gesis Leibniz Institute for the Social Sciences

Gewichtung in der Praxis

Matthias Sand & Tanja Kunz

April 2020, Version 1.0

Abstract

In diesem Dokument werden die theoretischen Grundlagen der Gewichtung von Umfragedaten behandelt sowie verschiedene Gewichtungsverfahren anhand konkreter Anwendungsbeispiele weiter ausgeführt. Das Dokument dient als Grundlage für eine vereinheitlichte Dokumentation der Gewichtungsprozesse in den Umfrageprogrammen der Integrierten Erhebungs- und Dateninfrastruktur (IEDI), sowie einem verbesserten Verständnis der Kriterien, nach denen die Gewichtung von Umfragedaten erfolgt. Weiterhin verdeutlicht es die Auswirkungen der (Nicht-) Gewichtung. Um einen möglichst umfassenden Überblick zu gewährleisten, werden unterschiedliche Arten der Designgewichtung und Anpassungsgewichtung genauer erläutert und das Vorgehen anhand ausgewählter Beispiele erklärt.

Citation

Sand, Matthias und Kunz, Tanja (2020). Gewichtung in der Praxis. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).

DOI: 10.15465/gesis-sg_030

1 Gewichtung von Umfragedaten

In der Umfrageforschung werden in der Regel nicht alle statistischen Einheiten einer Grundgesamtheit zu einem Thema oder Sachverhalt untersucht bzw. befragt, wie dies bei einem Zensus üblich ist, sondern nur ein Teil der Grundgesamtheit, der sogenannten Stichprobe. Dabei können verschiedene Stichprobenverfahren unterschieden werden (siehe hierzu Gabler & Häder, 2015; Häder, 2015).

Umfragedaten, die auf einer Zufallsstichprobe basieren, erlauben Rückschlüsse auf die zugrundeliegende Grundgesamtheit. Die „Gewichtung" von Umfragedaten bezieht sich auf das Vorgehen, durch das bei der Schätzung relevanter Statistiken der Grundgesamtheit, wie etwa dem Mittel- oder Totalwert einer Grundgesamtheit, einzelne Erhebungseinheiten (Zielpersonen oder Gruppen von Zielpersonen) in ihrer (relativen) Bedeutung verändert werden. Somit stellt ein Gewicht einen multiplikativen Faktor (kleiner, größer oder gleich 1) dar, der die relative Bedeutung der Erhebungseinheiten in Bezug auf die Schätzung verändert. Das Ziel einer Gewichtung ist es, für eine Erhebungseinheit (Zielperson) i ein Gewicht w_i zu entwickeln, das zum einen für jegliche Analysen des bestehenden Umfragedatensatzes verwendet werden kann und zum anderen geeignete (unverzerrte) Schätzwerte für die zugrundeliegende Grundgesamtheit der Erhebung ermöglicht (Valliant, Dever & Kreuter, 2013, S. 308). Ganz allgemein gesprochen sollen die gewichteten Schätzwerte näher an den tatsächlichen Parametern der Grundgesamtheit liegen als die ungewichteten Schätzwerte.

Die Hauptgründe für eine Gewichtung sind nach Gabler, Häder, Lehnhoff & Mardian (2012, S. 147) (a) die Reduktion potentieller Verzerrungen der Schätzwerte infolge ungleicher Auswahlwahrscheinlichkeiten, (b) die Reduktion potentieller Verzerrungen der Schätzwerte aufgrund von Antwortausfällen (Nonresponse) auf Zielpersonenebene, sowie (c) die Anpassung bestimmter (soziodemografischer) Merkmale an die entsprechende Verteilung in der Grundgesamtheit im Rahmen einer nachträglichen Schichtung (auch Poststratifikation genannt). Somit kann im Wesentlichen zwischen Gewichtungsverfahren unterschieden werden, welche die potenziell unterschiedlichen Auswahlwahrscheinlichkeiten einzelner Zielpersonen berücksichtigen (Designgewichtung) und denjenigen, welche mögliche Verzerrungen infolge von Antwortausfällen auf Zielpersonenebene verringern oder zur nachträglichen Schichtung (Kalibrierung oder Anpassungsgewichtung) angewendet werden.

Ad a) Im Falle ungleicher Auswahlwahrscheinlichkeiten aufgrund des Erhebungsverfahrens dient die Designgewichtung dazu, unverzerrte Schätzwerte für die Grundgesamtheit einer Erhebung zu erhalten. Beinhaltet der Erhebungsdatensatz bspw. häufiger bestimmte Erhebungseinheiten, weil diese wahrscheinlicher in die Erhebung gelangen, wird diese „Überrepräsentation" anhand der Designgewichtung revidiert (Lundström & Särndal, 2001, S. 17ff).

Abbildung 1 illustriert potenzielle Fehlerquellen, die im Zuge einer Erhebung aufgrund des Auswahlrahmens sowie des Auswahlprozesses entstehen und die Schätzung der Parameter der Grundgesamtheit beeinflussen können.

Eine Zielperson, die über den Auswahlrahmen zwar erreichbar ist, jedoch nicht der Grundgesamtheit angehört, kann als Karteileiche oder *Overcoverage* bezeichnet werden. Ein Beispiel hierfür wäre bei einer telefonischen Befragung der bundesdeutschen Bevölkerung eine Rufnummer, die zu einem gewerblich genutzten Telefonanschluss gehört. Andererseits beinhaltet der Auswahlrahmen nicht immer alle Elemente der Grundgesamtheit. Ein solcher Fehlbestand oder *Undercoverage* kann bspw. im Zuge einer Einwohnermeldeamtsstichprobe zur Befragung der bundesdeutschen Bevölkerung entstehen, wenn bestimmte Gemeinden nicht im Auswahlrahmen (aus dem die Gemeindestichprobe gezogen wird) enthalten sind.[1] Ebenso würde die Beschränkung auf eine reine Festnetzstichprobe bei jener Grundgesamtheit

[1] Dabei ist nicht gemeint, dass alle Gemeinden in einer Stichprobe enthalten sein müssen, sondern lediglich, dass jede Ge-

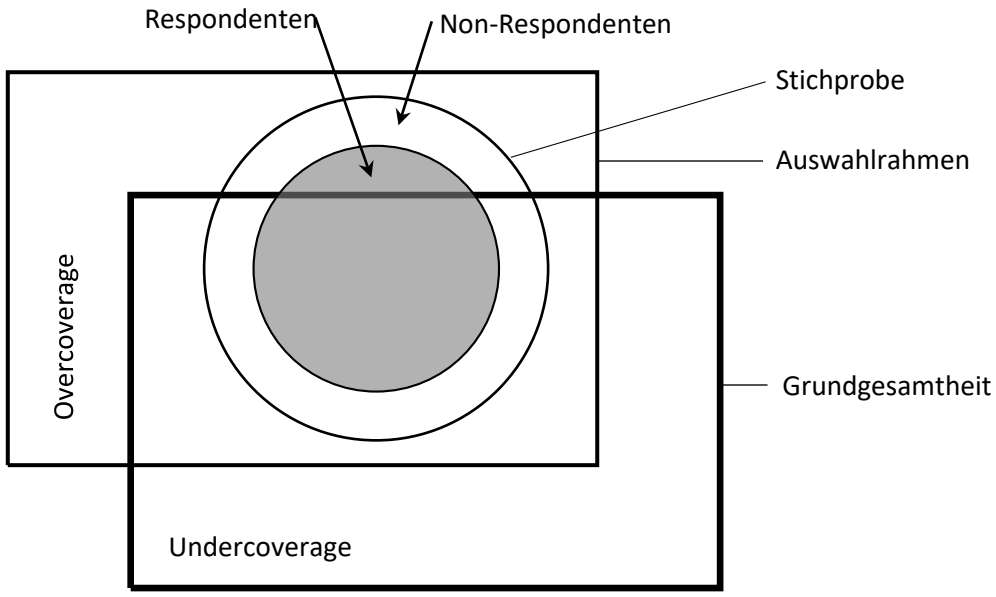

Abbildung 1: Stichprobe, Auswahlrahmen, Grundgesamtheit und mögliche Fehlerquellen (Särndal & Lundström, 2005).

zu Undercoverage führen, da Zielpersonen, die ausschließlich über einen Mobilfunkanschluss erreichbar sind, ausgeschlossen wären.

Ungleiche Auswahlwahrscheinlichkeiten sind ebenfalls im Stichprobenverfahren begründet. Viele Erhebungen wenden keine uneingeschränkte Zufallsauswahl an, sondern werden mit komplexen Auswahlverfahren erhoben. Komplexe Stichproben wiederum implizieren oftmals unterschiedliche Auswahlwahrscheinlichkeiten der Zielpersonen. Um möglichen Abweichungen der Schätzwerte von den jeweiligen Parametern in der Grundgesamtheit entgegenzuwirken, werden die Zielpersonen mit Designgewichten versehen (siehe hierzu Kap. 1.1 und 2.1).

Ad b) Es kann zwischen *Unit Nonresponse* im Sinne einer Nichtbeantwortung des kompletten Fragebogens (z.B. aufgrund von Verweigerung oder Nichterreichen der Zielperson) und *Item Nonresponse* im Sinne der Nichtbeantwortung einer oder mehrerer einzelner Fragen (z.B. Einkommensfrage) durch die Zielperson unterschieden werden. Abbildung 1 unterscheidet zwischen Respondenten und Non-Respondenten; Verzerrungen in den Schätzwerten aufgrund von Antwortausfällen (Nonresponse) können immer dann auftreten, wenn sich Respondenten und Non-Respondenten hinsichtlich der betreffenden Variablen systematisch unterscheiden (siehe hierzu Koch & Blohm, 2015). Mittels Anpassungsgewichtung können potenzielle Verzerrungen der Schätzwerte infolge von Antwortausfällen auf Zielpersonenebene verringert werden (siehe hierzu Kap. 1.2 und 2.2).

meinde (oder jede Gemeinde, in der auch Zielpersonen der Grundgesamtheit gemeldet sind) eine von Null verschiedene Auswahlwahrscheinlichkeit hat, um in die Stichprobe zu gelangen.

Ad c) Antwortausfälle können unabhängig davon, ob diese systematisch auftreten oder zufällig sind, die Varianz der Schätzwerte beeinflussen. Ein Antwortausfall führt grundsätzlich zu einer Reduktion der Stichprobengröße, was wiederrum zu einem Anstieg der Varianz führt. Auch ohne Antwortausfall kann eine nachträgliche Schichtung sinnvoll sein, um die Varianz der Schätzwerte zu reduzieren oder wenn eine Schichtung im Vorfeld der Erhebung bspw. aufgrund eines unvollständigen Auswahlrahmens nicht möglich war (siehe hierzu Kap. 1.2 und 2.2).

b) und c) können demnach als Begründung für eine Kalibrierung oder Anpassungsgewichtung angeführt werden. Die beiden Begriffe werden oftmals synonym verwendet, wobei die Bezeichnung Anpassungsgewichtung häufig dann verwendet wird, wenn aufgrund von b) gewichtet wird, während die Kalibrierung sowohl b) als auch c) umfassen kann.[2]

In welcher Form und unter welchen Bedingungen bestimmte Gewichtungsverfahren anzuwenden sind ist abhängig davon, wie die zugrundeliegenden Umfragedaten zustande gekommen sind und welche Aussagen der Forschende mittels der Daten treffen möchte. Wird bspw. eine deskriptive Statistik der Stichprobenzusammensetzung (Stichprobenpopulation) angestrebt, die nach der Erhebung (und nach dem Ausfallprozess einzelner Zielpersonen) in der Stichprobe vorzufinden ist, so ist diese auch ohne eine entsprechende Gewichtung möglich. Sobald jedoch generalisierbare Aussagen über die (Teile der) Grundgesamtheit getroffen werden sollen, muss eine Gewichtung der Umfragedaten zumindest in Erwägung gezogen werden (Sand & Gabler, 2019). Die unterschiedlichen Arten der Gewichtung sowie die Gründe für deren Anwendung werden im Folgenden näher erläutert. Abbildung 2 gibt eine Übersicht über verschiedene Gewichtungsverfahren, sowie deren Anwendungskontexte und Auswirkungen.

1.1 Warum Designgewichtung?

Der wichtigste Aspekt einer Designgewichtung besteht darin, dass das Stichprobenverfahren und die damit verbundene Auswahlwahrscheinlichkeit (Inklusionswahrscheinlichkeit) der Zielperson in der Schätzung von Parametern der Grundgesamtheit berücksichtigt werden. Eine Designgewichtung ist immer dann notwendig, wenn unterschiedliche Auswahlwahrscheinlichkeiten einzelner Erhebungseinheiten innerhalb der Stichprobe(n) bestehen und mittels der Umfragedaten generalisierbare Aussagen über die zugrundeliegende Grundgesamtheit getroffen werden sollen (Gabler et al., 2012). Basierend auf einer Zufallsstichprobe dienen Designgewichte, gelegentlich auch Base Weights genannt, somit der unverzerrten Schätzung der Parameter (bspw. Mittel- oder Totalwerte) einer Grundgesamtheit. Um mittels Designgewichtung unverzerrte Schätzer zu erzielen, ist grundsätzlich ein vollständiger Auswahlrahmen notwendig, der die gesamte Grundgesamtheit abdeckt, wobei potenzielle Verzerrungen der Schätzwerte aufgrund von Nonresponse an dieser Stelle vernachlässigbar sind (Valliant et al., 2013, S. 307). Die Designgewichtung steht im direkten Bezug zur Auswahlwahrscheinlichkeit der Erhebungseinheiten, weshalb bei der Bestimmung von Designgewichten in einem ersten Schritt versucht werden sollte, die Auswahl- bzw. Inklusionswahrscheinlichkeiten der Erhebungseinheiten unter Berücksichtigung aller Auswahlschritte möglichst exakt zu bestimmen.

Aufgrund eines fehlenden zentralen Melderegisters für Deutschland ist es bspw. für postalische und persönlich-mündliche Erhebungen oftmals notwendig, ein zweistufiges Verfahren anzuwenden, bei dem in einem ersten Schritt Gemeinden (größenproportional) und in einem zweiten Schritt innerhalb dieses Klumpens eine (gleiche) Anzahl von Zielpersonen durch das zuständige Einwohnermeldeamt gezogen werden. Wie aus Abbildung 1 jedoch deutlich wird, deckt ein Auswahlrahmen nicht zwingend alle

[2] Wichtig ist hierbei, dass es sich um unterschiedliche Begrifflichkeiten handelt, während die verwendeten Schätzer oftmals die gleichen sind. Auch soll hier ausgeschlossen werden, dass eine Anpassungsgewichtung nicht auch die Varianz reduzieren kann.

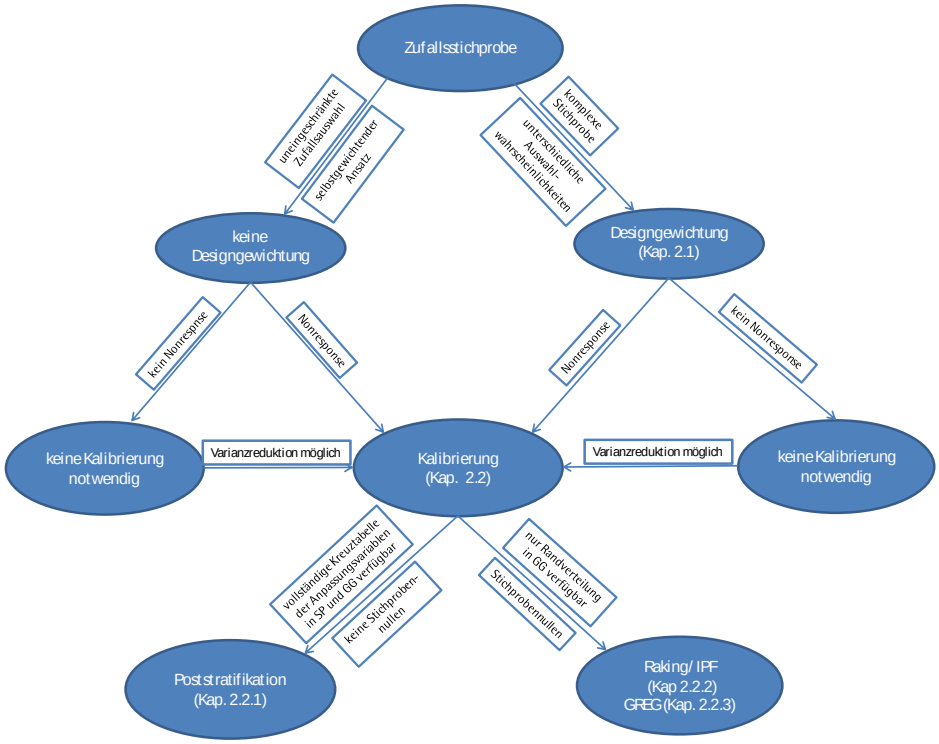

Abbildung 2: Übersicht über verschiedene Gewichtungsverfahren, deren Anwendung und Auswirkungen

Elemente der Grundgesamtheit ab. Bei deutschlandweiten telefonischen Erhebungen bspw. ist es notwendig, neben dem Auswahlrahmen für Festnetzstichproben zusätzlich einen für Mobilfunkrufnummern zu verwenden, um die gesamte Grundgesamtheit (annähernd) abzudecken.

Beim sog. Dual-Frame-Ansatz (Sand & Gabler, 2019) werden die Stichprobenelemente aus zwei verschiedenen Auswahlrahmen (Festnetz und Mobilfunk) gezogen. Für gewöhnlich werden unter Verwendung des Dual-Frame-Ansatzes nach Gabler und Häder (Gabler & Ayhan, 2007; Gabler & Häder, 2009) die Auswahlrahmen für die jeweiligen Telekommunikationsmedien so generiert, dass jede Rufnummer mit der gleichen Wahrscheinlichkeit gezogen werden kann. Dennoch können sich unterschiedliche Auswahlwahrscheinlichkeiten dadurch ergeben, dass die Zielpersonen über unterschiedlich viele Rufnummern in die Stichprobe gelangen können. Des Weiteren ist insbesondere bei Festnetzstichproben zu beachten, dass zunächst ein Haushalt kontaktiert wird und erst nach erfolgreichem Kontakt die Zielperson mittels Kish- oder Birthday-Verfahren ausgewählt wird, weshalb die Anzahl der erhebungsrelevanten Haushaltsmitglieder[3] bei der Bestimmung der Auswahlwahrscheinlichkeit berücksichtigt werden muss (Sand & Gabler, 2019). Die Auswahlwahrscheinlichkeit einer Zielperson berechnet sich dann aus der Wahrscheinlichkeit, dass der Haushalt kontaktiert wird und der Wahrscheinlichkeit, dass die Zielperson unter den erhebungsrelevanten Haushaltsmitgliedern ausgewählt wird.[4] [5] Die Berücksichtigung

[3] „Erhebungsrelevant“ sind diejenigen Haushaltsmitglieder, die Teil der Grundgesamtheit sind. Die Anzahl der erhebungsrelevanten Haushaltsmitglieder berechnet sich anhand der Haushaltsgröße, bereinigt um diejenigen Mitglieder, die nicht Teil der Grundgesamtheit sind (bspw. aufgrund des Alters).

[4] Die Gewichtung solcher Telefonstichproben wird in Kapitel 2.1 anhand des Beispiels der GLES-RCS-Erhebung näher erläutert.

[5] Ähnlich würde es sich für Random-Route-Verfahren gestalten. Bei einer Einwohnermeldeamtsstichprobe wiederum wer-

der Anzahl der erhebungsrelevanten Haushaltsmitglieder wird oftmals als „Transformationsgewicht“ bezeichnet, wobei es sich um einen Teil der Designgewichtung handelt. Das Designgewicht wird i. d. R. unter Berücksichtigung der „gesamten“ Auswahlwahrscheinlichkeit der Zielperson ermittelt. Diese beinhaltet dann ebenfalls die Anzahl der erhebungsrelevanten Haushaltsmitglieder, weshalb ein Transformationsgewicht nicht gesondert berechnet werden muss. Einzige Ausnahme ist, wenn die Erhebung sowohl eine Haushalts- als auch eine Personenbefragung beinhaltet.

1.2 Warum Kalibrierung/ Anpassungsgewichtung?

Eine Kalibrierung oder Anpassungsgewichtung kann zur *Reduktion der Verzerrung* aufgrund von Nonresponse sowie zur *Steigerung der Präzision der Schätzwerte* und demnach zur Reduktion der Varianz eingesetzt werden. Unabhängig von der Begründung einer Kalibrierung gilt es jedoch, die Aktualität der Angaben über die Grundgesamtheit, an die kalibriert wird, zu prüfen. So sollten die Angaben über die Grundgesamtheit, an die die Erhebungsdaten angepasst werden, zeitlich möglichst mit dem Zeitraum der Erhebung übereinstimmen bzw. nicht zu weit in der Zeit zurückliegen. Oftmals ist eine genaue zeitliche Übereinstimmung zwischen Erhebungsdaten und Angaben über die Grundgesamtheit aufgrund fehlender Verfügbarkeit nicht möglich. In solchen Fällen sollten die letzten verfügbaren (und somit aktuellsten) Angaben herangezogen werden. Die Angaben über die Grundgesamtheit müssen auf die zu untersuchende Population abzielen. So kann bspw. beobachtet werden, dass die Aufsummierung einer Randverteilung aus der Bevölkerungsfortschreibung basierend auf den Zensus 2011-Daten andere Schätzwerte ergibt als unter Verwendung der Mikrozensus 20XX-Daten aus einem anderen Jahr. Werden Verteilungen aus beiden Datenquellen zur Kalibrierung herangezogen und absolute Werte verwendet ist es zwingend erforderlich, eine Renormierung auf die Summe einer der beiden Datenquellen vorzunehmen (siehe hierzu Kap. 3.1).

Mittels Anpassungsgewichtung können potenzielle Verzerrungen der Schätzwerte infolge von Antwortausfällen auf Zielpersonenebene (Nonresponse) adressiert werden. Allerdings sollte eine Anpassungsgewichtung nicht als „Allheilmittel“ gegen Nonresponse angesehen werden. Vielmehr kann eine Anpassungsgewichtung die Verzerrung aufgrund von Nonresponse lediglich reduzieren bzw. die „Lücke“ zwischen den Populationsschätzwerten und dem wahren Wert eines Parameters in der Grundgesamtheit verkleinern. In diesem Zusammenhang ist darauf hinzuweisen, dass die zugrundeliegenden Modelle zur Erklärung des Ausfallmechanismus auf starken Annahmen beruhen. So wird i. d. R. angenommen, dass der Ausfall *Missing at Random* (MAR) entspricht. Das Zustandekommen der Antwortausfälle wird nicht mit dem Erhebungsgegenstand selbst, sondern mit weiteren Variablen, die mit der Teilnahme an der Befragung in Verbindung stehen, begründet. Diese Annahmen gelten jedoch als vereinfachte Begründung des Nonresponse und spiegeln meistens den wahren Sachverhalt für den Antwortausfall nicht bzw. nicht vollständig wider (Lohr, 2009).

Darüber hinaus kann eine Anpassungsgewichtung die Präzision der Schätzwerte erhöhen, indem die Varianz reduziert wird. So kann es unter Umständen auch für Erhebungen, in denen es keinen Nonresponse gibt, wünschenswert sein, eine Kalibrierung durchzuführen. Der Einsatz eines Verhältnis- oder Regressionsschätzers kann günstig sein, wenn ein starker linearer Zusammenhang zwischen Untersuchungs- und Hilfsvariablen besteht. Die Kalibrierung an Hilfsvariablen bringt dann eine Reduzierung des (Stichproben-) Fehlers mit sich (Särndal & Lundström, 2005, siehe hierzu Kap. 2.2.3).

Eine Anpassungsgewichtung kann auch dazu genutzt werden, eine nachträgliche Schichtung der Umfragedaten auf Basis der Stichprobe vorzunehmen. Dies ist dann sinnvoll, wenn bspw. der Auswahlrahmen

den die Zielpersonen direkt von den Einwohnermeldeämtern ausgewählt, weshalb ein solches Transformationsgewicht keine Anwendung findet.

einer bestimmten Grundgesamtheit die Bildung von Schichten anhand bestimmter Merkmale im Vorfeld der Stichprobenziehung nicht ermöglicht, da die Merkmale der Zielpersonen schlichtweg nicht vorhanden sind oder im Vorfeld keine Variablen identifiziert werden können, die zur Schichtung der Grundgesamtheit geeignet erscheinen.

Ein Beispiel ist eine telefonische Erhebung, bei der nur Vorwahl und Rufnummer bekannt sind, während sich Merkmale der Zielperson erst im Zuge der Befragung ermitteln lassen. Bei Festnetzstichproben ist lediglich eine regionale Schichtung (anhand der Vorwahl) möglich. Diese Möglichkeit besteht bei Mobilfunkrufnummern in Deutschland nicht. Bei einer nachträglichen Schichtung können die im Rahmen der Befragung gewonnenen Schichtinformationen in die Schätzung integriert werden (Lumley, 2010; Sand & Gabler, 2019).

Bei Befragungen von Schülern wird oft eine Klumpenstichprobe eingesetzt; der Zugang erfolgt über die Auswahl von Klassen innerhalb einer Schule. Als Auswahlrahmen steht hierbei häufig lediglich das Aggregat aus Schulen innerhalb einer Gemeinde sowie in der nächsten Auswahlstufe die Anzahl der Klassen innerhalb der Schule zur Verfügung. Die Schüler innerhalb der Klassen werden dann häufig alle befragt. Eine ex-ante Schichtung der Schüler ist dabei nur schwer möglich, da die Zusammensetzung bestimmter Merkmale der Schüler innerhalb der Klassen i. d. R. unbekannt ist.

2 Berechnung von Gewichten

2.1 Wie werden Designgewichte berechnet?

Eine Designgewichtung ist immer dann notwendig, wenn ein unverzerrter Schätzwert für einen Parameter der Grundgesamtheit bestimmt werden soll. Die Designgewichte selbst berücksichtigen indes die (unterschiedlichen) Auswahlwahrscheinlichkeiten der einzelnen Stichprobeneinheiten im Rahmen der Schätzung von Parametern der Grundgesamtheit (Gabler et al., 2012).

Beim Vorliegen einer Zufallsstichprobe S mit dem Umfang n aus einer Grundgesamtheit U wird das Designgewicht d_i der Erhebungseinheit i (mit $i = 1, \ldots, N$) als Inverse der Auswahl- bzw. Inklusionswahrscheinlichkeit π_i berechnet (unter Verwendung des Horvitz-Thompson-Schätzers). N entspricht dabei der Anzahl der Elemente innerhalb der endlichen Grundgesamtheit, sodass $U = \{1, \ldots, N\}$ gilt.

Ein Schätzwert unter Verwendung eines Designgewichts gilt genau dann als unverzerrt, wenn der designgewichtete Erwartungswert eines Parameters demjenigen der Grundgesamtheit entspricht. Durch die Verwendung des Horvitz-Thompson-Schätzers lässt sich ein Schätzer für den Totalwert Y einer Variablen y_i anhand von

$$\hat{Y} = \sum_{i \in S} \frac{1}{\pi_i} y_i = \sum_{i \in S} d_i y_i$$

bestimmen (Lohr, 2009, S. 240f).

Als erwartungstreue Schätzfunktion für die Gesamtsumme $Y = \sum_{i \in U} y_i$ verwendet man den erwartungstreuen Horvitz-Thompson-Schätzer[6]

$$\hat{Y}_{HT} = \sum_{i \in U} L_i \frac{y_i}{\pi_i}$$

mit $L_i = \begin{cases} 1 \text{ falls } i\text{-te Einheit ausgewählt wird} \\ 0 \text{ sonst} \end{cases}$ für $i = 1, \ldots N$.

Dabei wird vorausgesetzt, dass alle π_i von 0 unterschiedlich und positiv sind. Für die Varianz des Horvitz-Thompson-Schätzers erhält man

$$var(\hat{Y}_{HT}) = \sum_{i \in U} \sum_{j \in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j).$$

liegt ein Auswahlverfahren mit einem festen Stichprobenumfang n vor, so gilt

$$\sum_{j \in U} \pi ij = n\pi_i \text{ und } \sum_{i \in U} \sum_{j \in U} \pi_{ij} = n^2$$

und der sogenannte Yates-Grundy-Varianzschätzer

$$V_{YG} = \frac{1}{2} \sum_{i \in U} \sum_{j \in U} \frac{L_i L_j}{\pi_{ij}} (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2 (\pi_{ij} - \pi_i \pi_j)$$

schätzt $var(\hat{Y}_{HT})$ erwartungstreu, wenn alle π_{ij} positiv sind. Er ist offensichtlich nicht negativ, wenn $\pi_i \pi_j \geq \pi ij$ für alle i, j gilt.

Liegt eine uneingeschränkte (geschichtete) Zufallsauswahl (ohne Zurücklegen) zugrunde, so entspricht sowohl der Horvitz-Thompson-Schätzer als auch dessen Varianzschätzer den bekannten (Varianz-) Schätzformeln für uneingeschränkte (geschichtete) Zufallsauswahl. Der Horvitz-Thompson-Schätzer ist zwar immer erwartungstreu, seine Varianz kann aber sehr groß sein.[7]

Zufallsstichproben zeichnen sich dadurch aus, dass jeder möglichen Stichprobe S eine bekannte Wahrscheinlichkeit $P(S)$ zugeordnet ist. Die Menge aller Stichproben S mit $P(S) > 0$ heißt *Stichprobenraum*. Der Stichprobenraum gibt die Menge aller denkbaren Stichproben und somit auch Kombinationen bestimmter Erhebungseinheiten der Grundgesamtheit wieder, die (bei gegebenem Umfang und Design) eine von Null unterschiedliche, positive Eintrittswahrscheinlichkeit haben.

Die Auswahl- oder Inklusionswahrscheinlichkeiten

$$\pi ij = \sum_{Si, j \in S} P(S)$$

geben die Wahrscheinlichkeit an, dass die Einheiten i und j in die Stichprobe gelangen. π_i ist die Auswahl- oder Inklusionswahrscheinlichkeit erster Ordnung. Sie gibt die Wahrscheinlichkeit an, mit der die Erhebungseinheit i in die Stichprobe gelangt. Sind i und j verschieden, spricht man von Auswahl- oder Inklusionswahrscheinlichkeiten zweiter Ordnung. Diese gibt die Wahrscheinlichkeit an, dass sowohl i als auch

[6] Die folgenden Darstellungen finden sich in gleicher oder ähnlicher Weise auch in Gabler et al. (2015).

[7] Damit ist gemeint, dass sich aufgrund der Spannweite der Designgewichte Effizienzverluste einstellen können. Bspw. durch die Verwendung von Kalibrierungsgewichten, wie sie mittels Generalisiertem Regressionsschätzers (GREG) (siehe hierzu Kap. 2.2) berechnet werden, können Effizienzgewinne im Vergleich zu einer reinen Designgewichtung erwirkt werden.

j gemeinsam in der Stichprobe enthalten sind. Bei der *uneingeschränkten Zufallsauswahl* (ohne Zurücklegen) von n Einheiten aus einer Gesamtheit mit N Einheiten ist

$$\pi_i = \frac{n}{N}$$

$$\pi ij = \frac{n}{N}\frac{(n-1)}{(N-1)} \text{ für } i \neq j.$$

Bei einer *einfachen Zufallsauswahl* werden n Einheiten zufällig mit gleicher Wahrscheinlichkeit und ohne Zurücklegen aus der Gesamtheit ausgewählt. Eine *komplexe Stichprobe* kann sich in verschiedener Hinsicht von einer einfachen Zufallsstichprobe unterscheiden. Hierzu zählen beispielweise *Klumpenstichproben*, bei denen die Auswahlwahrscheinlichkeiten erster Ordnung für jedes Element der Grundgesamtheit identisch, die Auswahlwahrscheinlichkeiten zweiter Ordnung jedoch unterschiedlich sein können. Dies ist abhängig davon, ob Einheit i und j Teil desselben Klumpens sind oder nicht sowie davon, ob ein Klumpen vollständig oder nur teilweise erhoben wird.

Betrachten wir die *geschichtete Zufallsauswahl* als weiteres Beispiel für komplexe Stichproben. Gehen wir davon aus, dass die Gesamtheit U in H Schichten zerlegt ist, so würde bspw. eine Schichtung Deutschlands nach Bundesländern 16 Schichten ergeben. Werden aus der h-ten Schicht vom Umfang N_h genau n_h Einheiten uneingeschränkt zufällig ausgewählt, erhält man

$$\pi_i = \frac{n_h}{N_h} \text{ für } i \text{ aus Schicht } h$$

$$\pi_{ij} = \frac{n_h(n_h-1)}{N_h(N_h-1)} \text{ für } i \neq j \text{ und beide aus Schicht } h$$

$$\pi_i = \frac{n_h n_k}{N_h N_k} \text{ für } i \text{ aus Schicht } h \text{ und für } j \text{ aus Schicht } k, \text{ mit } h \neq k.$$

Häufig werden Einheiten der ersten Stufe, etwa Gemeinden, mit ungleichen Wahrscheinlichkeiten gezogen. Großstädte erhalten eine größere Auswahlwahrscheinlichkeit als kleine Gemeinden. Ein Element, das a priori jedoch eine sehr geringe Chance hat in die Stichprobe zu gelangen, ist, wenn es doch ausgewählt wird, "*gewichtiger*" als ein Element, das a priori eine sehr hohe Wahrscheinlichkeit hatte, gezogen zu werden. Dem Element mit einer *geringen Auswahlwahrscheinlichkeit* kommt daher ein *hohes Gewicht* zu, dem Element mit einer *hohen Auswahlwahrscheinlichkeit* dagegen ein *geringes Gewicht*.

Dem Horvitz-Thompson-Schätzer folgend werden Designgewichte als Inverse der Auswahlwahrscheinlichkeit für die ausgewählten Einheiten berechnet und an die Umfragedaten gespielt. Das Designgewicht d_i entspricht somit $\frac{1}{\pi_i}$.

Um Extremgewichte zu vermeiden, kann für die Gewichte eine Transformation vorgenommen werden, die die extremen Gewichte in ein vorgegebenes Intervall abbildet. Ein solches Vorgehen wird oftmals als „Trimmen" von Gewichten bezeichnet und wird hauptsächlich aufgrund der zuvor geschilderten „Probleme" einer evtl. hohen Varianz verwendet, die sich bei einer starken Spannweite der Gewichte einstellen kann. Die Gewichte werden auf Basis der vorgegebenen Intervallgrenzen (entweder als fixe Grenzwerte oder als Perzentil der Verteilung der Gewichte) renormiert, wobei die Summe aller Gewichte gleich bleiben muss. Ein solches Vorgehen lässt sich als Argument oder Funktion des *Mean Squared Errors (MSE)* verstehen. Der MSE ist die Summe der quadrierten Verzerrung und der Varianz eines Schätzers. Bei einem solchen Trimmen nimmt der Forscher eine größere Verzerrung (als beim nicht-getrimmten Schätzer) zugunsten einer geringeren Varianz in Kauf, unter der Annahme, dass die Einbußen der Genauigkeit (gemessen durch die Verzerrung) durch die Steigerung der Präzision (gemessen an der Varianz) zumindest ausgeglichen werden (Yu, 1994).

Ein Beispiel: Gewichtung von Telefonstichproben anhand der GLES-RCS 2017-Daten

Als Anwendungsbeispiel der Designgewichtung eignen sich Telefonstichproben, da sich hier die Bestimmung der Auswahlwahrscheinlichkeiten oftmals komplexer gestaltet und eine stärkere Variabilität zwischen einzelnen Erhebungseinheiten vorliegt, die nicht (ausschließlich) auf der Mehrstufigkeit des Auswahlprozesses beruht. Es wurde die *Rolling-Cross-Section-Wahlkampfstudie (RCS)* hierfür ausgewählt.

Die RCS-Studie ist eine Komponente der *German Longitudinal Election Study (GLES)*, die von GESIS in Zusammenarbeit mit der Deutschen Gesellschaft für Wahlforschung (DGfW) durchgeführt wird. Es handelt sich um eine Dual-Frame-Telefonbefragung mit Nachwahl-Panelwelle (Roßteutscher et al., 2019). Um die Komplexität des nachfolgenden Beispiels zu reduzieren beschränken sich die Ausführungen auf die Vorwahlbefragung aus dem Jahr 2017.

Dual-Frame-Erhebungen werden mittlerweile bei deutschlandweiten Erhebungen standardmäßig aufgrund der Mobile- und Landline-Only Haushalte empfohlen (ADM, Arbeitskreis Deutscher Markt und Sozialforschungsinstitute, 2017; Häder & Sand, 2019; Sand, 2018). Beide Stichproben - die Festnetz- und Mobilfunkstichprobe - basieren auf einem komplexen Auswahlverfahren, bei dem einzelne Erhebungseinheiten unterschiedliche Auswahlwahrscheinlichkeiten haben.

Gabler et al. (2012) beschreiben die Wahrscheinlichkeit einer Erhebungseinheit i über eine Festnetzstichprobe (π_i^F) gezogen zu werden mit

$$\pi_i^F = \frac{m^F}{M^F} * \frac{k_i^F}{z_i}$$

und diejenige für eine Mobilfunkstichprobe π_i^C mit

$$\pi_i^C = \frac{m^C}{M^C} * k_i^C$$

m sowie M entsprechen der Anzahl der Rufnummern in der Bruttostichprobe sowie der Anzahl der Rufnummern innerhalb des jeweiligen Auswahlrahmens. Der Hochindex (C, F) gibt an, ob es sich dabei um eine Angabe zu Festnetz oder Mobilfunk handelt. k_i entspricht der Anzahl der Rufnummern, über die die Erhebungseinheit i über das jeweilige Erhebungsmedium erreichbar ist. z_i gibt die Anzahl erhebungsrelevanter Haushaltsmitglieder der Einheit i an. $1/z_i$ wird indes oftmals zur Bestimmung von Transformationsgewichten für Personenauswertungen basierend auf Haushaltsstichproben herangezogen.[8]

Bei Dual-Frame-Erhebungen muss zusätzlich die gemeinsame Auswahlwahrscheinlichkeit derjenigen Erhebungseinheiten berücksichtigt werden, die sowohl über die Festnetz- als auch über die Mobilfunkstichprobe in die Erhebung gelangen können (Dual-User). Gabler et al. (2012) empfehlen die Verwendung eines Schätzers, der die gemeinsame Auswahlwahrscheinlichkeit anhand der Summe der beiden einzelnen Auswahlwahrscheinlichkeiten bestimmt und diese dann zur Berechnung der Designgewichte verwendet. Die hier getroffene Annahme ist, dass die Wahrscheinlichkeit, dass eine Erhebungseinheit sowohl über Festnetz als auch über Mobilfunk in die Erhebung gelangt ($\pi_i^F * \pi_i^C$), vernachlässigbar ist. Die gemeinsame Auswahlwahrscheinlichkeit π_i bestimmt sich dann nach

$$\pi_i \approx \pi_i^F + \pi_i^C.$$

Der designgewichtete Horvitz-Thompson-Schätzer wird dann wie in Kapitel 2.1 dargestellt berechnet.

[8] Gelegentlich wird in der Literatur empfohlen, die Anzahl erhebungsrelevanter Personen auch in der Bestimmung der Auswahlwahrscheinlichkeit für Mobilfunkstichproben heranzuziehen (bspw. ADM, Arbeitskreis Deutscher Markt und Sozialforschungsinstitute, 2017). Da jedoch weitgehend Konsens darüber besteht, dass ein Mobiltelefon ein personenbezogener Gegenstand ist (siehe bspw. Busse & Fuchs, 2013) wird an dieser Stelle darauf verzichtet.

Im Rahmen der GLES-RCS-Wahlkampfstudie 2017 wurden nach Angaben des Methodenberichts (Roßteutscher et al., 2019) 361.900 Festnetz- sowie 155.100 Mobilfunkrufnummern gezogen. Da die Größe des ADM-Auswahlrahmens für die jeweiligen Erhebungen nicht aus dem Methodenbericht hervorgeht wurden als Proxy die entsprechenden Größen aus der GESIS-Auswahlgrundlage für diesen Zeitraum verwendet. Der GESIS Auswahlrahmen für Festnetzstichproben beinhaltete 2017 178.831.800 Rufnummern, für Mobilfunkstichproben waren es 328.710.000 Rufnummern. Sowohl für die Festnetz- als auch für die Mobilfunkstichprobe wurden im Rahmen der Erhebung die Anzahl der Festnetz- und Mobilfunkrufnummern abgefragt. Diese sind in der aktuellen Version des Datensatzes[9] unter den Variablenbezeichnungen *pre128* (Anzahl der Festnetzrufnummern Festnetzbefragter), *pre129* (Anzahl der Festnetzrufnummern Mobilbefragter) sowie *pre131* (Anzahl der Mobilfunkrufnummern) zu finden. pre128 und pre129 sind hierbei disjunkte Angaben, weshalb in Abhängigkeit des jeweiligen Befragungsmodus der Wert einer der beiden Variablen auf *NA* gesetzt ist.[10] Zur Datenverarbeitung wurden *NAs* hier auf den Wert „0“ gesetzt. Die Variablenbezeichnung *pre116* gibt die Anzahl der erhebungsrelevanten Haushaltsmitglieder wieder. Fehlende Werte wurden (zu Veranschaulichungszwecken) auf den Wert „1“ gesetzt, da davon auszugehen ist, dass bei erfolgreicher Stichprobenziehung mindestens eine Person im Haushalt erhebungsrelevant ist. Die Auswahlwahrscheinlichkeit einer Erhebungseinheit i berechnet sich auf Basis des vorliegenden Datensatzes nach

$$\pi_i = \frac{361900}{1788318} \frac{(pre128_i + pre129_i)}{pre116_i} + \frac{155100}{3287100} pre131_i.$$

Für die laufende Nummer (ID) *10* würde sich so eine Auswahlwahrscheinlichkeit von

$$\pi_i = \frac{3619}{1788318} \frac{1}{3} + \frac{1551}{3287100} 1 = 0,00115$$

ergeben. Das Designgewicht d_i für den Horvitz-Thompson-Schätzer des Totalwertes würde dann durch die Inverse der Auswahlwahrscheinlichkeiten bestimmt werden. Eine Reskalierung dieser Designgewichte, sodass diese in ihrer Summe der Stichprobengröße entsprechen und einen Mittelwert von „1“ aufweisen, würde dann durch

$$d_i^* = \frac{d_i}{\sum_{i \in S} d_i} n$$

erfolgen, mit n als Stichprobengröße und S als Menge der Elemente innerhalb der Stichprobe. Für den Fall mit der laufenden Nummer *10* entspricht dies

$$d_i^* = \frac{872,29}{4659901} 7650 = 1,432.$$

Die Verteilung aller normierten Gewichte kann in Abbildung 3 abgelesen werden.

Die Auswirkungen der Designgewichtung auf die Schätzung können anhand Tabelle 1 und Tabelle 2 nachvollzogen werden.[11]

[9] Datensatz unter url{https://dbk.gesis.org/dbksearch/sdesc2.asp?no=6803&db=e&doi=10.4232/1.13213} abrufbar.

[10] Bei Befragung über das Mobiltelefon ist der Wert der Variablen pre128 = NA, während für Festnetzbefragte pre129 = *NA* gilt. Im supplementären Skript wurden die beide Variablen zu einem Vektor kombiniert.

[11] Um mit einer vollständigen Datenmatrix für die soziodemografischen Charakteristika arbeiten zu können, wurde eine nichtparametrische Imputation des um einen Großteil der Variablen reduzierten Datensatzes unter Zuhilfenahme des Random Forest Algorithmus verwendet (siehe hierzu das R-Paket *missForest*). Dieses Vorgehen wurde lediglich zu Demonstrationszwecken gewählt. Es empfiehlt sich, dass bei einer Imputation unter realen Bedingungen ein möglichst variablenreicher Datensatz verwendet wird. Weiterhin wurde in der Anwendung des besagten Pakets festgestellt, dass die Skalenniveaus der metrischen Variablen nicht eingehalten wurden. Dies sollte vermieden werden, weswegen die Autoren zur Imputation fehlender Werte in R ausdrücklich die Verwendung des Pakets *mice* empfehlen.

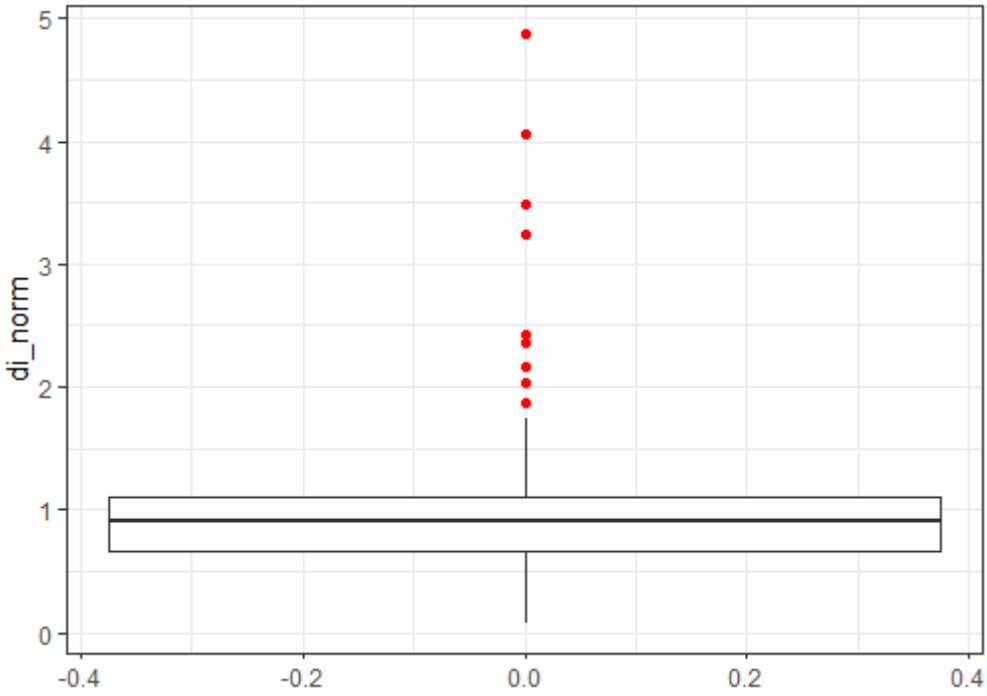

Abbildung 3: Boxplot der normierten Gewichte

Tabelle 1: Alter nach Gruppen in %

Alterskategorie	Ungewichtet	Designgewichtet	Bevölkerungsfortschreibung 2017
18 – 20	2,29	3,41	3,03
21 – 29	6,52	9,48	9,8
30 – 49	24,31	24,56	23,94
50 – 59	23,16	23	16,78
60 - 64	11,01	10,13	6,87
65 +	32,71	29,42	23,06

Tabelle 2: Geschlecht %

Geschlecht	Ungewichtet	Designgewichtet	Bevölkerungsfortschreibung 2017
Männlich	51,65	52,08	48,82
Weiblich	48,35	47,92	51,18

Aus beiden Tabellen geht hervor, dass eine Designgewichtung nicht immer zu einer Verringerung der Verzerrung führen muss. Dies kann bei Erhebungen in Folge von Antwortausfällen oder aufgrund eines unvollständigen Auswahlrahmens der Fall sein. Gerade in solchen Fällen ist eine Kalibrierung, wie sie nachfolgend erläutert wird, zu empfehlen.

2.2 Wie werden Kalibrierungs- und Anpassungsgewichte berechnet?

Wie bereits in Kapitel 1 erwähnt, werden die Begriffe der Kalibrierung und Anpassungsgewichtung häufig synonym verwendet. Jedoch kann die Anpassungsgewichtung als Spezialfall der Kalibrierung zur Reduktion von Verzerrungen aufgrund von Antwortausfällen betrachtet werden. Dass allerdings eine Kalibrierung auch ohne das Vorkommen solcher Antwortausfälle zur Steigerung der Präzision des Schätzers sinnvoll sein kann, wurde bereits in Kapitel 1.2 erläutert.

Unabhängig davon, ob eine Kalibrierung zur Steigerung der Präzision und/oder zur Verringerung der Verzerrung aufgrund von Antwortausfällen durchgeführt wird, bedarf es für deren Durchführung Hilfsvariablen, deren Verteilung sowohl für die Stichprobe als auch für die Grundgesamtheit bekannt ist. Durch die Anpassung der designgewichteten Schätzer an die Verteilung dieser Hilfsvariablen in der Grundgesamtheit soll dann die Genauigkeit und/oder Präzision der Schätzung verbessert werden.

Je nachdem, ob die gemeinsame Verteilung der Anpassungsvariablen (vollständige Kreuztabelle in der Stichprobe und Grundgesamtheit verfügbar) oder nur deren Randverteilungen in der Grundgesamtheit bekannt sind, unterscheidet man mehrere Fälle. Die gängigsten werden nachfolgend erläutert:

1. Wenn die gemeinsame Verteilung der Hilfsvariablen, auf die angepasst wird, bekannt ist und alle Kreuzkombinationen der Hilfsvariablen sowohl in der Grundgesamtheit als auch in der Stichprobe (hinreichend) vorliegen, wird oftmals eine ***Poststratifikation*** *(nachträgliche Schichtung)* angewendet. Die Kalibrierungsgewichte werden durch eine einfache Soll-/Ist-Anpassung ermittelt. Voraussetzung ist, dass alle Merkmalskombinationen der Hilfsvariablen für die Grundgesamtheit (als Kontingenztabelle) verfügbar sind. Die Gewichtung erfolgt für die jeweiligen Poststrata ($h = 1, \ldots, H$), sodass diese für die Grundgesamtheit und die Stichprobe hinsichtlich des (geschätzten) Totalwertes (oder des Anteils) der Subpopulation übereinstimmen. Die Schichten werden als Zellen bezeichnet. Nimmt für die Respondenten R_h der Schicht h aus der Stichprobe S die Variable ϑ_{hi} für die Erhebungseinheit i den Wert 1 an, so ist das Ziel dieser Gewichtung für die Menge der Respondenten aus Schicht h die Poststratifizierungsgewichte g_h^{POST} so zu entwickeln, dass

$$\sum_{i \in S} g_h^{POST} * d_i * \vartheta_{hi} = \sum_{i \in R_h} g_h^{POST} * d_i = N_h$$

   gilt. N_h entspricht hierbei der Anzahl an Einheiten der Grundgesamtheit in Schicht h. Diese wird anhand der Designgewichte durch

$$\hat{N}_h = \sum_{i \in R_h} d_i$$

   geschätzt, wonach sich das Kalibrierungsgewicht anhand von

$$g_h^{POST} = \frac{N_h}{\hat{N}_h}$$

   ergibt. Bei dieser Soll-/Ist-Gewichtung steht im Nenner die Summe der Designgewichte in Schicht h und im Zähler die Anzahl der Einheiten der Grundgesamtheit aus dieser Schicht. Ist der Ausfallprozess in jeder Zelle zufällig, so kann dieses Vorgehen gute Schätzungen hervorbringen.

Die Verfügbarkeit von Kontingenztabellen aller Hilfsvariablen in der Grundgesamtheit kann jedoch oftmals problematisch sein. Weiterhin können Zellen, die innerhalb der Stichprobe unbesetzt und innerhalb der Grundgesamtheit besetzt sind, die Anwendung dieser Gewichtung erschweren. Zuletzt kann bei der Verwendung mehrerer Hilfsvariablen das Problem entstehen, dass durch die vielen Merkmalskombinationen mehr Zellen gebildet werden müssen, als Einheiten in der Stichprobe sind oder die Zellen lediglich

mit sehr wenigen Einheiten in der Stichprobe besetzt sind. Wenn sich Aggregieren nicht anbietet, ist dies problematisch für die Verwendung der Poststratifikation.

In solchen Fällen wird oftmals lediglich auf die Randverteilungen der Hilfsvariablen in der Grundgesamtheit kalibriert. Dabei stimmen nach der Kalibrierung die einzelnen Verteilungen der Hilfsvariablen in Stichprobe und Grundgesamtheit überein, deren Kreuzkombination jedoch nicht unbedingt. Zwei unterschiedliche Gewichtungsansätze zu dieser Problemstellung werden folgend erläutert:

2. ***Raking***-Verfahren stellen eine Gruppe von iterativen Anpassungen auf die Randverteilung der Hilfsvariablen in der Grundgesamtheit dar. Das wohl bekannteste und am häufigsten angewendete Verfahren ist der sog. *Iterative Proportional Fitting (IPF) Algorithmus*, der auf Deming und Stephan (1941) zurückgeht (Gabler et al., 2015). Voraussetzung zur Anwendung des Verfahrens ist, dass für alle verwendeten Hilfsvariablen die Randverteilungen in der Grundgesamtheit bekannt sind. Die gemeinsame Verteilung der Hilfsvariablen kann mit Hilfe des Verfahrens zwar nicht geschätzt werden (außer diese sind vollständig unabhängig); jedoch können zumindest die Randverteilungen der Hilfsvariablen anhand der gewichteten Stichprobenergebnisse unverzerrt geschätzt werden. Dies ermöglicht eine Schichtung nach mehreren Anpassungsvariablen, ohne dass deren gemeinsame Verteilung in der Stichprobe und Grundgesamtheit bekannt sein muss.

   Das Vorgehen zur Entwicklung dieser Gewichte ist dem der Poststratifikation ähnlich. Im Unterschied zur Poststratifikation wird allerdings in jedem Schritt lediglich eine Soll-/Ist-Anpassung auf eine einzelne der Randverteilungen vorgenommen. Dieser Vorgang wird so lange wiederholt (Iterationen), bis sich die dadurch entwickelten Gewichte nicht mehr oder kaum noch verändern und die Verteilungen der Hilfsvariablen in der Stichprobe (annähernd) mit denjenigen in der Grundgesamtheit übereinstimmen (Lumley, 2010, S. 139).

   Sand und Gabler (2019) erläutern das Vorgehen bei dieser Gewichtung anhand von zwei Hilfsvariablen (Alter und Bildungsabschluss) wie folgt:

   - Stufe 1: Die Randverteilung der ersten Anpassungsvariablen (Alter) in der Grundgesamtheit wird durch die Randverteilung dieser Variablen in der Stichprobe dividiert. Die daraus berechneten Gewichte werden anschließend verwendet, um die Randverteilung der zweiten Anpassungsvariablen (Bildungsabschluss) in der Stichprobe neu zu aggregieren.
   - Stufe 2: Die Randverteilung der zweiten Anpassungsvariablen (Bildungsabschluss) in der Grundgesamtheit wird durch die Randverteilung dieser Variablen in der Stichprobe dividiert. Die daraus berechneten Gewichte werden anschließend verwendet, um die Randverteilung der ersten Anpassungsvariablen (Alter) in der Stichprobe neu zu aggregieren.
   - Die Stufen 1 und 2 werden solange wiederholt bis keine (nennenswerten) Änderungen in den Gewichten mehr sichtbar sind. Es gilt ferner zu beachten, dass nach Abschluss der ersten Iteration die Randverteilung der ersten Hilfsvariablen unter der Verwendung der bereits berechneten Kalibrierungsgewichte aggregiert wird.

   Das hier beschriebene Verfahren ist in aller Regel nicht auf zwei Hilfsvariablen beschränkt, sondern kann für beliebige Kreuzkombinationen einzelner Variablen verwendet werden.

3. ***verallgemeinerte Regressionsschätzer (Generalized Regression (GREG) Estimator)*** stellt ein Kalibrierungsverfahren dar, bei dem Designgewichtung und Anpassungsgewichtung kombiniert werden können.

   Bei komplexen Stichproben kann der Horvitz-Thompson-Schätzer so modifiziert werden, dass er einem Verhältnis- oder Regressionsschätzer ähnelt. Ähnlich wie bei anderen Kalibrierungsansät-

zen erfolgt eine Anpassung von Hilfsvariablen in der Stichprobe, dargestellt durch den Vektor x, anhand der Verteilung dieser Hilfsvariablen in der Grundgesamtheit. Jedoch wird hier das Stichprobendesign direkt in die Schätzung miteinbezogen. Särndal und Lundström (2005) verweisen auf eine Vielzahl positiver Eigenschaften, die dieser Schätzer aufweisen kann. So kann dessen Verwendung unabhängig von der Präsenz von Antwortausfällen eine Reduktion der Varianz des Schätzers $\hat{Y}$ eines Totalwertes der Grundgesamtheit herbeiführen. Das Ausmaß der Reduktion ist abhängig vom Zusammenhang der interessierenden Variablen y und der Matrix der Hilfsvariablen x.

Ein weiterer Vorzug dieses Schätzers liegt in der direkten Verwendung der Auswahlwahrscheinlichkeiten und damit der Berücksichtigung des Stichprobendesigns im Zuge der Kalibrierung. Darüber hinaus kann die (modellbasiert) geschätzte Antwortwahrscheinlichkeit (q_i) der Einheit i in die Kalibrierung miteinbezogen werden. Särndal und Lundström (2005) zufolge kann dies in zwei Phasen geschehen. Nach der Designgewichtung wird zunächst eine Anpassungsgewichtung zur Berücksichtigung von Nonresponse durchgeführt und anschließend durch den GREG-Schätzer kalibriert.

Der GREG-Schätzer eines Totalwertes der Grundgesamtheit $\hat{Y}_{GREG}$ wird anhand von

$$\hat{Y}_{GREG} = \sum_{i \in R} g_i^{GREG} * d_i * y_i$$

angegeben, wobei sich das Kalibrierungsgewicht g_i^{GREG} nach

$$g_i^{GREG} = \frac{1}{q_i}\Big(1 + c_i\Big(\sum_{i \in U} x_i - \sum_{i \in R} \frac{1}{\pi_i q_i} x_i\Big)' \Big(\sum_{i \in R} \frac{c_i}{\pi_i q_i} x_i x_i'\Big)^{-1} x_i\Big)$$

berechnen lässt. U entspricht dabei der Menge der Einheiten der Grundgesamtheit, q_i der Antwortwahrscheinlichkeit und c_i einem Gewicht für Einheit i, das vom Forschenden selbst festgelegt werden kann. Die Antwortwahrscheinlichkeit wird oftmals durch die Bildung von Antwortgruppen oder anhand der Schätzung (bspw. durch ein entsprechendes (logistisches) Modell) ermittelt (Gabler et al., 2015). Da jedoch unter der Erfüllung der MAR-Annahme ein kalibrierter Schätzer ebenfalls (annähernd) unverzerrt ist, wird von einer separaten Anpassungsgewichtung zur Berücksichtigung von Nonresponse oftmals abgesehen und q_i erhält den Wert 1. Gleiches gilt für das zusätzliche Gewicht c_i. Besteht kein besonderer Grund, einer oder mehreren Erhebungseinheiten ein (zusätzlich zur Designgewichtung) höheres oder niedrigeres Gewicht zu geben, wird dies oftmals ebenfalls auf 1 gesetzt (Sand und Gabler, 2019; Särndal und Lundström, 2005). Im einfachsten Fall kann g_i^{GREG} nach

$$g_i^{GREG} = 1 + \Big(\sum_{i \in U} x_i - \sum_{i \in R} \frac{1}{\pi_i} x_i\Big)' \Big(\sum_{i \in R} \frac{1}{\pi_i} x_i x_i'\Big)^{-1} x_i$$

berechnet werden.

Ein Beispiel zur Anwendung des GREG-Schätzers[12]

Betrachten wir ein einfaches Beispiel, bei dem man die Antwortwahrscheinlichkeiten kennt. In einer Firma werden von den 300 Männern und 1000 Frauen jeweils 100 zufällig ausgewählt. 30 Männer und 50 Frauen antworten davon auf Fragen der Zufriedenheit mit dem Arbeitsplatz. Von den 30 Männern sind 20 zufrieden, während 10 Frauen zufrieden sind. Will man die Zahl aller Mitarbeiter der Firma schätzen, die mit dem Arbeitsplatz zufrieden sind, würde man

$$20 \cdot \frac{1}{\frac{100}{300} \cdot 0,3} + 10 \cdot \frac{1}{\frac{100}{1000} \cdot 0,5} = 400$$

[12] Das folgende Beispiel wurde aus Gabler, Kolb, Sand & Zins (2015) unverändert übernommen.

als Schätzwert berechnen. Der Anteil der mit dem Arbeitsplatz zufriedenen Mitarbeiter beläuft sich daher schätzungsweise auf 31%. Dabei geht man davon aus, dass alle Männer jeweils mit Wahrscheinlichkeit 0,3 und alle Frauen jeweils mit Wahrscheinlichkeit 0,5 antworten. Weiß man jedoch aus Erfahrung, dass 24% der Männer und 50% aller Frauen antworten, hätte man

$$20 \cdot \frac{1}{\frac{100}{300} \cdot 0,24} + 10 \cdot \frac{1}{\frac{100}{1000} \cdot 0,5} = 250 + 200 = 450$$

als Schätzergebnis. Diese Schätzung hat allerdings einen Nachteil. Hätte man nämlich nach der Zahl der mit dem Arbeitsplatz unzufriedenen Mitarbeiter gefragt, hätte man

$$10 \cdot \frac{1}{\frac{100}{300} \cdot 0,24} + 40 \cdot \frac{1}{\frac{100}{1000} \cdot 0,5} = 125 + 800 = 925$$

erhalten und daher die Zahl aller Mitarbeiter auf 1375 geschätzt. Der geschätzte Anteil der mit der Arbeit zufriedenen Mitarbeiter plus den geschätzten Anteil der mit der Arbeit unzufriedenen Mitarbeiter addiert sich nicht zu eins. Diesem Umstand kann man dadurch Rechnung tragen, dass eine Kalibrierung an einen Vektor, der ausschließlich Einsen enthält, vorgenommen wird. Setzt man für die Kalibrierungsgewichte g_i^{GREG} die Werte $x_i = 1$, $c_i = 1$ für alle i und $\pi_i = 100/300 = 1/3$ für Männer bzw. $\pi_i = 100/1000 = 0,1$ für die Frauen, sowie $q_i = 0,24$ für Männer bzw. $q_i = 0,5$ für Frauen, und daher

$$g_i^{GREG} = \frac{1}{q_i} * \frac{N}{\sum_{i \in R} \frac{1}{\pi_i q_i}} = \frac{130}{11}$$

für Männer und $\frac{208}{11}$ für Frauen, so hätte man als Schätzer

$$20 \cdot \frac{130}{11} + 10 \cdot \frac{208}{11} = \frac{4680}{11} = 425,45$$

für die Zahl der mit dem Arbeitsplatz zufriedenen Mitarbeiter, d.h. 32,72% und

$$10 \cdot \frac{130}{11} + 40 \cdot \frac{208}{11} = \frac{9620}{11} = 874,55$$

für die Anzahl der mit dem Arbeitsplatz unzufriedenen Mitarbeiter, d.h. 67,28%. Die geschätzte Summe aller Mitarbeiter wäre dann 1300, was wiederum der tatsächlichen Mitarbeiterzahl entspricht.

Ein Beispiel: Kalibrierung anhand der Daten des ALLBUS 2016

Zur Demonstration der unterschiedlichen Kalibrierungsverfahren wurden die Daten des ALLBUS 2016[13] herangezogen. Diese wurden zunächst unter der Zuhilfenahme der bereits bereitgestellten Designgewichte zum Ausgleich der disproportionalen Aufteilung der Stichprobenpopulation zwischen neuen und alten Bundesländern gewichtet. Da es sich bei dem vorliegenden Stichprobendesign (unter Ausschluss der disproportionalen Aufteilung) um eine sog. selbstgewichtende Stichprobenstrategie handelt wurde lediglich die Populationsgröße geteilt durch die Stichprobengröße (N/n) als Designgewicht für alle Erhebungseinheiten verwendet. Diese wurden mit dem Ost-/West-Gewicht (*wghtpew*) multipliziert. Die daraus resultierenden Verteilungen wurden dann anhand der Bevölkerungsfortschreibung für das Jahr 2016 von Destatis[14] kalibriert.

[13] Siehe: GESIS - Leibniz-Institut für Sozialwissenschaften (2017): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2016. GESIS Datenarchiv, Köln. ZA5250 Datenfile Version 2.1.0, doi:10.4232/1.12796

[14] Siehe: https://www-genesis.destatis.de/genesis/online/data?operation=statistic&levelindex=0&levelid=1573387447833&code=12411

Im Folgenden werden die drei zuvor erläuterten Kalibrierungsverfahren dargestellt. Als Kalibrierungsvariablen wurden die Verteilungen von Alter und Bundesland verwendet (siehe Tabelle 3). Hierbei gilt zu beachten, dass ein Kalibrierungsmodell für gewöhnlich mehr als zwei Variablen beinhaltet. Da es sich bei den hier dargestellten Ergebnissen jedoch um eine Demonstration der Kalibrierungsverfahren handelt wurde aus Gründen der Übersichtlichkeit und der Verfügbarkeit der Daten ein reduziertes Modell verwendet.

Tabelle 3: Bevölkerungsverteilung nach Alter und Bundesland in Prozent

Bundesland/Alter	18-29	30-44	45-59	60-74	75-89	90+	Summe
BADEN-WÜRTTEMBERG	2,42	2,95	3,69	2,43	1,54	0,14	**13,17**
BAYERN	2,78	3,55	4,42	2,93	1,81	0,16	**15,65**
BERLIN	0,81	1,15	1,11	0,77	0,47	0,04	**4,34**
BRANDENBURG	0,36	0,63	0,93	0,67	0,44	0,03	**3,06**
BREMEN	0,16	0,19	0,21	0,16	0,10	0,01	**0,83**
HAMBURG	0,43	0,59	0,56	0,36	0,23	0,02	**2,19**
HESSEN	1,31	1,68	2,12	1,43	0,87	0,08	**7,50**
MECKLENB.-VORPOMMERN	0,26	0,41	0,58	0,44	0,28	0,02	**1,98**
NIEDERSACHSEN	1,63	1,98	2,76	1,90	1,21	0,11	**9,58**
NORDRHEIN-WESTFALEN	3,79	4,63	6,17	4,13	2,64	0,23	**21,59**
RHEINLAND-PFALZ	0,84	1,02	1,43	0,98	0,61	0,06	**4,94**
SAARLAND	0,20	0,24	0,36	0,26	0,17	0,01	**1,23**
SACHSEN	0,68	1,07	1,31	1,12	0,76	0,07	**5,01**
SACHSEN-ANHALT	0,36	0,54	0,79	0,64	0,41	0,03	**2,78**
SCHLESWIG-HOLSTEIN	0,56	0,71	1,02	0,71	0,45	0,04	**3,49**
THÜRINGEN	0,34	0,55	0,74	0,61	0,38	0,03	**2,66**
Summe	**16,93**	**21,87**	**28,19**	**19,55**	**12,37**	**1,09**	**1,00**

Tabelle 4 zeigt die Verteilung nach Alter und Bundesland auf Basis der Daten des ALLBUS 2016. Hier wird eine Complete Case Analyse durchgeführt, wodurch sich die Anzahl der Beobachtungen von 3.490 auf 3.486 reduziert. Angaben für ehemals Ost- und ehemals West-Berlin im ALLBUS 2016 wurden anhand der Variablen land zu einer Ausprägung „Berlin“ zusammengefasst.

Tabelle 4: Verteilung nach Alter und Bundesland im ALLBUS 2016
in Prozent

Bundesland/Alter	18-29	30-44	45-59	60-74	75-89	90+	Summe
BADEN-WÜRTTEMBERG	2,09	3,50	4,43	2,48	1,06	0,04	**13,59**
BAYERN	2,09	3,54	4,71	3,65	1,66	0,07	**15,72**
BERLIN	0,44	0,61	0,88	0,83	0,35	0,03	**3,13**
BRANDENBURG	0,34	0,63	1,15	0,89	0,46	0,02	**3,47**
BREMEN	0,07	0,11	0,07	0,21	0,07	NA	**0,53**
HAMBURG	0,32	0,64	0,67	0,53	0,18	NA	**2,34**
HESSEN	1,45	2,02	2,76	1,73	0,71	0,04	**8,71**
MECKLENB.-VORPOMMERN	0,17	0,28	0,75	0,44	0,35	NA	**1,99**
NIEDERSACHSEN	1,31	2,16	3,61	2,16	0,96	NA	**10,20**
NORDRHEIN-WESTFALEN	3,79	4,18	6,23	4,71	1,77	0,04	**20,71**
RHEINLAND-PFALZ	0,81	1,06	1,42	1,03	0,18	0,07	**4,57**
SAARLAND	0,18	0,11	0,25	0,28	0,14	NA	**0,96**
SACHSEN	0,54	0,89	1,33	1,35	0,70	NA	**4,81**
SACHSEN-ANHALT	0,41	0,44	1,01	0,93	0,43	NA	**3,23**
SCHLESWIG-HOLSTEIN	0,35	0,50	0,99	0,99	0,35	NA	**3,19**
THÜRINGEN	0,41	0,66	0,73	0,73	0,32	NA	**2,86**
Summe	**14,77**	**21,31**	**30,99**	**22,95**	**9,70**	**0,29**	**1,00**

Ein Problem, das sich bei einer *Poststratifikation* bzw. Zellgewichtung oftmals einstellt, wenn Variablen mit vielen Ausprägungen und/oder viele Kalibrierungsvariablen herangezogen werden: In der Spalte der mindestens 90-Jährigen werden einige der Kombinationen mit „*NA*“ ausgewiesen, was darauf hinweist, dass diese Merkmalskombinationen im Datensatz nicht vorhanden sind. Wenn Merkmalskombinationen in der Tabelle mit *NA* bezeichnet werden, aber in der Grundgesamtheit existieren, spricht man von sog. Nullzellen. Eine Soll-/Ist-Gewichtung ist dann nicht möglich (da man durch Null dividieren müsste). Es besteht jedoch die Möglichkeit, ein anderes Kalibrierungsverfahren zu wählen oder Zellen zu aggregieren. Letzteres wurde im obigen Beispiel umgesetzt, indem die Merkmalsausprägungen 75–89 und 90+ zur Ausprägung 75+ zusammengefasst wurden. Anschließend wurde die Poststratifikation durchgeführt. Tabelle 5 gibt die Poststratifizierungsgewichte wieder.

Durch Multiplikation der Verteilungswerte in Tabelle 4 mit den entsprechenden Poststratifizierungsgewichten in Tabelle 5 (z.B. 2,09 * 1,16 für 18-29 Jährige in Baden-Württemberg) erhält man die Verteilungen, wie sie in Tabelle 3 abgebildet sind (nachdem auch hier die Merkmalsausprägungen 75–89 und 90+ zusammengefasst wurden).

Tabelle 5: Poststratifizierungsgewichte im ALLBUS 2016

Bundesland/Alter	18-29	30-44	45-59	60-74	75+
BADEN-WÜRTTEMBERG	1,16	0,84	0,83	0,98	1,53
BAYERN	1,33	1,00	0,94	0,80	1,14
BERLIN	1,86	1,89	1,25	0,94	1,33
BRANDENBURG	1,06	1,00	0,81	0,76	0,98
BREMEN	2,26	1,75	3,04	0,73	1,54
HAMBURG	1,34	0,92	0,84	0,68	1,45
HESSEN	0,90	0,83	0,77	0,82	1,29
MECKLENB.-VORPOMMERN	1,54	1,48	0,77	1,00	0,85
NIEDERSACHSEN	1,24	0,92	0,76	0,88	1,37
NORDRHEIN-WESTFALEN	1,00	1,11	0,99	0,88	1,59
RHEINLAND-PFALZ	1,03	0,96	1,01	0,96	2,69
SAARLAND	1,12	2,25	1,44	0,92	1,27
SACHSEN	1,27	1,21	0,98	0,83	1,17
SACHSEN-ANHALT	0,86	1,23	0,78	0,69	1,04
SCHLESWIG-HOLSTEIN	1,59	1,42	1,03	0,72	1,39
THÜRINGEN	0,83	0,83	1,01	0,83	1,28

Im Folgenden wird anhand der Tabelle 6 bis Tabelle 8 das *Raking-Verfahren* demonstriert. Da dieses lediglich auf die Randverteilung, aber nicht auf die gemeinsame Verteilung der Kalibrierungsvariablen abzielt, musste keine Aggregation der Altersvariablen durchgeführt werden.

Tabelle 6: Raking Schritt 1

Bundesland/Alter	18-29	30-44	45-59	60-74	75-89	90+	Summe	Destatis
BADEN-WÜRTTEMBERG	2,40	3,60	4,03	2,11	1,36	0,13	**13,62**	**13,17**
BAYERN	2,40	3,63	4,28	3,11	2,12	0,26	**15,80**	**15,65**
BERLIN	0,50	0,62	0,80	0,70	0,45	0,11	**3,19**	**4,34**
BRANDENBURG	0,39	0,64	1,04	0,76	0,59	0,06	**3,47**	**3,06**
BREMEN	0,08	0,11	0,06	0,18	0,09	NA	**0,53**	**0,83**
HAMBURG	0,37	0,65	0,61	0,45	0,23	NA	**2,31**	**2,19**
HESSEN	1,66	2,07	2,51	1,48	0,90	0,13	**8,76**	**7,50**
MECKLENB.-VORPOMMERN	0,19	0,28	0,68	0,38	0,45	NA	**1,99**	**1,98**
NIEDERSACHSEN	1,50	2,22	3,28	1,84	1,22	NA	**10,06**	**9,58**
NORDRHEIN-WESTFALEN	4,34	4,29	5,67	4,01	2,26	0,13	**20,70**	**21,59**
RHEINLAND-PFALZ	0,93	1,09	1,29	0,87	0,23	0,26	**4,67**	**4,94**
SAARLAND	0,20	0,11	0,23	0,24	0,18	NA	**0,96**	**1,23**
SACHSEN	0,61	0,91	1,21	1,15	0,90	NA	**4,78**	**5,01**
SACHSEN-ANHALT	0,47	0,46	0,92	0,80	0,55	NA	**3,19**	**2,78**
SCHLESWIG-HOLSTEIN	0,41	0,51	0,90	0,84	0,45	NA	**3,11**	**3,49**
THÜRINGEN	0,47	0,68	0,67	0,63	0,41	NA	**2,85**	**2,66**
Summe	**16,93**	**21,87**	**28,19**	**19,55**	**12,37**	**1,09**	**1,00**	
Destatis	**16,93**	**21,87**	**28,19**	**19,55**	**12,37**	**1,09**		

In einem ersten Schritt erfolgt die Anpassung auf die Altersvariable sowie die anschließende Multiplikation der Ursprungsdaten mit den ermittelten Gewichten. Tabelle 6 zeigt, dass zwar nun die Altersverteilung mit der der Grundgesamtheit übereinstimmt, jedoch nicht mit der Verteilung der Bundesländer. Eine Anpassung auf die Bundeslandvariable erfolgt in einem zweiten Schritt. Die Ergebnisse der Multiplikation der auf Alter angepassten Daten mit den neu ermittelten Gewichten können anhand Tabelle 7 nachvollzogen werden.

Tabelle 7: Raking Schritt 2

Bundesland/Alter	18-29	30-44	45-59	60-74	75-89	90+	Summe	Destatis
BADEN-WÜRTTEMBERG	2,32	3,48	3,89	2,04	1,31	0,13	**13,17**	**13,17**
BAYERN	2,37	3,60	4,24	3,08	2,10	0,26	**15,65**	**15,65**
BERLIN	0,68	0,85	1,09	0,96	0,61	0,15	**4,34**	**4,34**
BRANDENBURG	0,34	0,57	0,92	0,67	0,52	0,05	**3,06**	**3,06**
BREMEN	0,13	0,17	0,10	0,28	0,14	NA	**0,83**	**0,83**
HAMBURG	0,35	0,62	0,58	0,43	0,21	NA	**2,19**	**2,19**
HESSEN	1,42	1,77	2,15	1,27	0,77	0,11	**7,50**	**7,50**
MECKLENB.-VORPOMMERN	0,19	0,28	0,68	0,38	0,45	NA	**1,98**	**1,98**
NIEDERSACHSEN	1,43	2,11	3,13	1,75	1,16	NA	**9,58**	**9,58**
NORDRHEIN-WESTFALEN	4,53	4,47	5,91	4,18	2,36	0,14	**21,59**	**21,59**
RHEINLAND-PFALZ	0,99	1,15	1,36	0,92	0,24	0,28	**4,94**	**4,94**
SAARLAND	0,26	0,14	0,29	0,31	0,23	NA	**1,23**	**1,23**
SACHSEN	0,64	0,95	1,27	1,20	0,94	NA	**5,01**	**5,01**
SACHSEN-ANHALT	0,41	0,40	0,80	0,69	0,48	NA	**2,78**	**2,78**
SCHLESWIG-HOLSTEIN	0,46	0,57	1,01	0,95	0,51	NA	**3,49**	**3,49**
THÜRINGEN	0,44	0,63	0,62	0,58	0,38	NA	**2,66**	**2,66**
Summe	**16,96**	**21,77**	**28,05**	**19,70**	**12,41**	**1,11**	**1,00**	
Destatis	**16,93**	**21,87**	**28,19**	**19,55**	**12,37**	**1,09**		

Es zeigt sich, dass zwar die Verteilung nach Bundesland mit der amtlichen Statistik übereinstimmt, die Verteilung nach Alter ist jedoch nicht mehr identisch mit derjenigen der amtlichen Statistik. Daher erfolgt nach dieser ersten Iteration eine erneute Anpassung auf die Verteilung nach Alter sowie eine Wiederholung des zweiten Schrittes und einer weiteren Iteration dieses Vorgehens. Nach 8 Iterationen, also 8 Wiederholungen der Schritte 1 und 2, ergibt sich die Verteilung wie sie in Tabelle 8 dargestellt ist.

Hier zeigt sich, dass beide Randverteilungen mit den Daten der amtlichen Statistik übereinstimmen. Eine nähere Inspektion der Werte in Tabelle 8 sowie der Vergleich mit Tabelle 3 zeigen jedoch, dass die einzelnen Merkmalskombinationen nicht übereinstimmen. Dies wäre lediglich bei einer reinen Poststratifikation der Fall.

Tabelle 8: Raking nach 8 Iterationen

Bundesland/Alter	18-29	30-44	45-59	60-74	75-89	90+	Summe	Destatis
BADEN-WÜRTTEMBERG	2,31	3,50	3,91	2,03	1,31	0,12	**13,17**	**13,17**
BAYERN	2,37	3,62	4,26	3,05	2,10	0,25	**15,65**	**15,65**
BERLIN	0,68	0,85	1,10	0,95	0,61	0,15	**4,34**	**4,34**
BRANDENBURG	0,34	0,57	0,92	0,66	0,51	0,05	**3,06**	**3,06**
BREMEN	0,13	0,17	0,10	0,28	0,14	NA	**0,83**	**0,83**
HAMBURG	0,35	0,62	0,58	0,43	0,21	NA	**2,19**	**2,19**
HESSEN	1,42	1,78	2,16	1,26	0,77	0,11	**7,50**	**7,50**
MECKLENB.-VORPOMMERN	0,19	0,28	0,69	0,38	0,45	NA	**1,98**	**1,98**
NIEDERSACHSEN	1,43	2,12	3,14	1,74	1,16	NA	**9,58**	**9,58**
NORDRHEIN-WESTFALEN	4,52	4,49	5,94	4,15	2,35	0,13	**21,59**	**21,59**
RHEINLAND-PFALZ	0,99	1,16	1,37	0,92	0,24	0,27	**4,94**	**4,94**
SAARLAND	0,26	0,14	0,29	0,31	0,23	NA	**1,23**	**1,23**
SACHSEN	0,64	0,96	1,27	1,19	0,94	NA	**5,01**	**5,01**
SACHSEN-ANHALT	0,41	0,40	0,80	0,69	0,47	NA	**2,78**	**2,78**
SCHLESWIG-HOLSTEIN	0,45	0,57	1,02	0,94	0,51	NA	**3,49**	**3,49**
THÜRINGEN	0,44	0,63	0,63	0,58	0,38	NA	**2,66**	**2,66**
Summe	**16,93**	**21,87**	**28,19**	**19,55**	**12,37**	**1,09**	**1,00**	
Destatis	**16,93**	**21,87**	**28,19**	**19,55**	**12,37**	**1,09**		

Zuletzt werden die Randverteilungen noch anhand des *GREG-Schätzers* kalibriert. Da sich hier allerdings keine einzelnen Zwischenergebnisse darstellen lassen, beschränken wir uns auf den (visuellen) Vergleich der Verteilung der Gewichte (siehe Abbildung 4).

Hierbei ist zur Kenntnis zu nehmen, dass alle drei Gewichtungsverfahren die Randverteilungen der Kalibrierungsvariablen abbilden. Es zeigt sich, dass die Verteilung der Gewichte nahezu identisch in ihrer Streuung sowie hinsichtlich der Median-Werte ist. Lediglich die Poststratifikation weicht leicht ab. So sind die Spannweite sowie der Median der Gewichte geringer. Es ist jedoch zu berücksichtigen, dass für die Kalibrierung unter der Verwendung einer Poststratifikation zwei Alterskategorien zusammengefasst wurden, was sowohl den geringeren Median als auch die geringere Streuung erklären kann. Letztendlich sind alle hier vorgestellten Kalibrierungsverfahren (approximativ und gegeben einer optimalen Modellierung des Nonresponse-Mechanismus) erwartungstreu. Demnach sollte die Entscheidung für ein bestimmtes Verfahren davon abhängig gemacht werden, welche Daten zur Anpassung verfügbar sind und was mit der Gewichtung bezweckt werden soll. Ist bspw. die gemeinsame Verteilung der Kalibrierungsvariablen in der Grundgesamtheit verfügbar und beinhaltet der Datensatz nicht zu viele „Stichprobennullen“, so stellt die Poststratifikation ein probates Mittel der Kalibrierung dar. Trifft dies jedoch nicht zu, so kann lediglich auf eines der beiden Verfahren über die Randverteilung zurückgegriffen werden. Wichtig ist jedoch, dass unter der Verwendung eines Raking-Ansatzes die Varianz des Schätzwertes aufwendiger zu ermitteln ist.

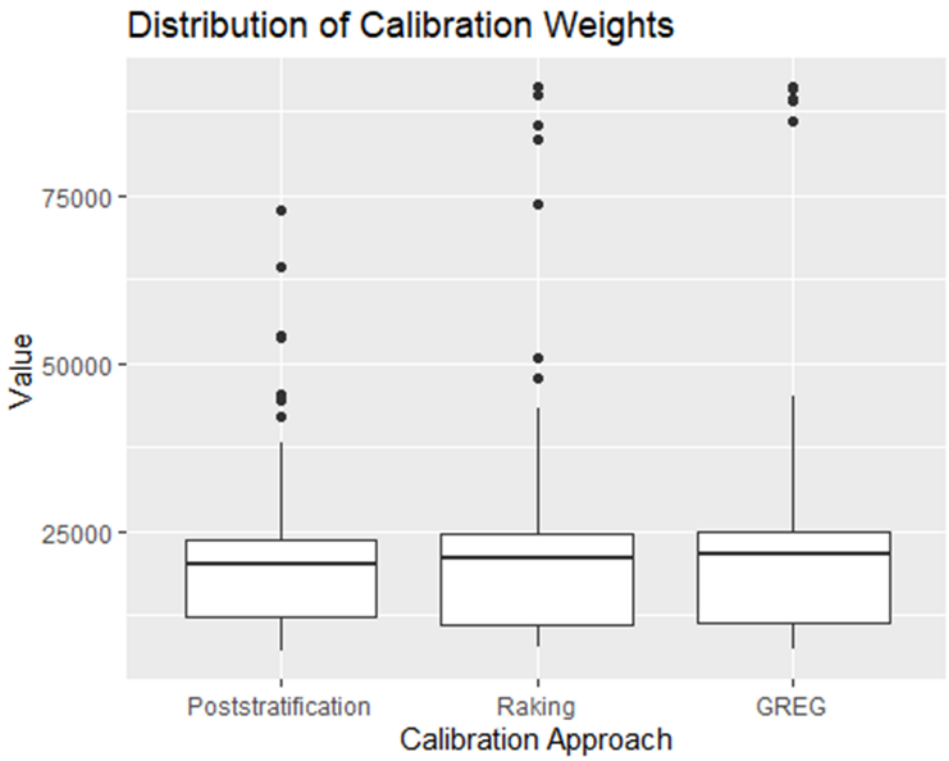

Abbildung 4: Verteilung der kalibrierten Designgewichte (unnormiert)

3 Weitere Formen der Gewichtung

Neben den soeben beschriebenen Gewichtungsverfahren begegnen dem Forschenden häufig noch weitere Formen der Gewichtung. Diese zielen oftmals darauf ab, bestehende Gewichte anzupassen oder zu verändern. Die wohl häufigsten Formen sind Verfahren zur Reskalierung bzw. Renormierung von Gewichten sowie die Verwendung von Längsschnitt- bzw. Panelgewichten.

3.1 Reskalierung/ Renormierung

Insbesondere bei international vergleichenden Studien, aber auch bei nationalen Erhebungen werden Designgewichte (und damit verbunden auch deren Kombination mit Kalibrierungsgewichten) oftmals bezüglich ihrer Basis verändert. Der Grund dafür liegt häufig in der Beschaffenheit des Horvitz-Thompson-Schätzers. So ist eine Schätzung auf dessen Basis, wie in Kapitel 2.1 bereits dargestellt, immer eine Schätzung eines Totalwertes der Grundgesamtheit. Die Summe aller Gewichte entspricht daher der geschätzten Größe der Grundgesamtheit. Da dies jedoch nicht immer wünschenswert ist, werden folgende Umformungen des Öfteren angewandt:

1) Reskalierung auf die Stichprobengröße
2) Reskalierung auf eine gemeinsame Basis (mehrerer Datensätze)

Das bereits berechnete Designgewicht d_i der Erhebungseinheit wird durch die Summe aller Designgewichte dividiert und im Anschluss mit dem numerischen Wert der Basis, die das Gewicht annehmen soll

(θ), multipliziert. Die Summe der reskalierten Designgewichte $\tilde{d}_i$ entspricht dann der neuen Basis. Formal entspricht dies:

$$\tilde{d}_i = \theta * \frac{d_i}{\sum_{i \in R} d_i}.$$

Bei einer Reskalierung auf Basis der Stichprobengröße (ad 1) entsprechen die neuen Gewichte der Anzahl der Respondenten der Erhebung. Dies wird jedoch oftmals aus rein „kosmetischen" Gründen angestrebt. Das Verhältnis der einzelnen Gewichte verändert sich (relativ) nicht. Eine solche Form der Gewichtung hat überdies den Vorzug, dass der Mittelwert aller Gewichte 1 entspricht, sodass es grundsätzlich einfacher sein kann, direkt festzustellen, ob eine bestimmte Erhebungseinheit tendenziell „runtergewichtet" (kleiner 1) oder „hochgewichtet" (größer 1) wird. Die Berechnung des Mittelwertes der Designgewichte stellt überdies für den Forschenden eine Möglichkeit dar, zu überprüfen, ob die in einem Datensatz enthaltenen Designgewichte auf die Stichprobengröße reskaliert wurden, da dies nicht immer eindeutig gekennzeichnet wird.

Eine Reskalierung auf die gemeinsame Basis mehrerer Datensätze (ad 2) hat oftmals folgende Beweggründe:

- die Auswertung und der anschließende (absolute) Vergleich mehrerer Länder einer transnationalen Erhebung sowie
- die gemeinsame Auswertung und Zusammenfassung von Daten mehrerer Länder zu einem kombinierter Datensatz.

So kann es bspw. im Rahmen des European Social Survey (ESS) interessant sein, für die Berechnung von Schätzwerten für bestimmte Regionen mehrere Länder und damit mehrere Datensätze mit eigenen Designgewichten zusammenzufassen. Da in den initialen Datensätzen jedes Designgewicht auf die Populationsgröße des entsprechenden Landes „hochgerechnet" wurde, kann eine Reskalierung (z.B. auf den Wert 10.000) das Zusammenfassen einzelner Länder erleichtern. Dies erlaubt es Schätzwerte bspw. für den skandinavischen Raum zu ermitteln. Durch das Designgewicht „Population Size Weight" kann die entsprechende Schätzung auf Basis der Kombination einzelner Länder vergleichsweise einfach durchgeführt werden. Weitere Beispiele auf Basis des ESS findet man bei Gabler & Ganninger (2010).

3.2 Längsschnitt-/ Panelgewichte

Bei wiederholter Befragung derselben Zielpersonen, sog. Panelbefragungen, werden neben Design- und Kalibrierungsgewichten oftmals auch Panelgewichte (der einzelnen Befragungswellen) angewendet. Panelgewichte zielen in erster Linie darauf ab, die Verbleibewahrscheinlichkeit einer befragten Zielperson in der Erhebungswelle zum Zeitpunkt t zu berücksichtigen.

Bei Panelbefragungen bleibt es nicht aus, dass einzelne Zielpersonen die Teilnahme an einer Welle einer Befragung temporär (für eine oder mehrere Wellen) oder endgültig verweigern (Blumenberg & Gummer, 2016). Da zum Zeitpunkt t bereits bekannt ist, ob eine Zielperson an dieser Welle teilgenommen hat und i. d. R. bereits Informationen über die Zielperson aus den vorherigen Befragungswellen vorliegen, kann die Verbleibewahrscheinlichkeit einer Zielperson für die aktuelle Erhebungswelle geschätzt werden. Das Panelgewicht kann analog zum Designgewicht anhand der Inversen der Verbleibewahrscheinlichkeit bestimmt werden.

Für eine bestimmte Welle ist demnach die Wahrscheinlichkeit, dass eine Zielperson in der Erhebung ist, abhängig von der initialen Auswahlwahrscheinlichkeit der Zielperson sowie ihrer jeweiligen Verbleibewahrscheinlichkeiten über die Wellen hinweg. Die Verbleibewahrscheinlichkeit für die aktuelle Erhebungswelle lässt sich als Produkt der Einzelverbleibewahrscheinlichkeiten der vorherigen Erhebungs-

wellen bestimmen. Analog dazu wird bei Verwendung der Panelgewichte das Designgewicht multiplikativ mit den (einzelnen) Designgewichten der vorherigen Wellen verknüpft.

Bei einer Verbleibewahrscheinlichkeit von φ_{it} zum Zeitpunkt t errechnet sich das verknüpfte Panelgewicht w_{it} einer Erhebungseinheit i nach

$$w_{it} = d_i \prod_{\substack{t \in T \\ i \in S}} \frac{1}{\varphi_{it}}.$$ [15]

Bei der zusätzlichen Verwendung von Kalibrierungsgewichten gilt zu beachten, dass diese für die aktuelle Erhebungswelle unter der Verwendung der Design- und Panelgewichte für die aktuelle Grundgesamtheit zu berechnen sind. Daher werden in einem ersten Schritt die Designgewichte der Basiswelle mit den jeweiligen Panelgewichten aller vorheriger Erhebungswellen bis zur betrachteten Welle multiplikativ verknüpft und in einem zweiten Schritt die Randverteilung der Hilfsvariablen in der Stichprobe, die für die Kalibrierung benötigt werden, anhand dieses Gewichts geschätzt. Die dadurch gewonnenen Ergebnisse werden dann erst an die entsprechenden Verteilungen der Grundgesamtheit des betrachteten Zeitraums angepasst.

Für den Anwender bedeutet dies, dass bei der Verwendung jeglicher Gewichtungsfaktoren in einem ersten Schritt die Designgewichte mit den einzelnen Panelgewichten multipliziert werden müssen und dann erst eine Multiplikation mit dem Kalibrierungs- oder Anpassungsgewicht der betrachteten Welle erfolgt. Eine reine Multiplikation des Kalibrierungs- oder Anpassungsgewichts der Basiswelle mit den Panelgewichten kann zu verzerrten Ergebnissen führen. Weiterhin gilt für den Anwender zu beachten, die Aktualität der Angaben über die Grundgesamtheit, an die kalibriert wird, sicherzustellen (siehe hierzu Kap. 1.2).

4 Empfehlungen und weitere Anmerkungen

Grundsätzlich sollte dem mit der Planung einer Erhebung betrauten Personenkreis bewusst sein, dass die Gewichtung von Umfragedaten kein „Allheilmittel“ gegen Fehler bei der Stichprobenerhebung und/oder eine minderwertig durchgeführte Erhebung sein kann. In solchen Fällen stellt die Gewichtung der Umfragedaten vielmehr eine kosmetische Maßnahme dar, bei der diese Mängel lediglich „verschleiert“ werden. Ferner besteht bei einer solchen (unzweckmäßigen) Anwendung von Gewichtungsverfahren das Problem, dass die Güte der Ergebnisse abhängig ist von dem Modell, das der Forschende für eine entsprechende Gewichtung verwendet. Besteht hier eine Fehlspezifikation im Sinne einer starken Abweichung des zugrundeliegenden Modells von den realen Verhältnissen, wirkt sich dies entsprechend negativ auf die Schätzergebnisse aus.

Was passiert, wenn ich nicht gewichte?

In der (sozialwissenschaftlichen) Umfragepraxis kann oftmals beobachtet werden, dass eine Anpassungsgewichtung zur Verringerung potenzieller Verzerrungen der Schätzwerte infolge von Antwortausfällen vorgenommen wird, ohne dass zuvor die entsprechenden Auswahlwahrscheinlichkeiten im Zuge einer Designgewichtung berücksichtigt werden. Gerade dies sollte aus statistischen Gründen vermieden werden, sodass bspw. bereits Schätzer für Werte innerhalb der Grundgesamtheit lediglich weiter angepasst werden (Gabler et al., 2015; Sand & Gabler, 2019).[16] Bei manchen Erhebungen ist jedoch aufgrund

[15] Es wird davon ausgegangen, dass ein „Reentry“ von bereits ausgeschiedenen Personen nicht möglich ist.

[16] Der grundlegende Gedanke hierbei ist, dass der Horvitz-Thompson-Schätzer unter Abwesenheit von Nonresponse ein unverzerrter Schätzer ist. Erst in Folge von Antwortausfällen wird eine Anpassungsgewichtung „notwendig“. Diese dient lediglich

mangelnder Informationen oder einem nicht (mehr) nachvollziehbaren Stichprobendesign eine Berechnung der Designgewichte nicht möglich. In einem solchen Fall wird häufig die Annahme getroffen, dass die Daten aus einer uneingeschränkten Zufallsstichprobe stammen und deren Auswahlwahrscheinlichkeiten verwendet. Es kann jedoch davon ausgegangen werden, dass die Nichtberücksichtigung der tatsächlichen Auswahlwahrscheinlichkeiten trotz der Verwendung der Kalibrierungsgewichte die Ergebnisse der Schätzung in ihrer Genauigkeit und Präzision negativ beeinflusst.

der Reduktion der Verzerrung innerhalb der Schätzung (und der Varianz). Das naive Verwenden dieser Gewichtung ohne vorherige Designgewichtung führt zu unterschiedlichen Anpassungsgewichten, die vom „Endgewicht“ (bspw. Produkt aus Design- und Anpassungsgewicht) abweichen. Im Fall von sog. „selbstgewichtenden“ Ansätzen wird als Designgewicht lediglich die Größe der Grundgesamtheit geteilt durch die Stichprobengröße verwendet, wenn Schätzungen auf der Ebene von Totalwerten erfolgen (Särndal & Lundström, 2005). Für Mittelwertschätzungen beträgt das Designgewicht für alle Einheiten 1.

Literaturverzeichnis

ADM, Arbeitskreis Deutscher Markt und Sozialforschungsinstitute. (2017). Jahresbericht 2016. Unter https://www.adm-ev.de/wp-content/uploads/2018/07/Jahresbericht-2016.pdf.

Blumenberg, M. S. & Gummer, T. (2016). Gewichtung in der German Longitudinal Election Study 2013. GESIS Papers 2016/1. Unter https://doi.org/10.21241/ssoar.47155.

Busse, B. & Fuchs, M. (2013). Prevalence of Cell Phone Sharing. Survey Methods: Insights from the Field. Unter https://surveyinsights.org/?p=1019.

Deming, E. & Stephan, F. (1941). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *The Annals of Mathematical Statistics*, *11*, 427–444.

Gabler, S. & Ayhan, Ö. (2007). Gewichtung bei Erhebungen im Festnetz und über Mobilfunk: Ein Dual-Frame Ansatz. In S. Gabler & S. Häder (Hrsg.), *Mobilfunktelefonie - Eine Herausforderung für die Umfrageforschung* (S. 39–45). ZUMA-Nachrichten Spezial Band 13, Mannheim.

Gabler, S. & Ganninger, M. (2010). Gewichtung. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 143–164). Wiesbaden: Springer VS.

Gabler, S. & Häder, S. (2009). Gewichtung für die Cella-Studie. In M. Häder & S. Häder (Hrsg.), *Telefonbefragungen über das Mobilfunknetz. Konzept, Design und Umsetzung einer Strategie zur Datenerhebung* (S. 51–55). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gabler, S. & Häder, S. (2015). Stichproben in der Theorie. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://doi.org/10.15465/gesis-sg_009.

Gabler, S., Häder, S., Lehnhoff, I. & Mardian, E. (2012). Weighting for Unequal Inclusion Probabilities and Nonresponse in Dual Frame Telephone Surveys. In S. Häder, M. Häder & M. Kühne (Hrsg.), *Telephone Surveys in Europe. Research and Practice* (S. 147–167). Heidelberg: Springer.

Gabler, S., Kolb, J.-P., Sand, M. & Zins, S. (2015). Gewichtung. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://doi.org/10.15465/gesis-sg_007.

Häder, S. (2015). Stichproben in der Praxis. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://doi.org/10.15465/gesis-sg_014.

Häder, S. & Sand, M. (2019). Telefonstichproben. In S. Häder, M. Häder & P. Schmich (Hrsg.), *Telefonumfragen in Deutschland. Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute* (S. 113–151). Wiesbaden: Springer VS.

Koch, A. & Blohm, M. (2015). Nonresponse Bias. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines). https://doi.org/10.15465/gesis-sg_004.

Lohr, S. L. (2009). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ: John Wiley & Sons.

Lundström, S. & Särndal, C.-E. (2001). *Estimation in the Presence of Nonresponse and Frame Imperfections*. Örebro: Scb-Tryck.

Roßteutscher, S., Schmitt-Beck, R., Schoen, H., Wessels, B., Wolf, C., Wagner, A. et al. (2019). Nachwahl-Querschnitt (GLES 2017). GESIS Datenarchiv, Köln. ZA6801 Datenfile Version 4.0.1, https://doi.org/10.

4232/1.13235.

Sand, M. (2018). Gewichtungsverfahren in Dual-Frame-Telefonerhebungen bei Device-Specific Nonresponse. GESIS-Schriftenreihe, 20: GESIS - Leibniz-Institut für Sozialwissenschaften. Unter: https://doi.org/10.21241/ssoar.60293.

Sand, M. & Gabler, S. (2019). Gewichtung von (Dual-Frame-) Telefonstichproben. In S. Häder, M. Häder & P. Schmich (Hrsg.), *Telefonumfragen in Deutschland. Schriftenreihe der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute* (S. 405–424). Wiesbaden: Springer VS.

Särndal, C.-E. & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York, NY: Wiley.

Valliant, R., Dever, J. A. & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York, NY: Springer.

Yu, K. F. (1994). Truncating Sample Weights Reduces Variance. *Statistics & Probability Letters*, *19*(4), 267–269.