, T. (2015). Prävention, Korrektur oder beides? Drei Wege zur Reduzierung von Nonresponse Bias mit Propensity Scores. In J. Schupp & C. Wolf (Hrsg.), *Nonresponse Bias. Qualitätssicherung sozialwissenschaftlicher Umfragen* (S. 13–44). Wiesbaden: Springer Fachmedien. doi:10.1007/978-3-658-10459-7

Chen, Q., Gelman, A., Tracy, M., Norris, F. H. & Galea, S. (2015). Incorporating the sampling design in weighting adjustments for panel attrition. *Statistics in Medicine,* 34(28), 3637–3647.

Cramer, J. S. (1999). Predictive Performance of the Binary Logit Model in Unbalanced Samples. *Journal of the Royal Statistical Society: Series D (The Statistician),* 48(1), 85–94.

Daniel, A., Hoffstätter, U., Huß, B. & Scheller, P. (2017). *DZHW-Studienberechtigtenpanel 2008. Daten- und Methodenbericht zu den Erhebungen des Studienberechtigtenjahrgangs 2008 (1. bis 3. Befragungswelle)*. Version 1.0.0. Hannover: FDZ-DZHW.

Ebel, T. & Meyermann, A. (2015). *Hinweise zur Anonymisierung von quantitativen Daten* (forschungsdaten bildung informiert Nr. 3). Verbund Forschungsdaten Bildung. doi:10.25656/01:21970

Föste-Eggers, D. (2021). *Migrationsbezogene Ungleichheiten bei Aufnahme und Abschluss von (dualen) MINT-Berufsausbildungen. Manuskript in Vorbereitung*.

Koberg, T. (2016). Disclosing the National Educational Panel Study. In H.-P. Blossfeld, J. von 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

Little, R. J. & Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine,* 22(9), 1589–1599.

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Survey Research Methods Section,* 225–230.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70(1), 41–55. doi:10.2307/2335942

Rubin, D. B. (1976). Inference and missing data. *Biometrika,* 63(2), 581–592.

Schneider, H. & Franke, B. (2014). *Bildungsentscheidungen von Studienberechtigten. Studienberechtigte 2012 ein halbes Jahr vor und ein halbes Jahr nach Schulabschluss* (Forum Hochschule Nr. 6). Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW).

Schupp, J. (2004). *Gewichtung in der Umfragepraxis. Das Beispiel SOEP*. Verfügbar unter http://eswf.uni-koeln.de/lehre/04/04_05/schupp.pdf

Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *The American Statistician,* 63(4), 366–372.