Wahlumfragen auf der Basis von Access-Panels

—

Wie „Propensity-Score“-Gewichtungen die Repräsentativität erhöhen können

Inauguraldissertation
zur Erlangung des Akademischen Grades
eines Dr. phil.,

vorgelegt dem Fachbereich 02 - Sozialwissenschaften, Medien und Sport
der Johannes Gutenberg-Universität
Mainz

von

Sven Vollnhals
aus Duisburg

Tag des Prüfungskolloquiums: 31.03.2017

Zusammenfassung

Eine sinkende Teilnahmebereitschaft an Bevölkerungsumfragen sowie zunehmende Probleme der Auswahlverfahren alle gesellschaftlich relevanten Gruppen zu erreichen: Die Umfrageforschung befindet sich in einer anhaltenden Krise. Die vorliegende Arbeit wird diese Herausforderungen analysieren und darauf aufbauend Lösungsstrategien aufzeigen. Hierzu wird die persönliche Befragung als ein Goldstandard der Erhebung herausgestellt, zu welchem die schnelleren und günstigeren, jedoch mit qualitativen Problemen behafteten Telefon- und Onlineumfragen kontrastiert und anschließend über Gewichtungsverfahren angeglichen und korrigiert werden können.

Basierend auf Datensätzen von Forsa, des Politbarometers und der Komponente 2 der „German Longitudinal Election Study“ (GLES) wird für die zufallsbasierten Telefonumfragen der immer selektiver werdende Teilnehmerkreis höher gebildeter und politisch interessierter Personen seit dem Jahr 1990 herausgestellt. Obwohl sich dies noch nicht auf das erfasste Wahlverhalten auswirkt, finden sich bereits systematische Probleme der korrekten Erfassung der Partizipationsbereitschaft. Für Onlineumfragen wird hingegen die Auswahlgrundlage als entscheidend skizziert, wozu Stichproben eines telefonisch offlinerekrutierten mit denjenigen eines rein onlinerekrutierten Access-Panels der Komponenten 3 und 8 der GLES verglichen werden. Ergeben sich durch ersteres Zufallsstichproben mit vergleichbaren Problemen zu denjenigen von zufallsbasierten Telefonumfragen, weist zweiteres dem gegenüber dem politischen System und seiner Akteure kritisch eingestellten Teils des Elektorats eine zu hohe Bedeutung zu, was sich folglich auch auf das erfasste Wahlverhalten auswirkt.

Der zweite Beitrag der Arbeit ist schließlich methodischer Natur und untersucht die Qualität verschiedener statistischer Verfahren zur Korrektur dieser aufgezeigten Verzerrungen durch die Anwendung von Gewichtungsverfahren auf der Basis von „Propensity Scores“ (PS). Ist ein inhaltlich gut spezifiziertes Korrekturmodell für die Qualität der Korrektur entscheidend, lässt sich ein zusätzlicher Erfolg durch eine überlegte Wahl der technischen Art der Angleichung an die jeweilige (persönliche) Referenzumfrage erreichen. Hierzu zeigt die Arbeit die systematische Überlegenheit der Anwendung differenzierterer Algorithmen im Rahmen eines „Propensity Score Matching“ (PSM) – insbesondere der Subklassifikation und des „Full-Algorithmus“ – gegenüber einer einfachen inversen PS-Gewichtung.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AAPOR „American Association for Public Opinion Research“

AB „Absoluter Bias“

ANB „Absoluten Nonresponse-Bias“

ADM „Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.“

AfD „Alternative für Deutschland“

Allbus „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“

ANES „American National Election Studies“

ASA „American Statistical Association“

BQA „Basic Question Approach“

BTW „Bundestagswahl“

BuWa „Bundeswahlleiter“

CAPI „Computer-Assisted Personal Interviewing“

CATI „Computer-Assisted Telephone Interviewing“

CBA „Callback Approach“

CEM „Coarsened Exact Matching“

DDR „Deutschen Demokratischen Republik“

Destatis „Statistische Bundesamt“

ELIPSS „Longitudinal Study by Internet for the Social Sciences“

EM „Exact Matching“

EPBR „Equal Percent Bias Reduction“

ESS „European Social Survey“

FGW „Forschungsgruppe Wahlen“

F2F „Face to Face“

FRM „Fixed Response Model“

FM „Full Matching“

GENESIS „Gemeinsames neues statistisches Informationssystem“

GM „Genetic Matching“

GESIS „Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen“

GLES „German Longitudinal Election Study“

GIP „German Internet Panel“

HT „Horvitz-Thompson“

infas „Institut für angewandte Sozialwissenschaft GmbH“

IPF „Iterative Proportional Fitting“

KQ „Kontrollquerschnitt“

LISS „Longitudinal Internet Studies for the Social sciences“

MAR „Missing at Random“

MCAR „Missing Completely at Random“

MLE „Maximum-Likelihood-Estimation“

MNAR „Missing Not at Random“

MSE „Mean Squared Error“

MZ „Mikrozensus“

NNM „Nearest Neighbour Matching“

OM „Optimal Matching“

OR Odds-Ratios

PBR „Percent Bias Reduction“

PEKS „Political Efficacy Kurzskala“

PPS „Probability Proportional to Size“

PS „Propensity Scores“

PSM „Propensity Score Matching“

PSU „Primary Sampling Unit“

RCM „Rubin Causal Model“

RCS „Rolling-Cross-Section“

RDD „Random Digit Dialing“

RB „Relative Bias“

RLD „Randomized Last Digit“

RNB „Relative Nonresponse-Bias“

RMSE „Root Mean Squared Error“

RRM „Random Response Model“

SB „Standardized Bias“

SD „Standardized Differences“

SUB „Subklassifikation“

SP „Sample Points“

SUTVA „Stable Unit Treatment Value Assumption“

TSE „Total Survey Error“

TSMV „Total Survey Measurement Variation

WKP „Wahlkampfpanel“

1. Einleitung

Am 23.06.2016 hat sich das Vereinigte Königreich für den Verbleib innerhalb der Europäischen Union entschieden. Die britische Unterhauswahl endete am 07.05.2015 in einem spannenden „Kopf-an-Kopf"-Rennen zwischen den „Conservatives" und „Labour". Auch Hillary Clinton wurde durch die US-Präsidentschaftswahl am 08.11.2016 zur Nachfolgerin von Barack Obama gewählt, ebenso wie die FDP im Deutschen Bundestag vertreten ist. In einem Referendum am 02.10.2016 wurde zudem durch die kolumbianische Bevölkerung der zur Abstimmung stehende Friedensvertrag zwischen der Regierung und der FARC ratifiziert. Der jahrzehntelange Konflikt ist hierdurch auch formal beigelegt worden. Dies ist die Realität, welche zentrale Befragungen des jeweiligen Elektorats im Vorfeld der aufgeführten Abstimmungen gezeichnet haben. Die Welt hat sich jedoch genau entgegengesetzt entwickelt. Großbritannien wird die Verhandlungen über den Austritt aus der EU aufnehmen, die Konservativen haben die Wahl zum britischen Unterhaus mit über sechs Prozentpunkten Vorsprung beendet und Donald Trump wurde am 20.12.2016 durch das „Electoral College" zum kommenden US-Präsidenten gewählt. Auch Kolumbien sucht weiter nach einem formalen Abschluss des seit Jahrzehnten andauernden Bürgerkriegs und die FDP hofft schließlich auf einen Wiedereinzug in den Bundestag 2017, nachdem die konstituierende Sitzung des 18. Deutschen Bundestages ohne eine FDP-Fraktion stattgefunden hat.[1]

Solche Vorhersagen über den Ausgang von Wahlen und Referenden sind mit einer gewissen Unsicherheit behaftet. Je dichter sich ein „Kopf-an-Kopf-Rennen" über eine Entscheidung abzeichnet, umso höher ist die Wahrscheinlichkeit einer Fehlprognose durch Umfragen – wie bei einem vorhergesagten „Remain". Ebenso sind durch „last-minute swings" Ergebnisse von Abstimmungen möglich, welche bereits eine andere Realität darstellen als die vorherigen Umfragen erfasst haben. So lässt sich die als Beispiel angeführte Wahl des britischen Unterhauses als die „[...] most volatile election since 1931 [...] " (Green/Prosser 2016: 1299) charakterisieren.

Der „Brexit" in Großbritannien wurde mit einem sehr knappen Vorsprung von

[1] Der Stichtag der berücksichtigten Entwicklungen ist der 07.06.2017.

3.8 Prozentpunkten beschlossen. 51.9 % der an der Abstimmung *beteiligten* Personen haben sich hierfür ausgesprochen.[2] Eine beispielhafte Umfrage im Vorfeld von 1000 zufällig befragten wahlberechtigten Personen weist für einen erreichten prozentualen Anteil eines „Leave“ von 51.9 % ein 95 %-Konfidenzintervall von ungefähr $\pm$ drei Prozentpunkten auf.[3] Innerhalb des Kreises der befragten Personen ist also eine Mehrheit für ein „Remain“ im Bereich des Wahrscheinlichen, obwohl das Elektorat sich insgesamt für ein „Leave“ entscheiden wird.

Eine Unabhängigkeit Schottlands vom Vereinigten Königreich durch das am 18.09.2014 durchgeführte Referendum lehnten 55.25 % der teilnehmenden Wahlberechtigten ab.[4] „Bevölkerungsrepräsentative“ Umfragen im direkten Vorfeld dieses Referendums haben daher nur eine äußerst geringe Wahrscheinlichkeit der Realisierung einer Stichprobe, welche *nicht* mehrheitlich eine Ablehnung der Unabhängigkeit signalisiert. Das Meinungsforschungsinstitut YouGov prognostizierte auf der Basis einer Erhebung jedoch eine kommende Forderung der Loslösung Schottlands.[5] Dieses Ergebnis ergab sich erst durch eine nachträgliche Veränderung der Ergebnisse der Umfrage: Hatten sich 53.8 % der befragten Personen für eine Ablehnung ausgesprochen, führte die Anwendung einer Gewichtungskorrektur schließlich zu einer Mehrheit für eine Unabhängigkeit von 51.4 %.[6] Solche Gewichtungskorrekturen können sinnvoll sein. Hierzu müssen sie adäquat berücksichtigen können, welche für die Umfrage relevanten Personenkreise sich eher für eine Teilnahme begeistern konnten als andere. Durch die Korrektur sind dann die Meinungen aller Gruppen wieder gemäß ihrer gesellschaftlichen Größe abgebildet. Ist jedoch nicht exakt bekannt welche Gruppen häufiger teilgenommen haben als andere, dann kann auch diese nachträgliche Korrektur eine fehlerhafte Erhebung nicht korrigieren.

Während der Aufarbeitung ihrer Fehlprognosen der britischen Unterhauswahl mussten die zentralen Meinungsforschungsinstitute in Großbritannien bekannt geben, dass Fehleinschätzungen der Bedeutung bestimmter Personengruppen für die fehlerhaften Prognosen der britischen Unterhauswahl die maßgebliche

[2] Quelle: http://www.electoralcommission.org.uk/find-information-by-subject/elections-and-referendums/past-elections-and-referendums/eu-referendum/electorate-and-count-information (Stand: 11. Mai 2018).

[3] Die Berechnung eines 95 %-Konfidenzintervalls ergibt sich in dem Fall als $1.96 \cdot \sqrt{\frac{0.519 \cdot (1-0.519)}{1000}} \approx 0.03$. Zur Berechnung siehe z.B. Gehring/Weins (2010: 267 f.).

[4] Quelle: http://scotlandreferendum.info/ (Stand: 11. Mai 2018).

[5] https://yougov.co.uk/news/2014/09/06/latest-scottish-referendum-poll-yes-lead/ (Stand: 11. Mai 2018).

[6] Siehe hierzu auch http://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/ywzyqmrf2u/Scotland_Final_140905_Sunday_Times_FINAL.pdf (Stand: 11. Mai 2018).

Ursache waren.[7] Der Kreis der durch die Institute befragten Personen hat sich grundlegend vom gesamten Elektorat unterschieden, die Anhänger von Labour wurden durch das angewandte Auswahldesign zur Befragung von Personen begünstigt. Die angewandten Gewichtungskorrekturen waren anschließend nicht in der Lage, diese „Ungleichbehandlung“ zu beheben (zur Darstellung siehe Sturgis et al. 2016). Ebenso wurden für diese Fehlprognosen „schlechte“ Umfragen in Form von „nicht-repräsentativen“ Erhebungen ausgemacht.[8]

Wahlumfragen können auch selbst eine Veränderung ihres Untersuchungsgegenstand aufweisen. Die medienwirksam verbreitete Prognose des Zweitstimmenanteils kann zu einer möglichen Verhaltensänderung des Wählers als Empfänger dieser Botschaft führen. Trifft dies zu, hat sich die reale Welt verändert und die vorher durchgeführte Umfrage ist bereits obsolet (z.B. Faas 2014, 2017). Insbesondere im direkten Vorfeld von Wahlen kann es zu kurzfristigen Änderungen des Verhaltens kommen. Die Folge ist ein „last-minute swing“ (Roth 2008: 174), wie das Scheitern der FDP an der 5 %-Hürde bei der Bundestagswahl 2013 gezeigt hat. Die Prognosen der Wahlumfragen im Vorfeld der Wahl ergaben für diese Zweitstimmenanteile von über 5 %. Dies suggerierte eine nicht mehr weitere notwendige Unterstützung dieser Partei, damit der Einzug in den Bundestag erreicht wird. Die FDP scheiterte mit 4.9 % an der 5 %-Klausel (Faas/Huber 2015). Solche Beispiele begründen die immer wiederkehrenden Fragen des Einflusses und mangelnder Neutralität von Wahlumfragen.[9]

Eine solche Fehleinschätzung durch telefonische Wahlumfragen in der Bundesrepublik ist keineswegs die Ausnahme, wie Schnell/Noack (2014) auf der Basis von 232 Vorwahlumfragen zwischen 1957 und 2013 zeigen. Lediglich 67 dieser Umfragen spiegeln die Zweitstimmenanteile *aller* angetretenen Parteien in einer solchen Genauigkeit wider, wie sie die Berechnung eines Konfidenzintervalls für die Anteilswerte einfacher Zufallsstichproben zu Grunde legt. Die Prognose auf der Basis von Wahlumfragen ist damit ungenauer, als sie kommuniziert wird. Das angegebene 95 %-Konfidenzintervall zeigt sich in der Realität nur als eine 29 %-ige Chance der Abdeckung des wahren Wahlergebnisses (Schnell/Noack 2014: 16). Wahlumfragen sind daher entweder ungenauer als

[7] Quelle: http://www.telegraph.co.uk/news/general-election-2015/12077405/Polling-companies-admit-they-got-2015-election-wrong-by-ignoring-the-views-of-pensioners.html (Stand: 11. Mai 2018).

[8] Z.B.: http://www.telegraph.co.uk/news/general-election-2015/12107167/Opinion-polls-failure-at-2015-election-due-to-unrepresentative-samples.html (Stand: 11. Mai 2018).

[9] Zum Beispiel der Zeit-Bericht http://www.zeit.de/politik/deutschland/2013-09/wahlumfragen-parteilichkeit-bundestagswahl (Stand: 11. Mai 2018) und https://www.welt.de/politik/wahl/bundestagswahl/article120027580/Die-unheimliche-Macht-der-Meinungsforscher.html (Stand: 11. Mai 2018).

sie suggerieren, oder sie sind systematisch mit einem Fehler belegt, welcher die Erfassung der korrekten Zweitstimmenanteile unmöglich macht (Schnell/Noack 2014: 21).

Doch wie lassen sich diese Beispiele fehlerhafter Einschätzung in den Bereich der methodischen Umfrageforschung einordnen? Handelt es sich hierbei lediglich um plakative Momentaufnahmen, oder lassen sich Umfragen auch wissenschaftlich zunehmende Herausforderungen bescheinigen? Das nachfolgende Unterkapitel wird diese Probleme herausstellen, die Relevanz der Arbeit hierdurch begründen und das Ziel dieser durch handlungsanleitende Fragen skiziieren.

1.1. Problembeschreibung, Relevanz und Ziel der Arbeit

Umfragen haben somit durch ihre Fehlprognosen der vergangenen Jahre an Reputation innerhalb der medialen Öffentlichkeit eingebüßt. Doch auch innerhalb der Umfrageforschung bestätigt sich die Diagnose zunehmender methodischer Probleme der korrekten Erfassung der (sozialen) Realität durch Umfragen (Brick 2011; Groves 2011; Baker/Brick et al. 2013; Brick 2013). Hat die Begründung der Anwendung der Inferenzstatistik durch eine zufallsbasierte Auswahl durch u.a. Neyman (1934), Deming (1950) und Cochran (1977) der Reichweite sozialwissenschaftlicher Analysen zum Durchbruch geholfen, drohen diese Entwicklungen die Gültigkeit zu konterkarieren. Die Annahme einer „[...] absence of selective forces in the sampling process.“ (Kruskal/Mosteller 1979c: 247) als Voraussetzung inferenzstatistischer Verfahren ist nicht mehr plausibel. Dies schränkt die (statistische) Repräsentativität von Umfragen als die Übereinstimmung der Randverteilungen auf allen als relevant angenommenen Variablen zwischen der vorliegenden Stichprobe und der ihr zu Grunde liegenden Grundgesamtheit ein (Bethlehem/Cobben/Schouten 2011: 16 f.).

Warum ein solcher systematischer Fehler bei Umfragen auftritt, hat verschiedenste Ursachen. Seit einigen Jahren steht insbesondere der Trend fallender Teilnahmeraten als Fehlerquelle im Fokus (Schnell 1997; Heer 1999; Atrostic et al. 2001; Steeh et al. 2001; Leeuw/Heer 2002; Stoop 2004; Brick/Williams 2012). Durch diese steigt die Gefahr einer verzerrten Erfassung der Realität, wenn sich bestimmte relevante Gruppen mit einer erhöhten Wahrscheinlichkeit der Befragung entziehen und hierdurch nicht mehr adäquat repräsentiert sind (Massey/Tourangeau 2012: 227). Ob sich eine niedrige Teilnahmerate negativ auswirkt, unterscheidet sich individuell für jedes Merkmal. Variablen auf die

dies zutrifft, sind durch ein „Missing Not at Random“ (MNAR) nach Little/Rubin (2002) gekennzeichnet. Die erhöhte Wahrscheinlichkeit der Nichtteilnahme von bestimmten Personengruppen ist abhängig von diesen Merkmalen, wie beispielsweise des Alters. Ist dies bekannt, kann über statistische Gewichtungskorrekturen der Ausfallmechanismus berücksichtigt werden. Als Folge ist dieser nun durch ein „Missing at Random“ (MAR) nach Little/Rubin (2002) charakterisiert. Der Ausfall hat weiterhin nicht zufällig stattgefunden, er lässt sich aber durch seine Aufdeckung kontrollieren. Sind Merkmale lediglich von einem zufälligen Ausfall betroffen, liegt ein „Missing Completely at Random“ (MCAR) bezeichnen. Eine niedrige Teilnahmerate hat in diesem Fall keinen Einfluss auf diese Gruppe von Variablen, da Personen über alle relevanten Gruppen rein zufällig nicht befragt werden konnten.

Sinkende Teilnahmeraten sind jedoch nicht die einzige Ursache für eine geringe Qualität von Umfragen. Telefonische Befragungen sind mit einem immer größeren Teil der Bevölkerung konfrontiert, welcher keine Festnetzanschlüsse mehr zu Gunsten von Mobilfunknummern nutzt (z.B. Lavrakas et al. 2007; Peytchev/Carley-Baxter/Black 2011; Busse/Fuchs 2012; Hox/Mohorko/Leeuw 2013). Diese Personen sind über ein auf Festnetznummern basierendem Auswahldesign nicht mehr kontaktierbar. Onlinebasierte Befragungen kämpfen mit ähnlichen Herausforderungen : Ein gewisser Teil der Bevölkerung hat keinen Zugriff auf einen Internetanschluss und steht somit für onlinebasierte Befragungen a priori nicht zur Verfügung (z.B. Dillman/Smyth/Christian 2009; Mohorko/Leeuw/Hox 2013). Zudem unterscheidet sich der Kompetenzgrad im Umgang mit dem Internet innerhalb der Bevölkerung, weshalb die Entscheidung zur Teilnahme von bestimmten Gruppen eher getroffen wird (Bethlehem 2009: 277). Diese Probleme der mangelnden Abdeckung der allgemeinen Bevölkerung durch Telefon- und Onlineumfragen ist in ihren Auswirkungen identisch zu denjenigen der geringeren Teilnahmeraten an Umfragen. Werden bestimmte Teile der Bevölkerung systematisch durch das Auswahldesign benachteiligt, ergibt sich ein mögliches MNAR für die erhobenen Merkmale durch eine selektive Erfassung. Um dies zu einem MAR zu negieren, muss ebenfalls bekannt sein, welche Variablen von dieser mangelnden Abdeckung durch das Erhebungsdesign betroffen sind. Auf deren verzerrte Erfassung kann dann wiederum über statistische Gewichtungsverfahren korrigiert werden.

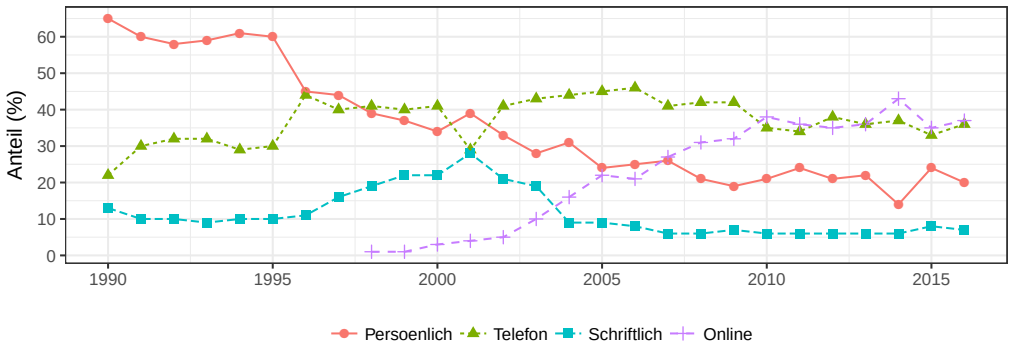
Für onlinebasierte Erhebungen ergibt sich ein weiteres grundlegendes Problem, sofern eine Stichprobe der allgemeinen Bevölkerung befragt werden soll. Eine Auswahlgrundlage für eine zufallsbasierte Ziehung steht nicht zur

Verfügung, sofern diese ebenfalls online geschehen soll. Inferenzstatistische Verfahren verlieren somit ihre Rechtfertigung (Couper 2000: 477; Bethlehem 2009: 284 f. Bethlehem/Biffignandi 2012: 45 ff.). Dies relativiert die Vorteile der Schnelligkeit und kostengünstigen Erhebung durch dieses Erhebungsformat (Schnell 2012: 290 f.). Access-Panels versuchen dieses Problem zu lösen, indem sie einen Pool von möglichen zu befragenden Personen als eine Grundgesamtheit konstruieren. Aus diesen ist dann eine zufallsbasierte Auswahl möglich, der Rückschluss auf die Gesamtheit aller Mitglieder des Access-Panel gegeben. Deren Nutzung zeigt sich auch in einem zunehmenden Maße in wissenschaftlichen Publikationen (Ansolabehere/Schaffner 2014: 285). Da der überwiegende Anteil der Access-Panel jedoch einen nicht-zufallsbasierten und online stattfindenden Rekrutierungsweg nutzt (Vehre 2012: 38 f.), lassen sich die zentralen Probleme der Durchführung von Onlineumfragen bezüglich der Anwendung inferenzsstatistischer Verfahren auch durch diese nicht lösen. Auch weiterhin bleibt ein nicht unwesentlicher Teil der allgemeinen Bevölkerung durch die Nutzung dieses Erhebungsformat ausgeschlossen und eine hohe Fehleranfälligkeit onlinebasierter Befragungen durch ein MNAR bei Bevölkerungsumfragen existiert. Auch für Erhebungen auf der Basis von Access-Panels gilt somit: Nur wenn die durch das nicht-zufallsbasierte Auswahldesign und den Ausschluss der Offliner erwartbaren systematischen Unterschiede eines MNAR aufgedeckt werden können, ist die Überführung in einen MAR-Ausfall möglich. Bevölkerungsumfragen mit einem Anspruch der repräsentativen Aussage sind dann onlinebasiert durch die Nutzung von Access-Panels möglich.

Die Beschäftigung und Lösung dieser aufgezeigten Probleme ist zentral für die Umfrageforschung. Standardisierte Umfragen sind und bleiben das mit Abstand zentrale Datenerhebungsinstrument der empirischen (Sozial-)Wissenschaften. Erst durch diese sind allgemeingültige Aussagen über eine größere Grundgesamtheit, wie die Bevölkerung eines Landes, überhaupt möglich (Schnell 2012: 28 ff.). Um Teilnahmeraten konstant zu halten zeigt sich jedoch auch immer häufiger: „[...] quality comes at a price: higher response rates require more expensive forms of data collection.“ (Bethlehem/Cobben/Schouten 2011: 84).

Die Abbildung 1.1 zeigt die Nutzung verschiedener Befragungsmethoden durch die ADM-Institute in der Bundesrepublik. Die zunehmende prozentuale Nutzung von Onlineumfragen in den letzten 15 Jahren ist klar zu Lasten der kostenintensiven persönlichen Befragung gegangen. Dies unterstreicht eine Priorisierung geringerer Kosten bei der Wahl der präferierten Erhebungsform bei einer Inkaufnahme der bekannten Einschränkungen der onlinebasierten Befragung

Abbildung 1.1.: Anteile Nutzung Erhebungsmodus der ADM-Institute 1990 bis 2015



Darstellung als Fortschreibung der Grafik von Schnell 2012: 308. Quelle der Daten: https://www.adm-ev.de/zahlen, Punkt 8 (Abruf: 11. Mai 2018).

(Schnell 2012: 308 f.). Der langjährige Verweis auf die qualitativen Schwächen von Onlineumfragen führt also nicht zu einer Abkehr von diesem Format. Die Arbeit wird sich daher auf den Aspekt der Möglichkeit der nachträglichen Korrektur von Onlineumfragen auf der Basis von Access-Panels eines spezifischen Bereichs der Umfrageforschung konzentrieren: Wahlumfragen. Hierbei stehen mehrere leitende Forschungsfragen im Raum: Was für ein qualitatives Niveau ist durch onlinebasierte Wahlumfragen auf der Basis von Access-Panels möglich? Welchen Einfluss hat die angewandte Rekrutierungsform in dieses? Lässt sich die eingeschränkte Qualität von dieser durch die Anwendung von statistischen Korrekturmethoden durch Gewichtungsverfahren steigern?

Ist dies der Fall, lassen sich auch deren Vorteile nutzen. Die Auswirkungen der Selektivität dieser Stichproben durch die mangelnde Möglichkeit der Abdeckung der Bevölkerung ohne Internetzugang und die nicht-zufallsbasierte Entscheidung zur Registrierung und Auswahl können dann abgemildert werden. Ebenfalls wird herausgestellt, dass bereits die Entscheidung für eine online oder offline stattfindende Art der Rekrutierung in Access-Panels einen zentralen Einfluss auf die Zusammensetzung und Qualität durch die ungleiche Verteilung der Internetnutzung innerhalb der Bevölkerung hat. Als Mittel der Steigerung der Datenqualität von Onlineumfragen ist die Angleichung an qualitativ hochwertige und zufallsbasierte Referenzumfragen eine prominente Möglichkeit. Der Schlüssel zum Erfolg ist ein gut spezifiziertes Korrekturmodell zur Herausstellung der Unterschiede von Onlineumfragen zu den jeweiligen Referenzumfragen. Dieses muss erklären können, welche Variablen für die qualitativen Einschränkungen verantwortlich sind *und* in einer Verbindung zu weiteren inhaltlich interessierenden und nicht korrekt erfassten Variablen stehen. Dann ist beispielsweise

die korrekte Erfassung der Wahlentscheidung durch Onlineumfragen über die statistische Angleichung an diese Referenzumfragen möglich.

Da sowohl die persönliche als auch telefonische Befragung zufallsbasierte Auswahldesigns ermöglichen, werden diese zueinander verglichen. Die leitende Fragestellung ist hierbei: Welche dieser beiden Befragungsformen garantiert zuverlässig die höhere Datenqualität bezüglich der Erfassung des Elektorats? Um dies herzuleiten, werden Erhebungen persönlicher und telefonischer Befragungen über einen langfristigen Zeitraum seit Anfang der 1990er Jahre betrachtet. Hierdurch ist die Entscheidung über einen „Goldstandard“ der zufallsbasierten Erhebung als eine qualitativ hochwertige Referenz möglich. Zu diesen können dann nicht-zufallsbasierte Erhebungen onlinebasierter Befragungen verglichen und hinsichtlich ihres Potentials der Erfassung der Wahlentscheidung und Beteiligungsabsicht eingeschätzt werden. Das nachfolgende Unterkapitel wird nun aufzeigen wie die Arbeit die Beantwortung dieser Fragen erreichen wird.

1.2. Weiterer Aufbau der Arbeit

Die Arbeit gliedert sich in drei große Bereiche: Einen theoretischen Rahmen, eine Diskussion der Erhebungsvariation von Bevölkerungsumfragen und einen empirischen Teil zur Überprüfung der Qualität von Wahlumfragen. Der *theoretische Rahmen* wird einleitend in Kapitel 2 den „Total Survey Error“ (TSE) als zentrales Konzept zur Erfassung von Fehlerquellen in Umfragen präsentieren. Hierdurch werden diese aufgearbeitet und ihre Bedeutung herausgestellt und abschließend in einem Gesamtmodell als Einschränkung der Qualität einer Umfrage visualisiert (Kapitel 2.4). Den gesamten Umfragefehler dann empirisch zu erfassen, widmet sich Kapitel 3. Hierzu wird der Begriff der Repräsentativität einer Umfrage diskutiert und die Erfassung einer Verzerrung durch das Konzept des statistischen Bias herausgestellt. Nach einer Diskussion der Rolle von Teilnahmeraten und der Erfassung der Wahrscheinlichkeit der Teilnahme an einer Umfrage, präsentiert das Kapitel die Möglichkeit der empirischen Erfassung einer Verzerrung durch zentrale Kennzahlen der Umfrage- und Wahlforschung. Darauf aufbauend kann das Kapitel 4 darlegen, wie diesen erfassten Verzerrungen durch Gewichtungsverfahren effektiv begegnet werden kann. Ein spezieller Fokus liegt auf der Präsentation der Modellierung von „Propensity Scores“ (PS) und der direkten Nutzung dieser als Bedeutungsgewichte, oder deren weitere Nutzung im Rahmen eines „Propensity Score Matching“ (PSM). Hierzu werden verschiedene Matchingalgorithmen präsentiert und die Balancierung als zentra-

les Gütekriterium herausgestellt. Da der Erfolg der Modellierung von PS zentral von den verfügbaren Informationen abhängt, wird das Kapitel 7 die zentralen Eigenschaften zur Korrektur geeigneter Variablen herausarbeiten. Anschließend werden soziodemografische Merkmale, die Persönlichkeit einer Person, der Interesse am und das Verständnis des Umfragethemas und die Einstellungen von Befragten als die Variablen herausgearbeitet, welche sowohl das Teilnahmeverhalten als auch zentrale Zielvariablen einer Wahlumfrage, wie das Wahl- und Beteiligungsverhalten, beeinflussen. Dies prädestiniert sie als „Wirkungskanäle der Korrektur".

Der anschließende *zweite Teil der Arbeit* wird die Erhebungslandschaft der Umfrageforschung diskutieren. Hierzu werden in einem ersten Schritt in Kapitel 6.1 die „zufallsbasierten Klassiker" der persönlichen und telefonischen Befragung präsentiert und hinsichtlich ihres qualitativen Niveaus eingestuft. Anschließend wird das Kapitel 6.2 den „neuen Erhebungsmodus" der onlinebasierten Befragung vorstellen. In einem ersten Schritt werden die allgemeinen Vor- und Nachteile präsentiert, welche mit dieser Form der Datenerhebung verbunden sind. Anschließend wird das Kapitel herausstellen, dass die Durchführung von Onlineumfragen mit einem nicht zu vermeidenden *systematischen* Noncoverage-Fehler belegt ist und zudem das hohe Potential eines *zusätzlichen* Nonresponse-Fehlers durch eine ungleiche Nutzungsverteilung innerhalb der allgemeinen Bevölkerung droht. Ob die Konstruktion von Access-Panels eine Möglichkeit des Umgangs der mangelnden Realisierbarkeit einer zufallsbasierten Auswahl darstellt, wird anschließend evaluiert. Als das zentrale Kriterium zur Einschätzung der Qualität eines solchen wird der Rekrutierungsprozess in das Access-Panel herausgestellt. Dieser Teil der Arbeit über die Erhebungsvariation der Umfrageforschung endet mit einer darauf aufbauenden Zusammenfassung der jeweiligen Bedeutung der Fehlerquellen für die betrachteten Erhebungsmodi.

Der nachfolgende *empirische Teil der Arbeit* wird in Kapitel 5 die Datengrundlage präsentieren. Hierzu werden die relevanten Komponenten der „German Longitudinal Election Study" (GLES), die persönlichen Befragungen der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (Allbus) und des „European Social Survey" (ESS), die telefonischen Befragungen des Politbarometer und Forsa sowie der Mikrozensus als (soziodemografische) Referenzverteilung der vorherigen Umfrageprogramme vorgestellt. Auf der Basis des letzteren wird zudem die Zusammensetzung der, im Rahmen der Arbeit genutzten, beiden Access-Panel von Respondi und LINK hinsichtlich ihrer Möglichkeit evaluiert, ein Abbild der volljährigen Bevölkerung darzustellen. Anschließend wird

das Kapitel 8 die Qualität zufallsbasierter Befragungen seit 1991 in Relation zu den Referenzverteilungen der amtlichen Statistik herausstellen. Hierdurch ist die Ermittlung eines „Goldstandards der Erhebung“ durch persönliche Befragungen möglich. Anschließend wird die Qualität der Erfassung des Wahlverhaltens und der Wahlbeteiligung auf der Basis telefonbasierter Wahlumfragen in Relation zu den persönlichen Umfrageprogrammen betrachtet. In einem ersten Schritt wird dies für den Zeitraum 1994 bis 2014 auf der Basis der Umfragen von Forsa und Politbarometer realisiert. In einem zweiten Schritt wird die telefonisch erhobene Komponente 2 der GLES, vor dem Kontext der Bundestagswahlen 2009 und 2013, untersucht.

Das letzte Kapitel 10 der Arbeit widmet sich schließlich der Qualität onlinebasierter Wahlumfragen auf der Basis von Stichproben aus Access-Panels. Hierzu wird in einem ersten Schritt der Kontext der Bundestagswahl 2013 beleuchtet: Lassen sich systematische Unterschiede zwischen einer persönlichen und onlinebasierten Befragungen aufzeigen? Anschließend wird die Selektivität der Stichproben der letzteren an Hand der im theoretischen Teil herausgearbeiteten Variablen modelliert und der Einfluss der Korrektur dieser systematischen Unterschiede auf das Wahl- und Beteiligungsverhalten analysiert. Die daraus gewonnen Erkenntnisse werden dann durch die langfristige Ausweitung auf den Zeitraum 2009 bis 2016 validiert. Hierdurch werden die aufgezeigten Verzerrungen onlinebasierter Befragungen auf der Basis von Access-Panels und ihre Möglichkeit der Korrektur als konstant herausgestellt. Durch die Möglichkeit eines Vergleichs eines online- und eines offlinerekrutierten Access-Panel wird zudem der Effekt dieser Form der Rekrutierung auf die resultierende Datenqualität der Stichproben aufgezeigt.

Schließlich endet die Arbeit in Kapitel 11 mit einem Fazit der Aufarbeitung der Ergebnisse. Hierdurch wird die Frage beantwortet: Welches Qualitätsniveau kann onlinebasierten Wahlumfragen auf der Basis von Access-Panels zugeschrieben werden und inwieweit lassen sich die Einschränkungen der Qualität durch eine Selektivität durch Gewichtungskorrekturen im Rahmen eins PSM begegnen? Hierzu werden die Erkenntnisse der telefonischen und onlinebasierten Befragung kritisch evaluiert und eine Prognose für die Entwicklung gezogen: Welcher Trend lässt sich für erstere ablesen und welche kommende Bedeutung impliziert dies für die zweitere Erhebungsart.

Teil I.

Theoretischer Rahmen

2. Fehlerquellen in Umfragen: Der „Total Survey Error“ (TSE)

Warum kommt es auf der Basis von Umfragen zu einer „[...] difference between an obtained value and the true value, usually the true value for the larger population of interest“ (Weisberg 2005: 18)? Welche Fehlerquellen können die Zusammensetzung von erhobenen Stichproben so beeinflussen, dass sich die Verteilungen von Variablen innerhalb der Stichprobe von deren Verteilungen in der anvisierten Grundgesamtheit unterscheiden? Als theoretischer Rahmen zur Erklärung eignet sich das Paradigma des „Total Survey Error“ (TSE) (Deming 1944; Andersen/Frankel/Kasper 1979; Groves 1989; Lessler/Kalsbeek 1992; Groves/Presser/Dipko 2004; Weisberg 2005).[1] Dieses versucht die Gründe zu systematisieren, warum Umfragen fehlerbehaftet sind. Es beantwortet warum sich die Umfragemessung unter tatsächlichen Bedingungen von denen unter idealen (also fehlerminimalen) Bedingungen unterscheidet und auf welche Ursachen sich diese Differenz zurückführen lässt (Faulbaum/Prüfer/Rexroth 2009: 50). Es ist damit ein Konzept, welches „[...] purports to describe statistical properties of survey estimates, incorporating a variety of error sources.“ (Groves/Lyberg 2010: 849) und stellt damit die „[...] accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data.“ (Biemer 2010: 817) dar.

Zur Erfassung des TSE existiert keine einheitliche Definition, was hierunter präzise verstanden werden soll und welche Fehlerquellen für Umfragen relevant sind (Groves/Lyberg 2010: 850). Was alle Klassifikationen jedoch gemeinsam haben, ist die Akzeptanz eines solchen „totalen Fehlers“ einer Umfrage als ein Indikator der Datenqualität dieser. Ebenso vereint er als Konzept, dass in Umfragen eine ganze Reihe von Gründen für diesen gesamten Fehler verantwortlich sind, welche über den reinen Stichprobenfehler hinausgehen (Groves/Lyberg 2010: 850 ff.). Die möglichen Fehlerquellen in Umfragen berühren daher ganz

[1] Für eine Übersicht über die Entwicklung des Konzepts siehe Groves/Lyberg (2010). Systematisierungsversuche existierten bereits vorher. Bereits Kish (1965: 519) führt eine solche Unterteilung verschiedener Fehlerquellen auf, ohne sie explizit als TSE zu bezeichnen. Erst Groves (1989) führte endgültig die Terminologie ein (Weisberg 2005: 13).

konkret die Qualität und Repräsentativität einer Umfrage.[2] Ob und in welchem Ausmaß Abweichungen zwischen den Schätzungen der Stichprobe und den zu Grunde liegenden Parametern der Grundgesamtheit auf welche Fehlerquellen zurückführbar sind, bleibt jedoch (meist) unklar. Sofern der wahre Wert eines Parameters bekannt ist, lässt sich nur die insgesamte Abweichung der Stichprobe diagnostizieren. Der zentrale Nutzen des TSE-Konzepts ist daher vielmehr den gesamten Fehler einer Umfrage überhaupt sinnvoll theoretisch unterteilen zu können und damit separiert anzugehen (Biemer 2010: 844 ff.).

Der prominenteste Fehler bezüglich der Nutzung von Umfragen ist der „Stichprobenfehler“. Dieser ergibt sich durch die Nutzung eines eingeschränkten Personenkreises an Stelle einer Vollerhebung der anvisierten Grundgesamtheit (z.B. Lohr 2010: 16). Neben diesem können weitere „Nicht-Stichprobenfehler“ entstehen. Eine Frage erfasst ein anderes als das eigentlich intendierte latente Konstrukt („Spezifikationsfehler“). Ein Fehler durch das genutzte Auswahldesign entsteht, wenn dieses nicht alle zur Grundgesamtheit gehörenden Personen berücksichtigen kann („Noncoveragefehler“). Wird eine Person erfolgreich zur Befragung ausgewählt, können zudem verschiedene Gründe eine Befragung verhindern („Nonresponsefehler“). Fehler entstehen ebenfalls durch die an einer Umfrage beteiligten Personen und deren Einfluss auf die Ergebnisse („Messfehler“), sowie Fehler, die während der Verarbeitung der erhobenen Daten („Verarbeitungsfehler“) auftreten (z.B. Biemer/Lyberg 2003).[3]

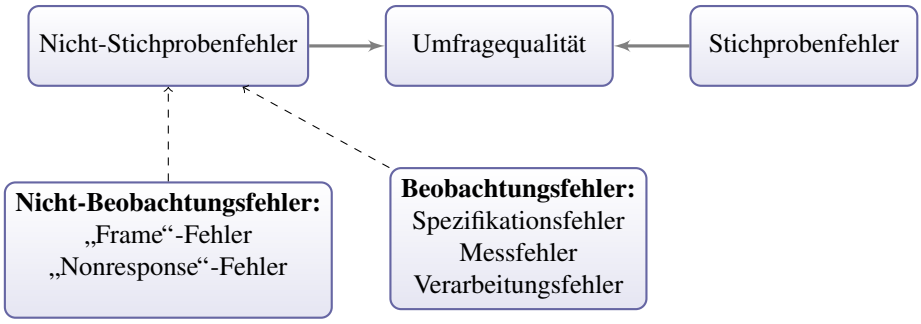
Abbildung 2.1 zeigt diese systematische Unterteilung der möglichen Fehlerquellen, welche die Qualität einer Umfrage im Rahmen des Paradigma des TSE beeinflussen.

Neben dem prominent in der Inferenzsstatistik ergründeten und quantifizierbaren Stichprobenfehler, lassen sich die verschiedenen Varianten des Nicht-Stichprobenfehlers zusätzlich auch konzeptuell in einen Beobachtungs- und Nicht-Beobachtungsfehler unterteilen (z.B. Bethlehem/Cobben/Schouten 2009: 180; Faulbaum 2014: 440). Diese Trennung ist analytisch hilfreich: Verhindert der Erhebungsrahmen („Frame“) bereits im Vorfeld die Auswahl einer Person

[2] Für eine ausführliche Diskussion des Begriffs der Repräsentativität und die empirische Erfassung des Qualitätsniveaus von Umfragen im Rahmen der Arbeit siehe Kapitel 3.

[3] Smith (2011) weist auch auf die zusätzliche Fehlerdimension der mangelnden Vergleichbarkeit innerhalb der vergleichenden Untersuchung („Comparison Error“) hin. Der wahre Wert ist daher selber nicht zwingend eindeutig und unterschiedlichen Interpretationen durch den jeweiligen Kontext ausgesetzt. Auch ist unklar, ob verschiedene Fragen desselben Konstrukts wirklich fehlerbelastet sind oder einfach nur unterschiedliche Nuancen erheben. Um dies zu berücksichtigen, wird ein etwas breiterer Rahmen des TSE als „Total Survey Measurement Variation (TSMV) vorgeschlagen (Smith 2011: 465), welcher jedoch im Rahmen der Arbeit nicht verfolgt wird.

Abbildung 2.1.: Fehlerquellen in Umfragen



Quelle: Eigene Darstellung.

als Teil der Stichprobe, oder ist eine Befragung einer durch den Erhebungsrahmen korrekt erfassten Person nicht möglich („Nonresponse“), dann findet auch keine Beobachtung dieser Person statt. Ist hingegen eine Person als Teil der Grundgesamtheit für die Befragung innerhalb der Stichprobe ausgewählt worden und hat sich entschieden zu partizipieren, existieren die aufgeführten Möglichkeiten des Beobachtungsfehlers. Der „Nicht-Beobachtungsfehler“ ist daher dem „Beobachtungsfehler“ hinsichtlich der Bedeutung vorgelagert. Aus diesem Grund lässt sich das TSE-Modell auch in die beiden Oberkategorien der „Repräsentation“ und „Messung“ zerlegen. Nicht-Beobachtungsfehler erzeugen als Resultat eine abweichende Repräsentation der Gruppen der anvisierten Grundgesamtheit. Beobachtungsfehler erzeugen anschließend eine Abweichung des gemessenen vom anvisierten wahren Wert dieser Messung. Dies wird durch das gewählte Messinstrument, den Einfluss der der beteiligten Personen oder die fehlerhafte Weiterverarbeitung von Daten verursacht (Groves/Fowler et al. 2009: 48). Da sich die Arbeit auf die Probleme der selektiven Teilnahme und Auswahl und ihre Korrektur durch Gewichtungsverfahren konzentriert, steht die „Repräsentation“ in Form der Nicht-Beobachtungsfehler im Fokus. Die Beobachtungsfehler werden hingegen nur subsumierend als weitere nachfolgende Fehlerquellen aufgeführt. Die nachfolgenden Unterkapitel werden nun diese Fehlerquellen hinsichtlich ihrer Bedeutung, Ursachen und Folgen beschreiben. Anschließend können die verschiedenen Fehlerursachen in einem Gesamtmodell zusammengeführt werden (Kapitel 2.4).

2.1. Der Stichprobenfehler

Der Stichprobenfehler tritt auf, da nur ein Teil des eigentlich interessierenden Personenkreises befragt wird und dies als eine mögliche Fehlerquelle berücksichtigt werden muss (Weisberg 2005: 18 f.). Er ist Folge der zufälligen Realisierungen der Schätzer von gezogenen Stichproben hinsichtlich der jeweiligen Parameter der anvisierten Grundgesamtheit. Das Ziel des Stichprobenfehlers ist die Berücksichtigung der Unsicherheit der Prognose auf der Basis der Stichprobe. Sofern kein zusätzlicher systematischer Fehler existiert, lässt sich eine Normalverteilung der Realisierungen eines Schätzers um den Populationswert der *anvisierten* Grundgesamtheit annehmen. Je höher die Streuung eines Schätzers in Form seines Stichprobenfehler ist, umso wahrscheinlicher werden auch augenscheinlich hohe Abweichungen von dem anvisierten Populationswert (z.B. Groves/Fowler et al. 2009: 56 ff.).

Der Stichprobenfehler ist – bei korrekter Berücksichtigung der Auswahlwahrscheinlichkeiten – für die Qualität einer Umfrage jedoch eine unproblematische Fehlerquelle (Bethlehem/Cobben/Schouten 2009: 180 f.). Er wird lediglich der Tatsache gerecht, dass eine Stichprobe und nicht alle möglichen Personen befragt werden (Thompson 2012: 5). Durch die Arbeiten von Neyman (1934) und Neyman (1937), Deming (1950) and Cochran/Rubin (1973), als maßgebliche Begründungen der Inferenzstatistik, lässt sich diese Unsicherheit in Form des Standardfehlers eines Schätzers angeben. Dieser ist abhängig von der Stichprobengröße und der Varianz einer Variablen. Die Konstruktion eines, von einem gewählten α-Niveau abhängigen, Konfidenzintervalls dient daher zur Angabe der Präzision einer Schätzung (z.B. Bethlehem/Cobben/Schouten 2009: 37). Für einen unbekannten Erwartungswert μ lässt sich auf Grundlage der Realisierung des Schätzers $\bar{x}$ innerhalb der Stichprobe das zugehörige $(1-\alpha)$-Konfidenzintervall bestimmen:

$$\bar{x} \pm t_{(n-1,1-\alpha/2)} \cdot \sigma_{\bar{x}} \text{ mit } \hat{\sigma}_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \text{ und } \hat{\sigma}_x = \frac{n}{n-1} s_x$$

Ebenso lässt sich zwischen arithmetischen Mittel und dem zu Grunde liegenden (angenommenen) Erwartungswert testen, ob eine statistisch signifikante Abweichung besteht. Hierzu muss lediglich der empirische t-Wert bestimmt werden:

$$T_n = \frac{\bar{x} - \mu}{s_x / \sqrt{n}} \, t(n-1)$$

Dieser kann dann mit seinem theoretischen Referenzwert verglichen werden.

Übersteigt der empirische den theoretischen Wert bei gegebenem α-Niveau:

$$|T_n| > t_{(n-1,1-\alpha/2)}$$

existiert eine statistisch signifikante Abweichung (z.B. Jann 2005: 145 f.) – die vorliegende Stichprobe mit der Realisierung von $\bar{x}$ kann damit nicht aus einer Grundgesamtheit stammen, wo μ gilt.

Da auf diesem Konzept auch komplexere Testverfahren, wie die Ermittlung von statistisch signifikanten Gruppenunterschieden der anvisierten Grundgesamtheit basieren, ist eine korrekte Berechnung der jeweiligen Standardfehler für die Bestimmung darauf aufbauender Konfidenzintervalle und Testverfahren elementar. Ansonsten droht eine zu optimistische Einschätzung der eigenen Ergebnisse durch zu geringe Standardfehler und zu enge Konfidenzintervalle. Statistische Tests weisen dann zu schnell signifikante Aussagen aus – die Realität ist aber unschärfer als suggeriert.

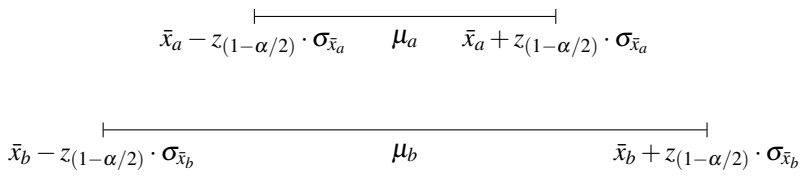
Die Fehleinschätzung des zufälligen Fehlers kann die Folge der fehlerhaften Berücksichtigung der Auswahlwahrscheinlichkeiten sein. Diese sind nicht konstant und bestimmte Gruppen können qua Auswahldesign (bewusst) über- oder unterrepräsentiert sein. Ebenfalls muss ein komplexeres mehrstufiges Auswahldesign bei der Bestimmung der Standardfehler berücksichtigt werden (z.B. Häder/Ganniner/Gabler 2009: 15 f.). Dies muss nachträglich bei der Schätzung korrigiert werden, um den Stichprobenfehler korrekt einzuschätzen (Groves/ Fowler et al. 2009: 77 ff.).[4]

Die Abbildung 2.2 veranschaulicht diesen Zusammenhang. Die beiden Stichproben a und b wurden aus derselben Grundgesamtheit nach einem zufallsbasierten Verfahren ausgewählt. Hierdurch sind die Erwartungswerte identisch ($\mu_a = \mu_b$). Die Stichprobe b zeigt gegenüber a jedoch eine größere Bandbreite des Konfidenzintervalls, gegeben α.

Dies kann drei Ursachen haben: Es werden in a schlicht mehr Person als in b befragt, wodurch sich die Präzision erhöht. Die Stichprobe b kann ebenfalls eine systematisch höhere empirische Varianz aufweisen – bei einer strikten Zufallsauswahl ist dies jedoch unwahrscheinlich. Eine weitere Ursache ist die Art der Auswahl: Handelt es sich bei a um eine einfache Zufallsauswahl, bei b hingegen um eine mehrstufige Zufallsauswahl, führt die korrekte Bestimmung

[4] Den Erhebungsrahmen als Fehlerquelle weist z.B. bereits Kish (1965: 519) als Fehler der Stichprobenziehung aus, da dieser die Auswahlwahrscheinlichkeiten berührt. Dieser ist aber nicht zu verwechseln mit dem allgemeineren „Frame“-Fehler, welcher ebenfalls durch den Erhebungsrahmen hervorgerufen wird (Groves/Lyberg 2010: 854).

Abbildung 2.2.: Konfidenzintervall zufälliger Fehler



Quelle: Eigene Darstellung.

(meist) zu einem größeren Standardfehler in b. Dies ist jedoch nur eine Frage der korrekten Angabe der Präzision und Genauigkeit der Prognose, nicht ob die Prognose überhaupt korrekt ist. Der schwerwiegendere Fall ist, wenn nicht (nur) ein zu gering eingeschätzter zufälliger Fehler existiert, sondern zusätzlich ein systematischer Unterschied zwischen der Schätzung auf der Basis der Stichprobe und der *eigentlich* anvisierten Grundgesamtheit besteht. Ein solcher systematischer Fehler kann durch Nonresponse und Noncoverage verursacht werden.

2.2. Nicht-Beobachtungsfehler

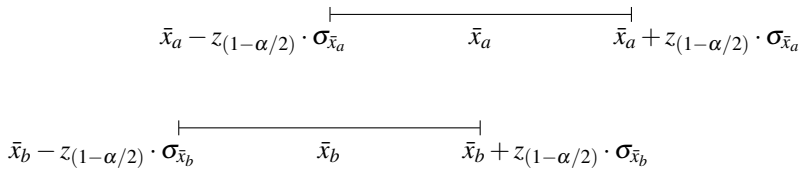
Ein adäquates Erhebungsdesign/‘‘Frame“ garantiert allen Personen/Gruppen der Grundgesamtheit durch die Art der Stichprobenziehung die gleiche oder bestimmbare Wahrscheinlichkeit, Teil der Stichprobe zu werden. Zusätzlich ist es wünschenswert, dass sich keine der Gruppen der Grundgesamtheit systematisch der Befragung entzieht. Beide Fehlerquellen erzeugen durch den „[...] failure to deliver the survey request [...] refusal to participate“ (Groves/Fowler et al. 2009: 192) einen Fehler durch die mangelnde Erfassung eigentlich wünschenswerter Teilnahmen von Personen. Diese bilden den Nicht-Beobachtungsfehler.

Existiert in einer Stichprobe b ein systematischer Fehler hinsichtlich einer interessierenden Variable, ist der zugehörige Erwartungswert μ_b nicht mehr deckungsgleich mit dem Erwartungswert μ_a einer Stichprobe a ohne einen solchen systematischen Fehler – obwohl beide Stichproben technisch aus derselben anvisierten Grundgesamtheit (z.B. die wahlberechtigte Bevölkerung) sein können. Sind hierfür Noncoverage- und Nonresponseprobleme verantwortlich, dann ist der Unterschied zwischen den beiden Stichproben durch eine selektive Befragung in b verursacht. Hierdurch ergibt sich eine größere Diskrepanz der arithmetischen Mittel beider Stichproben und damit auch ein Unterschied der Erwartungswerte ($\mu_a \neq \mu_b$. Durch die Selektivität der Auswahl oder/und Befragung unterscheiden sich die Grundgesamtheiten der beiden Stichproben.

Ist der Unterschied der beiden realisierten Stichproben a und b lediglich

geringfügig (Abbildung 2.3), also das Ausmaß einer systematischen Verzerrung eher gering, können die Auswirkungen hinnehmbar sein. Die Wahrscheinlichkeit

Abbildung 2.3.: Konfidenzintervall systematischer Fehler I

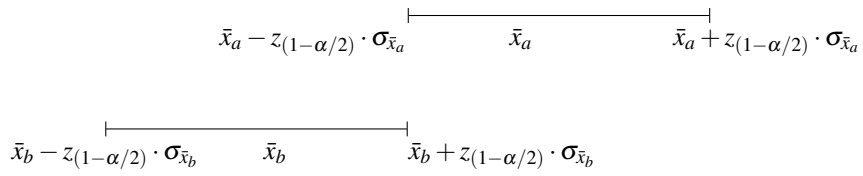


Quelle: Eigene Darstellung.

ist dann relativ hoch, dass beide Konfidenzintervalle den wahren Parameter μ_a enthalten. Für beide Stichproben ist es dann plausibel anzunehmen, dass sie aus derselben anvisierten Grundgesamtheit stammen.

Anders sieht dies aus, wenn es zu einem bedenklichen Ausmaß der Verzerrung kommt (Abbildung 2.4). Die Wahrscheinlichkeit ist sehr gering oder quasi Null, dass die vorliegende Stichprobe b überhaupt noch den wahren Wert für μ_a der eigentlich anvisierten Grundgesamtheit enthält.

Abbildung 2.4.: Konfidenzintervall systematischer Fehler II



Quelle: Eigene Darstellung.

Anzunehmen, dass beide Stichproben dann die gemeinsam anvisierte Grundgesamtheit repräsentieren können, ist utopisch. Der systematische Fehler in b sorgt daher für eine statistisch signifikante Abweichung zwischen $\bar{x}_b$ und μ_a. Ursächlich hierfür sind die Nicht-Stichprobenfehler. Neben der fehlerbelasteten Beobachtung/Messung sind dies vor allem die selektive Teilnahme an der Umfrage oder Probleme hinsichtlich der Auswahl durch das Erhebungsdesign. Diese Probleme werden nun als Nonresponse und Noncoverage dargestellt.

2.2.1. „Nonresponse“- Fehler

Nonresponse liegt vor, wenn über eine Person keine oder nur manche Informationen bekannt sind, welche sie nur selbst offenbaren kann. Entzieht sie sich komplett der Befragung, liegt ein Unit-Nonresponse vor. Weisen nur bestimmte Merkmale fehlende Wert auf, handelt es sich um ein Item-Nonresponse.

Eine grundsätzliche Teilnahme besteht, Antworten auf einzelne Fragen sind jedoch nicht vorhanden.[5] Durch Nonresponse sinkt erstens die Präzision der Schätzungen auf Basis der Stichprobe, da durch die sinkende realisierte Stichprobengröße der zufällige Fehler steigt. Tritt Nonresponse zusätzlich nicht rein zufällig auf, kommt es zu einem systematischen Fehler (Bethlehem/Cobben/ Schouten 2011: 3).

Die bereits einleitend skizzierte Entwicklung sinkender Teilnahmeraten bei Befragungen ist eine direkte Folge steigender Nonresponseraten (z.B. Schnell 1997; Heer 1999; Steeh et al. 2001; Leeuw/Heer 2002; Stoop 2004; Curtin/ Presser/Singer 2005; Holbrook/Krosnick/Pfent 2007; Brick/Williams 2012). Die zentralen Gründe für Nonresponse sind die explizite Verweigerung durch die ausgewählte Person, die mangelnde Möglichkeit der Teilnahme durch eine Erkrankung oder die Nichterreichbarkeit der identifizierten Person (Schnell 2012: 157). Als zentrale Gründe bezüglich Umfang und Fehleranfälligkeit durch Nonresponse haben sich schon früh die Verweigerung und Nichterreichbarkeit von Befragten herausgestellt (z.B. Leeuw/Heer 2002: 44). Bewusste Verweigerungen stellen hierbei traditionell den relativ größten Teil der Gründe für Nonresponse dar (z.B. Stoop et al. 2010: 158 f.; Bethlehem/Cobben/Schouten 2011: 6).

Zu einem systematischen Fehler kommt es, wenn Nonresponse abhängig von interessierenden Merkmalen der ausgewählten Personen ist. In dem Fall unterscheiden sich die Teilnehmer und Nicht-Teilnehmer einer Umfrage systematisch (Groves/Fowler et al. 2009: 196 ff.). Obwohl die Ursachen für eine bewusste Verweigerung der Teilnahme durch eine Person sehr vielfältig sein können, lassen sie sich häufig auf eine Kosten-Nutzen-Abwägung im Rahmen der „Rational-Choice“-Theorie zurückführen. Wird dies berücksichtigt, lässt sich Verweigerungen – mit einigem Aufwand – auch effektiv begegnen (Schnell 1997: 157-260). Hierzu muss identifiziert werden, welche Rahmenbedingungen einer Umfrage für eine Person eine hohe individuelle Bedeutung haben und somit für eine erfolgreiche Teilnahme verantwortlich sind. Dies sind beispielsweise Merkmale des Befragten, die Art und Gestaltung des Fragebogens sowie die Kontaktvorgaben durch die Konzeption der Studie sowie den Kontakt durch den Interviewer (Groves/Couper 1998: 31).

Im Rahmen der „Leverage-Saliency Theory“ (Groves/Singer/Corning 2000) wird diese relative Bedeutung einer Rahmensituation (z.B. eines Anschreibetext oder das Thema einer Umfrage) als „Leverage“ bezeichnet. Diese können sowohl

[5] Der Fokus der Arbeit liegt auf dem Unit-Nonresponse, welches nachfolgend sprachlich als Äquivalent zu Nonresponse genutzt wird.

positiv als auch negativ besetzt sein. Abhängig von den Merkmalen einer zu befragenden Person kann dann versucht werden, die Bedeutung („Salience“) bestimmter Faktoren mehr herauszuheben. Hierdurch wird die individuelle Wahrscheinlichkeit der Teilnahme einer Person erhöht (Groves/Singer/Corning 2000: 306 f.). Um eine hohe Teilnahmerate zu erreichen, sind daher die individuellen Kosten für jeden potentiellen Teilnehmer niedrig und der Nutzen durch eine spezifische Berücksichtigung seiner Präferenzen hoch zu halten (Dillman/Smyth/Christian 2009). Das situationsabhängige Zuschneiden der Rahmensituation auf die individuellen Bedürfnisse eines Befragten ist daher elementar, um seinen heuristischen Entscheidungsprozess zu Gunsten einer Teilnahme zu bewegen und Nonresponse zu verhindern (Groves/Couper 1998). Hierzu ist die Annahme plausibel, dass sich kein Personenkreis generell Befragungen um jeden Preis verweigert (Schnell 2012: 157 ff.). Praktisch geht eine solche „Individualisierung“ der Befragung aber mit höheren Kosten einher. Ebenso lassen sich nicht unbedingt alle Faktoren identifizieren, welche für eine erfolgreiche Teilnahme verantwortlich sind. Aus diesem Grund sind Umfragen immer mit einem gewissen Grad an Verweigerungen der Teilnahme konfrontiert. Ob dies in einer Umfrage zu einem systematischen Fehler führt, ist vom jeweiligen Kontext und Ziel der Befragung abhängig.

„Follow-up“-Studien von Nonrespondenten sind eine Möglichkeit, Informationen über diese Gruppe zu erlangen. Hierzu muss es geschafft werden, „[...] diejenigen zu befragen, die sich nicht befragen lassen [...]“ (Proner 2011: 30). Dies ermöglicht einen Einblick, ob sich diese Gruppe auch hinsichtlich eigentlich interessierender Merkmale (beispielsweise des Wahlverhaltens) von dem ursprünglichen Kreis der Respondenten unterscheidet. Diese Studien, im Sinne der Tradition des „Callback Approach“ (CBA) nach Hansen/Hurwitz (1946) und des „Basic Question Approach“ (BQA) nach Kersten/Bethlehem (1984), haben jedoch einen zentralen Nachteil: Sie sind sehr kostenintensiv (Bethlehem/Cobben/Schouten 2011: 290) und finden sich daher auch nur vergleichsweise selten (z.B. Keeter/Miller et al. 2000; Stoop 2004; Neller 2005; Proner 2011; Matsuo et al. 2010; Vandenplas et al. 2015). Hohe Datenschutzanforderungen machen zudem die erneute Kontaktierung von explizit verweigernden Personen auch praktisch problematisch, weshalb die Durchführung solcher Studien in der Bundesrepublik kaum stattfindet (Schnell 2012: 197). Werden sie doch durchgeführt, sorgen selbst hohe Anreize bei prominenten Bevölkerungsumfragen wie dem Allbus nur für überschaubare Resultate (z.B. zur Durchführung des Allbus+ Proner 2011: 288).

Neben der expliziten Verweigerung stellt die schwierige oder keine Erreichbarkeit einen weiteren zentralen Grund für einen Nonresponsefehler dar. Auch wenn ihr Anteil in Bevölkerungsumfragen nur einen geringeren Teil des Nonresponse ausmacht,[6] kann auch dieser bereits für einen systematischen Fehler ursächlich sein. Beispielsweise zeigen Lynn/Clarke et al. (2002): „[...] it is the difficult to contact that are most different from the easy to get.“ (Lynn/Clarke et al. 2002: 142), womit „Schwer Erreichbare [...] nicht einfach durch leichter Erreichbare ersetzt werden [können] [...]“ (Schnell 2012: 161). Als Folge korreliert die Nichterreichbarkeit häufig mit interessierenden Merkmalen einer Umfrage (z.B. Schnell 1998; Peytchev/Baxter/Carley-Baxter 2009). Daher finden sich Beispiele, wo die kleinere Gruppe der „Nicht-Erreichbaren“ in Relation zu der größeren der „Nicht-Kooperierenden“ in mündlichen (z.B. Heerwegh/Abts/Loosveldt 2007; Fuchs/Bossert/Stukowski 2013) und telefonischen Befragungen (z.B. Olson 2006) einen höheren systematischen Fehler verursachen, oder zu mindestens gleichberechtigt hinsichtlich der Bedeutung sein können (z.B. Busse/Laub/Fuchs 2015).

Nonresponse ist also ein weitverbreitetes Phänomen in Umfragen mit einem hohen Potential einen systematischen Fehler in Umfragen hervorzurufen. Es ist gleichzeitig jedoch fragwürdig, ob dies eine deterministische Prädisposition eines Befragten ist, oder vielmehr eine variable Entscheidung, welcher dann auch begegnet werden kann. Einerseits wird dies durch die Rahmenbedingungen der Befragung beeinflusst, andererseits kann ein bestimmter Personenkreise „anfälliger“ hierfür sein. Die zeitlichen, kapazitären und monetären Beschränkungen der Durchführung von Umfragen verhindern jedoch in der Regel eine effektive Begegnung der Auswirkungen von Nonresponse. Es kann daher davon ausgegangen werden, dass jeder umfragenbasierte Datensatz in einem gewissen Ausmaß von einem hierdurch bedingten systematischen Fehler betroffen ist. Nonresponse als Gefahr für die Qualität verschiedenster erhobener Variablen einer Umfrage kann daher nicht ignoriert werden. Dieser muss offensiv – auch während der Datenanalyse – begegnet werden.

2.2.2. „Noncoverage“-Fehler

Das angewandte Erhebungsdesign ist ebenfalls eine mögliche Fehlerquelle, indem es manche Personen der Grundgesamtheit bei der Auswahl nicht

[6] Im Allbus lassen sich seit 2008 konstant ca. 80 % der erfolgreich befragten Personen bereits nach maximal drei Kontaktversuchen zur Teilnahme bewegen (Quelle: Allbus kumuliert (ZA4578) und Allbus 2014 (ZA5240)).

berücksichtigen kann. Trifft dies mit einer erhöhten Wahrscheinlichkeit auf interessierende Gruppen von Personen zu, entsteht ein systematischer Fehler. Auch wenn der „Frame“/“Noncoverage“-Fehler damit andere Ursachen für einen systematischen Fehler als der „Nonresponse“-Fehler aufweist, sind die Konsequenzen identisch: Es existiert ein systematischer Fehler durch die abweichenden Anteile relevanter Gruppen innerhalb der Stichprobe gegenüber der Verteilung dieser in der anvisierten Grundgesamtheit, eine statistische Repräsentativität ist nicht mehr gegeben.

Personen, welchen das Erhebungsdesign keine Teilnahme an der Umfrage erlaubt, stellen einen Undercoverage dar. Die Auswahlwahrscheinlichkeit dieser beträgt Null. Gleichzeitig kann das Erhebungsdesign Personen erfassen, welche nicht zur Grundgesamtheit gehören (Overcoverage). Beide Fehlerquellen können ebenfalls die Genauigkeit von Schätzern beeinflussen (Biemer/Lyberg 2003: 66; Bethlehem/Biffignandi 2012: 12). Overcoverage kann jedoch leicht diagnostiziert werden, da diese Personen an der Umfrage teilgenommen haben. Durch die Definition der Grundgesamtheit lässt sich klar sagen, ob eine Person Teil ebendieser ist (Bethlehem 2009: 21). Eine Abwandlung von Overcoverage bezieht sich auf die mehrfache Listung einer Person, z.B. über mehrere erreichbare Telefonanschlüsse. Sofern dies bekannt ist, kann die erhöhte Auswahlwahrscheinlichkeiten korrigiert werden (Biemer/Lyberg 2003: 66). In Telefonumfragen ist ein weiteres Beispiel, wenn die kontaktierte Nummer nicht einem Haushalt, sondern beispielsweise einen Unternehmen zuzuordnen ist (Groves/Fowler et al. 2009: 76). Overcoverage ist daher kein grundlegendes Problem für die qualitativ hochwertige Durchführung einer Umfrage. Undercoverage stellt hier den relevanten entstehenden Fehler dar, welchen das Erhebungsdesign bedingt (Weisberg 2005: 206 f.). Da dies auf einen gewissen Anteil der anvisierten Grundgesamtheit bei jeder Erhebung zutrifft, scheitert an diesem Problem zudem technisch gesehen auch eine Vollerhebung aller Personen der Grundgesamtheit (Groves/Fowler et al. 2009: 54 f.).

Das Ausmaß des Undercoverage lässt sich nicht aus den erhobenen Daten heraus erkennen. Da bereits das Auswahlverfahren diese Personen nicht berücksichtigt, existiert keine Möglichkeit Informationen über diese Personen, beispielsweise als ein Äquivalent zu Nonresponse-Studien, zu erhalten. Häufig ist noch nicht einmal ihre Existenz bekannt. Aus diesem Grund der schwierigen Erfassbarkeit ist es ein Problem, welchem eine nur relativ geringe Aufmerksamkeit gegenüber anderen Fehlerquellen zugemessen wird (Eckman/Kreuter 2013: 266).

Die Wahl des Erhebungsmodus hat den größten Einfluss auf Probleme hinsichtlich der Abdeckung („Coverage“) der Grundgesamtheit (Dillman/Smyth/Christian 2009). Probleme treten bei Telefonumfragen auf, wenn die Auswahl über Festnetznummern gesteuert wird, eine Person jedoch nicht über eine solche erreichbar ist (Groves/Fowler et al. 2009: 72). In Zeiten der Tendenz rückläufiger registrierter Festnetzanschlüsse stellt dies eine zentrale Herausforderung für Telefonumfragen dar (z.B. Keeter/Kennedy/Clark et al. 2007; Busse/Fuchs 2012). Das Auswahlverfahren persönlicher Befragungen auf der Basis von Adressen kann ebenfalls ein Problem hervorrufen. Es muss sichergestellt werden, dass jede Adresse – und damit der zugehörige Haushalt – die Möglichkeit der Auswahl besitzt *und* alle zur Grundgesamtheit gehörigen Personen für die Befragung ausgewählt werden können (Eckman/Kreuter 2013: 265). Insbesondere onlinebasierte Erhebungsmethoden haben jedoch mit Undercoverage-Problemen zu kämpfen. Sie sind nicht in der Lage allen potentiell zur Grundgesamtheit gehörenden Personen die gleiche, eine bekannte oder sogar überhaupt eine Auswahlchance zuzugestehen (z.B. Bethlehem 2010: 43 ff.). Da der Internetzugang zugleich ungleich hinsichtlich grundlegender Merkmale von Befragten verteilt ist, haben bestimmte Gruppen mit einem erschwerten oder keinem Zugang zu einer onlinebasierten Erhebung zu kämpfen (z.B. Couper/Kapteyn et al. 2007; Bandilla et al. 2009).

Neben Nonresponse hat daher auch der designbedingte Ausschluss von Personen einen zentralen Einfluss auf die Qualität der Ergebnisse der Schätzungen auf der Basis einer Stichprobe. Es besteht daher per Erhebungsinstrument keine Möglichkeit teilzunehmen (Noncoverage) oder nach der erfolgten Auswahl und Übermittlung des Wunsches der Partizipation an der Umfrage kommt diese durch die ausgewählte Person nicht zu Stande (Nonresponse). Die beiden Varianten des Nicht-Beobachtungsfehlers haben unterschiedliche Ursachen und Gründe, jedoch eine identische Auswirkung: Es drohen systematische Fehler durch eine unterschiedlich starke Berücksichtigung bestimmter Gruppen der Grundgesamtheit. Ob dieses Problem durch das Erhebungsdesign (Noncoverage) oder durch die durch das Erhebungsdesign ausgewählten Personen (Nonresponse) verursacht wird, ist eine Frage der Ursache, nicht jedoch der Auswirkung auf die Qualität einer Umfrage.

Wenn alle Personen durch das Erhebungsdesign berücksichtigt werden, alle ausgewählten Personen kontaktierbar sind und keine dieser die Teilnahme an der Umfrage verweigert, drohen jedoch noch weitere Fehlerquellen während der Erhebungsphase und der korrekten Erfassung der Ergebnisse.

2.3. Beobachtungsfehler

Erst nach der Auswahl und erfolgreichen Kontaktierung einer Person erfolgt die eigentliche Erhebung. Hierdurch können weitere Fehler durch die Art und Weise der Beobachtung/Befragung entstehen und sich ebenfalls eine Abweichung der erfassten Werte von den zu Grunde liegenden wahren Werten ergeben. Die Wahrscheinlichkeit von Beobachtungsfehlern kann für jede Frage variieren: Kommt es überhaupt zu einer Antwort, ist diese korrekt erfasst und misst sie die korrekte Bedeutung. Entgegen der Frage der allgemeinen Berücksichtigung einer Person, stellen Beobachtungsfehler somit eine *variablenspezifische Frage der Qualität der Erfassung* dar. Sind manche Variablen von einem systematischen Fehler betroffen, gilt dies nicht qua Automatismus auch für andere.

Die Erfassung des Beobachtungsfehlers lässt sich durch die beiden Konzepte der Reliabilität („Zuverlässigkeit“) und der Validität („Gültigkeit“) eines Messinstruments erfassen. Validität impliziert dabei Reliabilität: Misst ein Messinstrument was es soll, ist es auch reliabel. Vice versa gilt dies nicht, eine Frage kann äußerst zuverlässig falsche Ergebnisse liefern. Eine hohe Reliabilität impliziert einen geringen zufälligen Fehler, eine hohe Validität einen geringen systematischen Fehler (Schnell/Esser/Hill 2013: 139-144). Ist das Messinstrument reliabel, produziert es zuverlässlich mögliche Realisierungen einer Zufallsvariable und erzeugt daher einen geringeren Standardfehler eines Schätzers. Validität hingegen zielt auf eine große Genauigkeit ab, also eine möglichst geringe Abweichung der Realisierungen vom unbekannten anvisierten Populationsparameter (Lessler/Kalsbeek 1992: 238).

Ein *Spezifikationsfehler* tritt auf, wenn „[...] the concept implied by the survey question and the concept that should be measured in the survey differ.“ (Biemer/Lyberg 2003: 38). Es wird daher ein falsches latentes Konstrukt durch die Operationalisierung einer Frage gemessen (Biemer 2010: 822 f.).[7] Als Folge dieses Problems der fehlerhaften Konzeptspezifikation ergibt sich ein substantieller Unterschied zwischen der eigentlich anvisierten und tatsächlich erfassten/gemessenen Variable (z.B. Schnell/Esser/Hill 2013: 118 f. Schnell 2012: 388). Die inhaltliche Validität einer Messung ist direkt berührt, da die Intention der Fragestellung von der inhaltlichen Interpretation durch die zu befragende Person abweicht. Schlimmstenfalls erfasst die Frage ein völlig anderes Konstrukt als eigentlich beabsichtigt (Faulbaum/Prüfer/Rexroth 2009: 47 f.). Ein Spezifika-

[7] Eine Verwechslung kann mit dem weiteren statistischen Konzept des Spezifikationsfehlers eines statistischen Modells auftreten, wenn nicht alle relevanten erklärenden Faktoren berücksichtigt wurden (z.B. Ramsey 1969; Weisberg 2005: 275).

tionsfehler hat zwingend einen systematischen Fehler zur Folge, da die Messung bereits im Voraus zum Scheitern verurteilt ist. Die inhaltliche Validität einer Frage für das theoretische Konstrukt ist nicht gegeben. Entspricht die Vorstellung der inhaltlichen Validität einer Frage nicht dem, was die Person während der Fragebogenkonstruktion hinsichtlich der inhaltlichen Bedeutung dieser zugemessen hat, dann kommt es zu einem Fehler durch die *andere* Interpretation der Frage durch die befragte Person (Faulbaum/Prüfer/Rexroth 2009: 48).

Der *Messfehler* wird von der Formulierung des Fragebogens, dem Verhalten des Respondenten und der Art der Datenerhebung beeinflusst (Leeuw/Hox/Dillman 2008: 11 ff.). Dieser zielt auf die theoretische Validität einer Frage ab (Faulbaum/Prüfer/Rexroth 2009: 46).[8] Da auch politikwissenschaftlich relevante Variablen (meist) operationalisiert werden müssen, sind auch diese möglicherweise fehlerbelastet (Arzheimer 2016: 34 f.). Tritt der Messfehler zufällig auf und steht nicht in einer Beziehung zu den wahren unbekannten Ausprägungen der Merkmale und sind die Messfehler unterschiedlicher Personen unabhängig voneinander, existiert auch keine Verzerrung (Gruijter/Kamp 2008: 13 f.). Ein Beispiel für einen Messfehler ist das „satisficing“ (Krosnick 1991). Eine befragte Person gibt innerhalb einer Umfrage nur eine verkürzte oder falsche Antwort auf eine Frage, um den für sich persönlichen Aufwand der Befragung zu reduzieren (Lynn/Kaminska 2012: 217). Beispiele sind die Tendenz zur Mitte von Bewertungsskalen oder die Wahl einer Kategorie einer bewusst enthaltenden Antwort (Krosnick 1999: 547 f.). Das Ausmaß des Problems steht in einer Abhängigkeit zur Schwierigkeit und Komplexität der gestellten Frage, der Kompetenz der jeweiligen Person solchen komplexen Fragestellungen kognitiv zu begegnen und die persönliche Motivation, dies auch zu tun (Krosnick 1991: 221 f.).

Sensitive Fragen sind ein weiteres prominentes Beispiel, warum ein solcher systematischer Messfehler entsteht. In diesem Fall ist der Inhalt der Frage die Ursache. Diese kann zu persönlich gestellt sein und es besteht die Gefahr einer sozialen Sanktionierung durch andere anwesende Personen oder sogar ein allgemeines persönliches Risiko, die wahre Antwort mitzuteilen (Lynn/Kaminska 2012: 218). Auch das generelle Thema einer Umfrage kann für bestimmte Personen bereits als zu sensibel wahrgenommen werden und daher bereits die Wahrscheinlichkeit der Teilnahme an einer Umfrage beeinflussen. In der Mehrheit handelt es sich jedoch hinsichtlich der Sensitivität um ein variablenspezifisches Problem (Tourangeau/Rips/Rasinski 2000: 261 ff.). Sensitive Fragen

[8] Für eine konzeptuelle Darstellung im Sinne der klassischen Testtheorie siehe (z.B. Gruijter/Kamp 2008: 10).

verursachen einen Effekt der „Sozialen Erwünschtheit“. Für die Beantwortung einer Frage ist dann eine soziale Norm handlungsanleitend und ursächlich und nicht mehr die eigenen wahren Einstellungen und Verhaltensweisen (Kreuter/ Presser/Tourangeau 2008: 848). Da Personen mit bestimmten soziodemografischen Rahmendaten eher zu diesem Verhalten neigen, ergibt sich das Potential für einen systematischen Fehler (z.B. Preisendorfer/Wolter 2014). Dem Interviewer kommt bei der Verhinderung dieses Fehlers bei sensitiven Fragen eine besondere Rolle zu. In selbst-administrierten Formen der Befragung fällt daher der Anteil als sensitiv wahrgenommener Aussagen und Positionen stark ab (Tourangeau/Yan 2007: 863). Gut trainierte Interviewer können den Effekt, welchen sie durch ihre Anwesenheit erst verursachen, minimieren. Hierfür ist ein Training der Interviewer erforderlich, welches aus aus Kostengründen jedoch selten durchgeführt wird. Moderierte Formen des Interviews, wie persönliche oder telefonische Befragungen, haben daher hier einen starken Nachteil gegenüber den selbst-administrierten Formen wie Online-Interviews (Weisberg 2005: 45).

Sensitivität ruft eine Verzerrung hervor, da der systematische Messfehler eine zu gering gemessene Zustimmungen zu Aussagen („underreporting“) hervorrufen kann, wenn die eigentlich durch den befragten präferierte Antwort sozial sanktioniert wird. Beispielsweise wird der Anteil an Personen mit rassistischen Einstellungen unterschätzt, wenn die Frage zu offensichtlich gestellt wird (Tourangeau/Rips/Rasinski 2000: 269). Ebenso kann der Anteil einer Zustimmung als zu hoch ausgewiesen werden („overreporting“). Da die Teilnahme an Wahlen eine sozial erwünschte Norm ist, gibt insbesondere die Gruppe der Nichtwähler, welche sich dieser Norm bewusst sind, systematisch falsche Antworten. Als Folge wird der prozentuale Anteil der Wähler überschätzt (z.B. Holbrook/Krosnick 2010; Tourangeau/Rips/Rasinski 2000: 273 f.). Der Fokus der Literatur zur Analyse abweichender reportierter Wahlbeteiligungen von offiziellen Wahlbeteiligungsraten nutzt maßgeblich dieses Phänomen zur Erklärung. Hier zeigt sich jedoch ebenfalls eine Kombination unterschiedlicher Ursachen. Die zu hohen Raten in Umfragen sind ebenfalls auf einen Nonresponsefehler zurückzuführen, da Personen mit einem erhöhten Interesse an Politik häufiger wählen und gleichzeitig auch häufiger an Umfragen partizipieren (Sciarini/Goldberg 2016: 120 f.).

Eine weitere Möglichkeit systematischer Messfehler liegt in der Zustimmungstendenz („Akquieszenz“) von Befragten zu Aussagen (Watson 1992; Billiet/Davidov 2008; Rammstedt/Goldberg/Borg 2010). Dies kann durch die Rolle des Interviewers verursacht werden, weshalb hier eine Sensibilisierung notwendig ist (Olson/Bilgen 2011). Sofern Zustimmungs- und Ablehnungsfragen im

selben Ausmaß von dem Problem betroffen sind, lässt sich diesem Problem durch eine Balancierung der Skalen begegnen. Die eine Hälfte wird als Zustimmungsfragen, die andere Hälfte als Ablehnungsfragen formuliert. Insgesamt negiert sich daher der Effekt. Diese Annahme lässt sich aber häufig nicht halten und das Problem lässt sich dann nur komplexer modellieren (Billiet/Davidov 2008: 545 f.).

In allen Fällen des „satisficing“ führt dieses zu dem identischen Problem: Der durch die befragte Person für wahr gehaltene und mitgeteilte Wert stimmt nicht überein. Da dies systematisch in einer Beziehung zur extern bedingten sozialen Erwünschtheit oder des Versuchs aus anderen Gründen den eigenen Aufwand durch die Befragung systematisch zu minimieren steht, kommt es neben dem zufälligen Messfehler nun zu einem zusätzlichen systematischen Messfehler.

Der *Verarbeitungsfehler („Data Processing Error“)* bezieht sich schließlich auf Fehlerquellen während der Aufbereitung der Daten *vor* statistischen Analysen. Beispiele hierfür sind fehlerhafte Codierungen von Fragen, falsch angewandte Imputationstechniken oder fehlerhafte Behandlung von Ausreißerwerten (Faulbaum/Prüfer/Rexroth 2009: 52; Schnell 2013: 420-428). Sie sind damit keine existentielle Fehlerquelle während einer Erhebung, sondern auf Fehler während der Datenaufbereitung zurückzuführen. Wie häufig dieses Problem allgemein auftaucht ist relativ unklar. Im Falle der Bereitstellung von Daten über die GESIS werden diese auftauchenden Fehler transparent dargestellt und korrigiert,[9] solange sie bemerkt werden. Da beinahe jeder Datensatz über das Zentralarchiv eine überarbeitete Version enthält, sind Fehler während der Datenverarbeitung jedoch nicht ungewöhnlich.

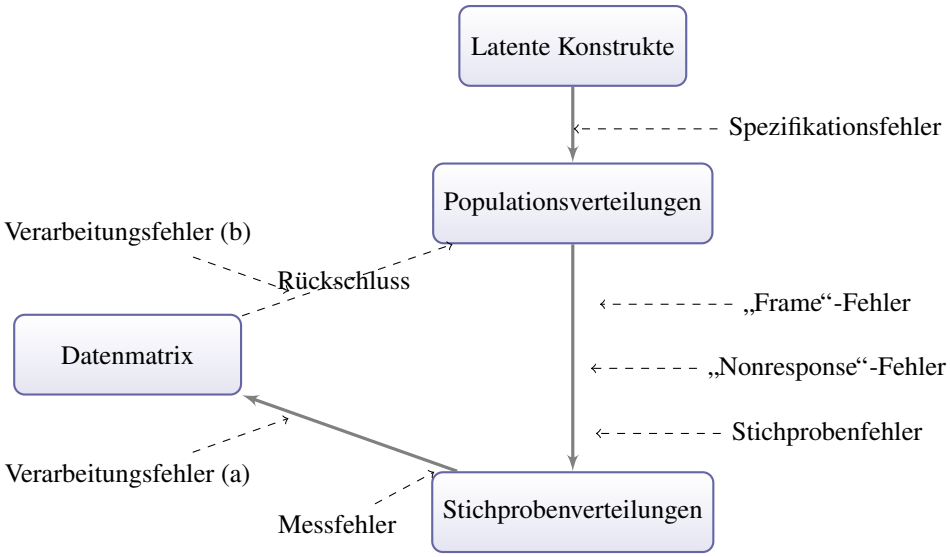
2.4. Zwischenfazit: Fehlerquellen in Umfragen

Das Ziel von Umfragen ist die systematische Erfassung einer Vielzahl von nicht direkt beobachtbaren Einstellungen und Werten von zu befragenden Personen aus einer größeren Grundgesamtheit. Bis zu einer abschließenden Analyse auf der Basis von Umfragedaten sind jedoch einige Hürden zu nehmen, um dies unverzerrt zu erreichen. Die Abbildung 2.5 stellt abschließend die möglichen diskutierten Fehlerquellen des vergangenen Kapitels dar.

Die latenten Konstrukte müssen durch eine theoriegeleitete Operationalisierung im Rahmen einer Konzeptspezifikation greifbar gemacht werden. Dann

[9] Beschreibung unter http://www.gesis.org/wahlen/gles/daten-und-dokumente/errata/. Alle verwendeten Datensätze werden von der GESIS verwaltet.

Abbildung 2.5.: Visualisierung „Total Survey Error“ (TSE)



Quelle: Eigene Darstellung.

können sich diese in Form einer Populationsverteilung manifestieren. Kommt es hierbei bereits zu einer fehlerhaften Umsetzung, liegt ein Spezifikationsfehler vor. Anschließend erfolgt die (zufallsbasierte) Auswahl des Personenkreises der Population, welche im Rahmen der Stichprobe befragt werden soll. Hier greifen bereits mehrere Fehlerquellen. Das Erhebungsdesign kann Probleme hinsichtlich der Repräsentativität erzeugen („Frame“-Fehler), indem es bestimmten Gruppen die Teilnahme an der Befragung erschwert oder unmöglich macht („Noncoverage“). Auch wenn dieses Problem überwunden ist, kann die selektive Verweigerung oder mangelnde Erreichbarkeit von eigentlich zur Befragung ausgewählten Personen eine weitere Fehlerquelle darstellen („Nonresponse--Fehler“). Als Folge dieser beiden Fehlerarten droht ein selektives Abbild der interessierenden Gruppen der Population, was sich in einem systematischen Unterschied zwischen den Verteilungen der beiden Ebenen widerspiegelt. Ebenfalls ist die Berücksichtigung eines nur kleinen Teils der Grundgesamtheit selber fehlerbelastet („Stichprobenfehler“), auch wenn diese Fehlerquelle bei einer zufallsbasierten Auswahl nur als Unsicherheitsfaktor mit berücksichtigt werden muss. Als Folge treten für unterschiedliche Stichproben auch unterschiedliche Realisierungen der Verteilung von interessierenden Variablen auf.

Selbst die Abwesenheit bewusster Verweigerung und die Möglichkeit der Kontaktierbarkeit aller ausgewählten Personen und deren adäquate theoriegeleitete Manifestierung der latenten Konstrukte garantiert noch nicht eine fehlerfreie Erfassung ebendieser Verteilungen. Personen können aus verschiedenen

Gründen ihre für sich korrekt erfassten wahren Werte bewusst oder unbewusst falsch angeben. Durch diesen Messfehler kommt ein erneuter Unsicherheitsfaktor der Prognose hinzu, welcher sich erst im Mittel ausgleicht. Dies gilt zudem nur, wenn er über alle Befragten hinweg zufällig verteilt auftritt. Tritt dieser Fehler der korrekten Erfassung der Daten zusätzlich systematisch auf, existiert auch hier eine weitere Fehlerquelle. Diese kann auftreten, da bestimmte Gruppen ein vom Rest der befragten Personen abweichendes Antwortverhalten aufweisen können. Sie sind zwar in der Lage, ihre wahren Werte korrekt durch die ihnen vorgetragene Operationalisierung eines Konstrukts zu identifzieren, weigern sich aber bewusst diesen Wert unverfälscht mitzuteilen („Sensitivität“). Ebenso können sie nicht in der Lage oder Willens sein, diese kognitiv zu erfassen und wählen ein von ihnen erwartetes Antwortverhalten. Ist schließlich auch diese Hürde genommen, muss noch sichergestellt werden, dass die Übertragung der Antworten auch korrekt durch den Interviewer oder die befragten Personen korrekt vorgenommen wird. In einem letzten Punkt müssen diese Daten dann noch in ein geeignetes Analyseformat übertragen werden („Verarbeitungsfehler“).

Sind diese Hürden genommen, ist ein Rückschluss auf die Populationsverteilungen und deren Parameter über die Schätzer der Stichprobe möglich, welcher lediglich von einem Stichprobenfehler behaftet ist. An diesem Punkt ist nun geklärt, welche Fehlerquellen während der Erhebung und Befragung einer Stichprobe, sowie deren Rückschluss auf die Grundgesamtheit entstehen können. Treten diese Fehlerquellen in einer gewissen Systematik auf, existiert offensichtlich eine qualitative Einschränkung der Ergebnisse auf Basis der Stichprobe. Das nachfolgende Kapitel 3 wird daher nun herausstellen, wie ausgehend von einer Diskussion und Einordnung des Begriffs der „Repräsentativität“ sich dieser Fehler erfassen lässt.

3. „Repräsentativität“? Erfassung der Qualität von Umfragen

Das vorherige Kapitel hat die Fehlerquellen in Umfragen aufbereitet und in Abbildung 2.5 dargestellt. Diese Fehler haben das Potential, eine Verzerrung der Ergebnisse statistischer Analysen auf der Basis von Umfragedaten hervorzurufen. Zu einer solchen kommt es, wenn für eine oder mehrere der dargestellten Fehlerquellen nicht nur eine rein zufällige, sondern eine zusätzliche systematische Komponente verantwortlich ist.

Existiert ein systematischer Fehler hinsichtlich einer interessierenden Variable, sind die auf diesen Werten realisierten Schätzungen nicht mehr valide. Der zu Grunde liegende Schätzer für den Populationsparameter wird nicht mehr unverzerrt erfasst. Zufällige Fehler hingegen berühren lediglich die Varianz eines Schätzers (Weisberg 2005: 22). Der begrifflich unscharfe Begriff der Repräsentativität (z.B. Diekmann 2007: 430 ff.) lässt jedoch eine hohe Banbreite an Definitionen zu,[1] und muss daher im Folgenden hinsichtlich seiner Bedeutung für die Arbeit präzisiert werden. Erst dann ist eine anschließende Definition des statistischen Konzepts einer Verzerrung sinnvoll.

Statistische Repräsentativität bezieht sich auf die Übereinstimmung der Randverteilungen auf allen als relevant angenommenen Variablen zwischen einer Stichprobe und der ihr zu Grunde liegenden Grundgesamtheit (Bethlehem/Cobben/Schouten 2011: 16 f.). Dies hat auch bereits die Abbildung 2.5 veranschaulicht. Unterschiede zwischen diesen beiden Ebenen der Verteilung ergeben sich, neben dem Stichprobenfehler, vor allem durch Nonresponse- und Noncoverage-Probleme. Repräsentativität als Begriff impliziert daher immer eine konditionale Aussage. Die Existenz dieser Eigenschaft muss für jede einzelne Variable ge-

[1] Zur Übersicht siehe (Kruskal/Mosteller 1979a,b,c). Gegenüber dem praktischen Problem der Umsetzung der zentralen Anforderung eines „Samplings“ im klassischen Sinne einer (einfachen) Zufallsstichprobe: „[...] difficulty is to ensure that every person or thing has the same chance of inclusion in the investigation.“(Bowley 1906: 553) Die erste grundlegende Formulierung eines „Probability Samplings“ als „Representative Method“ findet sich bei Neyman (1934). Weiter wegweisende Arbeiten sind Deming (1950) und Cochran (1977). als „[...] representative sampling is [...] the absence of selective forces in the sampling process.“ (Kruskal/Mosteller 1979c: 247)

zeigt werden, eine Umfrage kann nicht generell als repräsentativ angenommen werden (Schnell 2012: 173). Repräsentativität ist daher kein wissenschaftlicher Fachbegriff. Es geht vielmehr um die Frage der Möglichkeit der Realisierung einer Zufallsauswahl (Diekmann 2007: 430).

Wie in Abbildung 2.5 dargestellt, können systematische Verzerrungen einer Stichprobe durch unterschiedliche Fehlerquellen verursacht werden. Da insbesondere ein selektiver Zugang („Noncoverage“) oder eine selektive Teilnahme („Nonresponse“) Einschränkungen der Repräsentativität einer Stichprobe darstellen, steht das Potential dieser Fehlerquellen für eine Einschränkung der Qualität einer Umfrage im Fokus. Schließlich ist der Schluss von der Stichprobe auf die Grundgesamtheit nur zulässig ist, „[...] when nonsampling errors are small. [...]“ (Groves/Lyberg 2010: 850).

Insbesondere Nonresponse verursacht einen Verlust der Kontrolle über die Erhebung der Stichprobe. Nach der erfolgreichen Auswahl durch das Erhebungsdesign liegt die Entscheidung zur Teilnahme in der Hand der ausgewählten Personen. Nonresponse beeinflusst die Größe und Zusammensetzung einer Stichprobe und hat daher einen direkten Einfluss auf die Gültigkeit der Nutzung inferenzstatistischer Verfahren (Bethlehem/Cobben/Schouten 2011: 2). Die Auswirkungen des Nonresponse lassen sich aber bestimmen. Werden Personen zur Befragung ausgewählt und nehmen nicht teil, ist dies über die Dokumentation der Feldphase bekannt. Aus diesem Grund kann die Teilnahmerate als ein direkter Indikator für das Ausmaß von Nonresponse herangezogen werden. Kapitel 3.1 wird jedoch nachfolgend herausstellen, warum diese nur eine Aussage über das Ausmaß, jedoch nicht über die qualitative Einordnung des Problems zulässt.

Hierzu müssen weitere Kennzahlen genutzt werden. Das Kapitel 3.2 wird hierzu das statistische Konzept des Bias als eine besser geeignete Kennzahl der Umfragequalität beschreiben. Durch den „Mean Squared Error“ (MSE) lässt sich der „Total Survey Error“ (TSE) erfassen und in eine zufällige und systematische Komponente, den Bias, zerlegen. Ist der Fragebogen korrekt spezifiziert, die Messung fehlerfrei durchgeführt und die Daten korrekt verarbeitet, ist ein verbliebener systematischer Fehler dann durch die beiden Fehlerquellen Nonresponse und Noncoverage verursacht. Das Kapitel 3.3 wird daher aufzeigen, wie diese Fehlerquellen kombiniert erfasst werden können. Abschließend wird das Kapitel 3.4 die Erfassung der Verzerrung von Anteilswerten herausstellen.

3.1. Teilnahmeraten als Qualitätsmerkmal?

Durch die bereits länger anhaltende Diagnose sinkender Teilnahmeraten in Umfragen (z.B. Steeh 1981; Brehm 1994; Schnell 1997; Heer 1999; Steeh et al. 2001; Leeuw/Heer 2002; Stoop 2004; Curtin/Presser/Singer 2005; Holbrook/Krosnick/Pfent 2007; Brick/Williams 2012) steigt das Risiko einer Verzerrung der Ergebnisse durch Nonresponse. Teilnahmeraten in Umfragen und die Gefahr eines Fehlers durch Nonresponse konstant zu halten, wird hierdurch immer kostenintensiver (Chang/Krosnick 2009). Daher hat sich der Fokus der Umfrageforschung in den vergangenen Jahren sehr stark auf das Problem der sinkenden Teilnahmeraten und die Bedeutung des Erhebungsrahmens konzentriert (z.B. Couper 2011; Groves 2011).

Ab welchem Niveau eine Teilnahmerate jedoch bedenklich wird, lässt sich nicht generell festlegen, auch wenn manche Versuche in Bevölkerungsumfragen dahingehend gemacht werden. Beispielsweise definiert der „European Social Survey“ (ESS) in allen Runden der Erhebung ein Minimum von 70 % für die Teilnahmerate in den teilnehmenden Ländern (European Social Survey 2014: 9). Die tatsächlich realisierten Quoten unterscheiden sich jedoch sehr stark (Bethlehem/Cobben/Schouten 2011: 161). Zudem endet eine niedrige Teilnahmerate nicht kausal in einer Verzerrung der Ergebnisse (Peytchev 2012: 89). Es gilt vielmehr: „As the rate of nonresponse rises, [...] the potential for bias increases [...].“ (Massey/Tourangeau 2012: 227). Empirisch lässt sich daher auch nur ein sehr geringer Zusammenhang zwischen der Teilnahmerate und der Verzerrung einer Umfrage nachweisen (z.B. Curtin/Presser/Singer 2000; Keeter/Miller et al. 2000; Merkle/Edelman 2002; Groves 2006; Keeter/Kennedy/Dimock et al. 2006; Groves/Peytcheva 2008). Durch eine Fokussierung auf die Teilnahmerate als Qualitätsmerkmal einer Umfrage wird Nonresponse jedoch automatisch mit einer Verzerrung gleichgesetzt (Holbrook/Krosnick/Pfent 2007; Groves/Peytcheva 2008).

Ein weiteres Problem der Teilnahmerate als Indikator der Datenqualität einer Umfrage ist der sich daraus ergebende Anreiz. Eine Steigerung lässt sich am einfachsten durch eine Fokussierung auf die „soften“ Verweigerer erreichen. Der Aufwand der Überzeugung zur Umfrageteilnahme ist bei diesen nicht sehr hoch, da eine relativ hohe Teilnahmewahrscheinlichkeit vorliegt. Ist der qualitative Unterschied zwischen diesen ursprünglichen Verweigerern und den bereits befragten Personen jedoch gering, wird sich auch die Verzerrung trotz einer gestiegenen Teilnahmerate nicht reduzieren (Stoop 2005; Beullens/Loosveldt 2012). Für diese ist vielmehr eine Gruppe „hartnäckiger“ Verweigerer mit einer

nur geringen Teilnahmewahrscheinlichkeit ursächlich und hohe zeitliche und monetäre Kosten sind notwendig um diesen Personenkreis zu einer Teilnahme zu bewegen. Daher zeigen „Follow-up“-Studien von Nonrespondenten auch nur eine ernüchternde Steigerung der Datenqualität durch die nachträgliche Konvertierung von Nonrespondenten in Respondenten (z.B. an Hand des Allbus: Proner 2011). Eine Erhöhung der Teilnahmerate ist also nur ein Teil der Lösung des Problems der gestiegenen Fehleranfälligkeit von Umfragen (Stoop et al. 2010).

In der Praxis ist die Nutzung der Teilnahmerate als zentrales Qualitätsmerkmal jedoch weitverbreitet (z.B. Koch 1998; Merkle/Edelman 2002; Groves 2006) und steht damit in einem „[...] bemerkenswertem Gegensatz zur wissenschaftlichen Literatur.“ (Schnell 2012: 159). Speziell bei niedrigen Teilnahmeraten sollten daher weitere Indikatoren angegeben werden um die Qualität von Umfragen zu erfassen, wie es beispielsweise auch die „American Association for Public Opinion Research“ (AAPOR) empfiehlt.[2]

Zudem ist die Teilnahmerate an Umfragen durch sich unterscheidende Definition international häufig nicht vergleichbar (z.B. Schnell 1997; Groves/Presser/Dipko 2004). Die AAPOR empfiehlt deren Bestimmung als „[...] the number of complete interviews with reporting units divided by the number of eligible reporting units in the sample“ (AAPOR 2008: 34) und grenzt sie hierdurch von der Kooperationsrate ab als „[...] the proportion of all cases interviewed of all eligible units ever contacted.“ (AAPOR 2008: 36). Für die Bundesrepublik existiert keine vergleichbare Definition zur Einordnung von ausfallenden Personen und zur Bestimmung von Ausschöpfungsquoten (Fuchs 2010: 235 f.). Hier ist es stattdessen üblich, eine Brutto- und Nettostichprobengröße zu berechnen. Die Differenz zwischen beiden sind diejenigen Personen, welche nicht zur anvisierten Grundgesamtheit gehören (Schnell 1997: 27). Anschließend kann die Teilnahmerate dann als der prozentuale Anteil der durchgeführten Interviews der Nettostichprobe berechnet werden (Schnell 1997: 17).

Die fortlaufende Analyse des Anteils *und* der Qualität der Teilnahme versuchen während der Erhebungsphase „R(epresentativity)-Indikatoren“ zu erfassen (Bethlehem/Cobben/Schouten 2009; Wagner 2010; Schouten/Shlomo/Skinner 2011; Schouten/Bethlehem et al. 2012; Shlomo/Skinner/Schouten 2012).[3] Durch diese wird die Qualität der Teilnahme bereits während der Erhebungsphase fortlaufend gegenüber Referenzverteilungen, beispielsweise der amtlichen Statistik, evaluiert. Hierdurch können früh mögliche Probleme der Repräsentativität des

[2] Siehe hierzu die Darstellung unter http://www.aapor.org/Education-Resources/For-Researchers/Poll-Survey-FAQ/Response-Rates-An-Overview.aspx (Stand: 11. Mai 2018).

[3] Die Umsetzung erfolgte bisher in SAS und R durch Heij/Schouten/Shlomo (2014).

Rücklaufs auffallen. Diese Indikatoren sind „[...] a lack-of-association measure. The weaker the association the better, as this implies there is no evidence that non-response has affected the composition of the observed data.“ (Bethlehem/Cobben/Schouten 2009: 101) und stellen eine Alternative zur Nutzung der reinen Teilnahmerate dar. Repräsentativität des Rücklaufs ist hierbei konditional auf die verwendeten Variablen definiert, wie grundlegende soziodemografische Rahmendaten. Dies macht ihre Berechnung jedoch auch von dieser Auswahl abhängig. Werden hierzu nur leicht erreichbare Ziele gewählt, laden R-Indikatoren dann zu einer missbräuchlichen Nutzung ein. Modelliert werden sollte daher vor allem die Kombination an Variablen, welche den unangenehmsten Ausfallprozess darstellen (Schnell 2012: 174).

Diese Indikatoren zeigen aber bereits: Nicht nur die Quantität des Ausfalls ist entscheidend, sondern vor allem auch die Qualität. Teilnahmeraten eignen sich nur bedingt zur Diagnose dieser, sie berücksichtigen lediglich das Ausmaß der Teilnahme. Zudem eignet sich die Teilnahmerate nur zur Diagnose von Nonresponse und nicht Noncoverage. Um die Qualität einer Umfrage zu bestimmen, muss die Betrachtung also über die reine Teilnahmerate hinausgehen. Das nachfolgende Kapitel 3.3 wird nun konkretisieren, wann Nonresponse und Noncoverage einen Bias induzieren und damit zu einem qualitativen, und nicht nur quantitativen, Problem werden.

3.2. Das statistische Konzept eines „Bias“

Um den zufälligen und systematischen Fehler des Schätzers eines Merkmals einer Umfrage zu erfassen, eignet sich als Kennzahl der MSE. Er erfasst die Varianz und den Bias von diesem und verbindet somit die beiden Konzepte der Reliabilität und Validität (Weisberg 2005: 23). Die Varianz eines Schätzers ist der zufallsbedingte Teil und auf den Stichprobenfehler durch die Stichprobenziehung zurückzuführen (Kapitel 2.1). Der Bias ist hingegen der systematische Fehler, welcher einen Schätzer in eine bestimmte Richtung verzerrt. Ist ein Schätzer mit einem Bias belegt, zeigt sich eine signifikante Abweichung vom anvisierten Populationswert (Smith 2011: 466). Durch den MSE lässt sich somit der zufallsbedingte Teil der Varianz des Schätzers von der nicht-zufallsbedingten Komponente durch die Verzerrung differenzieren. Diese beiden Fehler ergeben den Gesamtfehler einer Schätzung, den „Total Survey Error“. Für einen Schätzer t und seinen Parameter θ ist der Bias $B(t)$ die Differenz zwischen dem Erwartungswert des Schätzers und dem Parameter der Population (nachfolgende

Notation nach Bethlehem 2009: 36 f.):

$$B(t) = E(t) - \theta \tag{3.1}$$

Der MSE kombiniert dies als den Erwartungswert der quadrierten Differenz zwischen Schätzer und Populationsparameter, welches sich in die beiden Komponenten der Varianz und der (quadrierten) Verzerrung eines Schätzers unterteilen lässt:[4]

$$MSE(t) = E(t-\theta)^2 = V(t) + B^2(t) \tag{3.2}$$

Die korrekte Berechnung der Unsicherheit der Prognose auf Basis der Stichprobe durch den Standardfehler implizieren also einen geringen Bias in Relation zur Varianz des Schätzers (Särndal/Swensson/Wretman 1992: 164).[5] Das Ziel des MSE ist hierdurch den gesamten Fehler einer Umfrage („Total Error“) zu erfassen (Kish 1965: 510)

In einem einfachen Fall handelt es sich um ein arithmetisches Mittel einer Stichprobe ($\bar{y}$), welches eine mögliche Abweichung von seinem wahren Wert in der anvisierten Grundgesamtheit ($\bar{Y}$) aufweist:

$$MSE(\bar{y}) = \underbrace{Var(\bar{y})}_{Varianz} + \underbrace{[E(\bar{y}) - \bar{Y}]^2}_{Bias} \tag{3.3}$$

Ist ein Schätzer unverzerrt, reduziert sich mit einer zunehmenden Stichprobengröße die Varianz eines Schätzers und der MSE konvergiert gegen Null. Der Erwartungswert des arithmetischen Mittels entspricht in diesem Fall dem Populationsparameter ($E(\bar{y}) = \bar{Y}$). Existiert jedoch ein Bias, zeigt sich ein systematischer Unterschied zwischen dem Erwartungswert und dem Populationsparameter und der MSE konvergiert auch approximativ nicht gegen Null (Kish 1965: 12 f.). Es lässt sich dann ein systematischer Unterschied zwischen der Kennzahl einer Stichprobe und dem Parameter der Grundgesamtheit, beispielsweise verfügbar über die amtliche Statistik, feststellen.

Dies lässt sich auch unter den Begriffen der Präzision („precision“) und Fehlerfreiheit („accuracy“) erfassen. Ein Schätzer ist präzise, wenn die Streuung der arithmetischen Mittel um ihren Erwartungswert möglichst gering ist. Stimmt

[4] Für andere Notationen und Herleitungen siehe z.B. Czado/Schmidt 2011: 104.

[5] Da der MSE als Varianz des Schätzers nicht in der ursprünglichen Skala vorliegt, wird zur Interpretation in der Regel der „Root Mean Squared Error“ (RMSE) als Standardabweichung des Fehlers durch die Wurzel des MSE herangezogen (z.B. Cobben 2009: 14; Kohler/Kreuter 2012: 200). Um den MSE weiterhin über mehrere Variablen vergleichen zu können, müssen diese zusätzlich identisch skaliert oder standardisiert sein (Chatfield 1988).

zusätzlich das arithmetische Mittel dieser Stichprobenverteilung mit dem wahren Wert überein, ist der Schätzer auch akkurat. Die Varianz des Schätzers ergibt sich nur aus dem Stichprobenfehler, ein systematischer Fehler liegt nicht vor (Kish 1965: 510). Der MSE kombiniert also die beiden Komponenten der Reliabilität und der Validität einer Schätzung. Sinkt die Sicherheit eines Schätzers, wird er unpräziser und die Reliabilität sinkt. Die Streuung der realisierten Schätzungen auf Basis von Stichproben variiert dann stärker um den wahren Wert. Steigt hingegen der systematische Fehler, nimmt die Validität ab. Die Abweichung zwischen dem Erwartungswert des Schätzers und dem wahren Wert der anvisierten Grundgesamtheit nimmt als Folge zu (Bethlehem 2009: 178).

Den beiden Komponenten der Varianz (*Var*) und des Bias (*B*) – und damit der Reliabilität und Validität – lassen sich die jeweiligen Fehlerquellen des „Total Survey Error“ (TSE) zuordnen (z.B. Schnell 2012: 387):

$$\begin{aligned} MSE &= \underbrace{(B_{\text{Spezifikation}} + B_{\text{Nonresponse}} + B_{\text{Coverage}} + B_{\text{Messung}} + B_{\text{Verarbeitung}})^2}_{\text{Validität}} \\ &\quad + \underbrace{Var_{\text{Stichprobe}} + Var_{\text{Messung}} + Var_{\text{Verarbeitung}}}_{\text{Reliabilität}} \end{aligned}$$

In der Realität müssen zur Bestimmung des MSE jedoch die Parameter der Grundgesamtheit bekannt sein. Daher ist dieser meist nicht quantifizierbar (Czado/Schmidt 2011: 104 f.). Der MSE dient daher vor allem dazu die additive Verbindung von Reliabilität und Validität aufzuzeigen (Weisberg 2005: 23).

Lässt sich der MSE jedoch bestimmen und eine eine systematische Abweichung zwischen der Schätzung der Stichprobe und dem Parameter der Grundgesamtheit existiert, dann ist dieser mit einem Bias belegt. Hierdurch ergibt sich der Rückbezug zum vorhergegangenen Kapitel. Ist die Fragebogenkonstruktion, die Messung und die Datenverarbeitung fehlerfrei durchgeführt worden, existiert kein Spezifikations-, Mess- und Verarbeitungsfehler. Eine verbliebene Abweichung zwischen einer Kennzahl oder einem Anteilswert der Stichprobe und dem jeweiligen Populationswert ist dann auf eine zufallsbedingte Abweichung durch die Stichprobenvarianz und auf den Nonresponse- und (Non)Coverage-Probleme zurückzuführen:

$$MSE = \underbrace{(B_{\text{Nonresponse}} + B_{\text{Coverage}})^2}_{\text{Validität}} + \underbrace{Var_{\text{Stichprobe}}}_{\text{Reliabilität}}$$

Die beiden letzteren definieren in ihrer Kombination als die beiden Fehlerquellen des Nicht-Beobachtungsfehlers den Bias. Die Abweichung von Schätzung und

Parameter der Grundgesamtheit ergibt sich, da lediglich ein selektiver Teil der anvisierten Grundgesamtheit befragt wurde.

Durch diese möglichen verzerrenden Wirkungen auf einen Schätzer hat sich insbesondere das Thema Nonresponse seit langer Zeit eine hohe Bedeutung im wissenschaftlichen Diskurs zur Sicherung der Qualität von Umfragen gesichert. In vielbeachteten Arbeiten (z.B. Groves/Couper 1998; Groves/Dillman et al. 2002), sich speziell diesem Thema widmenden Lehrbüchern (z.B. Bethlehem/ Cobben/Schouten 2011) und Spezialausgaben von Fachzeitschriften (z.B. Singer 2006), findet das Thema Nonresponse eine zentrale wissenschaftliche Beachtung im Bereich der Umfrageforschung. Seit 1990 wird zudem der durch Groves, Lyberg und Barnes ins Leben gerufene „International Workshop on Household Survey Nonresponse“ im jährlichen Turnus abgehalten.[6] Aus diesem Grund nimmt nachfolgend die Darstellung der statistischen Erfassung des Nonresponse- und Noncoverage-Fehlers eine zentrale Stellung ein. Spezifikations-, Mess- und Verarbeitungsfehler werden zur Erfassung des Bias daher konzeptionell als abwesend angesehen. Ob diese Fehlerquellen ebenfalls drohen, muss vielmehr aus dem Kontext einer Umfrage für jede Variable spezifische und an Hand theoretischer Überlegungen hergeleitet werden. Das Problem des Nonresponse lässt sich zudem für Umfragen sehr gut abschätzen, da es sich während der Erhebung offenbart. Das nachfolgende Kapitel wird nun aufzeigen, wie sich der Bias durch Nonresponse und Noncoverage auswirkt.

3.3. Nonresponse- und Noncoverage-Bias

Ein von Nonresponse (NR) und/oder Noncoverage (NC) berührter Schätzer ($\hat{Y}_{NR+NC}$) weicht von seinem Parameter Y ab. Es existiert neben dem Stichprobenfehler der Schätzung von Y ebenfalls ein systematischer Fehler durch diese beiden Fehlerquellen (Särndal/Lundström 2005: 45):

$$\hat{Y}_{(NR+NC)} - Y = \underbrace{(\hat{Y} - Y)}_{\text{Stichproben-Fehler}} + \underbrace{(\hat{Y}_{(NR+NC)} - \hat{Y})}_{\text{Nonresponse/Noncoverage-Fehler}} \tag{3.4}$$

Können diese beiden – und alle weiteren – Fehlerquellen hingegen ausgeschlossen werden, handelt es sich um einen unverzerrten und präzisen Schätzer, sofern er von allen möglichen unverzerrten Schätzern die minimalste Varianz aufweist (Bethlehem 2009: 178 f.).

[6] Zur Beschreibung des Workshops siehe die Homepage http://www.nonresponse.org/.

Damit Nonresponse eine Auswirkung auf die Umfrageergebnisse hat, müssen zwei Gründe vorliegen: Der Anteil der nicht teilnehmenden Personen muss prozentual groß sein, also eine verleichsweise niedrige Teilnahmerate vorliegen. Ebenfalls muss ein systematischer Unterschied zwischen diesen Nicht-Teilnehmern (NR) und den Teilnehmern (R) einer Umfrage existieren (z.B. Groves 2006: 220; Bethlehem/Cobben/Schouten 2009: 648):

$$B(\bar{y}_R) = \underbrace{\frac{N_{NR}}{N}}_{Q} \underbrace{(\bar{Y}_R - \bar{Y}_{NR})}_{K} = QK \tag{3.5}$$

Die Verzerrung ist umso stärker, je größer der Anteil der Nonrespondenten (Q) ist und je stärker sich ein Kontrast K der Schätzung zwischen diesen beiden Gruppen zeigt (Bethlehem/Cobben/Schouten 2011: 42). Existieren keine Nonrespondenten *oder* unterscheiden sich diese nicht systematisch von der Gruppe der Respondenten, dann existiert folglich auch kein Bias (Lohr 2010: 529). Handelt es sich um Noncoverage, ändert sich lediglich die Bezeichnung der beiden Gruppen (Biemer 2010: 840). Nonresponse- und Noncoverage-Fehler haben unterschiedliche Ursachen und Gründe (Kapitel 2.2.1 und 2.2.2). Die Wirkungsweisen ihres Einflusses auf die Ergebnisse der Schätzung von Parametern auf Basis einer Stichprobe sind jedoch identisch. Daher lassen sich auch beide Probleme additiv in einen gesamten Bias der jeweiligen Variable Y kombinieren:

$$B(\bar{y}_{R+C}) = \underbrace{\frac{N_{NR}}{N}(\bar{Y}_R - \bar{Y}_{NR})}_{\text{Nonresponse-Fehler}} + \underbrace{\frac{N_{NC}}{N}(\bar{Y}_C - \bar{Y}_{NC})}_{\text{Noncoverage-Fehler}} \tag{3.6}$$

Die Abweichung zwischen den Ergebnissen einer Stichprobe und den jeweiligen Kennzahlen der Grundgesamtheit sind bei einer fehlerfreien Beobachtung dann, neben dem Stichprobenfehler, auf den kombinierten Bias durch Nonresponse und Noncoverage zurückführbar.

Die bisherige Erfassung des absoluten (Nonresponse)-Bias ist jedoch abhängig von der zu Grunde liegenden Skalierung der Variable und lässt sich daher nicht für das Alter in Jahren und das Einkommen in Euro miteinander vergleichen. Identische Abstände implizieren ein unterschiedliches inhaltliches Ausmaß der Verzerrung. Neben dem „Absoluter Bias“ (AB) existiert daher noch eine weitere Kennzahl: Der „Relative Bias“ (RB). Dieser lässt eine direkte prozentuale Aussage der Verzerrung zu, da er den AB in Bezug zur Skalierung der zu Grunde liegenden Kennzahl setzt. Der Bezug ist hierbei das arithmetische Mittel der Respondenten (R) und Nonrespondenten (NR). Dieser wird ebenfalls

umso gravierender, je größer der relative Anteil der Nonrespondenten $(1-t_R)$ wird (Biemer/Lyberg 2003: 83):[7]

$$RB_{NR+NC} = (1-t_R)\frac{\bar{Y}_R - \bar{Y}_{NR}}{\bar{Y}_{R+NR}} + (1-t_C)\frac{\bar{Y}_C - \bar{Y}_{NC}}{\bar{Y}_{C+NC}} \tag{3.7}$$

Die Verzerrung wird angegeben als prozentualer Anteil des Parameters. Positive Werte bedeuten eine Überschätzung, negative Werte eine Unterschätzung des zu Grunde liegenden Parameters (Biemer/Lyberg 2003: 69 f.). Damit normiert der RB (Gleichung 3.7) den AB (Gleichung 3.5) am Parameter der Grundgesamtheit/Population $\bar{Y}_P$:

$$RB_{NR} = \frac{B_{\bar{y}_R}}{\bar{Y}_P} \tag{3.8}$$

Da zwischen dem Parameter und einer von Nonresponse und Noncoverage nicht betroffenen Stichprobe lediglich ein zufallsbedingter marginaler Unterschied besteht, lässt sich zudem auch das arithmetische Mittel einer unverzerrten Stichprobe nutzen. Zu einem identischen Ergebnis für den „Relative Nonresponse-Bias“ (RNB) kommt daher auch die weitverbreitete Bestimmung über den Unterschied zwischen der Kennzahl der Respondenten und der vollständig vorliegenden – also unverzerrten – Stichprobe n. Liegt diese Information der gesamten hypothetischen Stichprobe über externe Quellen vor, sind keine spezifischen Informationen über die Nonrespondenten nötig (z.B. Groves 2006: 658):

$$RB_{NR} = \frac{B_{\bar{y}_R}}{\bar{Y}_P} = \frac{B_{\bar{y}_R}}{\bar{Y}_{R+NR}} = \frac{(\bar{y}_r - \bar{y}_n)}{\bar{y}_n} \tag{3.9}$$

Wird der RB mit 100 multipliziert, gibt er an, um wie viel Prozent der wahre Wert durch die verzerrte Stichprobe unter- oder überschätzt wird. Die Richtung der Verzerrung wird dabei durch das Vorzeichen ausgedrückt (Biemer/Lyberg 2003: 71).

Für den Nonresponse- und Noncoverage-Bias existiert keine direkte Rechtfertigung statistischer Testverfahren. Ob eine Verzerrung relevant ist oder nicht, „[...] depends on the objectives of the survey and how data will be used. If the relative bias [...] drastically changes decisions or conclusions that analysts make from data, [it] is an important bias.“(Biemer/Lyberg 2003: 71). Zur Umsetzung der Messung der Verzerrung sind die Informationen über die Nonrespondenten

[7] Bethlehem (2009: 224) geben eine alternative Bestimmung des Konzepts eines relativen Bias durch die Division durch den Standardfehler des zugehörigen Konfidenzintervals des arithmetischen Mittels der Variable: $|B(\bar{y}_R)/S(\bar{y}_R)|$.

oder die Kennzahlen der unverzerrten Stichprobe notwendig.

Die bisherige Definition bezieht sich auf eine deterministische Sichtweise von Nonresponse. Die Teilnahme an einer Umfrage ist hier eine Prädisposition, welche anschließend beobachtet wird. Die Erfassung des Bias basiert daher auch nur auf den nach der Erhebung beobachteten Kennzahlen. Die Teilnahme an einer Umfrage kann jedoch auch einer gewissen Wahrscheinlichkeit unterliegen und ist damit kein vorab gegebenes Ereignis (z.B. Bethlehem 1988: 253; Bethlehem/Cobben/Schouten 2011: 40-45).[8]

Ob eine Variable unverzerrt erfasst wurde, hängt daher auch davon ab, welche Wahrscheinlichkeit der Teilnahme im arithmetischen Mittel über alle ausgewählten Personen existiert ($\bar{\rho}$) und wie stark sich diese Wahrscheinlichkeit über alle Personen hinweg gemäß ihrer Standardabweichung (S_ρ) und den Ausprägungen der betrachteten Variable Y (S_Y) unterscheiden. Zusätzlich ist relevant, wie stark der Zusammenhang zwischen der Wahrscheinlichkeit der Teilnahme an der Umfrage und der betrachteten Variable ist ($R_{\rho Y}$):

$$B(\bar{y}_R) = \frac{R_{\rho Y} S_\rho S_Y}{\bar{\rho}} \tag{3.10}$$

Dass ein Zusammenhang zwischen der Teilnahme und einer inhaltlichen Variable besteht, ist also eine Voraussetzung für eine Verzerrung. Existieren keine Unterschiede der Teilnahmewahrscheinlichkeiten, existiert auch kein Bias. Die Folge ist lediglich eine reduzierte Stichprobengröße. Bei gegebenen restlichen Werten sorgt weiterhin eine geringere durchschnittliche Teilnahmewahrscheinlichkeit für eine Verstärkung des Bias (Bethlehem/Cobben/Schouten 2011: 44).

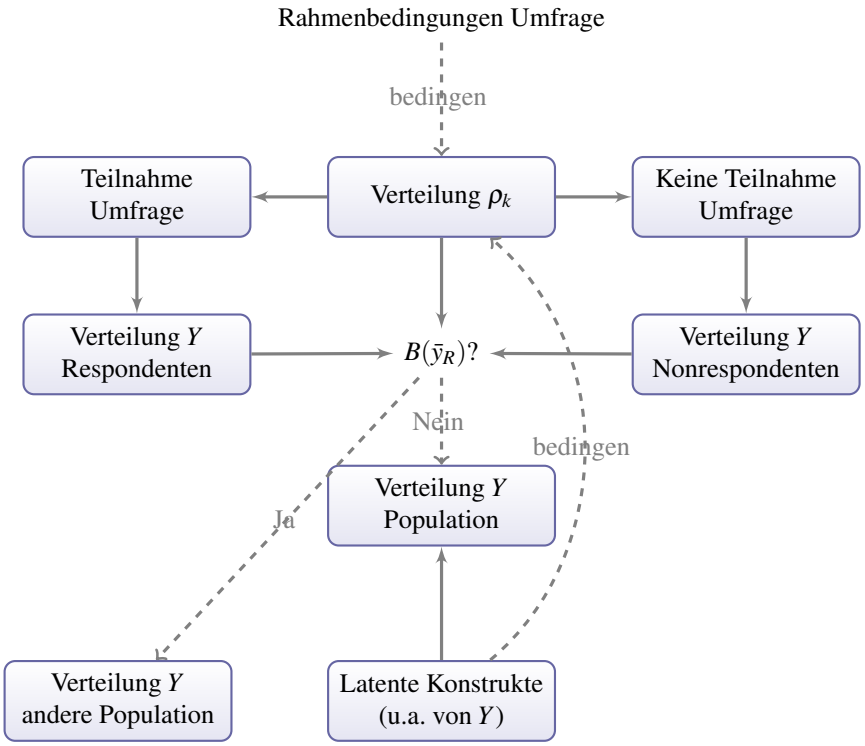
Der Bias ergibt sich nun daraus, dass steigende Nonresponse-Raten durch fallende durchschnittliche Teilnahmewahrscheinlichkeiten ein höheres Risiko für einen Bias implizieren (Kapitel 3.1) und gleichzeitig der Ausfallprozess über $R_{\rho Y}$ auch in einer Verbindung zur betrachteten Variable stehen muss. Dies ist eine analytische Erweiterung der Perspektive. Nicht nur der rein qualitative Unterschied der Respondenten zu den Nonrespondenten steht im Fokus (Gleichung 3.5). Vielmehr wird über die Erfassung der Teilnahmewahrscheinlichkeit der zu befragenden Personen die wahre Ursache dieser Verzerrung erfasst: Perso-

[8] Die nicht vorhandene Abdeckung einer Person durch das Erhebungsdesign lässt sich hingegen nur als ein dichotomes Ereignis erfassen. Entweder sie ist Teil der Auswahlgesamtheit oder nicht. Die Bestimmung einer individuellen Wahrscheinlichkeit durch das Erhebungsdesign abgedeckt zu sein, ergibt keinen Sinn. Wahrscheinlichkeiten des Noncoverage lassen sich vielmehr nur für Gruppen von Personen bestimmen, wenn bestimmte Gruppen beispielsweise in einem höheren Ausmaß von dem Problem betroffen sind. Daher bezieht sich dieser Teil lediglich auf den Nonresponse-Fehler.

nen(gruppen) haben eine unterschiedliche Wahrscheinlichkeit der Partizipation an der Umfrage, welche ebenfalls nicht unabhängig von bestimmten Kernmerkmalen ist. Als Folge kommt es zu einer Verzerrung interessierender Variablen, da durch diese niedrige Wahrscheinlichkeit der Teilnahme bestimmte Gruppen unterrepräsentiert sind und diese zudem ein abweichendes Antwortverhalten aufweisen.

Abbildung 3.1 stellt diesen Prozess der Modellierung in Umfragen noch einmal schematisch dar.

Abbildung 3.1.: Visualisierung Nonresponse-Bias



Quelle: Eigene Darstellung.

Damit eine Umfrage hinsichtlich einer Variable Y unverzerrt ist, sollte die Teilnahmewahrscheinlichkeit der ausgewählten Personen nicht in Verbindung zu den Rahmenbedingungen der Umfrage oder den Ausprägungen inhaltlicher Variablen der Erhebung stehen. Zweiteres erzeugt einen Bias, wenn es die Variable Y selber ist, welche auch die Teilnahmewahrscheinlichkeit beeinflusst. Dies gilt jedochauch, wenn andere Konstrukte diese Variation der Teilnahmewahrscheinlichkeit verursachen *und* diese Konstrukte ebenfalls die Ausprägungen von Y beeinflussen. Auch in diesem Fall ist dann $S_\rho > 0$ und $R_{\rho Y} \neq 0$ und als *Folge* äußern sich Unterschiede der Verteilung auf Y zwischen Respondenten und Nonrespondenten.

Um den Nonresponse-Bias adäquat zu bestimmen, ist die Ermittlung von ρ essentiell. Hierdurch bedingt sich die Teilnahme an der Umfrage, welche sowohl durch die Rahmenbedingungen der Erhebung als auch durch latente Konstrukte im Hintergrund gesteuert werden kann. Ist dies der Fall, ist die Gültigkeit des Inferenzschluss auf Basis der Respondenten nicht mehr valide und zielt dann fälschlicherweise auf eine andere Population ab als angedacht, da bestimmte Gruppen über- und andere unterrepräsentiert sind. Hierbei entsteht jedoch eine enorme Barriere: Um die Wahrscheinlichkeit der Teilnahme an einer Umfrage zu schätzen, müssen Daten über Personen verfügbar sein, welche *nicht* an der Umfrage teilgenommen haben. Das Kapitel 4 wird darstellen, über welche Verfahren und externe Quellen jedoch eine solche Verzerrung durch die Schätzung von „Propensity Scores“ (PS) diagnostiziert werden kann, ohne Informationen über die von Nonresponse und Noncoverage betroffenen Personen erlangen zu müssen.

3.4. Anteilswerte und ihre Verzerrung

Umfragen weisen ebenfalls kategoriale Variablen auf. Für die relativen Anteile der Kategorien lassen sich ebenfalls die Abweichungen von einer Referenzverteilung bestimmen. Nachfolgend werden zwei Varianten skizziert, welche dies leisten können. Der erste Fall ist schlicht die Übertragung des AB und RB auf Anteilswerte. Die zweite Möglichkeit sind Kennziffern, welche speziell für die Erfassung der korrekten Prognose von Wahlumfragen (z.B. Martin/Traugott/Kennedy 2005; Arzheimer/Evans 2014) konzipiert wurden. Hierdurch lassen sich prozentuale Anteile, wie beispielsweise Alterskategorien und Zweitstimmenanteile, in Bezug zu den Anteilen einer Referenzumfrage oder der amtlichen Statistik bringen. Auch ist es nicht zwingend notwendig, die Anteilswerte der Nonrespondenten als Werte vorliegen zu haben, wenn statt der Nonrespondenten diejenigen der Population als Referenz vorliegen.

In Übereinstimmung mit der Definition des RB in Gleichung 3.9, kann dieser auch für Anteilswerte einer bestimmten Gruppe innerhalb der Stichprobe (pr) und der Population (p) bestimmt werden (Busse/Laub/Fuchs 2015: 338):

$$RB_p = 100 \cdot \frac{(pr - p)}{p} \tag{3.11}$$

Hierdurch wird das Ausmaß der Verzerrung prozentualer Anteile zwischen der von einer selektiven Teilnahme betroffenen Stichprobe (pr) und der

vollständigen Population (p) in Bezug zur Bedeutung dieser jeweiligen Kategorien innerhalb der Population gesetzt. Absolute und relative Abweichungen zwischen zwei Gruppen sind dann wiederum auf Nonresponse und Noncoverage zurückführbar, unter der Annahme der fehlerfreien Operationalisierung und Messung der latenten Konstrukte und der Verarbeitung dieser Daten.

Die Gleichung 3.11 bezieht sich nur auf die Verzerrung einer Gruppe (z.B. einer Altersgruppe). Soll für die Anteilswerte verschiedener Gruppen einer Variablen (z.B. die definierten Alterskategorien) die Verzerrung insgesamt erfasst werden, lässt sich einfach über die k verschiedenen Kategorien das arithmetische Mittel bilden. Für den AB einer kategorialen Variable ist dies:[9]

$$\overline{|mANB_p|} = \frac{\sum_{i=1}^{k} 100 \cdot (|pr_k - p_k|)}{k} \tag{3.12}$$

und analog für den RB:

$$\overline{|mRNB_p|} = \frac{\sum_{i=1}^{k} 100 \cdot \left(\left| \frac{pr_k - p_k}{p_k} \right| \right)}{k} \tag{3.13}$$

Hierdurch ergibt sich auch für kategoriale Variablen die Möglichkeit einer Verzerrungskennziffer, welche zu den Gleichungen 3.6 und 3.9 vergleichbar ist.

Die Analyse der Genauigkeit geschätzter Anteilswerte in Referenz zu denen einer Population ist weiterhin auch ein Kerngebiet der Wahlforschung. Hier liegt die zentrale Herausforderung der Messung der Qualität der Erfassung von Anteilswerten verschiedener Parteien im Rahmen einer Umfrage. Um die Verzerrung von Wahlumfragen zu erfassen, haben Mosteller et al. (1949) einige Empfehlungen formuliert, welche durch Mitofsky (1998: 241) hinsichtlich zweier spezifischer Empfehlungen zur Nutzung empfohlen wurden. Hierbei handelt es sich erstens um das arithmetische Mittel der absoluten Abweichungen der in der Stichprobe vorliegenden Anteile und den tatsächlichen bei der Wahl erreichten Anteilen über alle Kandidaten/Parteien hinweg. Die zweite Empfehlung ist die Verwendung der Differenz des Abstands der beiden ersten Kandidaten/Parteien innerhalb der Stichprobe und dieser bei der tatsächlichen Wahl (Mitofsky 1998: 233). Diese Verwendung findet sich beispielsweise auch bei Traugott (2001: 394-397) als „Average Absolute Candidate Deviation“ und „Absolute Difference in the Differences“ (Traugott 2001: 396). Die erste auf Mosteller et al. (1949) zurückgehende Empfehlung ist identisch zu der Formulierung in Gleichung 3.12 aus dem Bereich der Surveyforschung. Diese Berechnung ist unabhängig von

[9] Die Darstellung basiert auf eigenen Notationen.

den Ergebnissen der jeweils anderen Kandidaten oder Parteien bei der betrachteten Wahl. Die zweite Empfehlung hingegen enthält eine kompetetive Betrachtung der ersten beiden Kandidaten.

Martin/Traugott/Kennedy (2005) berücksichtigen dies durch ihre Kennzahl A. Hierdurch wird einbezogen, dass absolute Abweichungen lediglich stichprobenbedingt aufreten können. Ebenso hat der Anteil der Unentschlossenen Personen und mehr als zwei antretende Parteien/Kandidaten einen Einfluss. Je nach Anteil einer Partei hat daher eine gegebene absolute Abweichung eine andere Bedeutung hinsichtlich der Bedeutung einer Verzerrung (Martin/Traugott/Kennedy 2005: 349 f.). Um dies zu erreichen betrachten sie anstatt der absoluten Abweichungen von einem amtlichen Ergebniss das Odds-Ratios (OR) zweier Parteien R und D[10] einer beliebigen Wahl j in einem Land k, sowie die Konstruktion des zu Grunde liegenden Konfidenzintervalls als Indikator einer Verzerrung über die Bestimmung der Varianz von A (Martin/Traugott/Kennedy 2005: 352):

$$A_{ijk} = log[(r_{ijk}/d_{ijk})/(R_{jk}/D_{jk})] \tag{3.14}$$

$$\text{Var}(A_{ijk}) = 1/n_i r_{ijk} d_{ijk} \tag{3.15}$$

Durch die Betrachtung der OR erleichtert sich deren Nutzung als zu erklärende Größe in Modellen. Diese ist weiterhin unabhängig von den Unentschlossenen einer Wahl, da A eine relative Größe darstellt. Eine Vergleichbarkeit über Wahlen und Länder hinweg ist ebenso gegeben, da es sich um eine relative Größe handelt (Martin/Traugott/Kennedy 2005: 366).

Ebenso wie der RNB gegenüber dem „Absoluten Nonresponse-Bias“ (ANB), versucht auch A die Relation zu berücksichtigen. Ein Nachteil ist jedoch die explizite Formulierung für lediglich zwei Parteien. Martin/Traugott/Kennedy (2005) formulieren jedoch bereits eine theoretische Weiterentwicklung von A zu A' (Martin/Traugott/Kennedy 2005: 367), um den Fall von mehr als zwei Parteien zu erfassen. Durand (2008) nutzen diese Vorgabe für eine Umsetzung im Vorfeld der französischen Präsidentschaftswahl 2007, um die OR des jeweiligen Kandidaten in Relation zu allen anderen zu erfassen (Durand 2008: 279 f.). Auch Wright/Farrar/Russell (2014) wandeln A zu A' zur Erfassung der Qualität von Wahlumfragen im neuseeländischen Kontext ab, um explizit auch kleinere Parteien zu berücksichtigen (Wright/Farrar/Russell 2014: 116).

[10] In der Terminologie von Martin/Traugott/Kennedy (2005) handelt es sich hier um die „Republican“ (R) und „Democrats“ (D), diese können jedoch ebenfalls für andere Parteien oder auch andere dichotome Gruppen stehen.

Erst Arzheimer/Evans (2014) haben jedoch die Berechung von A' konzeptionell auf ein Mehrparteiensystem verallgemeinert und die Berechnung in Stata implementiert (Arzheimer/Evans 2016). Sie zeigen, dass eine einfache Verallgemeinerung auf $k-1$ OR für k Parteien fehleranfällig ist, da sie die Wahl einer Partei als Referenzpartei impliziert. Ist die Erfassung dieser selbst fehlerbehaftet, sind es auch die OR aller anderen Parteien. Es liegt dann keine verallgemeinerbare Kennzahl vor. Aus diesem Grund setzen sie das Verhältnis einer Partei p_i in der Umfrage nach folgendem Maßstab in Relation zum Verhältnis des amtlichen Wahlergebnis v_i (Arzheimer/Evans 2014: 33):

$$A'_i = ln\left(\frac{p_i}{1-p_i} \cdot \frac{1}{O_i}\right) \text{ mit } O_i = \frac{v_i}{1-v_i} \tag{3.16}$$

Wenn somit eine Partei 5 % des Anteils der Zweitstimmen innerhalb der Umfrage *und* auf der Grundlage des amtlichen Endergebnis erhalten hat, wurde sie mit $A'_i = ln(\frac{0.05}{1-0.05} \cdot \frac{1}{0.05/(1-0.05)}) = 0$ unverzerrt durch die Umfrage geschätzt.

Dies lässt sich für eine Partei i verallgemeinern als das Verhältnis der logarithmierten Chancenverhältnise von Umfrage und amtliches Endergebnis der Wahrscheinlichkeit, sich für eine Partei i oder eine der anderen Parteien als Komplementärereignis zu entscheiden (Arzheimer/Evans 2016: 141):

$$A'_i = ln\left(\frac{\frac{p_i}{1-p_i}}{\frac{v_i}{1-v_i}}\right) = ln\left(\frac{\frac{p_i}{\sum_{j=1}^{k} p_j}}{\frac{v_i}{\sum_{j=1}^{k} v_j}}\right) \text{ mit } j \neq i \tag{3.17}$$

B ist dann das arithmetische Mittel des Vektors der absoluten Verzerrungen $A'_1, A'_2, ..., A'_k$ über alle k Parteien (Arzheimer/Evans 2014: 36):

$$B = \frac{\sum_{i=1}^{k} |A'_i|}{k} \tag{3.18}$$

Die Qualität der Erfassung der Genauigkeit von Parteien mit relativ gesehen großen Stimmenanteilen kann jedoch leiden, wenn viele Parteien mit einer geringen absoluten, aber starken relativen Abweichung vorliegen. Die Gewichtung mit dem Anteil der erlangten Stimmen kann dieses Problem jedoch beheben (Arzheimer/Evans 2014: 37). B_w gewichtet daher die Verzerrung über alle k Parteien hinweg mit ihrem relativen amtlichen Stimmenanteil (Arzheimer/Evans 2016: 141 f.):

$$B_w = \sum_{i=1}^{k} v_i \cdot |A'_i| \tag{3.19}$$

Wenn der Fokus der Evaluation auf Parteien mit einem relativ geringem Anteil

an den Gesamtstimmen liegt, ist B die geeignete Kennzahl. Liegt der Fokus auf den verhältnismäßig großen Parteien, ist B_w geeigneter, um der verzerrten Schätzung von kleineren Parteien eine niedrigere Bedeutung beizumessen (Arzheimer/Evans 2014: 37).

Da der Vektor $\mathbf{a} = (A'_1, A'_2, ..., A'_k)^t$ sich ohne weitere erklärende Größen als direkte Funktion der durch „Maximum-Likelihood-Estimation“ (MLE) geschätzten und approximativ normalverteilten Parameter $\Theta: \hat{\Theta}_1 = p_1, \hat{\Theta}_2 = p_2, ..., \hat{\Theta}_k = p_k$ ergibt (siehe Gleichung 3.17), lässt sich hierüber auch die Varianz-Kovarianzmatrix $\sum_{\hat{a}}$ und die Wahrscheinlichkeitsdichtefunktion $Pr_{\hat{a}}(\mathbf{y})$ von $\mathbf{a}$ bestimmen. Hierdurch ist der approximative Standardfehler für die Teststatistiken berechenbar. Für die gesamte Abweichung zwischen $\mathbf{p}$ und $\mathbf{v}$ in Form von B und B_w lässt sich hingegen im SRS-Fall ein adäquater χ^2- oder G^2-Anpassungstest durchführen. Muss ein komplexes Erhebungsdesign berücksichtigt werden, wird ein Wald-Test durchgeführt (Arzheimer/Evans 2016: 142 f.). Hierdurch ergibt sich die Möglichkeit der testbasierten Entscheidung, ob auf einer Variable hinsichtlich der k Kategorien generell eine statistisch signifikante Verzerrung gegenüber den Anteilswerten der Population existiert. Wenn dies der Fall ist, lassen sich zusätzlich die Kategorien bestimmen, welche für diese Verzerrung ursächlich sind. Die Möglichkeit der testbasierten Entscheidung der Verzerrung ist somit ein essentieller Vorteil der Nutzung von A_i in Relation zum AB und RB zur Erfassung einer Verzerrung bedingt durch Nonresponse und Noncoverage.

Nachdem an dieser Stelle nun die Definition und Messung einer Verzerrung in Form eines Bias geklärt ist, kann das nachfolgende Kapitel nun dazu übergehen, die Begegnung und Korrektur dieses Qualitätshemmnis einer Umfrage darzustellen.

4. Die Korrektur von Verzerrungen

Die vorausgegangenen Kapitel haben die möglichen Fehlerquellen in Umfragen im Rahmen des „Total Survey Error“ (TSE) herausgearbeitet (Kapitel 2) und den daraus resultierenden gesamten Fehlers definitorisch verortet und seine Messung quantifiziert und greifbar gemacht (Kapitel 3). Das nächste Kapitel wird nun Lösungsansätze präsentieren, um einer vorhandenen Verzerrung effektiv begegnen zu können. Die Lösungen des nachfolgenden Kapitels konzentrieren sich auf den Umgang der durch den Nonresponse- und Noncoverage-Fehler hervorgerufenen Probleme. Das Kapitel 3 hat aufgezeigt, dass trotz der unterschiedlichen Ursachen dieser beiden Fehlerquellen das Resultat identisch ist: Die Verteilungen innerhalb der Stichprobe weichen systematisch von denen in der anvisierten Grundgesamtheit ab, da ein lediglich selektiver Teil dieser befragt werden konnte. Ist durch Noncoverage keine Möglichkeit der Auswahl bestimmter Personen möglich, ist dies bei einer Verweigerung oder mangelnden Möglichkeit der Kontaktierung (Nonresponse) gegeben. Hierdurch ist eine Beeinflussung der ausgewählten Person hinsichtlich einer Beteiligung an der Befragung auch *während* der Kontakt- und Erhebungsphase noch möglich, oder die Kontaktaufnahme kann erneut geschehen.

Der Umgang mit einem möglichen Nonresponse-Fehler ist daher auch eine Qualitätsfrage, welche sich während aller Phasen der Planung, Durchführung, Dokumentation und Auswertung einer Umfrage stellt. Drohenden Verzerrungen durch Nonresponse lässt sich daher bereits während der verschiedenen Stufen der Erhebungsphase begegnen (Schnell 2012: 180 f.). Nonresponse kann daher als ein psychologisches und soziologisches Problem verstanden werden. Die Begegnung von Befürchtungen der zu befragenden Person oder monetäre Anreize sind eine Möglichkeit bereits während der Kontaktaufnahme die Wahrscheinlichkeit der Teilnahme an der Befragung zu steigern. Ebenso kann das Erhebungsdesign so entgegenkommend gestaltet werden, dass die Wahrscheinlichkeit von Ausfällen verringert wird. Eine Behandlung ist jedoch auch nachträglich möglich. Hierzu kann auf der Basis von Bedeutungsgewichten der erfolgreich befragten Personen die Verzerrung durch Nonresponse während der Analyse korrigiert werden (Brick 2013: 331).

Nachfolgend wird das Kapitel 4.1 die Begegnung einer drohenden Verzerrung einer Erhebung *während der Erhebungsphase* über die Verwendung von „Mixed Modes“ (Leeuw 2005), eines „Tailored Design“ (Dillman 1978; Dillman/ Smyth/Christian 2009) oder auch eines ‘Responsive Designs“, bzw. „Adaptive Designs“, vorstellen. Diese Ansätze versuchen aktiv Nonresponse zu verhindern, bevor es zu einer solchen Entscheidung durch die ausgewählte Person kommt.

Trotz dieser Vorkehrungen wird jedoch die Befragung bei einem bestimmten Teil der ausgewählten Personen nicht realisierbar sein, oder eine Auswahl ist im Vorhinein nicht möglich gewesen. Da mit zunehmendem Anteil und steigenden Unterschieden zu den erfolgreich befragten Personen eine umso stärkere Verzerrung hervorruft (Kapitel 3.3), kann eine *nachträgliche Korrektur* notwendig sein. Gewichtungsverfahren werden als eine solche Möglichkeit vorgestellt. Durch diese lassen sich bekannte Verzerrungen korrigieren. Hierbei kann es sich um eine bewusste Inkaufnahme dieser im Rahmen der Auswahl durch das Erhebungsdesign handeln. Ebenso ist die Umwandlung einer Haushalts- in eine Personenstichprobe notwendig, wenn erstere die Auswahlgrundlage bilden. Das Kapitel 4.2.1 wird daher Design- und Transformationsgewichte vorstellen, welche diesen bewussten „Ungleichgewichten“ in der Stichprobe begegnen.

Gewichtungsverfahren lassen sich jedoch ebenso zur nachträglichen Korrektur eines systematischen Ausfalls durch Nonresponse und Noncoverage nutzen. Hierzu werden auf der Grundlage bestimmter Merkmale Bedeutungsgewichte bestimmt, von denen eine Ursächlichkeit für den Ausfall angenommen wird. Hierzu wird die vorhandene Stichprobe immer relational zu einer Referenzquelle gesetzt. Dies können aggregierte externe Verteilungen der amtlichen Statistik sein, welche sich im Rahmen eines *„Rakings“* durch die Anwendung eines „Iterative Proportional Fitting“ (IPF)-Verfahrens nutzen lassen (Kapitel 4.2.2). Liegen geeignete Referenzdaten in Form von Individualdaten vor, können auch diese Informationen zur Korrektur genutzt werden. Hierzu lassen sich „Propensity Scores“ (PS) nach Rosenbaum/Rubin (1983) im Rahmen eines „Propensity Score Matching“ (PSM) nutzen (Kapitel 4.3). Neben der Qualität des angewandten Modells zur Erklärung des Ausfalls, ist auch die Wahl des zu Grunde liegenden technischen Matchingalgorithmus entscheidend, welche Kapitel 4.3.3 präsentiert. Dieser ist erfolgreich, wenn er die Verteilungen zwischen zwei Gruppen von Personen bestmöglich balanciert. Die Messung des Erfolgs dieser Balancierung stellt abschließend das Kapitel 4.3.4 dar.

4.1. Qualitätskontrolle während der Erhebungsphase

„Responsive Designs“ (Groves 2006; Kreuter/Olson et al. 2010; Olson 2012; Wagner 2012a; Kreuter 2013; Lundquist/Särndal 2013) und „Adaptive Designs“ (z.B. Peytchev/Riley et al. 2010; Wagner 2012b; Wagner et al. 2012; Lundquist/Särndal 2013; Calinescu/Bhulai/Schouten 2013; Calinescu/Schouten 2016) haben zum Ziel die Gefahr einer Verzerrung während der Erhebungsphase zu minimieren. Über die konstante Evaluation werden Änderungen bestimmter Eigenschaften des Erhebungsdesigns erreicht, welche den zu befragenden Personen entgegenkommen (Groves 2006: 440; Bethlehem/Cobben/Schouten 2011: 396 f.). Diese Konzepte versuchen Nonresponse bereits zu unterbinden, bevor es sich als ein systematisches Problem äußern kann. Es berücksichtigt „ Nonresponse als ein inhaltliches Problem, kein statistisches. [...] Jede sinnvolle Anwendung eines statistischen Nonresponse-Verfahrens setzt [somit] inhaltliches Wissen über den Erhebungsvorgang und eine inhaltliche Theorie über den Responseprozess voraus.“(Schnell 2012: 183).

Es ist der Königsweg der Vermeidung eines Nonresponse-Bias: Es wird aktiv das Qualitätsniveau über den gesamten Erhebungsprozess evaluiert und somit bereits während der Auswahl und Befragung sichergestellt (Kreuter 2012). Hierzu werden für verschiedene Gruppen spezifische Komponenten des Erhebungsdesigns angepasst (Calinescu/Schouten 2016: 35). Die Kontrolle und Evaluation der Qualität des Rücklaufs kann dabei über Paradaten erfolgen (Kreuter/Olson et al. 2010), also vor allem die Kontaktprotokolle der Interviewer mit erhebungstechnischen Merkmalen (Schnell 2012: 318). Da diese Informationen über den Interviewer erhoben werden, liegen sie auch für die Personen vor, welche nicht an der Befragung teilnehmen wollen. Lässt der Vergleich von Respondenten und Nonrespondenten bereits früh den Schluss eines systematischen Unterschieds zu, kann während der Erhebungsphase gegengelenkt werden. Diese Daten über die Probanden, oder auch die Erhebung selber, lassen sich „[...] bei einer geordnet durchgeführten Datenerhebung prinzipiell nahezu kostenneutral [...] gewinnen.“ (Schnell 2012: 183).

In der Praxis finden sich bisher noch sehr wenige Durchführungen der systematischen Evaluierung des Potentials von solchen Paradaten zur Nonresponsekorrektur (z.B. Krueger/West 2014). Eine technische Möglichkeit der Erfassung der Qualität der Teilnahme stellen vor allem die in Kapitel 3.1 präsentierten R-Indikatoren dar. Hierdurch ist die konstante Erfassung der qualitativen Veränderung des Rücklaufs während der Erhebungsphase in Relation zu aggregierten Rahmenverteilungen möglich.

Eine denkbare Folge der Realisierung einer entstehenden Verzerrung durch systematische Ausfälle wäre der komplette Wechsel des Erhebungsmodus, also die Verwendung von „Mixed Modes“ (Leeuw 2005). Das Ziel dieses Vorgehens ist Teile der Nichtabdeckung eines Erhebungsmodus durch die Berücksichtigung *dieser* ausgefallener Personen durch einen anderen Erhebungsmodus zu erreichen. Dies ist die Folge davon, dass „[...] der Coverage in der Summe der möglichen Modes und der ihnen eigenen Kommunikationskanäle zugenommen hat, aber für einzelne Modes rückläufig ist bzw. von jeher eingeschränkt war.“ (Fuchs 2012: 53). Es wird also versucht, dem Problem zunehmend mangelnder Erreichbarkeit und damit einhergehender sinkender Rücklaufquoten als mögliche Ursache einer Verzerrung der *gesamten* Bevölkerung durch lediglich *ein* Erhebungsverfahren während der Datenerhebungsphase bereits entgegenzusteuern. Da unterschiedliche Befragungsformen mit unterschiedlichen Graden an Reaktivität und sozialer Erwünschtheit zu kämpfen haben, sind die Ergebnisse hinsichtlich eines möglichen Untersuchungsgegenstands nicht ohne Probleme vergleichbar.[1]

Auch die Idee des „Tailored Design“ (Dillman 1978; Dillman/Smyth/Christian 2009) versucht dem Befragten durch eine individualisierte Art der Ausführung der Erhebung entgegenzukommen. Hierbei wird – längst nicht nur – bei postalischen Befragungen versucht durch eine Optimierung von einem personalisiertem Anschreiben, visuellem Erscheinungsbild, Layout und Kontaktablauf und insbesondere der Formulierung und Anordnung von Fragen eine höhere Rücklaufquote zu erreicheen (Schnell 2012: 249).

Insgesamt sind diese Ansätze als Teil der „Leverage-Salience-Theory“ (Groves/Singer/Corning 2000) zu sehen. Da verschiedene Komponenten einer Umfrage unterschiedlich auf verschiedene Gruppen von zu befragende Personen wirken, wird durch eine Personalisierung der Kontaktaufnahme der Versuch der Erhöhung einer individuellen Teilnahmebereitschaft der jeweiligen Person zu erreichen (Groves/Singer/Corning 2000: 306). Hierdurch werden die zu befragenden Personen „[...] in ihren Befürchtungen und individuellen Handlungsabsichten ernst genommen und als absichtvoll Handelnde behandelt [...]. Die Befragten sind weder beliebig austauschbar noch zu einer Kooperation verpflichtet.“ (Schnell 2012: 181). Wird dies als handlungsleitend für die Ausführung der Erhebung anerkannt und die Zeit und Kosten dieser fortlaufenden Evaluierung des Rücklaufs der Erhebung und die mögliche Anpassung getragen, lassen sich

[1] Dies triff beispielsweise bei Bevölkerungsumfragen zu, wenn Onlinebefragungen ein Teil der Erhebungsmethode sind (Leeuw/Hox 2011; Shin/Johnson/Rao 2012).

die Auswirkungen von Nonresponse – und auch eines möglichen Noncoverage durch die Variation des Erhebungsmodus – abmildern, indem das Problem gar nicht entsteht.

Als (momentanes) Resumée zeigt sich jedoch eine gewisse Ernüchterung. Auch durch den Versuch des Umgangs mit Problemen hinsichtlich Nonresponse während der Erhebungsphase wurde keine generelle Umkehr des sich fortschreibenden Problems durch steigende Nonresponseraten erreicht (Brick 2013: 332). Der anschließende *zusätzliche* Einsatz von Gewichtungsverfahren zur Korrektur ist daher eine wesentliche Komponente, um dem Problem erfolgreich zu begegnen. Von der anfänglichen Debatte einer einheitlichen Definition von Nonresponse, über den Versuch der Steigerung der Teilnahmeraten durch designtechnische Modifikationen während der Datenerhebungsphase hat der Fokus der Nonresponseforschung – aus mangelnder Wirksamkeit der Modifikationen – der nachträglichen statistischen Korrektur durch Gewichtungsverfahren eine Priorität eingeräumt (Singer 2006). Der nachfolgende Teil des Kapitel wird deren Anwendung nun vorstellen.

4.2. Die nachträgliche Korrektur durch Gewichtungsverfahren

Die Verwendung von Gewichtungsverfahren hat innerhalb der Umfrageforschung eine lange Tradition. Hierzu steht ein breites Instrumentarium der Anwendung zur Verfügung (zur Übersicht z.B. Kalton/Flores-Cervantes 2003). Gängige Einsatzgebiete sind die Korrektur unterschiedlicher Auswahlwahrscheinlichkeiten, die Behebungen der Auswirkungen durch Nonresponse und Noncoverage und die Korrektur bekannter Abweichung von der Grundgesamtheit durch die Stichprobenziehung (Brick/Kalton 1996). Gewichtungsverfahren ordnen Personen individuelle Bedeutungsgewichte w_i in der Form zu, dass sich bezüglich der Verteilungen der an der Bestimmung der Gewichte beteiligten Variablen wieder eine konditionale Repräsentativität der vorliegenden Stichprobe einstellt. Für eine interessierende Variable Y kann dies zum Beispiel das arithmetische Mittel sein. Über die Gewichtung ist dies, konditional der an der Berechnung beteiligten Variablen, unverzerrt (Bethlehem 2009: 250):

$$\bar{y}_W = \frac{1}{N} \sum_{i=1}^{n} w_i y_i$$

Personen aus Gruppen mit zu geringen Anteilen erhalten eine höhere Gewichtung, solche aus Gruppen mit zu hohen eine niedrigere (Bethlehem 2015: 165). Gewichtungsverfahren versuchen also die Verzerrung von Variablen durch eine nachträgliche inhaltliche Korrektur zu beheben. Die Anwendung dieser erhöht jedoch automatisch die Varianz des jeweiligen Schätzers. Ist die Reduktion des Bias daher nur gering, kann die Anwendung einer Gewichtung den „Mean Squared Error“ (MSE) (Kapitel 3.2) sogar erhöhen. Die Datenqualität reduziert sich dann durch die Anwendung (z.B. Frankel 2010: 122), Gewichtungsverfahren zur Begegnung von Verzerrungen sind daher nur unter bestimmten Umständen effektiv (für ein simuliertes Beispiel siehe z.B. Arzheimer 2009). Zudem korrigieren Gewichtungsverfahren nur diejenigen verzerrten Variablen, welche in einem Zusammenhang mit dem zur Bestimmung der Bedeutungsgewichte angewandten Korrekturmodell stehen. Die Beseitigung der Verzerrung auf allen interessierenden Variablen ist utopisch (Lohr 2010: 345). Gewichtungsverfahren sind daher kein „Allheilmittel“, um eine Umfrage wieder als vollkommen „repräsentativ“ einzustufen. Wie das Kapitel 3 bereits herausgestellt hat, ist der Begriff der Repräsentativität immer konditional der an der Korrektur beteiligten Variablen zu verstehen. Werden Ungleichgewichte hinsichtlich der Abweichungen der Verteilung des Alters einer Stichprobe von denjenigen der amtlichen Statistik angewandt, stellt sich auf weiteren inhaltlichen Variablen nur dann eine statistische Repräsentativität ein, wenn der Bias nur durch die verzerrte Erfassung des Alters verursacht wurde.

Eine zentrale Gefahr von Gewichtungsverfahren ist daher schlicht ihre naive Anwendung. Die Bereitstellung in einem Datensatz birgt die Gefahr der Auffassung als eine „natürliche Eigenschaft“ der Befragten. Diese muss nicht hinterfragt werden und die Anwendung dieses Repräsentativ-Gewichts folglich konsequent. Da deren Bestimmung jedoch immer konditional gesehen werden muss, sind sie künstliche aggregierte Eigenschaften eines Datensatzes und nicht individuelle Eigenschaften einer Person (Gelman 2007: 155 f.).[2] Ein weiteres Problem sind sich technisch ergebende hohe Bedeutungsgewichte einzelner Personen, welche ab einem bestimmten Niveau auf einen fixen Wert gedeckelt werden können („trimming/truncation“). Dies führt häufig zu besseren Ergeb-

[2]Ein Beispiel ist der „Faktor Repräsentativgewicht“ in den Politbarometer-Datensätzen. Wird dies angewandt, stellt sich anscheinend Repräsentativität ein. Dies wird laut Methodenbeschreibung der Erhebungen des Politbarometers konditional hinsichtlich Geschlecht, Alter und Bildung definiert. Es sei jedoch „die gewichtete Umfrage [...] unter Berücksichtigung der wahrscheinlichkeitstheoretischen Grundlagen von Stichproben repräsentativ für die wahlberechtigte Bevölkerung Deutschlands [ist].“ (http://www.forschungsgruppe.de/Umfragen/Politbarometer/Methodik/, Abruf: 11. Mai 2018).

nissen (Lee/Lessler et al. 2011). Hierdurch wird jedoch eine gewisse Willkür der Erstellung der berechneten Gewichte induziert. Zudem sind hohe Gewichte ebenfalls häufig ein Indiz für eine fehlerhafte Bestimmung dieser (Stuart 2010: 10). Die Reduktion großer Gewichte auf einen Maximalwert reduziert daher die Varianz eines Schätzers, erhöht aber gleichzeitig die Verzerrung (Elliot/Little 2000).

Vor diesem Hintergrund ist es wichtig einzuordnen, wann eine Gewichtung eine logische Folge und Notwendigkeit der angewandten Erhebung ist und wann sie eine hochspekulative Möglichkeit der Behebung etwaiger nicht bekannter Verzerrungen darstellt. Das nachfolgende Kapitel 4.2.1 stellt Design- und Transformationsgewichte vor. Deren Anwendung ergibt sich logisch aus der Art der Stichprobenziehung. Die breite Diskussion des Pro und Contra der Verwendung von Gewichtungsverfahren dreht sich nicht um diese Form der Gewichtung, sondern vor allem um die Korrektur von Nonresponse und Noncoverage durch „Redressement-Gewichte“ (Arzheimer 2009: 363), welche das Kapitel 4.3 anschließend vorstellen wird.

4.2.1. Design- und Transformationsgewichte

Die Anwendung von „Design“- und „Transformations“-Gewichten stellt eine weitverbreitete Anwendung von Gewichtungsverfahren dar. Ihr Ziel ist die Korrektur unterschiedlicher *bekannter* Auswahlwahrscheinlichkeiten der Befragten Teil der Stichprobe zu werden (Kalton 1983: 69). Allgemein lassen sich Design-Gewichte unter der Idee des „Horvitz-Thompson“ (HT)-Schätzers (Horvitz/ Thompson 1952) erfassen. Jede Person hat eine durch das Design bedingte Wahrscheinlichkeit π_k. Über alle N Elemente entspricht dies der Summe der befragten Personen $\left(\sum_{i=1}^{N} \pi_k = n\right)$. Die Inverse dieser Auswahlwahrscheinlichkeit einer Person definiert das ihr zugeordnete Design-Gewicht $d_k = 1/\pi_k$. Personen mit einer erhöhten Auswahlwahrscheinlichkeit werden daher hinsichtlich ihrer Bedeutung herabgestuft. Mit der Anwendung des HT-Schätzers lässt sich dann das arithmetische Mittel einer Variable y wieder unverzerrt hinsichtlich der unterschiedlichen Auswahlwahrscheinlichkeiten schätzen (z.B. Quatember 2015: 20 ff. Bethlehem 2009: 249; Särndal/Lundström 2005: 7 f.):

$$\bar{y}_{HT} = \frac{1}{N} \cdot \sum d_k \cdot y_k$$

(Notation nach Quatember (2015)). Bei bevölkerungsweiten Umfragen in der Bundesrepublik wird häufig der ostdeutsche Bevölkerung eine erhöhte Auswahl-

wahrscheinlichkeit zugestanden. Durch ihren erhöhten Anteil in der Stichprobe lassen sich separate Analysen in Ostdeutschland effizienter durchführen. Ein Design-Gewicht korrigiert die ungleichen Auswahlwahrscheinlichkeiten, wenn eine Aussage für die gesamte Bundesrepublik getroffen werden soll. Dieses lässt sich einfach bestimmen, da die Referenzanteile für Ost und West über die amtliche Statistik bekannt sind (Terwey/Baltzer 2014: iv f.).[3] Transformationsgewichte lassen sich ebenfalls sehr einfach bestimmen. Im Rahmen der Erhebung des Allbus berechnet sich das Transformationsgewicht $w*_i$ als die individuelle reziproke reduzierte Haushaltsgröße w_i in Relation zum arithmetischen Mittel dieser: $w*_i = \frac{w_i}{\bar{w}}$. Für die gesamtdeutsche Stichprobe wird dieses noch mit der Design-Korrektur multipliziert: $w_i^{*(Design)} = w*_i \cdot w_i^{Design}$(Wasmer et al. 2014: 52 f.).

Solche Korrektur-Gewichte lassen sich generell über simple Zellgewichtungen („Poststratification“) erreichen.[4] Hierzu werden die Befragten in h Gruppen stratifiziert. Der individuelle Gewichtungsfaktor g_i eines Befragten i ergibt sich dann als das Verhältnis des Anteils seiner zugehörigen Gruppe h in der Grundgesamtheit (N_h/N), in Relation zum Anteil in der Stichprobe (n_h/n): $g_i = \frac{N_h/N}{n_h/n}$. Gruppen mit zu geringen Anteilen erhalten hierdurch ein Gewicht über eins, überrepräsentierte Gruppen hingegen einen Wert unterhalb von eins (Bethlehem/Cobben/Schouten 2011: 214 ff.). Poststratifizierung ist auf die simultane Korrektur vieler Variablen verallgemeinbar. Es kann sich jedoch das Problem zu geringer Gruppengrößen zur stabilen Bestimmung der Gewichte zeigen (Frankel 2010: 101). Poststratifizierung hat umso stärkere Auswirkungen, je homogener die Personen einer gebildeten Gruppe hinsichtlich *weiterer* inhaltlich interessierender Variablen sind. Dann kann die Gewichtung sowohl den Standardfehler von Schätzern als auch deren Verzerrung verringern, der MSE sinkt. Korreliert hingegen nur die Wahrscheinlichkeit der Teilnahme *innerhalb* einer Gruppe mit dem generellen Untersuchungsziel der Umfrage, kann auch die Korrektur der Verhältnisse *zwischen* den Gruppen die Verzerrung nicht beheben (Bethlehem/ Cobben/Schouten 2011: 214 ff.).

Auch für mehrstufige Zufallsziehungen lassen sich die Auswahlwahrscheinlichkeiten der unterschiedlichen Stufen kombinieren und berücksichtigen. Ein Beispiel hierfür sind komplexe räumliche Haushaltsstichproben mit anschließender Zufallsziehung einer Person in diesem (Quatember 2015: 119 f.). Lässt sich bereits während der Stichprobenziehung eine Auswahlwahrscheinlichkeit

[3] Siehe hierzu auch z.B. das FAQ des Allbus unter http://www.gesis.org/allbus/allgemeine-informationen/faqs/, Abruf: 11. Mai 2018.

[4] Little (1986) bietet eine Übersicht über verschiedene Varianten der Zellgewichtung.

proportional zum Anteil der interessierenden Größen realisieren („Probability Proportional to Size“ (PPS)-Ziehung), ist anschließend keine Gewichtung mehr im Rahmen eines HT-Schätzers notwendig. Da jedoch meist elementare Informationen über die Erhebungseinheiten vorab nicht bekannt sind, kann in der Praxis nur eine konditionale Berücksichtigung hinsichtlich bestimmter bekannter Merkmale realisiert werden (Quatember 2015: 129-135). Soll auf der Grundlage einer Haushaltsstichprobe Aussagen auf der Individualebene getroffen werden, muss diese über Transformationsgewichte in eine Personenstichprobe transformiert werden. Falls jedoch Einpersonen- gegenüber Mehrpersonenhaushalten tendenziell schwieriger zu erreichen sind, konterkariert dieser Effekt die möglichen ungleichen Auswahlwahrscheinlichkeiten. Die Wirkung von Transformationsgewichten ist daher keineswegs immer eindeutig positiv (Arzheimer 2009: 363 f.). Zeigen sich Auswirkungen, liegt dies bei Transformationsgewichten an einer Verzerrung bedingt durch die Haushaltsgröße. Inhaltliche Variablen, welche mit dieser in einem Zusammenhang stehen, sind daher ohne Korrektur fehlerhaft erfasst (Hartmann/Schimpl-Neimanns 1992: 323 f.).

Die bis hierhin dargestellten Gewichtungskorrekturen der Design- und Transformationsgewichte sind jedoch prinzipiell unstrittig, sie ergeben sich logisch aus dem Stichprobendesign. Stärkere Auswirkungen finden sich jedoch selten bei deren Anwendung (Arzheimer 2009: 363). Die dargestellten Zellkorrekturen der „Postratifzierung“/“Proportionalisierung“ können jedoch auch über reine Designkorrekturen hinausgehend genutzt werden. Dann sind sie ein Mittel, um durch den Einbezug inhaltlicher Variablen deren Verzerrungen bedingt durch Nonresponse zu korrigieren. Hierzu können Referenzverteilungen über die Population herangezogen werden oder Informationen über die Nonrespondenten genutzt erden (Kalton/Maligalig 1991: 412 ff.). Das nachfolgende Kapitel 4.2.2 wird ersteres in Form eines „Raking“ durch ein IPF-Verfahren aufzeigen und sowohl deren Stärken, aber auch Einschränkungen, darstellen. Hierdurch sind Gewichtungsverfahren beispielsweise in der Lage Verzerrungen soziodemografischer Variablen durch eine Angleichung an die Referenzverteilungen der amtlichen Statistik zu korrigieren.

4.2.2. „Raking“ – Die Angleichung an eine Referenzverteilung

Notwendige Design-Korrekturen können mit weiteren Korrekturen durch Zellgewichtungen kombiniert werden, indem die jeweilig getrennt bestimmt Gewichte multiplikativ verkettet werden. Als Folge stimmen auf allen beteiligten Variablen die marginalen Verteilungen mit den vorgegebenen Referenzverteilungen

überein (Särndal/Lundström 2005: 12). Eine solche Bestimmung von Gewichten durch Poststratifizierung weist jedoch einige Nachteile auf: Wenn eine Zelle nicht durch mindestens eine Beobachtung besetzt ist, kann logischerweise auch kein Gewicht berechnet werden. Finden sich nur sehr wenige Personen in einer Zelle, ist die Berechnung von Gewichten technisch realisierbar, ihre Bestimmung aber sehr instabil. Eine steigende Berücksichtigung an Variablen und/oder Variablenausprägungen ist daher zwar prinzipiell ein Informationsgewinn für die Bestimmung der Bedeutungsgewichte. Mit jeder weiteren Variable müssen jedoch auch mehr Zellen zur Korrektur konstruiert werden, die Wahrscheinlichkeit gering besetzter Zellen und daraus folgender instabiler Berechnung der Gewichte steigt (Bethlehem/Cobben/Schouten 2011: 217). Stellen die Personen gering besetzter Zellen zudem auch die unterrepräsentierten Gruppen dar, werden für diese Personen hohe Bedeutungsgewichte bestimmt. Andere Befragte stark vertretener Gruppen können hingegen analog nur eine sehr geringe Bedeutung für die gewichteten Endergebnisse erhalten. Auch erfordert die Korrektur die Verfügbarkeit der externen *multivariaten* Verteilung aller Variablen, welche für eine Verzerrung ursächlich sind (Frankel 2010: 124).

Insbesondere das letzte Problem versucht das IPF-Verfahren (Deming/Stephan 1940)[5] zu umgehen. Für alle an der Korrektur beteiligten Variablen wird iterativ ein Gewichtungsvektor in der Form bestimmt, dass sich die Randverteilungen der vorliegenden Stichprobe nicht mehr von den zur Bestimmung herangezogenen Referenzverteilungen der Population (z.B. über die amtliche Statistik) unterscheiden. Iterativ bezieht sich auf die nachfolgende Optimierung der Gewichtung aller Variablen: Mit jedem Schritt wird die beste Lösung für die Angleichung der univariaten Verteilung der vorherigen Variable als Basis genommen. Auf diese kann sich die aktuell betrachtete Variable als Startvektor konzentrieren und hier nun eine Angleichung der Verteilungen durch die Modifikation der Gewichte zu erreichen. Für jede Variable versucht der Algorithmus mehrmals, auf Grundlage des bis dahin bestmöglichen Vektors an Gewichten, diese weiter zu optimieren. Sukzessive wird mit jedem Durchlauf die Angleichung *insgesamt* besser. Schließlich konvergiert der Algorithmus und die multivariate Verteilung *aller* Variablen stimmt mit den Vorgaben der Randverteilung der anvisierten Grundgesamtheit überein (z.B. Bethlehem/Cobben/Schouten 2011: 232 ff. Frankel 2010: 125 f.).

In der Regel kommt es bereits nach sehr wenigen Iterationen zu einer zu-

[5] In der Literatur existieren auch die Synonyme „raking“, „raking ratio estimation“ oder auch „multiplicative weighting“ (Bethlehem/Cobben/Schouten 2011: 231).

friedenstellenden Lösung und der Iterationsprozess kann mit einer bereitgestellten Gewichtungsvariable erfolgreich beendet werden. Logischerweise nimmt die Anzahl der Iterationen aber mit zunehmender Anzahl zu korrigierender Variablen oder ihrer Merkmalsausprägungen zu. Auch eine größere Anzahl befragter Personen, die Anzahl der Zellen mit einer prozentual nur sehr geringen Besetzung sowie Stärke der Differenzen vor Start der iterativen Angleichung erhöhen die Komplexität der Angleichung (Battaglia/Hoaglin/Frankel 2009). Das „Raking“ ist damit ebenfalls ein Verfahren, um eine Zufallsstichprobe nachträglich zu schichten. Ein elementarer Vorteil gegenüber der vorher Stratifizierung während der Erhebungsphase liegt darin, dass nachträglich für unterschiedliche Zielvariablen auch unterschiedliche Schichtungsvariablen angewandt werden können (Quatember 2015: 101).

Zur Korrektur werden häufig soziodemografische Variablen herangezogen, da hierfür die Rahmendaten über die amtliche Statistik verfügbar sind. Der Effekt der Nutzung dieser Art von Variablen zur Reduzierung einer Verzerrung kann jedoch als nur sehr begrenzt eingeschätzt werden (Shadish/Clark/Steiner 2008; Peytcheva/Groves 2009). Diese mangelnde Effektivität ergibt sich, wenn nicht alle die Verzerrung inhaltlicher Variablen beeinflussenden Faktoren berücksichtigt werden. Die Nonresponserate einer Umfrage zu bestimmen, ist jedoch um einiges einfacher als die Gründe herauszustellen, welche für eine zu hohe Rate ursächlich sind. Die Teilnahmerate ist bekannt, die Unterschiede zwischen Respondenten und Nonrespondenten auf interessierenden Größen – und damit der eigentliche Nonresponse-Bias – hingegen nicht (Wagner 2012b: 556 f.). Steht die betrachtete Zielvariable jedoch nicht mit diesen Faktoren in einer Beziehung, führt die Anwendung der Gewichtung in jedem Fall zu einer Erhöhung der Varianz eines Schätzers. Als Folge steigt der MSE. In der praktischen Anwendung werden solche bereitgestellten Korrektur-Gewichte jedoch als willkommene Möglichkeit des Ausschluss einer generellen Verzerrung genutzt (Lohr 2010: 345). Durch die Angleichung an die „offiziellen“ Verteilungen der Rahmendaten der amtlichen Statistik suggeriert die Anwendung dieser oftmals als „Repräsentativgewichte“ die Beseitigung aller Fehlerquellen bedingt durch Nonresponse und Noncoverage.

Damit die Korrektur durch IPF-Gewichte erfolgreich ist, muss diese also über soziodemografische Faktoren hinausgehen. Hierzu sind diese Informationen jedoch in irgendeiner Form notwendig, wozu es möglich sein müsste als eine Referenzquelle „[...] diejenigen [zu] befragen, die sich nicht befragen lassen, um etwas über ihre politischen Einstellungen und ihr Verhalten zu erfahren.“ (Pro-

ner 2011: 30). Diese Informationen versuchen „follow-up“/“Nonrespondenten“-Studien in der Tradition des „Callback Approach“ (CBA) nach Hansen/Hurwitz (1946) und des „Basic Question Approach“ (BQA) (Kersten/Bethlehem 1984) zu erreichen, indem sie ursprüngliche Nonrespondenten in die Umfrage im Nachhinein „konvertieren“. Das Problem solcher Studien ist jedoch die Annahme der Repräsentativität der durch den „follow-up“ erfassten Nonrespondenten für die Nonrespondenten insgesamt (Bethlehem/Cobben/Schouten 2011: 292 ff.). Es muss also für diese davon ausgegangen werden muss, „[...] dass sie den endgültigen Nonrespondenten ähnlicher sind als den Kooperativen.“ (Proner 2011: 40). Gilt dies nicht, sind auch nicht die Faktoren erfasst, welche die Verzerrung verursachen.

Es existiert jedoch eine weitere Möglichkeit zu diagnostizieren, ob eine Umfrage hinsichtlich über die Soziodemografie hinausgehende Faktoren verzerrt ist: Den Vergleich der interessierenden Verteilungen der vorliegenden Umfrage mit den Verteilungen der identischen Variable einer anderen, als qualitativ hochwertiger angenommenen, Umfrage. Das nachfolgende Kapitel 4.3 wird nun zeigen, wie über die Modellierung von PS und ein darauf aufbauendes PSM für eine vorliegende Stichprobe über die Angleichung an eine höherwertigere Referenzumfrage Gewichte in der Form bestimmt werden können, dass die Unterschiede der genutzten Variablen zwischen diesen beiden Umfragen minimal werden.

4.3. „Propensity Score Matching“ (PSM) und Referenzumfragen

Der Vorteil der Nutzung von Referenzumfragen gegenüber Referenzverteilungen ist die Möglichkeit der Nutzung der Individual- statt Aggregatebene zur Angleichung. Dies erhöht einerseits die Informationsvielfalt, da in Umfragen eine Vielzahl an möglichen Variablen themenspezifisch abgefragt werden kann. Zudem liegen diese Informationen als multivariate Verteilungen vor, da alle Eigenschaften und Antworten eindeutig Personen auf der Individualebene zugeordnet werden können. Dies ermöglicht eine Korrektur durch die Angleichung einer von Verzerrungen betroffenen Umfrage an eine als besser eingestufte Referenzumfrage über die Modellierung von PS als „[...] the conditional probability of assignment to a particular treatment given a vector of observed covariates.“(Rosenbaum/Rubin 1983). Auf Grundlage der Logik des „Neyman-Rubin-Models“ (Neyman 1923; Rubin 1974, 1977, 1978, 2006) ist das (ursprüngliche) Ziel der Modellierung dieser die nachträgliche Herstellung einer Experimentalsituation.

Das Ziel der Modellierung von PS ist die Herstellung einer konditionalen Repräsentativität für eine vorliegende Umfrage. Die PS sind die Wahrscheinlichkeit für die befragten Personen der Vergleichsumfrage, auch (hypothetisch) Teil der Referenzumfrage zu sein. Je wahrscheinlicher dies ist, umso bedeutender sind diese Personen zur Erlangung statistischer Repräsentativität. Je höher die PS kalkuliert sind, umso größer wird also das jeweilige Bedeutungsgewicht für diese Person. Repräsentativität ist erreicht, wenn die Verteilungen aller beteiligten Variablen zwischen den beiden Umfragen durch die Gewichtung negiert werden können. In diesem Fall lässt sich die Gruppenzugehörigkeit zur Referenz- (1) und Vergleichsumfrage (0) nicht mehr durch die zur Korrektur genutzten Variablen erklären, die Unterschiede sind durch die Anwendung der Bedeutungsgewichte negiert. Konditionale Repräsentativität ist hierbei also eng mit den Verteilungen und der Qualität der gewählten Referenzumfrage verbunden. Es ist daher essentiell das letztere für die gewählte Refernzumfrage als hoch eingeschätzt werden kann.

Die Kalkulation von PS ist eng mit der Idee der Modellierung der Teilnahmewahrscheinlichkeit an Umfragen verbunden. Diese wird nun nachfolgend als Idee der Modellierung logistsicher Regressionen in Kapitel 4.3.1 dargestellt. Das nachfolgende Kapitel 4.3.2 wird dann aufzeigen, wie diese PS definitorisch von ihrer ursprünglichen Konzeption auf den Bereich der Umfrageforschung zur

Nutzung von Referenzumfragen zur Angleichung weiterer Umfragen verstanden werden können, wie diese technisch bestimmbar sind und als Gewichtungsvariable genutzt werden können. Anschließend wird Kapitel 4.3.3 aufzeigen, wie über statistische Matchingverfahren die PS weiterverarbeitbar werden können und welche verschiedenen Möglichkeiten der Nutzung eines Matchingalgorithmus dazu bestehen. Damit ein solcher Algorithmus als „gut“ im Sinne des Ergebnis beurteilt werden kann, ist eine bestmögliche „Deckungsgleichheit“ der am Matching beteiligten Gruppen essentiell. Das Kapitel 4.3.4 stellt daher heraus, welche Kennzahlen und Möglichkeiten existieren, um den Erfolg in Form der erreichten „Balancierung“ zu bestimmen.

4.3.1. Teilnahmewahrscheinlichkeit an Umfragen

Die Teilnahme an Umfragen lässt sich an Hand zwei verschiedener theoretischer Modelle erfassen: Dem „Fixed Response Model“ (FRM) und dem „Random Response Model“ (RRM). Beide verfolgen eine unterschiedliche Auffassung, wieso manche der ausgewählten Personen letztendlich an der Umfrage erfolgreich teilnehmen, andere jedoch nicht. Im FRM ist die Teilnahme ein bereits vorab festgelegtes Ereignis, da sich die anvisierte Grundgesamtheit bereits klar in Respondenten und Nonrespondenten unterteilen lässt. Jeder Versuch der Beeinflussung der Teilnahme einer ausgewählten Person ist wirkungslos. Einzig ist *insgesamt* die Wahrscheinlichkeit bestimmbar, einen Respondeten/Nonrespondenten aus der anvisierten Grundgesamtheit zu erlangen, sofern die Aufteilung in dieser bekannt ist. Unter der Annahme des FRM ist die Teilnahme/Response R einer Person k daher eine rein dichotome Entscheidung ($R_k = \{1,0\}$). Im RRM hingegen stellt sich die Entscheidung zur Teilnahme ($R_k = 1$) als Funktion verschiedener Faktoren dar: Sie ist mit einer individuellen *Wahrscheinlichkeit des Eintretens* behaftet. Wird eine Person durch das angewandte Erhebungsdesign ausgewählt, erfolgt daher im RRM mit der Wahrscheinlichkeit $P(R_k = 1) = \rho_k$ die Teilnahme. Mit $P(R_k = 0) = 1 - \rho_k$ liegt hingegen ein Fall von Nonresponse vor. Unterschiede der Teilnahmewahrscheinlichkeit zwischen Personen sind auf die Rahmenbedingungen der Umfrage und eine Kombination verschiedener Merkmale X dieser Person zurückzuführen. Die Teilnahmerate ergibt sich daher nicht als der Anteil der Respondenten an den Nonrespondenten in der Grundgesamtheit, sondern als das arithmetische Mittel der individuellen Wahrscheinlichkeiten der Teilnahme innerhalb des ausgewählten Personenkreises aus der Grundgesamtheit. (Bethlehem/Cobben/Schouten 2011: 42; Bethlehem/Cobben/Schouten 2009: 221 f.).

Diese Wahrscheinlichkeiten der Teilnahme verteilen sich für alle Personen der

Grundgesamtheit konditional hinsichtlich deren Eigenschaften auf bestimmten Variablen X (Bethlehem/Cobben/Schouten 2011: 179 ff.):

$$\rho_X = (\rho_X(x_1), \rho_X(x_2), ..., \rho_X(x_N))^T \text{ mit: } \bar{\rho} = \frac{1}{N}\sum_{k=1}^{N} \rho_X(x_k) \tag{4.1}$$

Wenn über alle Kombinationen von X die Teilnahmewahrscheinlichkeit konstant ist und somit für alle $\rho_X(x_N) = \bar{\rho}$ gilt, ist der beste Grad an Repräsentativität erreicht. Dies korrespondiert „[...] to a missing-data mechanism that is MCAR with respect to X, as MCAR states that we cannot distinguish respondents from nonrespondents based on knowledge of X.“ (Bethlehem 2009: 103).

Die Teilnahmewahrscheinlichkeit lässt sich durch binär-logistische Regressionen bestimmen. Als abhängige Variable wird die Entscheidung zur Partizipation, also ein Respondent zu sein, erklärt. Umfasst das Modell alle relevanten Faktoren als erklärende Variablen, lässt sich der wahre Wert der Teilnahmewahrscheinlichkeit erfassen (z.B. Bethlehem 2009: 266; Guo/Fraser 2015: 137). Hierbei handelt es sich für eine Person k also um seine geschätzte Wahrscheinlichkeit $\hat{\rho}_k$ zur Gruppe der Respondenten ($R_k = 1$) – gegeben seiner Eigenschaften X_k *und* das sie aus der Population für die Stichprobe ausgewählt wurde ($a_k = 1$) – zu gehören (Bethlehem/Cobben/Schouten 2011: 329):

$$\hat{\rho}_k = \hat{\rho}(X_k) = P(R_k = 1 | a_k = 1, X_k) \tag{4.2}$$

Sind alle relevanten Informationen über die Gruppe der Respondenten und Nonrespondenten bekannt, lässt sich die individuelle Teilnahmewahrscheinlichkeit für jede Person in Abhängigkeit der genutzten Merkmale abschätzen. Diese können dann in einem weiteren Schritt als Inverse zur Gewichtungskorrektur als Transformatonsgewicht genutzt werden.

Dies impliziert jedoch zwei Herausforderungen: Nonrespondenten lässen sich nur sehr schwer umfassend beobachten/befragen. Alle relevanten Informationen lassen sich daher also nicht berücksichtigen. Für eine Herstellung statistischer Repräsentativität ist es zudem ebenfalls notwendig Daten über die von Noncoverage betroffenen Personen heranzuziehen. Da diese, im Gegensatz zu den Nonrespondenten, noch nicht einmal bekannt sind, ist dies unmöglich.

Die Nutzung einer geeigneten qualitativ hochwertigen Referenzumfrage ist eine Lösung dieses Problems. Diese ist im Vergleich zur vorliegenden Umfrage auf einer Vielzahl von Variablen mit einem geringeren systematischen Fehler belegt. Sie stellt hierdurch einen Idealzustand für die anzugleichende Umfrage

dar. Über eine Gewichtung muss dann erreicht werden, dass sich die Verteilungen der an der Korrektur beteiligten Variablen innerhalb der vorliegenden Umfrage möglichst stark an diejenigen der Referenzumfrage angleichen. An die Stelle des Vergleichs von Respondenten (1) zu Nonrespondenten (0) tritt damit der Vergleich der Teilnehmer der Referenzumfrage (1) und derjenigen der anzugleichenden Umfrage (0). Die nachfolgenden Kapitel werden nun das Vorgehen darstellen um hierdurch eine Steigerung der Datenqualität letzterer durch eine Angleichung an die Referenzumfrage zu erreichen.

HIER NOCH: Mit Bezug auf die Teilnahme an Umfragen sind diese PS „[...] the conditional probability that an individual with observed characetristics X responds in a survey when invited to do so [...]“ (Bethlehem/Cobben/Schouten 2011: 53).

4.3.2. „Propensity Scores“ – Definition und Schätzung

Die Idee der PS nach Rosenbaum/Rubin (1983) wurde ursprünglich für den Bereich der Kausalanalyse entwickelt, um nachträgliche Experimentalsituationen herzustellen. Hierzu wird ebenfalls eine logistische Regressionsgleichung modelliert. Diese schätzt für jeden Befragten, konditional einer einfließenden Zahl an Kovariaten $\mathbf{X}$, die Wahrscheinlichkeit der Zugehörigkeit zur Treatmentgruppe (1) (zur technischen Darstellung z.B. Guo/Fraser 2015: 137 ff. Pan/Bai 2015: 6; Bethlehem 2009: 266):

$$Pr(T_i = 1|\mathbf{X}_i) = \frac{exp(\mathbf{X}_i\beta)}{1 + exp(\mathbf{X}_i\beta)} = e(\mathbf{X}_i) \tag{4.3}$$

Die Schätzung der PS ist notwendig, sofern es sich nicht um eine echte Experimentalsituation handelt, da nur dort die Wahrscheinlichkeit der Treatmentzuweisung bekannt oder durch eine Randomisierung zufällig ist (Luellen/Shadish/Clark 2005: 532). Was den Einsatz der PS im als Gewichtungsfaktor vorteilhaft macht, ist die Komplexitätsreduktion durch die Schätzung dieser. Unabhängig der Anzahl an k Variablen zur Bestimmung der PS, reduziert sich das k-dimensionale Problem der Erfassung auf eine eindimensionale Fragestellung: Wie lassen sich die PS zwischen den beiden beteiligten Gruppen balancieren (z.B. Guo/Fraser 2015: 134; Smith/Todd 2005: 313 f.). PS sind konzipiert worden, um eine kausale Analyse gegeben der beteiligten Variablen nach dem „Potential Outcome Framework“ (Rubin 1974; Robins 1986; Holland 1988) innerhalb des „Neyman-Rubin-Models“ (Neyman 1923; Rubin 1974, 1977, 1978, 2006) zu erreichen, durch Holland (1986) prominent auch als „Rubin Causal

Model“ (RCM) bezeichnet. Durch die zunehmende „Experimentalisierung der Umfrageforschung“ seit Sniderman/Grob (1996), ist die Anwendung dieser Form der Kausalanalyse auch innerhalb von Bevölkerungsdatensätzen immer bedeutender.[6]

Die PS bringen in einer Experimentalsituation die Wahrscheinlichkeit einer Treatmentzuweisung z_i, gegeben ihrer Ausprägungen auf den betrachteten Kovariaten $\mathbf{X}_i$ zum Ausdruck. Sofern die Entscheidung zur Teilnahme $(r = 0, 1)$ und die Zuordnung zur Treatment- und Kontrollgruppe, gegeben der Kovariaten $\mathbf{X}_i$, unabhängig voneinander sind („Ignorability Assumption“), gilt dies auch automatisch für die PS ($e(\mathbf{X}_i)$):

$$(r_{1i}, r_{0i}) \perp z_i | \mathbf{X}_i \Rightarrow (r_{1i}, r_{0i}) \perp z_i | e(\mathbf{X}_i) \tag{4.4}$$

Der unverzerrte Treatmenteffekt kann dann konditional auf $\mathbf{X}_i$ bestimmt werden, sofern ebenfalls gemäß der „Stable Unit Treatment Value Assumption“ (SUTVA)-Annahme die Beobachtung eines Probanden unabhängig von der Treatmentzuweisung anderer Probanden ist *und* alle Kombinationen von $\mathbf{X}$ eine Wahrscheinlichkeit der Treatmentzuweisung haben ($0 < p(t = 1|X) < 1$) (Rosenbaum/Rubin 1983: 45; Notation nach Pan/Bai 2015: 5 f.). Die PS lassen sich daher in Abhängigkeit der Gruppenzugehörigkeit als Gewichtungsfaktor nutzen: Die Treatmentgruppe erhält den Faktor $1/e(x_i)$ und die Kontrollgruppe $1/(1 - e(x_i))$ (Guo/Fraser 2015: 241). Durch diese Art der Gewichtung wird versucht „[...] to remove [all] bias due to all observed covariates.“ (Rosenbaum/Rubin 1983). Eine PS-Gewichtung erreicht somit eine Negierung der Unterschiede zwischen den beiden Experimentalgruppen über die bestimmten Gewichtungsfaktoren an Hand der Wahrscheinlichkeit der Treatmentgruppe (1) der logistischen Regression zu sein. Dies lässt sich auf den Bereich der Umfrageforschung übertragen. Eine PS-Gewichtung negiert dort die Unterschiede zwischen zwei Datensätzen durch die Bestimmung der WS Teil der gewählten Referenzumfrage (1) zu sein. Über die Gewichtung können dann die Personen der anzugleichenden Umfrage (0) höher gewichtet werden, welche auch eine höhere Wahrscheinlichkeit haben Teil der Referenzumfrage zu sein.

Praktisch hat sich durch die Idee des RCM der abgeleitete Trend der Angleichung von (meist günstig) erhobenen Datensätzen an (meist teurere) Referenzumfragen ergeben. Durch dieses Vorgehen lassen sich Unterschiede beheben, indem die Angleichung einer als nicht repräsentativ angenommenen Stichprobe

[6] Für eine detaillierte Einführung Luellen/Shadish/Clark (siehe z.B. 2005).

an eine repräsentative Referenzstichprobe vorgenommen wird (Schnell 2012: 303 f.).

Seit den Pionierstudien durch „Harris Interactive“ (Terhanian/Bremer 2000; Terhanian/Bremer et al. 2000) und weiteren Versuchen aus der Marktforschung (z.B. Börsch-Supan et al. 2004; Schonlau et al. 2004; Duffy et al. 2005) steht im Fokus von PS-Modellierungen die Angleichung von Onlineumfragen an Referenzdatensätze durch Telefonumfragen (z.B. Lee/Valliant 2009: 320 ff.) oder persönliche Interviews (z.B. Duffy et al. 2005), insbesondere wenn die Onlineumfragen nicht auf einer Zufallsauswahl basieren (z.B. Yeager et al. 2011).[7] Das Ziel ist die Reduzierung der Auswirkungen der mangelnden Möglichkeit onlinebasierter Befragungen zur Erlangung zufallsbasierter Stichproben der allgemeinen Bevölkerung durch ihre häufig vorhandenen Abdeckungsprobleme der anvisieren Population sowie ihr nicht zufallsgesteuertes Erhebungsverfahren (das Kapitel 6.2 wird diese Probleme im Rahmen von Onlineumfragen noch darstellen). Die verwendete Referenzumfrage zeichnet sich hingegen durch eine vergleichsweise kleine und hochwertig erhobene zufallsbasierte Stichprobe aus. Das Ziel der Angleichung kann auch darin liegen, die Fallzahl dieser qualitativ hochwertigen, aber teuren Umfrage zu erhöhen. Hierdurch lässt sich die Varianz der erwartungstreuen Schätzer auf Basis dieser reduzieren (Bethlehem/Cobben/Schouten 2009: 293). Es ändert sich also lediglich die Terminologie: Die PS stellen nicht mehr die Wahrscheinlichkeit der Treatmentzuweisung gegeben $\mathbf{X}$ dar, sondern die Wahrscheinlichkeit einer Person, innerhalb der Referenzumfrage ($T_i = 1$) befragt worden zu sein.

Der Vorteil dieser Herangehensweise der Modellierung der Wahrscheinlichkeit der Teilnahme an einer Referenzumfrage ist die Flexibilität. Diese Wahrscheinlichkeit der Teilnahme wird nicht als statisch begriffen, sondern ist dynamisch von bestimmten Faktoren abhängig und daher als logistisches Regressionsmodell der Teilnahme modellierbar (Olson 2006; Peytchev/Carley-Baxter/Black 2010; Fricker/Tourangeau 2011). Das inhaltliche Problem des Nonresponse lässt sich letztendlich hierdurch in ein statistisches Modellierungsproblem überführen. Qualitätsverbesserungen können hierdurch auch *nach* der Datenerhebungsphase durch die Bestimmung statistischer Gewichtungsverfahren ansetzen. Hier liegt der eigentliche Fokus und Diskussion der Anwendung statistischer Korrekturen zur Begegnung der Auswirkungen von Nonresponse auf die Schätzer von Umfragen (z.B. Kalton/Flores-Cervantes 2003; Särndal/Lundström

[7] Es können andererseits auch Web-Surveys genutzt werden, um die Qualität eines Referenzsurveys hierdurch noch weiter zu erhöhen. Elliot/Haviland (2007) zeigen einen solchen Fall.

2005; Arzheimer 2009; Valliant/Dever/Kreuter 2013).

Die Bestimmung der PS liefert die Möglichkeit der Schätzung der eigentlich unbekannten Wahrscheinlichkeit der individuellen Teilnahme einer Person. Sofern die *wahre* Wahrscheinlichkeit der Teilnahme für alle befragte Personen mit identischen Kombinationen auf **X** konstant ist, handelt es sich dann um einen „Missing at Random" (MAR)-Prozess des Ausfall (Bethlehem/Cobben/Schouten 2009: 266 f.). Auswirkungen bedingt durch Nonresponse und Noncoverage der anzugleichenden Umfrage lassen sich dann durch die Nutzung der PS als Gewichtungsfaktor beheben, da sie bekannt und hierdurch über Gewichtungskorrekturen behebbar sind. Je stärker sich hingegen die *geschätzte* Verteilung der Teilnahmewahrscheinlichkeit zwischen den beiden beteiligten Umfragen unterscheidet, umso stärker sind auch die Unterschiede zwischen diesen. Kann der Referenzumfrage ein hohes qualitatives Niveau bescheinigt werden, ist dies auf Verzerrungen der anzugleichenden Umfrage zurückzuführen.

Damit sich die Qualität einer verzerrten Umfrage also steigern lässt, müssen *drei Annahmen* zutreffen: Der Referenzdatensatz muss *erstens* der Anforderung der Durchführung inferenzstatistischer Verfahren genügen, also für die anvisierte Grundgesamtheit ein repräsentativer Teilausschnitt durch eine zufallsbasierte Auswahl darstellen sein, oder dies zu mindestens konditional der einbezogenen Variablen erlauben. Der Referenzdatensatz muss *zweitens* alle notwendigen Variablen enthalten, welche für den Teilnahmeprozess relevant sein können. *Drittens* müssen diese Informationen auch in dem anzugleichenden Datensatz vorliegen. Gilt dies, wäre der anzugleichende Datensatz über die Nutzung der berechneten Gewichtungsvariable anschließend repräsentativ für die zu Grunde liegende Grundgesamtheit des Referenzdatensatzes, konditional der genutzten Variablen zur Berechnung der PS.

Die zentrale Herausforderung liegt jedoch in der korrekten Modelspezifizierung zur Bestimmung dieser. Fehlspezifikationen können starke negative Auswirkungen haben (Smith/Todd 2005; Schafer/Kang 2008). Dies trifft meist zu, „[...] wenn die Wahrscheinlichkeit des Zugangs [...] von Variablen abhängt, über die im Referenzsurvey keine Informationen vorliegen." (Schnell 2012: 304). „Ob solche Verfahren [...] eine nachträgliche Rechtfertigung einer Stichprobe [...] erlauben, erscheint [...] für unabsehbare Zeit [daher] fraglich." (Schnell 2012: 305). Um ein geeignetes Modell zu finden, wird meist ein manuelles iteratives Vorgehen zum Auffinden eines solchen gewählt (Imai/Ratkovic 2014: 244). Gütekriterien in Form verschiedener Pseudo R^2 sind eine Möglichkeit herauszufinden, ob das Modell zur Erklärung der Unterschiede zwischen einer

vorliegenden Umfrage und einem Referenzdatensatz adäquat bestimmt wurde oder noch weitere Modifikationen notwendig sind (Eltinge 2002: 434).

Eine weitere Voraussetzung der erfolgreichen Modellierung von PS ist, dass sich überhaupt Überschneidungen der PS zwischen den beiden Gruppen übergeben, was praktisch nicht immer über die gesamte Bandbreite der Variablen der Fall ist. Dies kann eine Verzerrung hervorrufen, da diese Bereiche der Verteilungen nicht zusammengebracht werden können und häufig die Personen dieser Bereiche aus dem Matching ausgeschlossen werden (Guo/Fraser 2015: 209). Eine bessere Lösung kann sein, die an den Enden der Verteilungen platzierten Personen auszuschließen, wozu sich adäquate Regeln herleiten lassen, um einen Schwellenwert des Ausschluss in Abhängigkeit der vorliegenden Verteilung der PS zu erreichen (Crump et al. 2009: 188). Treffen diese Bedingungen zu, kann jedem Befragten die Inverse der vorhergesagten Wahrscheinlichkeit (also der PS) an der Referenzumfrage teilzunehmen, als Gewichtungsfaktor zugewiesen werden (z.B. Kalton/Flores-Cervantes 2003: 89). Dies stellt eine Alternative zur Nutzung von Zellkorrekturen dar (Little 1988: 293).

Die PS können aber auch die Grundlage sein für eine weitergehende Optimierung mittels der Anwendung statistischer Matchingverfahren (z.B. Luellen/ Shadish/Clark 2005: 532; Stuart 2010). Hierdurch sind die PS die Grundlage, um Befragte beider Umfragen in der Form zueinander bringen, dass die Unterschiede zwischen den beiden Gruppen sich noch stärker verringern lassen, bzw. die Verteilungen der beteiligten Variablen noch besser angeglichen werden können, als durch die Nutzung der PS als direkte Gewichte.

4.3.3. Von den „Propensity Scores“ zu den Matchingverfahren – Die Bedeutung des Algorithmus

Die Idee von Matchingverfahren ist für jede Person der Treatmentgruppe eine oder mehrere Personen der Kontrollgruppe zu finden, welche hinsichtlich der betrachteten Eigenschaften möglichst deckungsgleich mit dieser Person sind. Wird dies über den gesamten Datensatz durchgeführt, ist eine Abschätzung der kontrafaktischen Aussagen über den Treatmenteffekt möglich (Guo/Fraser 2010: 75 f.). Bei statistischen Matchingverfahren handelt es sich somit um „[...] broadly to be any method that aims to equate (or „balance“) the distribution of covariates in the treated and control groups.“ (Stuart/Lalongo 2010: 1). Je besser diese Angleichung gelingt, umso unabhängiger wird die Treatmentzugehörigkeit zwischen den Gruppen hinsichtlich der am Matching beteiligten Kovariaten.

Wenn die Balancierung optimal ist, dann lässt sich der Effekt der Treatmentzugehörigkeit schließlich als reine Mittelwertsdifferenz zwischen den beiden Gruppen erfassen. Dies wäre auch durch eine explizite Modellierung möglich, aber durch den erhöhten Inferenzanspruch dort weitaus fehleranfälliger (Ho et al. 2007). Eine Gewichtungsvariable kann ebenfalls die Folge der Anwendung eines Matchingverfahrens sein und deren Verwendung ist als ein analoges Vorgehen zur Gewichtung auf Basis der Auswahlwahrscheinlichkeiten zu verstehen (Guo/Fraser 2010: 131).

Die Anwendung einer Weiterverarbeitung der PS im Rahmen eines PSM hat eine exponentielle Verbreitung in den letzten Jahren erlebt (Diamond/Sekhon 2013: 932), insbesondere in der Medizin (Austin 2008a,b) und den Sozialwissenschaften (Thoemmes/Kim 2011). Sie ist daher ein äußerst beliebtes statistisches Verfahren zur möglichst guten Angleichung zweier Gruppen von Beobachtungen hinweg. Aus diesem Grund sind statistische Matchingverfahren auch in den gängigen Softwarepaketen implementiert, wobei neuere Verfahren tendenziell in R zuerst auftauchen (zur Übersicht z.B. Schuler 2015).[8]

Matchingverfahren unterscheiden sich danach, welcher Matchingalgorithmus genutzt wird (zur Übersicht z.B. Clark 2015: 115 f.). Diese variieren hinsichtlich der Definition, wann zwei Probanden/Umfrageteilnehmer sich „möglichst ähnlich“ sind. Die einfachste Variante liegt im „Exact Matching“ (EM): Bringe diejenigen Personen zusammen, welche hinsichtlich der gewählten Kovariaten eine perfekte Übereinstimmung zeigen und somit statistische Zwillinge sind. Hierdurch werden alle Unterschiede zwischen zwei Gruppen hinsichtlich der für das Matching verwendeten Variablen perfekt eliminiert (Ho et al. 2007: 217). Diese simple, einfache und bestmögliche Herangehensweise zur Balancierung hat aber eine starke Einschränkung, welche sie praktisch (häufig) unbrauchbar macht: Die Nummer der gefunden Paare reduziert sich sehr schnell, wenn das Matchingmodell komplexer wird. Hierdurch droht die Verzerrung wiederum zu steigen, da die Matchinglösung auf zunehmend weniger Paaren basiert. Bereits die Abweichung auf lediglich einer der beteiligten Variablen negiert eine exakte Übereinstimmung zwischen zwei Personen (Rosenbaum/Rubin 1985: 34). Gerade bei sozialwissenschaftlichen Fragestellungen zeigt sich bereits bei wenigen erklärenden Größen eine extrem große Bandbreite an möglichen Paarkonstellationen (Gangl 2010: 936 f.).

„Subklassifikation“ (SUB) ist ein möglicher Weg, diesem Dilemma zu ent-

[8] Ebenso bietet die Homepage von Elizabeth R. Stuart eine Übersicht über Implementierungen in statistische Softwarepakete: http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html, Abruf: 11. Mai 2018.

kommen. Personen werden hierzu in verschiedene Subklassen auf Basis verschiedener Bereiche der PS eingeordnet. Das Ziel ist nicht mehr die exakte individuelle Übereinstimmung zu erreichen, sondern die Personen so in Klassen einzuordnen, dass zwischen den beiden Gruppen die Verteilungen innerhalb der jeweiligen Subklassen so ähnlich wie möglich sind (Ho et al. 2011: 6). Dies ist analog zur Poststratifikation zu verstehen, welche nun aber auf Basis der PS an Stelle der ursprünglichen Variablen durchgeführt wird. Die Subklassifikation ist umso erfolgreicher, je geringer die Variation der Teilnahmewahrscheinlichkeit an der Referenzumfrage innerhalb der Subklassen ist (Bethlehem 2009: 267). Gegenüber der Poststratifikation wird durch die Subklassifikation auf Basis der PS jedoch das Problem der exponentiell wachsenden Zellen bei zunehmender Anzahl an Variablen gelöst, da die PS das Problem der hohen Dimensionalität auflösen (Guo/Fraser 2015: 204 f.).

Die Anzahl der Klassen unterliegt jedoch einem Dilemma: Mit zunehmender Anzahl dieser (also geringerer Intervallbreiten der Einordnung) wird die Homogenität innerhalb der Klassen immer besser erfüllt, die PS sind konstanter und die Verzerrung von Schätzern wird reduziert. Gleichzeitig steigt jedoch die Varianz wieder durch die erhöhte Anzahl an Klassen, da die resultierende Gewichtungsvariable mehr Ausprägungen aufweist. Die optimale Anzahl an Klassen ist daher studienabhängig und nicht klar vorgebbar (Guo/Fraser 2015: 207 f.). In Experimentalsituationen können aber bereits fünf Klassen ein sehr gutes Resultat sein (Cochran/Chambers 1965; Rosenbaum/Rubin 1984). Ergeben sich hinsichtlich der PS-Verteilungen der beiden Gruppen zudem zu geringe Übereinstimmungen, resultiert dies in ungleich großen Gruppen durch das SUB, was wiederum die Teststärke reduziert (Clark 2015: 116).

„Nearest Neighbour Matching“ (NNM)[9] versucht ebenfalls das Problem einer sich reduzierenden Anzahl möglicher Paare durch komplexere Matchingmodelle bei Anwendung eines EM zu umgehen. Hierzu werden einfach die beiden Probanden zusammengebracht, welche sich möglichst ähnlich – aber nicht zwingend vollkommen identisch – sind. Um zu verhindern, dass die beiden möglichst ähnlichen Personen sich trotzdem sehr stark unterscheiden, lassen sich „caliper“ definieren (Cochran/Rubin 1973: 432). Die beiden ähnlichsten Personen werden dann lediglich von einem vordefinierten Pool sich ähnlicher Personen innerhalb des jeweiligen „caliper“ gewählt. Als gängiges Kriterium des NNM hat sich hierfür das 0.25-fache der Standardabweichung der bestimmten PS bewährt (Guo/Fraser 2015: 146 f. Bai 2015: 81).

[9] In der Literatur auch oft als „greedy matching“ bezeichnet (z.B. Stuart 2010: 10).

„Optimal Matching“ (OM) wurde von Rosenbaum (1989) entwickelt, um das NNM zu verbessern. Dies unterscheidet sich nur in einem Aspekt von diesem: Anstatt lediglich zwei Objekte mit der geringsten Distanz zueinander zusammenzubringen, optimiert dieser Algorithmus auch die gesamte Matchinglösung fortlaufend. Das Ziel ist eine Matchinglösung zu finden, welche zusätzlich auch die Distanz aller Paare am Ende minimalisiert hat (Hansen 2004: 611 f.). Gu/Rosenbaum (1993) haben in ihrer Simulationsstudie bereits gezeigt, dass OM dem NNM überlegen ist, wenn die Größe der Kontrollgruppe ein vielfaches der Treatmentgruppe beträgt. Es scheint daher zu sein „[...] that optimal and greedy matching select more or less the same controls but assign them to different treated units, so distance within pairs is affected but balance is not.“ (Gu/Rosenbaum 1993: 418).

Sowohl NNM als auch OM erlauben eine mehrfache Verwendung der Kontrolleinheiten („Mit Wiederholung“). Gleichzeitig lässt sich die maximale Anzahl der Kontroll- zu Treatmentmatches begrenzen (Ho et al. 2011). Mit einer höheren Zahl an erlaubten Personen pro Person der Treatmentgruppe lässt sich die Varianz eines resultierenden Schätzers reduzieren. Wenn jedoch die nähesten Personen hierdurch immer unähnlicher werden, konterkariert dies den Effekt (Smith/Todd 2005: 315). Die Verwendung eines Matchens mit Wiederholung, also mehrfacher Nutzung der Kontrolleinheiten, scheint in Anwendungen einen höheren Grad der Bias-Reduktion zu erreichen, auch wenn hierdurch die Verwendung der Personen nicht mehr unabhängig voneinander ist (z.B. Dehejia/Wahba 2002; Bai 2015). Die mehrfache Verwendung impliziert, dass ein zusätzliches Bedeutungsgewicht angewandt wird. Gleichzeitig droht die Gefahr, dass die Realisierung von Schätzern sehr abhängig von einem kleinen Personenkreis aus der Kontrollgruppe ist, da diese mehrfach zugeordnet werden können (Stuart/Lalongo 2010: 11). Ob nun NNM, OM oder „Caliper Matching“ die bessere Variante ist, ist jedoch fallabhängig und nicht generalisierbar beantwortbar (Austin 2014).

„Full Matching“ (FM) ist eine flexiblere Variante des OM. Hierdurch lassen sich mehrere Personen der Kontrollgruppe einer der Personen der Treatmentgruppe *oder* vice versa zuordnen. Die exakte Zahl, welche zugeordnet wird, kann je nach vorliegenden Realisierungen der PS variieren. Praktisch verbessert dies die Matchinglösung, wenn sich die Wahrscheinlichkeit zur Treatmentgruppe zu gehören klar zwischen der Treatment- und Kontrollgruppe – gegeben der Kovariaten – variiert (Hansen 2007: 20 f.). FM ist daher eine Kombination von OM und einer automatisierten SUB-Variante, welche alle Beobachtungen im

Matchingprozess halten kann. Das einzig praktisch auftretende Problem ist die hohe Relation von Kontrollpersonen zu *einer* Treatmentperson, was sich in sehr hohen Gewichten der Angleichung äußern kann (Stuart/Lalongo 2010: 398 f.). Hansen (2004) schlagen daher vor, die Nummer der maximal *und* minimal zu matchenden Kontrollpersonen auf eine (Start)Wert von $(1/2,2) \cdot (1-\hat{p})/\hat{p}$ zu begrenzen, wobei $\hat{p}$ der relative Anteil der Treatmentgruppe ist (Hansen 2007: 21). Von dieser Kombination ausgehend kann dann in einem iterativen Vorgehen die beste Matchinglösung gefunden werden.

Auch das „Coarsened Exact Matching" (CEM) (Iacus/King/Porro 2009; Iacus/King/Porro 2011; Iacus/King/Porro 2012) versucht das Komplexitätsproblem des EM zu lösen und gleichzeitig den Idealzustand des 1:1 Matchens möglichst gut zu erreichen. Hierzu werden in einem zweistufigen Verfahren erst die Merkmalsausprägungen zu sinnvollen Klassen („strata") zusammengefasst („coarsened"-Schritt). Anschließend kann auf diese „strata" ein EM ausgeführt und Personen ohne einen sinnvollen Match aus dem Prozess durch ein Gewicht von Null ausgeschlossen werden. Dies geschieht, wenn in einer Klasse lediglich eine Person einer der beiden beteiligten Gruppen vorhanden ist. Alle anderen Personen erhalten an Hand ihrer Gruppenzugehörigkeit und der relativen Anteile dieser Gruppe in ihrer Klasse ein individuelles Gewicht. Auf Grundlage der zusammengefassten Klassen ergeben sich anschließend keine Unterschiede mehr zwischen den Verteilungen der beiden Gruppen. Ein solches Vorgehen ist gerade deshalb sinnvoll, da angewandte Likert-Skalen in Umfragen häufig auf ihre essentiellen Aussagen zusammengefasst werden, das „coarsening" kann also auf der Grundlage theoretischer Überlegungen passieren. Nach dem Matchingprozess werden wiederum die ursprünglich Variablen an der Stelle der zusammengefassten genutzt. Das CEM nutzt daher nicht die ursprünglichen Individualwerte, sondern Intervalle dieser für den Matchingprozess (Iacus/King/Porro 2012: 8 f.).[10]

Das Ziel ist die Lösung eines zentralen Problems der bisherigen Algorithmen zu umgehen: Matchingverfahren sind erfolgreich, wenn das korrekte PS-Modell spezifiziert wurde, also keine Beobachtungen ungewöhnlich weit von den restlichen entfernt an Hand der PS liegen und daher keine Zuordnung finden. Ob das Modell korrekt ist, kann jedoch nur auf Grundlage der vorhandenen

[10]Ein Beispiel sind die abgefragten Skalometer-Variablen in Wahlumfragen. Diese weisen typischerweise einen Bereich von -5 der kompletten Ablehnung bis zu +5 der kompletten Zustimmung auf. Diese Information kann von 11 möglichen Ausprägungen auf die drei Antwortmöglichkeiten „unsympathisch", „neutral" und „sympathisch" zusammengefasst werden, was ein EM sehr stark erleichtert. Die einzige Entscheidung hierbei ist die Definition der Grenzen der drei Kategorien an Hand der ursprünglichen Skala.

Möglichkeit der Balancierung durch die vorliegenden PS bestimmt werden. Ob die Balancierung jedoch erfolgreich ist, lässt sich wiederum nur auf Grundlage des korrekt spezifizierten PS-Modells sagen, da nur dann außergewöhnliche Beobachtungen bekannt und ausschließbar sind. Da dies nicht möglich ist, wird der Schritt meist übersprungen, ein PS-Modell geschätzt und die Balancierung angegeben. Der Ausschluss weit entfernter Beobachtungen sorgt jedoch für eine geringere Modellabhängigkeit des Matchingprozesses und kann daher das Ergebnis verbessern (Iacus/King/Porro 2012: 4 f.).

Das CEM stellt daher einen einfachen und intuitiven Weg bereit, sich zu stark von allen anderen Beobachtungen unterscheidende Personen durch die Anwendung des EM auf Basis klassierter Intervallbereiche der beteiligten Variablen durchzuführen. Dies schließt sehr außergewöhnliche und das PS-Modell verzerrende Personen automatisiert aus und benötigt daher nicht einen solchen notwendigen Vorausschluss, welcher praktisch kaum angewandt wird (Iacus/King/Porro 2012: 11). CEM erweist sich auch bei einer hohen Anzahl an Variablen praktisch umsetzbar, da häufig mehrere zur Entscheidung genommene Variablen stark korrelieren und daher einen ähnlichen Bereich der PS generieren. Ebenfalls ist das CEM äußerst schnell in der Berechnung (Iacus/King/Porro 2012: 13).[11]

Das „Genetic Matching“ (GM) geht auf Diamond/Sekhon (2013) zurück, welches „[...] eliminates the need to manually and iteratively check the propensity score. GenMatch uses a search algorithm to iteratively check and improve covariate balance [...]“ (Diamond/Sekhon 2013: 932). Um dies zu erreichen, wird in einem iterativen Vorgehen nach einer Kombination von einem gewichteten Eingehen der Kovariaten selbst gesucht. Hierzu wird eine festgelegt „Verlustfunktion“ auf Basis des spezifizierten Distanzmaß zweier Personen minimiert (Diamond/Sekhon 2013: 934 f.). Zwar führt das iterative Vorgehen zu einer nicht-konstanten Berechnung der resultierenden Gewichte bei einmaliger Berechnung und kann daher variieren, der Algorithmus konvergiert jedoch asymptotisch in einer stabilen Lösung (Mebane/Sekhon 2011). Die Implementierung der „GenMatch“-Funktion in R bietet dazu mehrere mögliche Funktionen, welche minimiert werden können.[12] Das Ziel von GM ist den iterativen Part der Identifizierung eines guten PS-Models zu automatisieren, um das häufig wenig transparente Vorgehen der manuellen Korrektur zu umgehen (z.B. Sekhon 2011; Iacus/King/Porro 2012). Das GM ist eine Verallgemeinerung des Matchens auf der Basis der Mahalanobis-Distanz auf Grundlage einer Matrix an Kovariaten $\mathbf{X}$.

[11] Nähere Beschreibung unter http://gking.harvard.edu/cem (Abruf: 11. Mai 2018).

[12] Nähere Beschreibung: http://sekhon.berkeley.edu/matching/GenMatch.html (Stand: 11. Mai 2018.

Für zwei Personen wird die Distanz zueinander daher klassisch ausgedrückt als:

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)}$$

Sie folgen dann als ersten Schritt der Empfehlung von Rosenbaum/Rubin (1985) und nutzen *zusätzlich* geschätzte PS als „Vorabmodell" in **X**. Der zweite Schritt ist dann der iterative Schritt: Die MD wird durch eine Gewichtungsmatrix W verallgemeinert:

$$GMD(X_i, X_j, W) = \sqrt{(X_i - X_j)^T (S^{-1/2})^T W S^{-1/2} (X_i - X_j)}$$

Die Gewichte der Kovariaten werden an Hand der relativen Bedeutung dieser Variablen für die Balancierung bestimmt und starten alle mit einer Null. Bekommt nur der Vektor der PS ein von Null abweichende Gewichtung, liegt das „klassische" PSM vor. Bekommen alle außer den PS einen von Null abweichenden Wert, liegt das MDM vor. In der Praxis wird die Lösung in der Regel zwischen diesen beiden Enden liegen (Diamond/Sekhon 2013: 934). Ein zentraler Nachteil des GM ist jedoch die sehr hohe Rechenintensität, welche die Anwendung dieses Algorithmus für den empirischen Teil der Arbeit unmöglich machen.[13]

Aber welcher Matchingalgorithmus ist nun die adäquate Wahl? Wo findet sich die beste Balancierung der beiden Umfragen durch die Angleichung einer vorliegenden Umfrage an eine qualitativ hochwertigere Referenzumfrage? Hierzu werden Qualitätskennzahlen benötigt, welche diese Balancierung messen. Diese werden nachfolgend vorgestellt.

4.3.4. Balancierung als Gütekriterium für Matchingalgorithmen

Zentral für die Güte eines Matchingprozesses ist, ob hinsichtlich der gewählten Kovariaten eine ausreichende Balancierung zwischen den Gruppen durch den Zuordnungsprozess erreicht wird. Auf diesen Variablen sollten also möglichst wenige Unterschiede hinsichtlich der Verteilungen zwischen diesen vorliegen. Ein zweites Kriterium ist, dass möglichst viele Beobachtungen auch nach dem durchgeführten Matchingprozess über die resultierenden Gewichte in der Analyse gehalten werden können und nicht ohne Gegenpart der jeweiligen anderen Gruppe durch den Algorithmus ausgeschlossen werden. Dann ist eine Redu-

[13] Die Durchführung eines Matchingprozess auf der Basis von 100 „Generationen" dauert teilweise mehrere Stunden. Da innerhalb des empirischen Teils der Arbeit ca. 1500 verschiedene Umfragen hinsichtlich ihrer Abweichungen zu den jeweiligen Referenzdatensätzen analysiert werden, lässt sich ein GM daher zeitlich nicht umsetzen.

zierung der Modellabhängigkeit möglich und ein geringeres Potential für einen Bias, eine geringe Varianz von Schätzern und ein niedriger MSE erreicht. Die eigentlich interessierende Zielvariable hat während des Matchingprozess keine Relevanz. Die Auswahl der besten Matchinglösung durch verschieden denkbare Korrekturmodelle ergibt sich rein auf der Evaluation der Balancierung der beteiligten Kovariaten (Ho et al. 2007: 216). Daher gilt: „The goal of matching is to create a data set that looks closer to one that would result from a perfectly blocked (and possibly randomized) experiment. [...] we need the distribution of covariates to be the same within the matched treated and control groups.“ (Ho et al. 2011: 3).

Mit den „Standardized Differences“ (SD) (Rosenbaum/Rubin 1985: 34):

$$SD = 100 \cdot \frac{\bar{x}_{(t=1)} - \bar{x}_{(t=0)}}{\sqrt{(s_{(t=1)} + s_{(t=0)}) \cdot 0.5}} \tag{4.5}$$

existiert ein populärer Indikator zur Messung der Güte der Balancierung auf den jeweiligen Variablen – manchmal auch direkt als „standardisierter Bias“ bezeichnet (z.B. Pan/Bai 2015: 9; Gangl 2010: 946). Ein Unterschreiten dieser Kennzahl unterhalb eines Schwellenwert von nur noch 5 % verbliebener standardisierter Unterschiede durch den Matchingprozess kann als erfolgreich betrachtet werden (Caliendo/Kopeinig 2008). Für eine dichotome Variable mit den beiden Anteilswerten $p_1(X_k)$ und $p_0(X_k)$ ergibt sich für den „Standardized Bias“ (SB) als Bestimmung :

$$SB = \frac{p_1(X_k) - p_0(X_k)}{\sqrt{\frac{p_1(X_k)(1-p_1(X_k)) + p_0(X_k)(1-p_0(X_k))}{2}}} \cdot 100\ \% \tag{4.6}$$

Die Kennzahl der „Percent Bias Reduction“ (PBR) (Cochran/Rubin 1973: 420) wiederum bietet die Möglichkeit der Angabe einer prozentualen Reduzierung der Verzerrung B einer Variable durch das Matching (Pan/Bai 2015: 9):

$$PBR = \frac{B_{\text{before matching}} - B_{\text{after matching}}}{B_{\text{before matching}}} \cdot 100\ \% \tag{4.7}$$

Die eigentliche Debatte der Nutzung von Balancierungskennziffern dreht sich darum, ob die jeweils genutzte Kennzahl maximiert oder als statistischer „Goodness of Fit“-Test formuliert werden soll (Hansen/Bowers 2008). Statistische Tests auf Unterschiede hinsichtlich der arithmetischen Mittel stellen jedoch kein geeignetes Mittel dar, um die erreichte Balancierung zu evaluieren (z.B.

Imai/King/Stuart 2008). Die Balancierung ist vielmehr eine Eigenschaft der vorliegenden Stichproben und keiner theoretischen Population die anvisiert wird. Daher gilt: Je besser die Balancierung ist, umso besser ist sie. Eine definierte Schwelle (z.B. durch $\alpha = 0.05$) suggeriert hingegen eine künstliche Barriere, welche nicht notwendig ist. Weiterhin ist die Anwendung eines statistischen Tests zur Definition einer adäquaten Schwelle der Balancierung abhängig von der Anzahl der befragten Personen (Ho et al. 2007: 220 ff.). Zur Evaluierung müssten auch aufwendigere Testverfahren, wie der „Kolmogorov-Smirnov"-Test, angewandt werden, wenn die Verteilung insgesamt im Fokus stehen soll (Abadie 2002: z.B.).

Dies führt zu der praktischen Lösung der zusätzlichen Bestimmung der Evaluierung der Balancierung über beispielsweise „empirical quantile-quantile plots" (eQQ) (Ho et al. 2007: 220). Die standardisierte Differenz ist dabei das eigentlich gängige Mittel um die Balancierung zu bestimmen (Ho et al. 2011: 21 f.). Jedoch sollte die Evaluierung darüber hinausgehen und neben der Differenz der standardisierten arithmetischen Mittel der beiden Gruppen auch weitere Aspekte der Verteilung einbeziehen (Iacus/King/Porro 2012: 7). Ist eine Balancierung der relevanten Kovariaten erreicht, kann die Modellabhängigkeit überwunden werden. Diese liegt vor, da das Standardvorgehen der Modellschätzung in den empirischen Sozialwissenschaften rein parametrisch orientiert ist: Viele mögliche Modellierungen liegen vor, die letztendliche Entscheidung präsentiert aber lediglich ein Modell (Ho et al. 2007: 214).

Matchingverfahren geht es also nicht primär um das Zusammenbringen von Untersuchungseinheiten: „[...] matching does not require pairing observations [...]" (Ho et al. 2007: 212). Zentral ist das Ziel die Verteilung ausgewählter Kovariaten der beiden definierten Gruppen möglichst gut anzugleichen. Werden Matchingverfahren in empirischen Studien genutzt, ist die mangelnde Offenlegung der erreichten Balancierung in vielen Bereichen ein Problem (z.B. Austin 2008a; Diamond/Sekhon 2013: 932). Fehlanwendungen und -schlüsse hinsichtlich der Nutzung von Matchingverfahren sind daher weitverbreitet (z.B. Imai/King/Stuart 2008). Dies ist eine Folge der immer weiteren Verbreitung und Nutzung von Matchingverfahren in populären Wissenschaftsbereichen außerhalb der mathematischen Statistik, bei einer nicht unerheblichen Variation an Handlungsempfehlungen zur Durchführung und Anwendung von solchen Verfahren (Stuart/Lalongo 2010). Gleichzeitig existieren für andere zentrale Herausforderungen kaum bisher gesicherte Empfehlungen. Für die Berücksichtigung komplexer Erhebungsdesign, welche eine Design-Korrektur notwendig machen,

finden sich lediglich erste Vorschläge, wie hier mit den ursprünglichen Design-Gewichten während des Matchingvorgangs umgegangen werden sollte (DuGoff/ Schuler/Stuart 2014).

Ist das PS-Modell jedoch korrekt spezifiziert, lässt sich auch die „Equal Percent Bias Reduction“ (EPBR) Annahme halten (Sekhon 2011: 3). Ist die Annahme der EPBR nach Rubin (1976a,b) erfüllt, dann erhält man durch das Matching „[...] the same percent reduction in bias for any linear function of the **X** [...]“ (Rubin 1976a: 110). Jedoch wenn eine „[...] matching method is not EPBR for X, then the matching substantially increases the bias for some linear functions of X, even if all univariate means are closer [...]“ (Rubin 1976a: 110) nach der Anwendung des Matchings.

Die EPBR ist bei ellipsoiden Verteilungen gegeben (Rubin/Thomas 1992b), beispielsweise kann eine multivariate Normalverteilung von **X** bei ausreichend hoher Anzahl an Kontrolleinheiten genauso effizient in der Biasreduktion sein, wie das (eigentlich unbekannte) wahre PS-Modell der Population (Rubin/Thomas 1992a). Wenn sich die Annahme der EPBR aber nicht halten lässt, ergeben sich Kombinationen der Kovariaten von **X** mit einem höheren Bias als vorher. Praktisch lässt sich diese Annahme aber meist nicht halten. Ein PSM sorgt dann aber nur für keine zusätzlich mögliche Verzerrung, wenn das wahre PS-Model bei ausreichend großer Stichprobengröße geschätzt wird (Sekhon 2011: 2 f.).

Für einen Erfolg des empirischen Teils der Angleichung von Umfragen über ein PSM an vorligende Referenzumfragen ist somit zentral von dem angewandten Korrekturmodell abhängig. Ebenso ist essentiell, dass der angewandte Algorithmus auch in der Lage ist, eine Balancierung zwischen diesen beiden Umfragen zu erreichen. Ist das Ziel die Angleichung von onlinebasierten Befragungen, ist dies eine Herausforderung, da da die in solchen Surveys vorliegenden Kovariaten „[...] anscheinend bisher weder die Verfügbarkeit der Internetanschlüsse noch die tatsächliche Teilnahme an diesen Surveys.“ (Schnell 2012: 304 f.) erklären können. Aus diesem Grund wird das nachfolgende Kapitel 7 nun dazu übergehen herausstellen, was eine gute Information zur Korrektur in Form von Variablen ausmacht. Ebenso werden Varianten von Variablen herausgestellt, welche für Wahlumfragen essentiell sind, durch eine hohe Stabilität gekennzeichnet und somit für eine Korrektur prädestiniert sind. Zudem dürfen diese Variablen nicht selbst eine inhaltlich interessierende Information darstellen (wie z.B. das Wahlverhalten), da eine Beteiligung am Matchingprozess die Analyse dieser bei einer Anwendung der resultierenden Gewicht unmöglich macht.

5. Datengrundlage der Arbeit

Die nachfolgenden Unterkapitel beschäftigen sich mit der Datengrundlage der Arbeit. In Kapitel 5.1 werden zunächst die, im Zentrum der Analyse stehenden, Komponenten der „German Longitudinal Election Study“ (GLES)-Studie der Erhebungen 2009 und 2013 vorgestellt. Hierdurch wird der Bereich der wissenschaftlichen Wahlforschung abgedeckt; die GLES offeriert Datensätze aller drei zentralen Befragungsformen der Umfrageforschung. Für die langfristige Einschätzung der Qualität verschiedener Erhebungsverfahren sind Datensätze weiterer bevölkerungsweiter Umfragen notwendig, welche bereits seit einem längeren Zeitraum erhoben werden. Das Kapitel 5.3 stellt als Beispiele persönlicher Interviews die beiden prominenten bevölkerungsweiten Umfragen der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (Allbus) und des „European Social Survey“ (ESS) vor. Anschließend werden die Umfragen von Forsa und Politbarometer, als Vertreter kommerzieller Wahlumfragen durch telefonische Bevölkerungsumfragen, in Kapitel 5.2 präsentiert. Schließlich stellt das Kapitel 5.4 den Mikrozensus und die „Repräsentative Wahlstatistik“ des „Bundeswahlleiter“ (BuWa) vor. Im Laufe der Arbeit dienen diese Datensätze der amtlichen Statistik hinsichtlich ihrer reportierten aggregierten Randverteilungen als eine vielfache Referenzverteilung für die betrachteten Umfrageprogramme. Das Kapitel schließt mit einer Gesamtübersicht der verwendeten Datensätze und der Herausarbeitung ihrer zentralen Charakteristika.

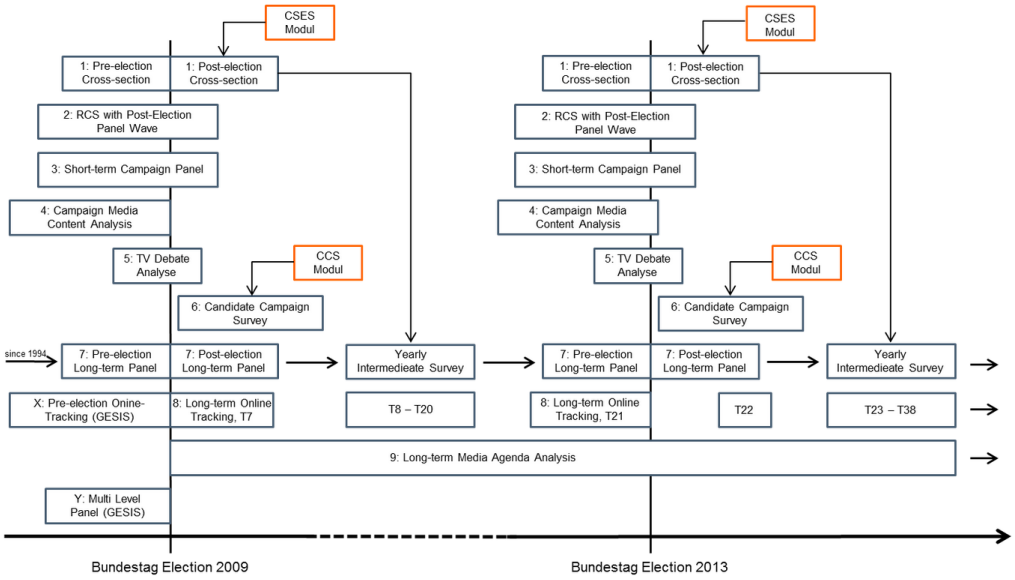
5.1. Multikomponentenstudie: Die „German Longitudinal Election Study“ (GLES)

Das Design der GLES setzt sich aus diversen Komponenten zusammen,welche auf unterschiedlichsten Erhebungsmodi und spezifischen Zielen basieren. Sie haben jedoch gemeinsam, dass sie die (mehr oder weniger stark abgegrenzte) wahlberechtigte Bevölkerung der Bundesrepublik im Kontext der drei Bundestagswahlen 2009, 2013 und 2017 (kommend) erfassen wollen. Die Komponenten weisen (meist) einen gemeinsamen Kernfragenkatalog und einen ähnlichen zeit-

lichen Erhebungsrahmen auf, welche eine Vergleichbarkeit der Komponenten untereinander ermöglicht.[1] Sofern ein identischer Erhebungszeitraum und eine vergleichbare Frageformulierung zwischen den jeweils betrachteten Komponenten vorliegt, oder mögliche Abweichungen als nur geringfügig herausgestellt werden können, sind Unterschiede der Zusammensetzungen der jeweiligen Stichproben auf unterschiedliche Einflüsse der in Kapitel 2 angesprochenen Fehlerquellen zurückführbar.

Abbildung 5.1 zeigt den Aufbau der verschiedenen Komponenten der GLES. Genutzt werden die Komponenten 1, 2, 3 und 8, welche nachfolgend vorgestellt werden.

Abbildung 5.1.: Aufbau der Komponenten der GLES



Quelle: http://gles.eu/wordpress/design/, Abruf: 11. Mai 2018.

Der **Vor- und Nachwahl-Querschnitt** der Komponente 1 – der Kern der GLES-Studie – basiert auf persönlich-mündlichen Befragungen. 2003 Personen wurden im Vorfeld der Bundestagswahl 2013 erfolgreich befragt, 2173 sind dies in 2009 gewesen. Auch im Anschluss an die jeweilige Wahl wurden erneut Stichproben aus der wahlberechtigten Bevölkerung ausgewählt und befragt.[2] Durch das ADM-Design wurde eine dreistufige räumliche Stichprobenziehung angewandt; die anvisierte Grundgesamtheit ist die wahlberechtigte Bevölkerung der Bundesrepublik ab 16 Jahren. Die Bevölkerung der neuen Bundesländer ist

[1] Quelle und Beschreibung des Designs: http://gles.eu/wordpress/design/ (Abruf: 11. Mai 2018). Für eine ausführlichere Projektbeschreibung siehe auch die Homepage der GLES unter gles.eu (Abruf: 11. Mai 2018.).

[2] Beschreibung unter: http://gles.eu/wordpress/design/querschnitt/ (Abruf: 11. Mai 2018).

hierbei bewusst überquotiert worden, eine Design-Korrektur bei einer Aussage über das gesamte Elektorat muss daher angewandt werden (Rattinger et al. 2011, 2014a). Nachfolgend werden die Erhebungen als *Komponente 1* bezeichnet. Die Feldzeit der Vorwahlbefragung 2009 ist der 10.08.-26.09.2009, für 2013 der 29.07.-21.09.2013. Die Nachwahlbefragungen wurden 2009 vom 28.09.-23.11.2009 und 2013 vom 23.09.-23.12.2013 durchgeführt.[3]

Die **RCS-Wahlkampfstudie** der Komponente 2 ist als telefonische Befragung konzipiert und wird als ein zwei Wellen umfassendes Paneldesign im Vor- und Nachgang der jeweiligen Bundestagswahl durchgeführt. Zur Erhebung wurde das „Rolling-Cross-Section" (RCS)-Design (z.B. Johnston/Brady 2002) genutzt. Für die Bundestagswahl 2009 wurden 6008 Interviews gleichmäßig auf 60 Zufallstagesstichproben im Rahmen der Vorwahlbefragung aufgeteilt, 2013 wurden 7882 Personen in 76 Tagesstichproben befragt. Zur Auswahl der zu befragenden Personen wurde das ADM-Telefonstichprobensystem nach dem Gabler/Häder-Modell angewandt, welches eine mehrstufige Zufallsziehung gewährleistet. Kommen mehrere Personen eines ausgewählten Haushalts für die Erhebung in Frage, wurde die „Last-Birthday-Methode" zur Auswahl genutzt. Die Definition der Grundgesamtheit umfasst die jeweils wahlberechtigte Bevölkerung der Bundesrepublik, welche in Privathaushalten mit mindestens einem Festnetzanschluss lebt (Rattinger/Roßteutscher/Schmitt-Beck/Weßels 2013; Rattinger et al. 2014b). Personen ohne einen solchen konnten bei der Erhebung also nicht berücksichtigt werden. Die Vorwahlbefragungen wurden 2009 vom 29.07. bis zum 26.09.2009 durchgeführt, 2013 im Zeitraum 08.07. bis zum 21.09.2013. Für spätere Analysen wird die jeweilige Erhebung im Vorfeld der Bundestagswahl genutzt und als *Komponente 2* bezeichnet.

Das **Kurzfrist-Wahlkampfpanel** der Komponente 3 ist als standardisierte Onlineumfrage konzipiert. Hierdurch lässt sich flexibel und schnell über einen kurzen Zeitraum eine Panelbefragung von mehreren Wellen im Vorfeld der jeweiligen Bundestagswahl realisieren. 2009 wurden 4552 Personen in der Rekrutierungswelle des Panels erfolgreich erfasst, 2013 sind dies 5200 Personen.[4] 1030 der Teilnehmer der Befragung 2013 wurden zudem bereits in 2009 befragt, welche von der Analyse ausgeschlossen wurden. Dies soll Effekten einer möglichen Professionalisierung durch die erneute Befragung vorbeugen, da diese Personen bereits mit einem vergleichbaren Fragebogen in 2009 konfrontiert wurden. Die Grundlage der Auswahl der Komponente 3 sind die registrierten

[3] Quelle: http://www.gesis.org/wahlen/gles/daten/ (Stand: 11. Mai 2018).
[4] Quelle: http://gles.eu/wordpress/design/wkp/ (Abruf: 11. Mai 2018).

wahlberechtigten Mitglieder des Access-Panel der Respondi AG.[5] Die Panelisten werden überwiegend aktiv über onlinebasierte Quellen rekrutiert, die Teilnahme an (eingeladenen) Umfragen wird incentiviert. Respondi betreibt ein eigenes Qualitätsmanagement und eine geringe Einladefrequenz der registrierten Mitglieder. Dies soll möglichen Paneleffekten und einer Professionalisierung der Teilnehmer vorbeugen (Rattinger et al. 2015c: 10 ff.). Ein fehlendes Antwortverhalten über zwölf Monate, eine Doppelanmeldung oder bewusste Falschangaben führen zu einem Ausschluss aus dem Access-Panel. Hierdurch soll, neben einer nicht weiter angegebenen massvollen Einladefrequenz, die Qualität verbessert werden. Ca. 15 % der registrierten Panelisten scheiden jährlich, nach einer durchschnittlichen Verweildauer von 1.5 Jahren, aus dem Panel aus. Die aktive Rekrutierung soll Selbstselektionseffekten vorbeugen, findet jedoch (beinahe) ausschließlich über Onlinewerbung, Suchmaschinen und die Facebook-Homepage des Betreibers statt (Rattinger et al. 2015b: 8). Genutzt wird zudem die Rekrutierung über bestehende Teilnahmen an anderen Online-Umfragen. Hierdurch soll eine am Thema motivierte Teilnahme gegenüber der Perspektive auf eine rein monetäre Aussicht dominieren (Rattinger et al. 2016a: 8).

Für das Jahr 2009 umfasst das Respondi-Panel ca. 65 000 Personen, 2013 sind es bereits 96 445. Die Stichproben werden über eine Quotenauswahl auf Grundlage der Merkmale Geschlecht, Alter und Bildung bewusst gesteuert.[6] Im Vorfeld der Bundestagswahl 2013 wurden zusätzlich drei Kontrollquerschnitte nach identischen Kriterien durchgeführt. Originäres Ziel dieser Kontrollgruppen (jeweils ca. n=1000) aus dem Respondi-Panel ist die Analyse von Paneleffekten der Befragten innerhalb des eigentlichen Wahlkampfpanels.[7] Die Kontrollquerschnitte der dritten (ZA5753) und fünften Welle (ZA5754) wurden vor der Bundestagswahl erhoben, derjenige der siebten Welle (ZA5755) nach der Wahl. Für spätere Analysen wird die jeweilige erste Rekrutierungswelle des Wahlkampfpanels 2009 und 2013 als ein eigenständiger Querschnitt aufgefasst. Der Panelcharakter hat somit im Rahmen der Arbeit keine Bedeutung. Die späteren Kontrollquerschnitte des Jahres 2013 stellen ebenfalls eigenständige Querschnitte dar.

2009 wurde die erste Welle der Befragung vom 10.07. bis zum 20.07.2009 durchgeführt (Steinbrecher/Roßmann/Bergmann 2015: 9), 2013 vom 20.06. bis zum 07.07.2013. Der „Kontrollquerschnitt“ (KQ) I wurde vom 01.08. bis

[5] Selbstbeschreibung durch Respondi unter: http://www.respondi.com/de/access-panels-leistungen (Abruf: 11. Mai 2018).

[6] Beschreibung der Datensätze ZA5704 (2013) und ZA5305 (2009) der GESIS.

[7] Beschreibung: http://gles.eu/wordpress/design/wkp/, Abruf: 11. Mai 2018.

zum 11.08.2013, der KQ II vom 02.09 bis zum 12.09.2013 erhoben. Die Durchführung des KQ III wurde vom 24.09. bis zum 03.10.2013 als Nachwahlbefragung erhoben (Rattinger et al. 2016a: 6). Für 2009 ergibt sich eine gesamte Feldzeit vom 10.07-20.07, für 2013 vom 20.06-12.09, exklusive der Nachwahlbefragung des KQ III. Die Durchführung 2009 hat somit gegenüber den Komponenten 1 und 2 der GLES den Nachteil einer noch zwei Monate entfernten Bundestagswahl zum Ende der Feldlaufzeit. Für 2013 ergibt sich in Kombination mit den KQ I+II eine vergleichbare Feldlaufzeit zu den anderen beiden Komponenten. Hier ist zudem durch die Unterteilung ermittelbar, ob die Entfernung der Wahl einen Einfluss auf die Qualität der Erfassung inhaltlicher Zielvariablen hat. Nachfolgend wird diese Datenquelle als *Komponente 3* bezeichnet. Die Rekrutierungswelle des eigentlichen Wahlkampfpanels wird mit *W1* kenntlich gemacht, die Kontrollquerschnitte in 2013 mit *KQ I*, *KQ II* und *KQ III*.

Auch das **Langfrist-Online-Tracking** der Komponente 8 basiert auf standardisierten Onlineumfragen. Diese Komponente wird, im Gegensatz zu den bisherigen, fortlaufend seit 2009 erhoben.[8] Auch die Komponente 8 wird als Stichprobe aus einem Access-Panel gezogen. Für die Analyse der Qualität von Onlineumfragen auf der Basis von Access-Panels ergibt sich eine interessante Aufteilung: Bis einschließlich des „Tracking“ (T) T16[9] im Dezember 2012 (ZA5349), wurde zur Rekrutierung das Access-Panel von Respondi genutzt. Seit der Umfrage T17 basieren die Stichproben jedoch auf dem Access-Panel von LINK.[10] Innerhalb der ersten 16 Erhebungen variiert die Art der Rekrutierung in das zu Grunde liegende Access-Panel von Respondi ebenso. Die ersten acht Erhebungen nutzen mit einem Anteil von 87 % am häufigsten die Rekrutierung über die eigenen Themenportale „sozialand/demandi“. Die Erhebungen 9 bis 12 setzen zu 61 % auf Online-Werbung und zu 31 % auf die Themenportale. Ab der Erhebung 13 bis 16 kommt es, neben der konstanten Nutzung der Online-Werbung, zu einer Nutzung der Empfehlungen bestehender registrierter Personen an weitere Personen (10 %), Rekrutierungen über die Facebook-Seite (5 %) und die Nutzung des „Mingle Trend Blogs“ (15 %). CATI-Rekrutierungen

[8] Durchschnittlich wird viermal pro Jahr eine Befragung von ca. $n = 1000$ durchgeführt. Lediglich in 2009 wurden acht und in 2012 nur zwei Befragungen durchgeführt. Zur Übersicht: http://www.gesis.org/wahlen/gles/daten/, Abruf: 11. Mai 2018.

[9] Die Bezeichnung der Nummern der Erhebung als „Tracking“ wird in Übereinstimmung mit der GLES-Projektseite übernommen.

[10] Quelle: Jeweilige Studienbeschreibungen der „Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen“ (GESIS).

machen lediglich 2 % aus, wobei sie auch nur bis T12 genutzt werden.[11] Respondi stellt somit Stichproben aus einem onlinerekrutierten Access-Panel bereit. Personen ohne Internetzugang haben keine Chance der Berücksichtigung, Personen mit einem solchen müssen diesen zudem auch aktiv nutzen, um Ziel der Rekrutierung über die geschaltete Onlinewerbung oder das Themenportal in das Access-Panel zu werden.

Die Grundgesamtheit der Erhebungen von T17 bis T33 (aktuell) seit 2012 sind die ca. 40 0000 Personen des „LINK Internet Panel“. Diese müssen volljährig sein und die deutsche Staatsbürgerschaft besitzen. Der Pool aus ca. 45 000 Personen in der Bundesrepublik wird, im Gegensatz zu Respondi, ausschließlich aktiv offline über eine zufallsbasierte Auswahl durch Telefonstichproben ausgewählt; eine Selbstregistrierung ist nicht möglich. Hierdurch soll ebenfalls Selektionseffekten vorgebeugt werden.[12] Die Auswahl erfolgt auf der Grundlage des ADM-Designs für Telefonstichproben, teilweise in Form eines Dual-Frame-Ansatzes. Die Qualität des Panels soll zusätzlich durch den Ausschluss von Personen mit widersprüchlichen oder zu schnellen Antworten, oder bei einer Diskrepanz zwischen der Angabe von soziodemografischen Merkmalen bei der Registrierung und späteren Angabe in Umfragen, sichergestellt werden. Mitglieder des Panels werden höchstens einmal im Monat zu Befragungen eingeladen, wobei wiederholte Befragungen zu identischen Themen ausgeschlossen werden. Auch geben lediglich 10 % der Mitglieder an, auf anderen Access-Panels registriert zu sein. Insgesamt kommt es pro Jahr zu einer Fluktation von ca. 20 % des Panels, bedingt durch den Ausschluss oder Ausfall von Personen und die Ersetzung durch neue Panelmitglieder (Rattinger et al. 2015a: 6). Die Teilnahme an einer Umfrage wird monetär entlohnt und für die Stichprobenziehung eine, dem Access-Panel von Respondi vergleichbare, Quotierung nach Geschlecht, Alter und Bildungsabschluss angewandt (Rattinger et al. 2015c: 8 ff.). In den beiden vorliegenden Access-Panels werden die Incentivierungen ohne Bedingungen ausgezahlt und sind im Vorfeld bekannt. Im Respondi-Panel entspricht dies ca. 0.06 € pro Minute, auszahlbar ab einem Betrag von 20 € (Rattinger et al. 2015b: 4). Im LINK Internet-Panel wird der Betrag von ca. 3.50 € für eine halbstündige Befragung direkt im Anschluss an die Befragung als Amazon-Gutschein offeriert (Rattinger et al. 2016b: 6). Die Incentivierung im Access-Panel von LINK ist somit ca. doppelt so hoch und wird zudem schneller ausgezahlt.

Durch die Verwendung der beiden Access-Panel von Respondi und LINK

[11] Quelle der Angaben: Jeweilige Studienbeschreibungen der Erhebungen, beziehbar über die GESIS. Stand der Zahlen: 11. Mai 2018.

[12] Quelle: http://tinyurl.com/jn4lawu (Abruf: 11. Mai 2018).

durch die Komponente 8 bietet sich somit die Möglichkeit der Analyse des Einflusses der Rekrutierungsart (Offline vs. Online) auf die Zusammensetzung und Qualität onlinebasierter Befragungen auf der Basis von Access-Panels. Werden durch die Art der Rekrutierung durch beide Betreiber Personen ohne einen Internetzugang nicht berücksichtigt, ist im von Respondi betriebenen Access-Panel zudem die Wahrscheinlichkeit der Rekrutierung abhängig von der aktiven Nutzung des Internets. Personen mit einer hohen Nutzungsfrequenz haben eine höhere Wahrscheinlichkeit das Ziel der online geschalteten Werbung zu werden. Diese ist zudem nicht bekannt. Für das LINK-Panel trifft dies nicht zu, da durch die zufallsbasierte Auswahl auf der Basis von Telefonstichproben allen Personen eine Auswahlwahrscheinlichkeit zugeordnet werden kann.

Insgesamt deckt die Multikomponentenstudie der GLES bereits in einem breiten Spektrum persönliche, telefonische und onlinebasierte Befragungen ab, welche sich auf Grund ähnlicher Erhebungszeiträume vergleichen lassen. Mit der Komponente 8 bietet sich zudem eine kontinuierliche Abdeckung der Zeiträume zwischen den beiden Bundestagswahlen 2009 und 2013. Die Erhebungen der GLES stellen jedoch nicht die einzigen frei verfügbaren Wahlumfragen innerhalb der Bundesrepublik dar, die Forschungsgruppe Wahlen und Forsa stellen ebenso die Daten ihrer telefonisch befragten Personen zur freien wissenschaftlichen Nutzung über das Datenarchiv der GESIS bereit.

5.2. Telefonische Wahlumfragen: Politbarometer und Forsa

Das von der „Forschungsgruppe Wahlen“ (FGW) erhobene **Politbarometer** hat bereits eine lange Tradition. Die erste Erhebung erfolgte im Jahr 1977. Wurden bis 1988 persönliche Befragungen durchgeführt, erfolgt seitdem die Erhebung telefonisch.[13] Für den Zeitraum ab 1990 sind über das Datenarchiv der GESIS die kumulierten Jahresstichproben beziehbar.[14] Bis auf wenige Ausnahmen (1996, 1997 und 1998) sind diese nach Ost und West separiert, seit 2012 jedoch auch als gesamtdeutsche Stichprobe verfügbar.[15] Die Stichproben der alten Bundesländer sind konstant durch ein mehrstufiges Zufallsverfahren

[13] Aufgrund der langen Tradition ergibt sich bereits eine lange Zeitreihe der Daten seit 1977 über die GESIS (*ZA2391*), welche jedoch nur die alten Bundesländer umfasst. Beziehbar über http://www.gesis.org/wahlen/politbarometer/aktuelle-zeitreihe/ (Abruf: 11. Mai 2018).

[14] Das Politbarometer wird ca. einmal pro Monat telefonisch erhoben.

[15] Alle Jahreskumulationen sowie die gesamte partielle Kumulation sind über http://www.gesis.org/wahlen/politbarometer (Abruf: 11. Mai 2018) beziehbar.

nach dem „Randomized Last Digit“ (RLD)-Verfahren ausgewählt worden. Bis zum Jahr 1997 und einmalig in 1999, visierten diese telefonischen Befragungen mittels standardisiertem Fragebogen als Grundgesamtheit „Wahlberechtigte, die in Privathaushalten mit Telefonanschluss leben“[16] an. Von da an wird diese als „Wahlberechtigte Wohnbevölkerung“[17] definiert. 1998 und seit 2000 ist konstant die Beschreibung einer mehrstufigen Zufallsauswahl von Haushaltsadressen durch das RLD-Verfahren und eine anschließende Personenauswahl nach den Geburtstagen der relevanten Haushaltspersonen als angewandte Methoden im Datenarchiv der GESIS erfasst.[18] In den neuen Bundesländern wurde noch bis einschließlich 1993, teilweise auch noch in 1995, eine persönliche Befragung durchgeführt.[19] Seit 1994 werden auch dort telefonische Interviews über eine mehrstufige Zufallsauswahl nach dem RLD-Verfahren mit anschließender Auswahl über das Geburtsdatum angewandt.

Für den empirischen Teil der Arbeit werden die jeweils separaten Jahreskumulationen zu einem gesamtdeutschen Datensatz, hinsichtlich der berücksichtigten Variablen, zusammengeführt. Bei eventuellen Abweichungen der Kodierung werden diese zwischen unterschiedlichen Jahrgängen angepasst.[20] Der Zeitraum des gesamten Datensatzes umfasst die Jahre 1991 bis einschließlich 2014, die Angabe der erhobenen kalendarischen Woche definiert dabei eine erhobene Umfrage im Rahmen der Arbeit. Ist diese Information für ein Jahr nicht verfügbar, bildet der Erhebungsmonat die Basis. Innerhalb der zusammengespielten Datensätze ergibt sich zudem ein zu großer Anteil der ostdeutschen Bevölkerung, da diese durch die getrennte Erhebung der Politbarometer-Daten überquotiert ist. Über die bereitgestellten Gewichtungsvariablen ist eine alleinige Korrektur dieser Überquotierung allerdings nicht möglich, da nur ein gesamtes „Repräsentativ-Gewicht“ angeboten wird.[21] Für die Jahre ohne eine gesamtdeutsche Kummulation, welche der amtlichen Verteilung der Ost-West-Zugehörigkeit entspricht, wird daher eine eigene „Ost-West-Korrektur“ bestimmt. Als Referenzdaten dienen hierzu die Verteilungen hinsichtlich der regionalen Zugehörigkeit der

[16] Z.B. Methodologie von ZA3045: Politbarometer 1997 (Kumulierter Datensatz).

[17] Z.B. Methodologie von *ZA3160: Wahlstudie 1998 (Politbarometer)*.

[18] Quelle: Jeweilige Methodologie-Beschreibungen.

[19] Beschreibung Methodologie von ZA2114, -2287, -2390 und -2777. Für 1995 wurde bis einschließlich Juli eine mündliche Befragung durchgeführt.

[20] Maßgabe für die Anpassung ist die Vergleichbarkeit mit den Datensätzen der GLES. Sollte dies nicht möglich sein, wird die Kodierung der ostdeutschen Kumulation an die westdeutsche angepasst.

[21] Quelle: http://www.forschungsgruppe.de/Rund_um_die_Meinungsforschung/Methodik_Politbarometer/ (Stand: 11. Mai 2018).

neuen und alten Bundesländer.[22] Für eine Darstellung der Verteilung der berechneten Gewichte und der jeweiligen Referenzverteilunge siehe Abbildung A1 in Anhang A.

Die Stichproben des **Forsa-Bus** sind ebenfalls über die GESIS als Jahreskumulationen frei verfügbar. Diese umfassen die durch Forsa befragten Personen eines Kalenderjahres, wobei durch Forsa Tagesstichproben von ca. 500 Personen erhoben werden.[23] Bis einschließlich 2000 handelt es sich um standardisierte telefonische Befragungen von Stichproben der wahlberechtigten Bevölkerung, wobei Haushalte nach dem RLD-Verfahren ausgewählt wurden. Seit 2001 basiert die Stichprobenziehung auf dem ADM-Telefonstichprobensystem. Damit ergibt sich auch durch den Forsa-Bus eine konstante Analysemöglichkeit der Zusammensetzung von Telefonstichproben seit 1991, lediglich die Art der Stichprobenerhebung wurde leicht modifiziert. Ebenfalls handelt es sich bei den Jahreskumulationen seit 1991 um gesamtdeutsche Stichproben, eine Design-Korrektur ist daher nicht nötig. Für den empirischen Teil wurden alle später berücksichtigten Variablen der Jahreskumulationen von 1991 bis 2014 zu einem Gesamtdatensatz zusammengefügt. Eine einzelne Umfrage wurde dabei an Hand der verfügbaren kalendarischen Kalenderwoche definiert.[24]

Die Umfrageprogramme der FGW und des Forsa-Bus sind somit standardisierte telefonische Erhebungen mit einem expliziten Bezug zur Wahlforschung, auch wenn insbesondere durch Forsa auch zusätzlich andere Themen während der Erhebung abgefragt werden. Die Stichproben beider Umfrageinstitute basieren auf einer mehrstufigen Zufallsauswahl. Erst wird über das RLD-Verfahren oder über die Nutzung des ADM-Designs für Telefonstichproben ein Haushalt ausgewählt, um anschließend mit einer zufällig ausgewählten und zur Grundgesamtheit gehörigen Person ein Interview durchzuführen.[25] Die Abbildung 5.2 führt die Anzahl der, im Rahmen der Arbeit, definierten Umfragen der beiden Institute pro Jahr auf. Für Forsa ergibt sich eine Umfrage pro Woche seit 1991 (insgesamt handelt es sich um 1201 Umfragen), für das Politbarometer variiert die Anzahl der an Hand von Variablen *identifizierbaren* jährlichen Erhebungen zwischen 10 und 25. Ebenso sind die im arithmetischen Mittel erhobenen Stich-

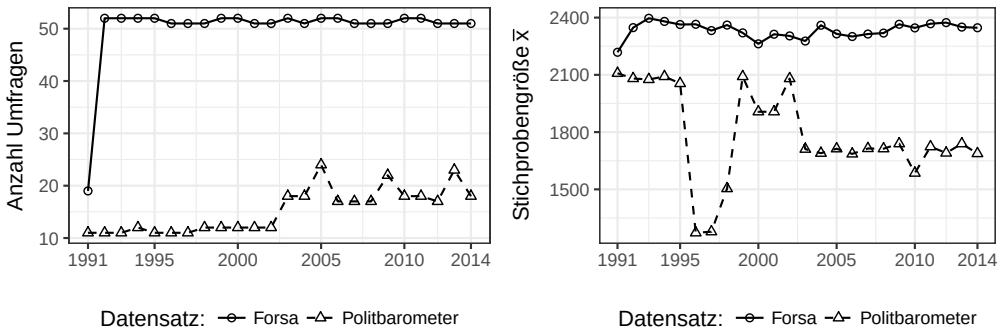
[22] Genutzt wurde der Anteil der alten und neuen Bundesländer für das jeweilige Jahr gemäß Indikator *12411-0009* der „Gemeinsames neues statistisches Informationssystem“ (GENESIS)-Datenbank.

[23] Siehe z.B. http://www.gesis.org/wahlen/laufende-erhebungen/#c54911 (Stand: 11. Mai 2018).

[24] Innerhalb der über die GESIS bereitgestellten Datensätze sind dies die Variablen *woc* und *WOC*.

[25] Vergleichbare Umfragen werden auch durch weitere Insitute durchgeführt, diese legen jedoch nicht ihre ermittelten Rohdaten offen. Siehe wahlrecht.de für eine Übersicht.

Abbildung 5.2.: Umfragen und Stichprobengröße Forsa und Politbarometer 1991 bis 2014



Angegebene Zahlen stellen die durchschnittliche Stichprobengröße eines Jahres dar, Anzahl Umfragen bezieht sich auf die zu Grunde liegenden Umfragen des jeweiligen Jahres.

probengrößen verzeichnet. Für die Forsa-Erhebungen beläuft sich dies auf konstante Werte um $n = 2400$, im Politbarometer variiert die Zahl der identifizieren Umfragen zwischen 1400 und 2100 Personen.

Neben den bisherigen expliziten Wahlumfragen existieren jedoch noch langjährige Umfrageprogramme der Sozialwissenschaften, welchen auch für die Wissenschaft eine zentrale Bedeutung der Analyse und Nutzung als Referenzdaten für weitere eigene erhobene Umfragen zukommt: Der Allbus und der ESS.

5.3. Persönliche Bevölkerungsumfragen: Allbus und ESS

Der seit 1980 durchgeführte Allbus und der seit 2002 erhobene ESS weisen einige Gemeinsamkeiten hinsichtlich Konzeption und Zielsetzung auf: Es handelt sich bei beiden um langjährige wissenschaftliche Projekte persönlich-mündlicher Bevölkerungsumfragen, welche aus fest finanzierten Mitteln zur Dauerbeobachtung der Gesellschaft durchgeführt werden. So soll durch den ESS „Stabilität und Wandel in der gesellschaftlichen Struktur [...] aufgezeigt werden. [...] höhere Standards in der international vergleichenden Sozialforschung erreicht [...] werden, [...] Indikatoren für die Entwicklung nationaler Gesellschaften [eingeführt werden], [...] Training europäischer Sozialforscher [...] durchgeführt und gefördert werden. [...] der Zugang zu den Daten über sozialen Wandel verbessert werden.“.[26] Auch der Allbus setzt sich zum Ziel „[...] eine langfristig

[26]Selbstbeschreibung ESS: http://www.europeansocialsurvey.org/about/country/germany/index.html, Abruf: 11. Mai 2018.

angelegte, multithematische Umfrageserie zu Einstellungen, Verhaltensweisen und Sozialstruktur der Bevölkerung in der Bundesrepublik Deutschland.“[27] zu sein.

Der **Allbus** erhebt in einem zweijährigen Turnus seit 1980 Stichproben der erwachsenen Wohnbevölkerung der Bundesrepublik. Die Definition der anvisierten Grundgesamtheit war von 1980 bis einschließlich 1990 die „[...] wahlberechtigte Bevölkerung in Privathaushalten der alten Bundesrepublik inkl. West-Berlins.“ (Wasmer et al. 2014: 6). Durch den Beitritt der „Deutschen Demokratischen Republik“ (DDR) zum Geltungsbereich des Grundgesetz änderte sich diese Definition: „Seit 1991 umfasst die Grundgesamtheit der ALLBUS-Studien damit die gesamte erwachsene Wohnbevölkerung (d.h. Deutsche und Ausländer) in Privathaushalten in West- und Ostdeutschland.“ (Wasmer et al. 2014: 6). Seit 2000 wird die Befragung zudem als „Computer-Assisted Personal Interviewing“ (CAPI) durchgeführt. Bis 1992 und einmalig 1998 wurde das ADM-Design zur Stichprobenziehung angewandt, wodurch es sich für diese Jahre um Haushaltsstichproben handelt. Die restlichen Erhebungen sind als Personenstichproben aus den Einwohnermelderegistern gezogen. Die Stichprobenziehung überquotiert – wie die Komponente 1 der GLES – bewusst die Bevölkerung der neuen Bundesländer.[28] Für die Stichprobenziehung auf Basis der Einwohnermeldeämter wird ein zweistufiges Verfahren angewandt. Auf der ersten Ebene werden die nach verschiedenen geografischen Merkmalen stratifizierten, Gemeinden der Bundesrepublik ausgewählt. In diesen werden dann Personenadressen durch eine systematische Zufallsauswahl aus den Adressen der Einwohnermeldeämter ausgewählt (Wasmer et al. 2014: 47 ff.).

Auch der **ESS** ist eine registergestützte Umfrage. Er wird ebenfalls als CAPI-Studie durchgeführt und visiert als Grundgesamtheit die Bevölkerung der Bundesrepublik ab 16 Jahren an, welche in einem Privathaushalt lebt. Analog zum Allbus werden ebenso Gemeinden der Bundesrepublik ausgewählt und anschließend in den jeweiligen Einwohnermeldeämtern personenbezogene Adressen gezogen. Auch der ESS überquotiert über separate Ziehungen die ostdeutsche gegenüber der westdeutschen Bevölkerung,[29] wodurch eine Design-Korrektur ebenfalls notwendig ist.[30] Hinsichtlich der anvisierten Grundgesamtheit versucht

[27] Quelle: http://www.gesis.org/allbus/allgemeine-informationen/, Abruf: 11. Mai 2018.

[28] Quelle: http://www.gesis.org/allbus/allgemeine-informationen/stichproben-und-erhebungsdesign/ (Abruf: 11. Mai 2018).

[29] Quelle: http://www.europeansocialsurvey.org/about/country/germany/methods.html (Abruf: 11. Mai 2018).

[30] Abrufbar unter: http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf (Abruf: 11. Mai 2018).

der ESS eine Aussage über „[...] die in Deutschland lebenden Personen, die älter als 15 Jahre sind (keine obere Altersgrenze) und in in einem Privathaushalt leben, ungeachtet ihrer Nationalität, Staatsbürgerschaft oder Sprache.“[31] zu treffen. Der ESS wird zudem als Umfrageprogramm in einer Vielzahl von Ländern durchgeführt und das Auswahldesign muss daher bei einem Vergleich der befragten Personen verschiedener Länder vergleichbar bleiben. Die länderspezifischen Herangehensweisen, bedingt durch unterschiedliche mögliche Auswahlrahmen und der daraus resultierenden Stichprobenziehungen in den teilnehmenden Länder, können dem jedoch entgegenstehen (z.B. Häder/Ganniner/Gabler 2009: 183 ff.). Dies berührt das Ziel der Arbeit jedoch nicht, da der Fokus der Nutzung lediglich auf den in der Bundesrepublik durchgeführten Interviews liegt. Die deutsche Teilstudie des ESS wurde bisher bei allen zweijährigen Runden der Durchführung seit 2002 durch das „Institut für angewandte Sozialwissenschaft GmbH“ (infas) erhoben.[32]

Sowohl der Allbus, als auch der ESS, stellen somit persönlich-mündliche Interviews mit einer langjährigen Tradition und Erfahrung der Erhebung dar. Ihre anvisierte Grundgesamtheit enthält gegenüber der Komponente 1 und 2 der GLES auch den nicht wahlberechtigten Teil der (volljährigen oder über 15-jährigen) Bevölkerung. Ob sich diese beiden anvisierten Grundgesamtheiten auf grundlegenden soziodemografischen Variablen unterscheiden, wird das folgende Unterkapitel 5.4 durch den Vergleich des Mikrozensus und der „Repräsentativen Wahlstatistik“ durch den BuWa beleuchten.

5.4. Amtliche Referenz: Mikrozensus und amtliche Wahlstatistik

Bezüglich der Bereitstellung amtlicher Referenzdaten dominiert als zentrale Umfragequelle der jährlich durch das „Statistische Bundesamt“ (Destatis) durchgeführte **Mikrozensus**. Die hohe Qualität der Befragung von jährlich ca. 800 000 zufällig ausgewählten und persönlich befragten Personen (ca. 1 % der Bevölkerung der Bundesrepublik) ergibt sich durch die gesetzlich geregelte Auskunftspflicht. Der „Mikrozensus“ (MZ) visiert als repräsentative Haushaltsbefragung die Bevölkerung in privaten Haushalten und Gemeinschaftsunterkünften

[31] Quelle: http://www.europeansocialsurvey.org/about/country/germany/methods.html (Abruf: 11. Mai 2018).

[32] Zur Übersicht: http://www.uni-bielefeld.de/soz/ess/forschungundlehre/wellen.html, Abruf: 11. Mai 2018. Lediglich für den Durchgang 2010/2011 hat TNS Infratest die Erhebung durchgeführt.

an.[33] Damit visiert der Mikrozensus die in Deutschland lebende Bevölkerung an, zu der auch Personen ohne die deutsche Staatsbürgerschaft und somit ohne Wahlrecht bei einer Bundestagswahl gehören. Werden die Ergebnisse des Mikrozensus auf die volljährige Bevölkerung eingeschränkt, sind diese mit der Grundgesamtheit des Allbus und – auf die volljährigen befragten Personen eingeschränkten – ESS deckungsgleich. Durch die Auskunftspflicht sind (hypothetisch) Nonresponse-Probleme im Mikrozensus nicht möglich. Für die Wahlstudien der GLES ergibt sich jedoch eine Einschränkung der Nutzung des Mikrozensus als Vergleichsquelle, da die allgemeine volljährige und die wahlberechtigte Bevölkerung keine deckungsgleichen Populationen sind. Ca. 8 % der Bevölkerung ab 18 Jahren besitzt nicht die deutsche Staatsbürgerschaft und ist somit bei einer Bundestagswahl nicht wahlberechtigt.[34] Unterscheiden sich diese beiden Gruppen hinsichtlich relevanter Rahmendaten der amtlichen Statistik, ist diese als Referenz für die *wahlberechtigte* Bevölkerung ebenfalls verzerrt.

Ein solcher Vergleich ist über eine zweite Datenquelle der amtlichen Statistik möglich: Die **„Repräsentative Wahlstatistik“**. Hierbei handelt es sich um eine Erhebung der Gemeinden in ausgewählten Wahlbezirken im Rahmen einer Wahl. Erhoben werden die erfassten Daten in 2 500 Urnenwahlbezirken und 350 Briefwahlbezirken.[35] Als Merkmale werden die „Anzahl Wahlberechtigte“, „Wahlscheinvermerke“, „Beteiligung an einer Wahl“, „Geburtsjahresgruppe“ sowie „Geschlecht“ als Merkmale in der jeweiligen Gemeinde erhoben. Die gesetzliche Grundlage ergibt sich aus dem Wahlstatistikgesetz (WStatG).[36] Da diese Daten jedoch nur bei Bundestagswahlen und Wahlen des Europäischen Parlaments erhoben werden, lassen sich hieraus keine eigenständigen jährlichen Zeitreihen bilden. Für das Alter, das Geschlecht und die regionale Zuordnung der neuen und alten Bundesländer lässt sich jedoch das Ausmaß der fehlerhaften Erfassung der wahlberechtigten Bevölkerung durch den Mikrozensus über die „Repräsentative Wahlstatistik“ des Bundeswahlleiters abschätzen.

Abbildung 5.3 zeigt den Vergleich für die Jahre 2005, 2009 und 2013 hinsichtlich des „Relativen Bias“ (RB) aus Kapitel 3.3. Durch die Kennzahl wurden

[33] Quelle: https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Mikrozensus.html (Abruf: 11. Mai 2018)

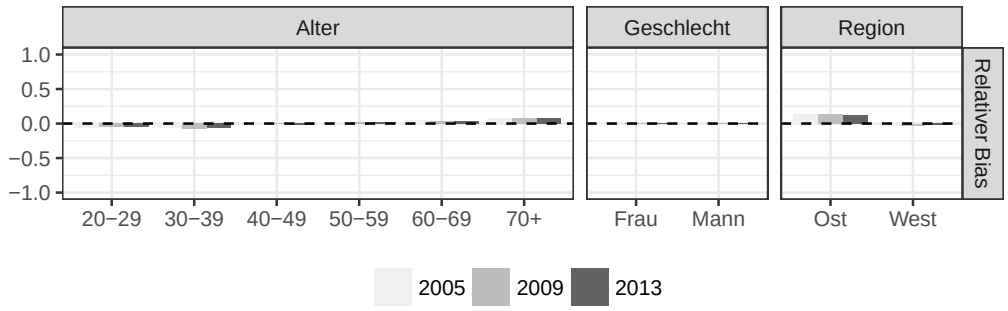
[34] Quelle: Daten Zensus 2011, abrufbar über die GENESIS-Datenbank Indikator 12111-0002. Von insgesamt 67 085 343 volljährigen Personen in der Bundesrepublik weisen 5 398 973 keine deutsche Staatsbürgerschaft im Rahmen der Erhebung des Zensus auf.

[35] Quelle: https://www.bundeswahlleiter.de/de/glossar/texte/Wahlstatistik.html, Abruf: 11. Mai 2018.

[36] Quelle: http://www.bundeswahlleiter.de/de/bundestagswahlen/downloads/rechtsgrundlagen/wahlstatistikgesetz.pdf (Abruf: 11. Mai 2018).

Abweichungen der „Repräsentativen Wahlstatistik“ von den Verteilungen des Mikrozensus zu bestimmt. Dargestellt ist also, wie stark die jeweiligen Kategorien der wahlberechtigten Bevölkerung relativ von denen der breiter definierten volljährigen Bevölkerung abweichen.

Abbildung 5.3.: Vergleich Mikrozensus und Wahlstatistik



Quelle: Eigene Darstellung. Daten: GENESIS-Datenbank und bundeswahlleiter.de. Erfasst wurden Abweichung der Wahlstatistik von den Rahmendaten des Mikrozensus.

Es zeigen sich nur marginale Unterschiede: Die wahlberechtigte Bevölkerung ist im Vergleich zur volljährigen Bevölkerung etwas älter und überproportional in den neuen Bundesländern vertreten. Die prozentuale Abweichungen hinsichtlich des Alters liegen laut dem RB unter 10 %. Auch für die neuen Bundesländer liegen die relativen Abweichungen unterhalb von 15 % des Vergleichswerts des MZ. Es sind daher hinsichtlich grundlegender soziodemografischer Rahmenverteilungen über die amtliche Statistik keine großen Ungenauigkeiten erwartbar, wenn für Wahlumfragen ebenfalls der MZ an Stelle der Repräsentativen Wahlstatistik verwendet wird.

Nachdem an diesem Punkt die verwendeten Datensätze unterschiedlichster Erhebungsmodi vorgestellt sind (siehe für eine abschließende Übersicht die nachfolgende Tabelle 5.1), können mit Hilfe der aufgeführten langfristigen telefonischen und persönlichen Umfrageprogramme diejenigen Faktoren herausgestellt werden, welche sich zur Korrektur onlinebasierter Erhebungen eignen können.

Tabelle 5.1.: Übersicht verwendete Datensätze

Name	Jahre	Umfragen	Modus	N Gesamt	Turnus	N mean	Institut	Auswahlverfahren	Grundgesamtheit
GLES Komp. 1	2009/2013	2	CAPI	4 176	BTW	2 088	MARPLAN	Mehrst. Zufalls. (ADM)	Bevölkerung ab 16 Jahren mit deutscher Staatsbürgerschaft in Privathaushalten
GLES Komp. 2	2009/2013	2	CATI	13 890	BTW	6 945	Ipsos	Mehrst. Zufalls. (RLD) als RCS-Design	Bevölkerung ab 18 Jahren mit deutscher Staatsbürgerschaft in Privathaushalten
GLES Komp. 3 (+KQ 2013)	2009/2013	5	CAWI	11 220	BTW	4 382	Respondi	Quotenvorgabe (Alter, Geschlecht und Bildung)	Bevölkerung mit deutscher Staatsangehörigkeit und Wahlrecht sowie Mitglied Respondi-Panel
GLES Komp. 8	2009-2016	33	CAWI	30 443	ca. 4 pro Jahr	1 049	Respondi (T1-16)/LINK (T17-33)	Quotenvorgabe (Alter, Geschlecht und Bildung)	Bevölkerung mit deutscher Staatsangehörigkeit und Wahlrecht sowie Mitglied Respondi/LINK-Panel
Allbus	1980-2014	19	CAPI	59 011	alle 2 Jahre	3 106	TNS Infratest	Mehrst. Zufalls. (ADM) oder Registerstichprobe	Bevölkerung Bundesrepublik ab 18 Jahren in Privathaushalten
ESS	2002-2014	7	CAPI	20 490	alle 2 Jahre	2 927	infas	Registerstichprobe	Bevölkerung Bundesrepublik ab 15 Jahren in Privathaushalten
Politbarometer (Ost-West)	1992-2014	357	CATI	625 519	Wöchentlich/-Monatlich	1 752	FG Wahlen	Mehrst. Zufalls. (RLD)	Wohnhafte und wahlberechtigte Bevölkerung der Bundesrepublik
Politbarometer (West-Kumulation)	1977-1991	157	CAPI (bis 06-1988) / CATI (ab 08-1988)	495 558	Wöchentlich/-Monatlich	3 156	FG Wahlen	Mehrst. Zufalls. (bis 1988: ADM-Design, von da an RLD)	Bis Juni 1988: „Personen, die zur Bundestagswahl wahlberechtigt sind.“, von da an: „Wahlberechtigte, die in Privathaushalten mit Telefonanschluss leben.“
Forsa	1991-2014	1203	CATI	2 810 597	Wöchentlich	2 336	FORSA	Mehrst. Zufalls. (RLD)	Wohnhafte und wahlberechtigte Bevölkerung der Bundesrepublik
Mikrozensus	1990-2015	25	CAPI	Randverteilung	Jährlich	Randverteilung	Destatis	Mehrst. Zufalls.	Wohnhafte Bevölkerung der Bundesrepublik

Quelle: Eigene Darstellung. Information auf Grundlage des vorangegangen Kapitels oder auf Grundlage der Auswertung der Datensätze (z.B. Anzahl befragte Personen).

6. Erhebungsvariation in der Umfrageforschung

An diesem Punkt ist klar, wie man den im Kapitel 2 dargestellten „Total Survey Error“ (TSE) (Kapitel 2) empirisch erfassen (Kapitel 3) und der zu Grunde liegenden Verzerrung effektiv begegnen kann (Kapitel 4). Ebenso wurden die zentralen Anforderungen an die zu Grunde liegenden Variablen eines Korrekturmodells herausgearbeitet und verschiedene Gruppen von Merkmalseigenschaften von Personen vorgestellt, welche für eine Korrektur prädistiniert sind (Kapitel 7). Der theoretische Rahmen der Arbeit ist somit gegeben. Der nachfolgende Teil der Arbeit wird sich nun auf die Variation der Erhebungsmethoden von Umfragen konzentrieren. Hierzu müssen zwei Merkmale definitorisch voneinander abgegrenzt werden: Das *Auswahlverfahren*, oder auch *Stichprobendesign*, einer Erhebung. Durch dieses erfolgt die Auswahl der zu befragenden Personen aus der anvisierten Grundgesamtheit. Durch das *Erhebungsverfahren* wird hingegen bestimmt, wie die ausgewählten Personen praktisch befragt werden. Die Auswahl von Personen ist somit vorgelagert, jedoch nur auf den ersten Blick von der gewählten Befragungsform unabhängig. Die Wahl des Stichprobenverfahrens bestimmt vielmehr bereits die gesamte Erhebung und für jede Befragungsform lässt sich daher eine enge Verbindung zu einem angewandten Auswahlverfahrens der zu befragenden Personen finden (Kalton 1983: 6; Groves/Fowler et al. 2009: 163).

Ein *Auswahlverfahren* wird angewandt, da Befragungen (beinahe immer) auf Stichproben und nicht auf Vollerhebungen basieren. Die Auswahlgesamtheit ist derjenige Teil der Grundgesamtheit, welcher eine Chance hat, Teil der Stichprobe zu werden (Groves/Fowler et al. 2009: 44 f.).[1] Auf Basis der erfolgreich befragten Stichprobe ist dann der Schluss auf die Inferenzpopulation möglich. Existiert kein systematischer Unterschied zwischen diesen drei Gruppen, ist das Auswahlverfahren zur Stichprobenziehung geeignet gewesen. Das zentrale Gütekriterium für ein angewandtes Stichprobendesign ist die Möglichkeit

[1] Unterschiede zwischen der Grund- und Auswahlgesamtheit sind auf Noncoverage zurückzuführen. Personen sind Teil der anvisierten Grundgesamtheit, haben aber keine Chance der Auswahl durch das Erhebungsdesign.

einer zufallsbasierten Auswahl (Schnell/Esser/Hill 2013: 257-263). Erst hierdurch kann jeder Person der Auswahlgesamtheit eine gewisse und bekannte Wahrscheinlichkeit der Teilnahme an der Umfrage zugeordnet werden. Ist die Auswahl nicht zufallsbasiert, können selektive Zugänge in die Stichprobe nicht vermieden werden. Die korrekte Anwendung statistischer Inferenzverfahren ist nicht gewährleistet (Kalton 1983: 7 f.). Haben alle Personen der anvisierten Grundgesamtheit eine identische Wahrscheinlichkeit der Teilnahme, liegt eine einfache Zufallsstichprobe vor (Groves/Fowler et al. 2009: 103 ff.). Ein Stichprobenverfahren sollte ebenfalls alle Personen der Auswahlgesamtheit nur einmalig berücksichtigten, Personen ohne Zugehörigkeit zur Grundgesamtheit hingegen keine Wahrscheinlichkeit der Ziehung in die Stichprobe erhalten (Kish 1965: 53-59). Existiert für die Anwendung des Auswahlverfahrens nur eine Liste mit gruppierten Einträgen von Personen (z.B. Haushalte), ist für die Rechtfertigung der Anwendung inferenzsstatistischer Verfahren die Kenntniss der sich daraus ergebenden unterschiedlichen Auswahlwahrscheinlichkeiten essentiell. In diesem Fall können diese berücksichtigt werden, wie bereits in Kapitel 4.2.1 dargestellt.

Ist das Stichprobendesign erfolgreich angewandt worden und eine zufallsbasierte Auswahl realisiert, muss die Art der Durchführung der Befragung festgelegt werden. Die Wahl dieses *Erhebungsverfahren* beeinflusst die Art der Administration, den zentral eingesetzten Kommunikationskanal und die Art der technischen Infrastruktur (z.B. Weisberg 2005: 30; Faulbaum/Prüfer/Rexroth 2009: 30 f.). Insbesondere die Entscheidung für eine interviewer- oder selbstadministrierte Erhebung hat einen Einfluss auf den möglichen Grad der Reaktivität des gewählten Erhebungsverfahrens, wobei Erhebungen der empirischen Sozialforschung unabhängig davon von einem hohen Standardisierungsgrad gekennzeichnet sind (Schnell/Esser/Hill 2013: 311 ff.).

Zwei Trends lassen sich in den letzten Jahren feststellen: Die durch Interviewer administrierten Erhebungsverfahren der persönlich-mündlichen („Face to Face“ (F2F)) und telefonischen Befragung haben sich durch die elektronische Unterstützung zu „Computer-Assisted Personal Interviewing“ (CAPI) und „Computer-Assisted Telephone Interviewing“ (CATI) weiterentwickelt. Durch das Aufkommen der onlinebasierten Befragung wurde zudem die Erhebungslandschaft technisch erweitert. Wie die postalische Befragung handelt es sich hierbei um ein selbst-administriertes visuelles Format ohne eine Interviewerbeteiligung (Bethlehem/Biffignandi 2012: 37). Eine Typologie der Erhebungsverfahren lässt sich daher auch entlang der persönlichen, telefonischen und postalischen Befragung als „Klassiker“ der Datenerhebung treffen, welche durch

den neu aufkommenden Typus der onlinebasierten Befragung ergänzt wurden (Bethlehem 2009: 153).

Die postalische Befragung hat historisch einen schweren Stand gegenüber der schnelleren und günstigeren Form der telefonischen Befragung gehabt. Auch schließt ihre Selbstadministration systematisch den nicht ausreichend alphabetisierten Teil der Bevölkerung aus (Schnell 2012: 244-247). Dieser Kostennachteil hat sich durch die zunehmende Nutzung der, ebenfalls selbstadministrierten, onlinebasierten Befragung noch einmal verstärkt und die postalische Befragung in ein Nischendasein geführt (Dillman/Smyth/Christian 2009). Innerhalb der Bundesrepublik hat sie daher als eigenständige Befragungsform kaum noch eine Relevanz (ADM 2014: 22).[2]

Aus diesem Grund wird sich der nachfolgende Teil der Arbeit auf die Darstellung der „Trias“ der Datenerhebung der persönlichen, telefonischen und onlinebasierten Befragung konzentrieren. Für diese drei Erhebungsverfahren lassen sich bestimmte komparative Vorteile aufführen. Sie haben daher alle ihre Berechtigung (z.B. zur Übersicht Weisberg 2005: 29 f.). Um dies zu zeigen, werden in einem ersten Schritt die spezifischen Vor- und Nachteile der persönlichen (Kapitel 6.1.1) und der telefonischen (Kapitel 6.1.2) Befragung herausgestellt und die jeweils charakteristischen Varianten der (zufallsbasierten) Stichprobenziehung beschrieben. Die Möglichkeit eines „Design-Effekts“ wird anschließend diskutiert. Dieser bedingt sich durch das Erhebungsdesign oder die intervieweradministrierte Durchführung der Befragung. Wie ein solcher Effekt berücksichtigt werden kann, wird das Kapitel 6.1.3 darstellen.

Anschließend kann der dritte Erhebungstypus näher betrachtet werden: Onlinebasierte Erhebungen nehmen in jeder Hinsicht in der Erhebungslandschaft (noch) einen Sonderstatus ein. Kapitel 6.2 wird die Gründe hierfür aufzeigen. Ebenso werden die Unterschiede in Form von komparativen Vor- und Nachteilen dieses neuen Typus der Befragung zu den beiden anderen Varianten herausgestellt (Kapitel **??**). Dass ein spezifischer Teil der Bevölkerung über keinen Internetzugang verfügt und über seine systematischen Unterschiede einen Noncoverage-Fehler qua Design bei Onlineumfragen bedingt, wird Kapitel 6.2.1 zeigen. Die qualitativen Unterschiede dieser verbleibenden „Offliner“ werden im Zuge der Ausweitung des Internetzugangs immer weiter zunehmen. Der mögliche Noncoverage-Bias gestaltet sich daher größer, als es die rein quantative Entwicklung des Internetzugangs vermuten lässt.

[2]Dies bedeutet nicht, dass postalische Befragung obsolet sind, gerade in der Kombination einer postalischen Kontaktierung mit einer Onlineumfrage lässt sich signifikant die Teilnahmerate von Onlineerhebungen steigern (z.B. Messer/Dillman 2011).

Auch hier wird das Auswahldesign als die zentrale Eigenschaft zur qualitativen Einstufung der Qualität von Datensätzen onlinebasierter Erhebungen herausgestellt. Kapitel 6.2.2 widmet sich daher ganz der Frage „Zufall oder kein Zufall“ als entscheidendes Zugangskriterium zu einer onlinebasierten Erhebung. Anschließend wird das Kapitel 6.2.3 Access-Panels vorstellen. Durch diese Bereitstellung eines großen „Pools“ von möglichen zu befragenden Personen ergibt sich die Möglichkeit der Ziehung von Stichproben aus ebendiesen. Hierzu werden einleitend allgemeine Begriffe definiert und die gängigen Rekrutierungs- und, meist nicht zufallsbasierten, Auswahldesigns dieser vorgestellt. Darüber hinaus wird die Rolle der Nutzung von Incentivierungen diskutiert. Abschließend ist die Arbeit in der Lage eine qualitative Einschätzung persönlicher, telefonischer und onlinebasierter Befragungen an Hand der Literatur vorzunehmen.

6.1. Zufallsbasierte Klassiker

Nachdem nun die Begriffe des Auswahl- und Erhebungsdesigns definiert und die zentral verwendeten Erhebungsverfahren der persönlichen, telefonischen und onlinebasierten Befragungen vorgestellt sind, werden nachfolgend deren spezifische Stärken und Schwächen herausgearbeitet. Hierzu wird erst die persönliche (Kapitel 6.1.1), anschließend die telefonische (Kapitel 6.1.2) und schließlich die onlinebasierte Befragung beleuchtet (Kapitel 6.2). Ein zentrales Ziel des Kapitels ist es, eine enge Verbindung zwischen diesen Befragungsformen und dem jeweils verwendeten Auswahldesigns aufzuzeigen.

6.1.1. Die persönliche Befragung

Persönliche Befragungen nehmen eine zentrale Stellung innerhalb der Umfrageforschung ein, ihre Befragungsform wird häufig mit hohen Teilnahmeraten assoziiert (z.B. Goyder 1987; Hox/Leeuw 1994). Die anvisierte Grundgesamtheit persönlich-mündlicher Bevölkerungsumfragen unterscheidet sich dabei nur unwesentlich und umfasst eine mehr oder weniger klare Definition der allgemeinen Bevölkerung (Schnell 2012: 204). Die beiden in Kapitel 5 vorgestellten Umfrageprogramme der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (Allbus) und „European Social Survey“ (ESS) belegen dies. Visiert der ESS die Bevölkerung der Bundesrepublik ab 16 Jahren an, definieren die Durchführungen des Allbus die Grundgesamtheit ab 18 Jahren. Persönlich erhobene Wahlumfragen treffen eine zusätzliche Eingrenzung auf den wahlberechtigten Teil der Bevölkerung.

Ein Unterschied zwischen persönlichen Befragungen ist jedoch das verwendete Auswahlverfahren. Die zufallsbasierte Auswahl auf Basis der Einträge der Einwohnermelderegister stellt die in der Praxis bestmögliche Variante zur adäquaten Abdeckung der Grundgesamtheit dar (Schnell 2012: 205 f.). Die Nutzung dieser ist jedoch nur möglich, wenn die durchzuführende Umfrage ein klares öffentliches Interesse aufweist. Die Koordination mit den Einwohnermeldeämtern bringt auch einen zusätzlichen Zeit- und Kostenaufwand mit sich und daher wird diese Variante der Auswahl nur relativ selten angewandt (Koch 1997). Ist der Rückgriff auf die Einwohnermelderegister nicht möglich oder gewollt, kommt in der Bundesrepublik (meist) das ADM-Design zur Anwendung. Über regional stratifizierte Ziehungspunkte[3] lässt sich durch diese eine mehrstufige Zufallsauswahl aus der gesamten Bevölkerung der Bundesrepublik realisieren. Der „Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V." (ADM) stellt hierzu vordefinierte Netze für seine Mitgliedsinstitute zur Verfügung. Diese stellen jeweils ein eigenes Abbild der Bevölkerung bereit und die daraus resultierenden Stichproben zwischen verschiedenen Netzen/Instituten sind überschneidungsfrei. Innerhalb dieser Netze können dann Interviewer zur Auswahl von Haushalten und der Befragung von darin wohnenden Personen eingesetzt werden (Heckel/Hofman 2014).[4]

Werden in einem ersten Auswahlschritt durch die Interviewer nur Adressen erhoben, handelt es sich um ein „Adress-Random". Wird nach der Auswahl der Adresse auch direkt durch diese die Befragung durchgeführt, handelt es sich um das „Standard-Random", welches aus Kostengründen das vorherrschende Verfahren in der Praxis ist (Schnell 2012: 206 ff.). Eine zufallsbasierte Auswahl wird durch eine Stratifizierung der Erhebung nach den Merkmalen Bundesland, Regierungsbezirk und der Gemeindetypisierung gewährleistet. Durch diese Schichtung fließen die stratifizierten Merkmale proportional zu ihrer Bedeutung innerhalb der Bundesrepublik in die Stichprobe ein. Eine Zufallsauswahl der Elemente der ersten Ziehungsebene hinsichtlich dieser Merkmale ist hierdurch garantiert (Häder 2015: 151 f.). Die Auswahl an Ziehungspunkten wird dann proportional zur Anzahl der Haushalte in den Schichten vorgenommen (Kauermann/Küchenhoff 2011: 201).

Durch die Berücksichtigung der unterschiedlichen Auswahlwahrscheinlich-

[3] Bis 2003 stellten diese die durch die jeweilige Bevölkerung gewichteten Stimmbezirke dar. Seit 2003 werden die Samplepoints auf Basis digitalisierter Straßenabschnitte erstellt, was die Präzision gegenüber Stimmbezirken erhöht (Häder 2015: 151 f.).

[4] Ebenfalls zur Übersicht über das ADM-Design siehe z.B. Kauermann/Küchenhoff (2011: 200 f.).

keiten der ersten Ziehungsebene über die Stratifizierung liegt ein „Probability Proportional to Size"-Design vor. Dieses verhindert eine „zufallsbedingte" starke Variation der möglichen Realisierungen von Stichproben durch variierende Größen der Cluster, welche die erste Ziehungsebene nutzt. Ein „Probability Proportional to Size" (PPS)-Design stellt somit sicher, dass die Zufallsauswahl der zweiten Stufe nicht abhängig von der Auswahl der ersten Stufe ist (Kish 1965: 217). ADM-Stichproben sind somit Haushaltsstichproben, da diese die eigentliche Auswahlgrundlage darstellen. Soll eine Aussage auf Basis der Individualebene getroffen werden, muss daher eine Transformationsgewichtung angewandt werden (siehe hierzu Kapitel 4.2.1). Dies bedingt sich durch die unterschiedlichen Auswahlwahrscheinlichkeiten in Abhängigkeit der Anzahl der zur Grundgesamtheit gehörenden Personen („reduzierte Haushaltsgröße"). Bei Registerstichproben handelt es sich hingegen um Personenstichproben, da diese direkt durch die Einwohnermeldeämter ausgewählt werden. Ein Transformationsgewicht ist nicht notwendig, sofern Aussagen auf der Individual- und nicht Haushaltsebene getroffen werden sollen.

Ein Vorteil des ADM-Designs gegenüber registergestützten Stichproben ist die hohe Aktualität der Adressen durch die Begehung und gleichzeitige Befragung durch ein Standard-Random. Da der Anteil nicht erfolgter An- und Abmeldungen bei den Einwohnermeldeämtern durch umgezogene Personen in Großstädten mehrere Prozent der Bevölkerung ausmachen kann, sind diese Register nicht immer aktuell (Schnell 2012: 205 f.). Die Interviewer bei einem Standard-Random durch das ADM-Design sind jedoch auch eine Fehlerquelle. Ignorieren diese systematisch bestimmte Häuser oder Straßenzüge während der Begehung, haben die darin wohnenden Personen keine Chance der Auswahl. Die Folge ist ein Noncoverage-Fehler (z.B. Eckman/Kreuter 2013). Dieses Problem weist die Stichprobenziehung auf der Basis der Einwohnermeldeämter wiederum nicht auf, die Interviewer keine Möglichkeit der Substitution existierender, aber schwer zu begehende, Adressen durch andere haben. Dieses Vorgehen wird von Interviewern angewandt, um künstlich ihre Ausschöpfungsquote zu erhöhen (Koch 2002: 33 f.).

Noncoverage lässt sich jedoch, auf Grundlage der verwendeten Auswahldesigns persönlicher Befragungen, als ein im Vergleich zu anderen Auswahlgrundlagen geringes Problem bezeichnen (z.B. Busse/Fuchs 2012; Joye et al. 2012). Aus diesem Grund haben sich räumliche Stichproben persönlich-mündlicher Befragungen auch als ein qualitativer Erhebungsstandard etabliert, sofern die Vermeidung eines Noncoverage-Fehlers das zentrale Ziel ist (Groves/Fowler et al.

2009: 163). Dieser Vorteil durch die räumliche Stratifizierung ist jedoch mit einer Einschränkung verbunden: Die regionale Stratifizierung kann einen „Design-Effekt“ verursachen (Kish 1965: 255-259), welcher bei einem Schluss von der vorliegenden Stichprobe auf die Grundgesamtheit berücksichtigt werden muss (Kish 1995: 58). Dieser Effekt tritt dadurch auf, dass die Antworten der befragten Personen innerhalb eines Klumpens der ersten Ziehungsebene homogener zueinander, als zu allen restlichen befragten Personen sind. Die wahre Variation innerhalb der Stichprobe ist daher nicht so groß, wie die Stichprobengröße dies suggeriert (Häder/Ganniner/Gabler 2009: 15 f.). Auch die Aufteilung der Befragung auf unterschiedliche Interviewer kann einen solchen Effekt hervorrufen und den der regionalen Stratifizierung übersteigen (z.B. Schnell/Kreuter 2005). In der Praxis findet sich diese Berücksichtigung, trotz klarer Auswirkungen auf die Genauigkeit der Ergebnisse, selten (Beullens/Loosveldt 2016).

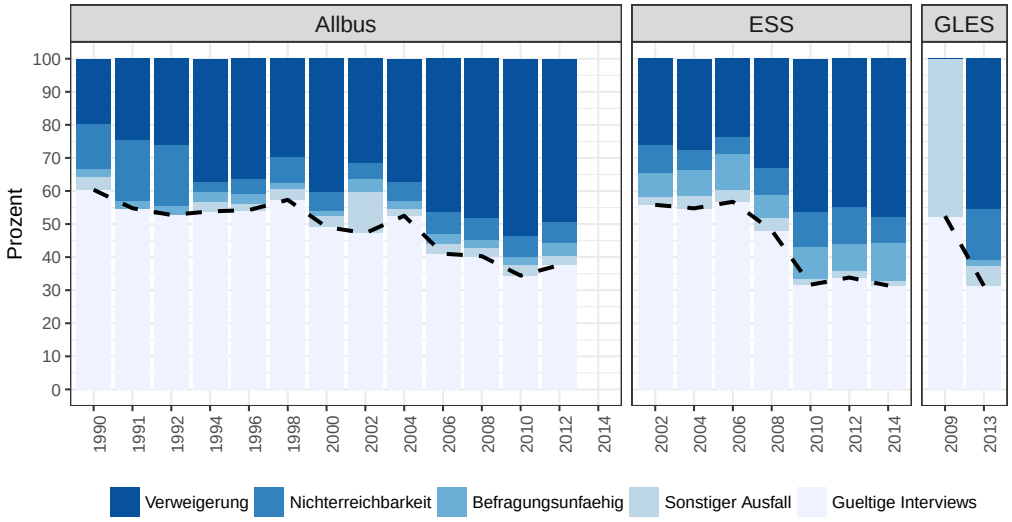
Ist das Auswahldesign entschieden und aus der Auswahlgesamtheit eine Stichprobe zur Befragung bestimmt, haben persönliche Befragungen eine vergleichsweise gute Prognose hinsichtlich ihrer Teilnahmeraten. Das Auswahldesign und die Durchführung dieser weisen im Gegensatz zu anderen Befragungsformen traditionell höhere Teilnahmeraten auf (z.B. Goyder 1987; Hox/Leeuw 1994), die Gefahr eines Nonresponse-Bias ist dadurch niedriger.[5] Auch persönliche Befragungen sind jedoch durch den Trend fallender Teilnahmeraten betroffen.

Abbildung 6.1 skizziert die Entwicklung der Teilnahmeraten des Allbus, ESS und der ebenfalls persönlich-mündlichen Erhebung der Komponente 1 der GLES[6] als Anteil der realisierbaren Interviews an der Gesamtanzahl der ausgewählten Personen. Über alle drei aufgeführten Umfrageprogramme zeigt sich, trotz unterschiedlich langer Zeitreihen der Erhebung, ein gemeinsamer Trend: Auch die Teilnahmeraten der traditionell als hochwertig angesehenen persönlichen Befragungen sinken durch Verweigerungen. Weist der Allbus von 1990 bis 1998 noch Werte von nahe 60 % aus, sinkt der Anteil der erfolgreich ausgewählten *und* befragten Personen auf unter 40 % seit 2006. Auch der ESS zeigt diese Tendenz, die Quote fällt von 55 % in 2002 auf nur noch knapp über 30 % seit 2010. Dies lässt sich nahezu identisch vom ESS auf die Komponente 1 der GLES übertragen. Weist die erste Erhebung in 2009 noch eine Rate von über

[5] Dieser hohe Anspruch der Teilnahmeraten bei der Durchführung persönlicher Interviews zeigt sich auch in der Zielformulierung einer Ausschöpfungsquote von 70 % der teilnehmenden Länder innerhalb des ESS (European Social Survey 2014: 9), auch wenn die realisierten Quoten teilweise deutlich darunter liegen (z.B. Bethlehem/Cobben/Schouten 2011: 161).

[6] Hierbei handelt es sich wie die anderen beiden Umfragen um eine persönliche Befragung, welche durch das ADM-Design die wahlberechtigte Bevölkerung anvisiert. Für eine detailliertere Beschreibung siehe Kapitel 5.1.

Abbildung 6.1.: Entwicklung Ausschöpfungsquote und Ausfallgründe bei persönlichen Befragungen



Quelle: Methodenberichte Allbus und Komponente 1 der „German Longitudinal Election Study“ (GLES), sowie Angaben unter http://www.europeansocialsurvey.org/essdoc/ (Abruf: 11. Mai 2018) für den ESS. Einordnung Ausfallgründe ESS an die Klassifikation des Allbus angepasst. Für die Komponente 1 ist für 2009 nur der allgemeine Anteil der systematischen Ausfälle angegeben, weshalb diese insgesamt als „Sonstige“ kodiert wurden.

50 % auf, fällt diese in 2013 bereits unter 40 %.

Die Verweigerung durch erfolgreich ausgewählte Personen dominiert in allen Erhebungen. Auch die Nichterreichbarkeit kann in einigen Erhebungen der aufgeführten Umfrageprogramme einen Anteil von 10 Prozent aller zu befragenden Personen ausmachen, mit Bezug auf die deutlichen Sprünge der Erhebungen des Allbus bis einschließlich 2000 ist dies jedoch mit dem angewandten Auswahldesign verbunden. Wurde in den Jahren bis 1992 und 1998 das ADM-Design angewandt,[7] ist dies in den restlichen Jahren eine Auswahl auf der Basis der Einwohnermeldeämter gewesen. Das ADM-Design weist somit höhere Anteile der Nichterreichbarkeit auf, auch wenn sich der Wechsel des Auswahldesigns nicht auf die Höhe der Teilnahmeraten auswirkt.

Die aufgeführten Entwicklungen der Teilnahmeraten an persönlichen Befragungen zeigen eine klare Tendenz: Weniger als jede zweite Person entscheidet sich nur noch zur Partizipation an den prominenten Umfrageprogrammen der Sozialwissenschaften. Wird die Verweigerung durch bestimmte interessierende gesellschaftliche Gruppen häufiger getroffen, ist ein Nonresponse-Fehler die Folge.

[7] Siehe hierzu die Beschreibung unter: http://www.gesis.org/allbus/allbus/allgemeine-informationen/stichproben-und-erhebungsdesign/ (Stand: 11. Mai 2018).

Eine bedeutende Rolle bei persönlichen Befragungen haben zudem die Interviewer. Deren Einsatz hat einige Vorteile, wie eine längere Befragungsdauer, mögliche Hilfestellungen für die zu befragende Person und die damit verbundene Möglichkeit der Nutzung komplexerer Fragen oder auch der motivierende Einfluss auf die Teilnahme einer Person (Schnell 2012: 308 f.). Auch kann der Interviewer bereits Daten wie das Geschlecht erheben, ohne die befragte Person hiernach fragen zu müssen (Bethlehem/Cobben/Schouten 2009: 95 f.). Der Einsatz von Interviewern bringt jedoch auch Nachteile mit sich: Die regionale Stratifizierung der Erhebung erhöht den Zeit- und Kostenaufwand durch Reisetätigkeiten, die Kosten persönlicher Erhebungen sind daher sehr hoch.[8] Auch ist die Qualitätskontrolle der Interviewer hinsichtlich möglicher Fehler und Manipulationen aufwendig (Bethlehem/Cobben/Schouten 2009: 94). Der zentrale Nachteil von Interviewern ist jedoch ihre hervorgerufene Reaktivität auf die zu befragende Person. Hierdurch kann ein Messfehler entstehen, wenn die gegebenen Antworten allgemeine Zustimmungstendenzen ohne Bezug zum Inhalt ausdrücken oder aus sozial erwünschten Gründen geäußert werden (Schnell/Esser/Hill 2013: 345 ff.).[9] Dieser Einfluss variiert zudem zwischen den Interviewern und reduziert hierdurch die effektive Stichprobengröße durch einen hervorgerufenen Design-Effekt (z.B. Kish 1962; Mangione/Fowler/Louis 1992; O'Muircheartaigh/Campanelli 1999; Schnell/Kreuter 2005). Ebenso wie der Effekt der räumlichen Stratifizierung führt daher der Einsatz von Interviewern zu einer zu geringen Einschätzung der Variation einer Variablen. Je stärker der interviewerspezifische Effekt und je höher der Anteil eines individuellen Interviewers an der Gesamtzahl an Interviews ist, umso stärker ist dieser Einfluss (Kish 1962: 94 f. Schnell 2012: 208 f.). Die erzeugte Homogenität der Interviewer ist zudem nicht selten stärker als der Einfluss der räumlichen Stratifizierung (O'Muircheartaigh/Campanelli 1999; Schnell/Kreuter 2005).

Mehrere Erklärungen sind hierfür denkbar. Verschiedene Interviewer motivieren unterschiedliche Arten von Personen zur Teilnahme an der Umfrage (z.B. O'Muircheartaigh/Campanelli 1999; West/Olson 2011; West/Kreuter/Jaenichen 2013). Eine andere Möglichkeit ist der Effekt unterschiedlicher Erfahrungen, Ansichten und psychologischer Eigenschaften der Interviewer, welche wiederum einen Effekt auf die Kooperationsbereitschaft haben (Jäckle et al. 2012). Ebenso kommt ein zu geringer Standardisierungsgrad der Umfrage als Ursache in Be-

[8] Für eine Beispielkalkulation einer Durchführung einer bevölkerungsweiten persönlichen Befragung kommt beispielsweise Schnell (2012: 192) auf einen Betrag von mehreren Millionen €.

[9] Siehe hierzu auch die Darstellung in Kapitel **??**.

tracht. Interviewer gehen in einer unterschiedlichen Art und Qualitität mit der ihnen gewährten Handlungsfreiheit um und erzeugen daher eine beobachtbare Heterogenität (z.B. Fowler/Mangione 1990; Mangione/Fowler/Louis 1992). Praktisch ist dies darauf zurückzuführen, dass Interviewer Fragen anders als vorformuliert präsentieren oder die zu befragende Person aktiv bei einer Antwortfindung unterstützen (z.B. Kish 1962; Schnell/Kreuter 2005). Ob hierdurch eher sensitive Fragen beeinflusst werden und dadurch ein Messfehler entsteht, lässt sich auf Grund der hohen Divergenz an empirischen Resultaten jedoch kaum sagen (siehe zur Diskussion z.B. Schnell/Kreuter 2005: 391 f.). Denkbar ist auch eine gemeinsame Ursache: Interviewer mit schlechten Kooperations- und hohen Nonresponseraten sind ebenfalls die Interviewer, welche für einen hohen Teil dieser Heterogenität verantwortlich sind (z.B. Brunton-Smith/Sturgis/Williams 2012). Der Grund liegt in unterschiedlichen Erfahrungen mit Umfragen, der Zufriedenheit mit der Bezahlung sowie psychologischen Merkmalen der Interviewer (Malgorzata et al. 2015: 83).

Resümierend lässt sich für persönliche Befragungen festhalten, dass sie eine enge Verknüpfung zu einem regional stratifzierten und mehrstufigen Auswahlverfahren aufweisen. Die registergestütze Ziehung und die Anwendung des ADM-Designs stellen die maßgeblichen Varianten dieser Form der Auswahl in der Bundesrepublik dar. Ihre Anfälligkeit hinsichtlich eines Noncoverage-Fehlers ist, trotz möglicher Ungenauigkeiten der Melderegister oder der fehlerhaften Beachtung von Adressen durch Interviewer im ADM-Design, gering. Ebenfalls lässt sich ein Trend fallender Teilnahmeraten der prominenten Umfrageprogramme nachweisen. Realisierbare Teilnahmeraten von ca. 40 % erlauben ein Potential für einen Nonresponse-Bias. Zudem stellt der maßgebliche Teil des Ausfalls die bewusste Verweigerung der Teilnahme an der Umfrage dar. Hierzu wird nachfolgend die Analyse der Teilnahmeraten an telefonischen Umfragen als Vergleichsbasis zur Einschätzung der Qualität dieser Raten dienen. Neben der Fehlerquelle durch eine selektive Teilnahme an der Umfrage, hat das Kapitel zusätzlich den Einfluss des Interviewers bei persönlichen Befragungen herausgearbeitet. Hinsichtlich sensibler Fragen kann sich eine soziale Erwünschtheit oder Zustimmungstendenz durch die befragten Personen ergeben. Das nachfolgende Kapitel wird diese Erkenntnisse nun an Hand des Forschungsstands telefonischer Befragungen kontrastieren.

6.1.2. Telefonische Bevölkerungsumfragen

Ist das Ziel die Befragung großer und flächendeckender Stichproben, hat die telefonische Befragung durch ihre Schnelligkeit und Kosteneffizienz einen komparativen Vorteil gegenüber der persönlichen Variante. Daher sind sie der gängige Modus zur Befragung bevölkerungsweiter Stichproben (Schnell 2012: 245), was bereits die Abbildung 1.1 im einleitenden Teil herausgestellt hat. Auch Wahlumfragen werden in der Regel telefonisch durchgeführt. Von den auf wahlrecht.de gelisteten Instituten[10] nutzt lediglich Allensbach die persönliche Befragungsform und INSA eine onlinebasierte Variante. Alle anderen Institute befragen hingegen das Elektorat in regelmäßigen Abständen telefonisch.

Auch bei telefonischen Befragungen lässt sich eine enge Verknüpfung zwischen dem gewählten Erhebungsmodus und dem verwendeten Auswahldesign herstellen. Da für telefonische Umfragen nur unzureichende Listen von Telefonnummern (z.B. über ein Telefonbuch) als Auswahlgrundlage einer einfachen Zufallsstichprobe herangezogen werden können, impliziert die Entscheidung einer durchzuführenden bevölkerungsweiten Telefonumfrage bereits ein bestimmtes Stichprobendesign. Zur Zufallsauswahl von Anschlüssen kommt in den USA meist das „Random Digit Dialing“ (RDD)-Verfahren zum Einsatz (z.B. Lavrakas 2010: 475 f.). Hierbei werden nach einem Schema zufällig Telefonnummern generiert und hierdurch auch nicht eingetragene Telefonnummern auf der Basis modifizierter vorhandener Telefoneinträge zu erreichen (Bethlehem/Cobben/Schouten 2009: 155). In der Bundesrepublik ist für Telefonumfragen das abgewandelte „Randomized Last Digit“ (RLD)-Verfahren des ADM die häufigste Wahl. Auf Basis einer eingetragenen und verfügbaren Telefonnummer wird die letzte Zahl durch eine Zufallsauswahl verändert. Auch hierdurch lassen sich nicht eingetragene Anschlüsse erreichen, da Festnetznummern in der Bundesrepublik in Blöcken organisiert sind (Schnell/Esser/Hill 2013: 281 f.). Der Nachteil ist lediglich, dass die Auswahlwahrscheinlichkeit einer Festnetznummer dann von der Anzahl der Nummern im zugehörigen Nummernblock abhängig ist. Eine perfekte Zufallsauswahl kann also auch das RLD-Verfahren nicht garantieren (Schnell 2012: 269).

Das zentrale Problem von Telefonumfragen sind jedoch zunehmende Probleme der Abdeckung der gesamten Bevölkerung auf der Basis dieser aufgeführten Auswahldesigns. Durch die zunehmende alleinige Nutzung von Mobilfunknummern existiert ein größer werdender Personenkreis, welcher nicht

[10] Siehe zur Übersicht der Institute die Darstellung auf http://www.wahlrecht.de/umfragen (Abruf: 11. Mai 2018).

mehr zur Auswahlgesamtheit dieser Verfahren auf der Basis von Festnetzanschlüssen gehört (z.B. Ehlen/Ehlen 2007; Lavrakas et al. 2007; Peytchev/Carley-Baxter/Black 2010, 2011; Busse/Fuchs 2012; Hox/Mohorko/Leeuw 2013). Diese Entwicklung zeigt sich sowohl im europäischen (z.B. Busse/Fuchs 2012; Heckel/Wiese 2012) als auch US-amerikanischen Raum (z.B. Blumberg/Luke 2014). In den USA sind bereits zwei von fünf (erwachsenen) Personen nur noch über eine Mobilfunknummer telefonisch erreichbar (Blumberg/Luke 2014: 5). Unterscheidet sich dieser Personenkreis alleiniger Mobilfunknutzer vom Rest der Bevölkerung, kommt es zu einer Verzerrung durch einen Noncoverage-Fehler (Bethlehem/Cobben/Schouten 2011: 100 ff.). Auch innerhalb der EU ist mittlerweile jede dritte Person nicht mehr über einen Festanschluss erreichbar, wobei sich das Ausmaß als sehr heterogen zwischen den Mitgliedsstaaten zeigt: In der Bundesrepublik trifft die alleinige mobile Erreichbarkeit nur auf ca. 10 % der Bevölkerung zu. Ist ein ebenso hoher Anteil nur über das Festnetz erreichbar, hat der überwiegende Anteil der Bevölkerung auf beide Formen der Telekommunikation Zugriff (European Commission 2014: 15 ff.). Auch die Zahlen des ADM weisen einen Anteil von 12 % reiner Mobilfunknutzer an der gesamten Bevölkerung ab 14 Jahren aus (ADM 2012). Relevant für einen Noncoverage-Bias ist jedoch nicht (nur) das Ausmaß des Problems, sondern auch der mögliche Unterschied der reinen Mobilfunknutzer zu denjenigen mit einem Festnetzanschluss (siehe hierzu auch die Darstellungen in Kapitel 3.3). Dieser Unterschied findet sich: Die „Mobile-Onlys“ sind vor allem jünger als 30 Jahre und leben häufig in Einpersonenhaushalten (European Commission 2014: 31). Ebenfalls zeigt die reinen Mobilfunknutzer ein geringeres Interesse an Politik (z.B. Joye et al. 2012).

Der Anteil der Bevölkerung der Bundesrepublik mit einem vorhandenen, oder sogar exklusiven, Festnetzanschluss ist also hoch. Trotzdem dürfte die Tendenz durch die soziodemografischen und psychologischen Unterschiede dieser Gruppe zu den alleinigen Mobilnutzern die Notwendigkeit der Nutzung von „Dual-Frame“-Erhebungen in Zukunft steigen lassen (Schneiderat/Schlinzig 2012). Umfrageinstitute bevölkerungsweiter telefonischer Wahlumfragen gehen in der Bundesrepublik bisher unterschiedlich mit diesem Problem um. Der Deutschlandtrend bezieht bereits seit Januar 2013 zu 30 % Mobilfunknummern in sein Erhebungsdesign durch ein „Dual-Frame“ ein. Laut Eigeneinschätzung verbessert dies die korrekte Abbildung der Altersstruktur der Bevölkerung

und die Prognose der Parteipräferenzen zu Gunsten der kleineren Parteien.[11] Auch die Erfahrung der Stichprobenqualität der CELLA-Studie zeigt, dass die Bereicherung reiner Festnetzstichproben durch alleinige Mobilfunknutzer die Qualität steigern kann (Graeske/Kunz 2009). Die Forschungsgruppe Wahlen hat sich hingegen (vorläufig) dazu entschieden, keine „Dual-Frame"-Erhebung zur Befragung im Rahmen des Politbarometers umzusetzen. Der Anteil und die Zusammensetzung der über das Festnetz erreichbaren Bevölkerung wird als noch adäquat eingeschätzt (Hunsicker/Schroth 2014). Auch Forsa begründet die weiterhin stattfindende alleinige Nutzung von Telefonstichproben auf der Festnetzbasis mit den bisherigen guten Resultaten der fortlaufenden Evaluation der Qualität der eigenen Wahlprognosen.[12]

Werden Erhebungen der gleichzeitigen Befragung über Festnetz- und Mobilfunkanschlüsse zum Standard, bedroht dies jedoch ebenfalls die Datenqualität von Telefonumfragen. Zunehmende Versuche der Kontaktaufnahme über das Mobilfunknetz, in unterschiedlichsten Lebenslagen von potentiellen Umfrageteilnehmern, werden schlicht die Akzeptanz dieser Form der Befragung weiter verringern. Im ungünstigsten Fall droht telefonischen Befragungen die Rolle einer Befragungsmethode spezieller Teilbereiche der Bevölkerung, welche diese Form der Kontaktaufnahme (noch) tolerieren (Schnell 2012: 285). Geringe Rückgänge der Teilnahmeraten an Telefonumfragen können hierbei bereits entscheidend sein. Durch die intensive Nutzung telefonischer Befragungen weisen diese schon eine längere Zeit eine geringere Teilnahmebereitschaft als persönliche Befragungen auf (Schnell 2012: 290 f.). Ausschöpfungsquoten von lediglich 20 % sind daher auch für professionell angelegte wissenschaftliche Telefonumfragen keine Seltenheit (z.B. Häder/Ganniner/Gabler 2009: 72; Schneiderat/Schlinzig 2012: 124 f.). Neben der bewussten Verweigerung der Teilnahme sind hierfür vor allem auch zunehmende Kontaktprobleme verantwortlich (Fuchs 2010: 235), welche im Rahmen der dargestellten persönlichen Interviews nur eine geringe Bedeutung als Ausfallgrund hatten (Abbildung 6.1).

Eine Abschätzung der Ausfallgründe ist über die transparente Aufschlüsselung innerhalb der telefonischen Befragung der Komponente 2 der GLES möglich. Für 2013 verzeichnet der Studienbericht einen Anteil von 64,4 % der systematischen Ausfälle, welche durch die Verweigerung der Teilnahme durch die Kontaktperson entstehen. Hierzu kommen noch einmal 9,4 % der Fälle, in denen die erreichte Zielperson die Teilnahme verweigert. Kontaktprobleme

[11] Quelle: http://www.research-results.de/fachartikel/2014/ausgabe-7/bei-anruf-frage.html (Abruf: 11. Mai 2018)

[12] Quelle: https://www.forsa.de//dialog-mit-forsa/ (Abruf: 11. Mai 2018).

(„Teilnehmer hebt nicht ab/Anrufbeantworter“ und „Anschluss besetzt“) machen einen weiteren Anteil von 6,6 % der Fälle aus. Verweigerungen machen insgesamt einen Anteil von beinahe 90 % an der Gesamtzahl der als *systematisch klassifizierten* Ausfälle aus. Lediglich 15,8 % der Nettostichprobe können schließlich noch als Interview realisiert werden (Rattinger et al. 2014b: 7). Dies ist ein nochmals niedrigerer Wert im Vergleich zur Durchführung der Komponente 2 in 2009, bei der die Ausschöpfungsquote noch bei 20 % lag. Verweigerungen durch die Kontakt- (52,3 %) oder Zielperson (12,6 %) machen jedoch auch hier bereits einen Anteil von ca. 80 % der als systematisch klassifizierten Ausfälle aus. Mit 11,3 % ist zudem der Anteil der Kontaktprobleme in 2009 noch größer gewesen (Rattinger/Roßteutscher/Schmitt-Beck/Weßels 2013: 7). Auch in den USA zeigt sich dieser Trend. Seit Mitte der 1990er verzeichnet beispielsweise das „Pew Research Center“ klar sinkende Raten der Kontakt-, Kooperations- und Teilnahmeraten. Letztere sank von 36 % auf lediglich 9 % in 2012.[13] Da in der Bundesrepublik jedoch keine einheitlichen Standards zur Erfassung von Gründen für eine Nichtteilnahme eigentlich ausgewählter Personen existiert,[14] ist eine Einordnung der Gründe der Nichtbefragung zwischen verschiedenen Studien nur mit Einschränkungen möglich (z.B. Fuchs 2010: 235 f.). Erschwerend kommt hinzu, dass kommerzielle Umfragen ihre Teilnahmeraten und Klassifikation der Ausfälle erst gar nicht offen legen. Es ist nicht unwahrscheinlich, dass diese mit einer noch geringeren Kooperationsbereitschaft konfrontiert sind und daher lediglich Ausschöpfungsquoten von weit unter 20 % aufweisen. Die telefonische Kooperationsbereitschaft liegt damit eindeutig niedriger als bei einer persönlichen Kontaktaufnahme, was auch die Literatur belegt (z.B. Blohm 2006). Die Probleme des Noncoverage durch das Auswahldesign bei telefonischen Befragungen und die, im Vergleich zur persönlichen Befragung, noch einmal stark erhöhten Verweigerungsraten ergeben ein hohes Potential für eine verzerrte Erfassung auch inhaltlich interessierender Zielvariablen. Die Wahrscheinlichkeit von einem Bias betroffen zu sein ist klar erhöht.

Da Telefonumfragen ebenfalls interviewerbasiert durchgeführt werden, existieren auch bei diesen die Vor- und Nachteile durch den Einsatz eines Interviewers. Gegenüber persönlichen Befragungen lässt sich jedoch ein höherer Grad der Standardisierung und eine bessere Interviewerkontrolle erreichen, da

[13] Quelle: http://www.people-press.org/2012/05/15/assessing-the-representativen-ess-of-public-opinion-surveys/. Abruf: 11. Mai 2018.

[14] Im Gegensatz beispielsweise zum us-amerikanischen Raum, wo mit den „American Association for Public Opinion Research“ (AAPOR)-Standards (AAPOR 2008, 2011) eine solche Klassifikation ausgearbeitet wurde.

telefonische Befragungen in speziell dafür eingerichteten Räumen durchgeführt werden. Auch die erneute Kontaktierbarkeit nicht erreichter Personen gestaltet sich einfacher, schneller und kostengünstiger. Zudem kommt es nur zu einem mündlichen Kontakt. Dies schwächt die Beeinflussung der zu befragenden Person gegenüber der visuellen Kontaktaufnahme bei persönlichen Befragungen ab (z.B. Häder/Ganniner/Gabler 2009: 14 ff. Lavrakas 2010: 471f.). Der Einsatz der Interviewer kann jedoch auch bei telefonischen Befragungen einen Design-Effekt verursachen. Die negativen Auswirkungen können hier ein Mehrfaches derjenigen in persönlichen Befragungen ausmachen, da hier die Anzahl der Interviews pro Interviewer viel höher ist. Hierdurch ist die Auswirkung des individuellen Effekts eines Interviewers sehr viel stärker (Schnell 2012: 279).

Dass Interviewer einen Effekt auf die Genauigkeit der Ergebnisse haben, liegt an dem bereits angesprochenen Design-Effekt. Dieser kann durch die räumliche Stratifizierung der Erhebung im Rahmen persönlicher Befragungen hervorgerufen werden. Das nachfolgende Kapitel 6.1.3 wird nun darstellen, welche allgemeinen Auswirkungen solche Design-Effekte generall auf die Genauigkeit von Schätzern auf der Basis von komplexen Stichprobenverfahren haben können und wie deren Auswirkungen bei einem Schluss auf die Grundgesamtheit berücksichtigt werden müssen.

6.1.3. Design-Effekte bei komplexen Erhebungsdesigns

Mehrstufige komplexe Stichprobendesigns können von einem „Design-Effekt“ durch die Stratifizierung der Erhebung betroffen sein (Kish 1965: 255 f.). Dieser tritt bei der Anwendung eines mehrstufigen Stichprobenverfahrens auf, da sich die befragten Personen innerhalb eines Erhebungsclusters ähnlicher sind als zu dem Rest der befragten Personen (Häder/Ganniner/Gabler 2009: 15 f.). Der Effekt ist umso stärker, je ausgeprägter diese Homogenität innerhalb der Cluster ist und je weniger Cluster insgesamt zur Erhebung einer festen Zahl an Personen genutzt werden. Als Folge der Homogenität wird die Varianz von interessierenden Variablen, welche von dieser Homogenität betroffen sind, als zu gering geschätzt. Der zunehmende Design-Effekt führt zu einer geringeren effektiven Stichprobengröße, als die Anzahl der befragten Personen suggeriert (Kish 1965: 161 ff.). Bei regional stratifizierten Stichproben tritt dieser Effekt auf, weil die soziodemografischen Rahmendaten, Einstellungen und Ansichten von Personen in der Regel ähnlicher zu ihren direkten Nachbarn als zu denjenigen benachbarter Stadtteile oder sogar anderer Regionen sind. Die wahlberechtigte Bevölkerung des Wahlkreises „Düsseldorf I“ wird im Mittel andere Einstellun-

gen zu verschiedenen Themen aufweisen, als die Bevölkerung der Wahlkreise „Mecklenburgische Seenplatte“ oder des „Rhein-Hunsrück-Kreises”.

Das Prinzip regional stratifizierter komplexer Stichprobendesigns nutzt genau diesen Effekt. Statt einer Vollerhebung innerhalb der ausgewählten Cluster durch eine Klumpenstichprobe, ist eine einfache Zufallsauswahl der relativ homogenen Personen innerhalb eines Clusters ausreichend (Bethlehem/Cobben/Schouten 2009: 112). Da innerhalb des ADM-Designs ein Ziehungspunkt durchschnittlich 600 bis 700 Haushalte enthält (Häder 2015: 151 f.), würde bereits die Ziehung von drei Punkten die ungefähre Stichprobengröße vieler Befragungen von ca. 2000 Personen realisieren. Effektiv reicht jedoch bereits eine viel geringere Zahl aus, um die Variation *innerhalb* jedes Clusters ausreichend zu erfassen. Ob insgesamt ein „Design-Effekt“ vorliegt, hängt deshalb davon ab, welcher Effekt stärker ist: Der Präzisionsgewinn durch die Stratifizierung oder der Präzisionsverlust durch die zu starke Homogenität innerhalb der Cluster.

Diese mögliche Unsicherheit der Prognose bezüglich der Schätzung der Parameter der Grundgesamtheit auf der Grundlage von Schätzern eines komplexen Stichprobendesigns muss berücksichtigt werden. Ansonsten drohen zu enge Konfidenzintervalle und zu optimistische Testaussagen, da einfache Anwendungen inferenzstatistischer Verfahren für einfache Zufallsstichproben konzipiert sind (z.B. Bacher 2009: 255 f.; Lohr 2010: 309 ff.; Schnell/Esser/Hill 2013: 284 f.). Ist die Genauigkeit der Prognose in Form eines Konfidenzintervalls nicht von Interesse oder die Annahme einer annähernden einfachen Zufallsziehung gegeben, ist hingegen keine Korrektur notwendig (Kish 1995: 58). Die Bestimmung des „Design-Effekts“ $deff$ ergibt sich als das Verhältnis der Varianz eines Schätzers bei Anwendung des komplexen Designs im Verhältnis zur Varianz dieses Schätzers bei einer einfachen Zufallsziehung identischen Umfangs (Kish 1965: 258). Die Wurzel von $deff$ ergibt das Verhältnis der Standardfehler des jeweiligen Schätzers ($deft$) und kann als Korrekturfaktor für die Berechnung von Konfidenzintervallen, zum Beispiel für das arithmetische Mittel einer Variable ($\bar{x}$), genutzt werden (Kish 1995: 55):

$$\sqrt{deff_{\bar{x}}} = deft_{\bar{x}} = \frac{\sigma_{\bar{x}_{sample}}}{\sigma_{\bar{x}_{SRS}}} = \sqrt{\frac{var(\bar{y})}{s^2/n}} \tag{6.1}$$

Der Faktor $deft$ erfasst durch das Verhältnis der beiden Varianzen somit den Effekt des Erhebungsdesigns auf die Prognosesicherheit eines Schätzers einer Variable, unabhängig von der jeweiligen Skala und Stichprobengröße (Kish 1995: 55). Für einen t-Test von zwei unabhängigen Stichproben lässt sich $deff$

analog bestimmen (Kish 1995: 64):

$$deff(\bar{y}_1 - \bar{y}_2) = deft^2(\bar{y}_1 - \bar{y}_2) = \frac{Var(\bar{y}_1) + Var(\bar{y}_2)}{SRS\ Var(\bar{y}_1) + SRS\ Var(\bar{y}_2)} \tag{6.2}$$

Ein Design-Effekt ist damit keine generelle Eigenschaft eines erhobenen Datensatzes, sondern tritt für jede Variable in einem unterschiedlichen Ausmaß auf. Je stärker die Cluster der Befragten einen Einfluss auf die Antworten einer Variable haben, umso stärker ist auch der dort betrachtete „Homogenisierungseffekt". Als Folge fällt die effektive Stichprobengröße bezüglich der Erfassung dieser Variable geringer aus. Für jede Variable muss der Design-Effekt daher getrennt bestimmt werden.[15] Neben dieser Homogenität innerhalb der Cluster, gemessen durch die Intraklassenkorrelation ρ, ist weiterhin der Grad der genutzten Clusterung relevant. Dieser bestimmt wie viele Personen (m) im Mittel pro Ziehungspunkt befragt werden (Schnell/Esser/Hill 2013: 284):

$$deft = \sqrt{1 + \rho \cdot (m - 1)} \tag{6.3}$$

Bei zunehmender Nutzung der Clusterung eines Erhebungsdesigns (steigendes m) wirkt sich ein gegebener Design-Effekt umso stärker auf die Prognosequalität von Schätzern aus. Dies zeigt der Vergleich der angegebenen Sympathien verschiedener Parteien innerhalb der Komponente 1 der GLES in Tabelle 6.1.

Tabelle 6.1.: Design-Effekt: Parteiensympathien Komponente 1 (GLES 2013)

Partei	$\bar{x}$	$\hat{\sigma}_{\bar{x}}^{SRS}$	$\hat{\sigma}_{\bar{x}}^{ADM}$	$\hat{\sigma}_{\bar{x}}^{INT}$	$deft_{ADM}$	$deft_{INT}$	ρ_{MSRS}	ρ_{INT}
CDU	6.72	0.0756	0.1016	0.1194	1.34	1.58	0.11	0.14
CSU	5.89	0.0799	0.1139	0.1380	1.43	1.73	0.15	0.18
SPD	6.66	0.0618	0.0834	0.0955	1.35	1.55	0.12	0.13
FDP	5.01	0.0633	0.0868	0.1012	1.37	1.60	0.12	0.14
Linke	4.67	0.0721	0.1036	0.1281	1.44	1.78	0.15	0.20
Grüne	5.95	0.0712	0.1052	0.1139	1.48	1.60	0.17	0.14

Quelle: Eigene Berechnungen. Datengrundlage: Komponente 1 der GLES für 2013 (ZA5700). Angewandtes Gewicht: *w_trow*. Standardfehler bei einer einfachen Zufallsziehung (SRS), Berücksichtigung der regionalen Stratifizierung (MSRS; „Primary Sampling Unit" (PSU): vpoint, strata: ostwest) oder der Interviewerclusterung (INT; PSU: intnum, strata: ostwest). *deft* nach Formel 6.1. 8.11 befragte Personen pro „Sample Points" (SP) im Mittel, 10.95 pro Interviewer. Für ρ wurde Gleichung 6.3 umgestellt.

Aufgeführt sind die im arithmetischen Mittel gemessenen Bewertungen und

[15] Über Statistikprogramme lässt sich dieser für eine Variable bestimmen, indem man dem jeweiligen Programm die Anwendung eines mehrstufigen Auswahlverfahrens mit Clusterung mitteilt. Für R lässt sich dies über den *svydesign*-Befehl definieren, welcher auch für alle späteren Anwendungen genutzt wurde. Siehe Kauermann/Küchenhoff 2011: 214 ff. für eine praktische Umsetzung.

die zugehörigen geschätzten Standardfehler bei Unterstellung einer einfachen Zufallsziehung (SRS). Alternativ wurde das mehrstufige ADM-Design oder die Berücksichtigung der Aufteilung auf verschiedene Interviewer (INT) bei der Bestimmung berücksichtigt. Ein Design-Effekt der räumlichen Stratifizierung findet sich über die Bewertungen aller Parteien hinweg in einem vergleichbaren Ausmaß, die Grünen sind marginal am stärksten betroffen. Der Effekt der Aufteilung auf die Interviewer ist über alle Parteien hinweg stärker, die Homogenität nach ρ jedoch nur leicht erhöht, bei den Grünen sogar geringer. Der stärkere Design-Effekt durch die Interviewerauftеilung ist also nicht durch einen größeren Einfluss dieser auf die Homogenität der gegebenen Antworten zurückzuführen. Mit durchschnittlich 10.95 befragten Personen pro Interviewer an Stelle von 8.11 pro Ziehungspunkt ist die Clusterung schlicht geringer genutzt worden, was gemäß Gleichung 6.3 die Homogenität ρ ebenso bedingt. Der Effekt der räumlichen Stratifzierung und die Aufteilung auf die Interviewer lässt sich analytisch jedoch nur in einem „interpenetrierenden Design“ trennen. Da Interviewer auf ausgewählte regionale Ziehungspunkte zugewiesen werden, stellen diese ansonsten ebenfalls regionale Einheiten dar (Schnell/Kreuter 2005: 393 f.). Es ist also möglich, dass die Interviewerclusterung der Komponente 1 der GLES ebenso durch die räumliche Clusterung verursacht wird. Diese Fragestellung lässt sich mit den vorliegenden Daten nicht auflösen.

Im Fall von Tabelle 6.1 wäre die Berechnung eines α-Konfidenzintervalls für eine einfache Zufallsstichprobe für alle Parteien zu optimistisch hinsichtlich der wahren Genauigkeit der Prognose. Die Berücksichtigung eines Design-Effekts mit einem $deft > 1$ führt daher zu einer breiteren Schätzung eines Konfidenzintervalls, da hierdurch die Unsicherheit der Prognose durch die vorhandene Homogenität der Clusterung berücksichtigt wird (z.B. Lohr 2010: 310 f.):

$$\hat{\bar{y}} \pm 1.96\sqrt{deff}\sqrt{\frac{\hat{S}^2}{n}} \quad \text{mit:} \quad \sqrt{deff} = deft(\bar{y}) = \frac{SE(\hat{\bar{y}}_{\text{plan}})}{\frac{s}{\sqrt{n}}} \tag{6.4}$$

In der Praxis ist die Notwendigkeit einer solchen Berücksichtigung eines komplexen Stichprobendesigns durch die Korrektur der Standardfehler die Regel (z.B. Schnell 1997: 272-284). Auch Konfidenzintervalle eines Anteilswertes p bedürfen einer Korrektur. Ansonsten wird eine zu optimistische Genauigkeit der Prognose für den unbekannten Anteilswert θ der Population angegeben. Die Verwendung der lehrbuchhaften Vermittlung berücksichtigt dies jedoch nicht (Schnell/Noack 2014: 12). So weisen beispielsweise Gehring/Weins (2010: 267

f.) die Bestimmung aus als:

$$p \pm z_{1-\frac{\alpha}{2}} \cdot \sigma_p \quad mit\ \hat{\sigma}_p = \sqrt{\frac{p \cdot (1-p)}{n}} \tag{6.5}$$

Zur korrekten Bestimmung muss jedoch bezüglich der Schätzung der Varianz des Anteilswerts der Effekt der Anwendung eines komplexen Erhebungsdesigns berücksichtigt werden (Lohr 2010: 310):

$$\hat{V}(\hat{p}) = deff \cdot \frac{\hat{p} \cdot (1-\hat{p})}{n} \tag{6.6}$$

Dann ist eine valide Bestimmung des Konfidenzintervalls eines Anteilswertes möglich (Schnell/Noack 2014: 12):

$$p \pm z_{1-\frac{\alpha}{2}} \cdot \sigma_p \cdot deft \quad mit\ deft = \frac{\hat{\sigma}_{\theta,complex}}{\hat{\sigma}_{\theta,SRS}} \tag{6.7}$$

Handelt es sich nicht um ein binomiales, sondern um k simultane multinomiale Ereignisse, wird der ursprüngliche $z_{1-\alpha/2}$-Wert alternativ durch $z_{1-\alpha/(2 \cdot k)}$ bestimmt (Schnell/Noack 2014: 16). Da dies bei Wahlumfragen auf Basis komplexer Telefonstichprobendesigns in der Bundesrepublik nicht geschieht, suggerieren diese eine Genauigkeit, welche nicht gegeben ist (Schnell/Noack 2014).[16]

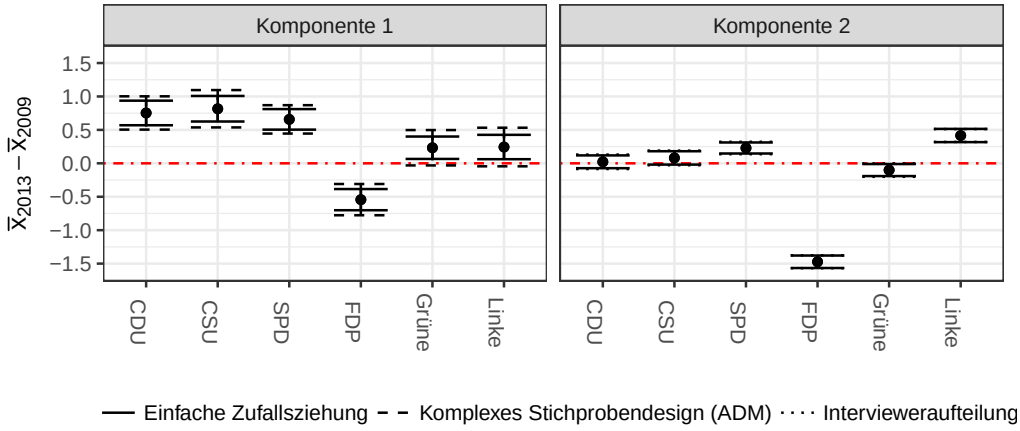
Ein Design-Effekt komplexer Stichprobendesigns berührt über falsch berechnete Standardfehler somit auch die Testentscheidungen auf Basis einer Stichprobe. Abbildung 6.2 zeigt die Veränderungen der im arithmetischen Mittel gemessenen Bewertungen der Parteien aus Tabelle 6.1 zwischen 2009 und 2013 innerhalb der Komponenten 1 und 2 der GLES. Wird an Stelle der Annahme einer einfachen Zufallsziehung das komplexe ADM-Design (Komponente 1) angewandt,[17] lässt sich die Annahme einer statistisch signifikanten Verbesserung der Bewertung der Parteien „Bündnis 90/Die Grünen“ und „Die Linke“ von 2009 auf 2013 für das persönliche Interview nicht mehr aufrecht erhalten.[18] Das Telefoninterview zeigt sich hingegen gegenüber des Einflusses der Interviewerauf-

[16] Beispielsweise verwendet das Politbarometer diese Berechnung der Prognose, auch wenn die Unsicherheit entgegenkommend gerundet wird (Quelle: http://www.forschungsgruppe.de/Umfragen/Politbarometer/Methodik/, Abruf: 11. Mai 2018).

[17] Die Komponente 1 enthält nur für 2013 die Angabe der Nummer des Interviewers.

[18] Die Veränderung von 2009 auf 2013 mit 0.23 im arithmetischen Mittel ist für „Die Grünen“, ohne Einbezug des Design-Effekts, mit $t = 2.73$ und $p = 0.0065$ für $\alpha = 0.05$ statistisch signifikant. Wird dieser berücksichtigt, ergibt sich mit $t = 1.73$ und $p = 0.0846$ für $\alpha = 0.05$ keine Verbesserung. Auch für „Die Linke“ ergibt sich dies (Differenz=0.24). Ohne Berücksichtigung findet sich eine statistisch signifikante Verbesserung für $\alpha = 0.01$ ($t = 2.62$ und $p = 0.0088$). Mit Berücksichtigung lässt sich diese für $\alpha = 0.05$ nicht mehr halten ($t = 1.66$ und $p = 0.0984$).

Abbildung 6.2.: Design-Effekt und die Schätzung der Veränderung von Parteiensympathien



Quelle: Eigene Darstellung. Dargestellt sind die jeweiligen 95 %-Konfidenzintervalle für ein multinomiales Ereignis. Korrekturfaktor $deft$ nach Formel 6.2 bestimmt. Komponente 1 ist gewichtet mit *w_trow* und das regional stratifizierte ADM-Design berücksichtigt. Für die Komponente 2 wurde die Intervieweraufteilung berücksichtigt.

teilung robust. Interviewereffekte sind möglicherweise vorhanden, variieren aber zumindestens nicht stark *zwischen* den einzelnen Interviewern. Die Unsicherheit der Prognose muss lediglich marginal korrigiert werden und die höhere Anzahl der Befragten pro Interviewer äußert sich hier nicht in einem starken Design-Effekt.[19][20]

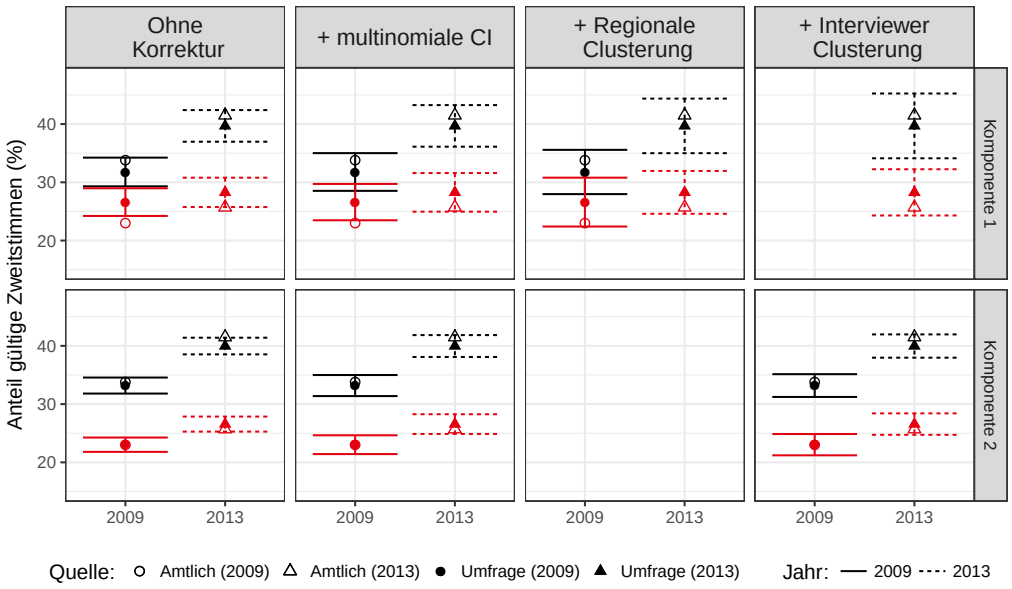
Als ein weiteres Beispiel zeigt die Abbildung 6.3 den prozentualen Anteil der Zweitstimmen von CDU/CSU und SPD der beiden Komponenten 1 und 2 der GLES. Auf Grundlage der Formel 6.5 für eine einfache Zufallsstichprobe ergibt sich sowohl für 2013 als auch für 2009 auf Grundlage des persönlichen Interviews ein statistisch signifikanter Vorsprung der CDU/CSU vor der SPD. Die Unsicherheit dieser Prognose ist jedoch zu optimistisch, das Konfidenzintervall von p_{SPD} umfasst noch nicht einmal den wahren Wert der nachfolgenden Wahl.

Durch die Berücksichtigung des multinomialen Charakters der Wahlentscheidung der abgegebenen Zweitstimmen (+multinomiale CI), des komplexen Stichprobendesigns (+Regionale Clusterung) oder der Aufteilung der zu befragenden Personen auf die Interviewer (+Interviewer Clusterung) ändert sich das Bild. Die

[19] Formal wird hierdurch mit $p = 0.0509$ bereits die Nullhypothese keiner Veränderung beibehalten und stellt eine Gradwanderung dar mit berechtigter Kritik an solchen Schwellen. Siehe hierzu z.B. auch die Stellungnahme der „American Statistical Association“ (ASA) zu dem Thema durch Wasserstein/Lazar 2016: 131 oder auch Trafimow/Marks 2014.

[20] Die auf der Basis der Standardfehler berechneten $deft$ betragen für das komplexe ADM-Stichprobendesign im Mittel $\bar{x} = 1.47$ mit $s_x = 0.0959$ und für die Interviewerauftteilung der Komponente 2 nur $\bar{x} = 1.08$ mit $s_x = 0.0636$.

Abbildung 6.3.: Design-Effekt und die Schätzung von Zweitstimmenanteilen



Quelle: Eigene Darstellung. Komponente 1 wurde mit *w_trow* gewichtet. Dargestellt sind die jeweiligen 99 %-Konfidenzintervalle.

korrigierte Bestimmung der Konfidenzintervalle auf der Basis von Formel 6.7 enthält nun den wahren Wert des Anteils der SPD (θ_{SPD}) und spiegelt somit die wahre Sicherheit der Prognose wider. Welche der beiden Parteien als stärkste Kraft aus der Bundestagswahl 2009 hervorgehen wird, ist keineswegs ein sicher prognostizierbares Ereignis auf Grundlage der vorliegenden Stichprobe der Komponente 1 der GLES. Für die Konfidenzintervalle der – relativ gut treffenden – Punktschätzer der Komponente 2 ändert sich das Bild jedoch auch nicht, wenn multinomiale Konfidenzintervalle sowie die Aufteilung auf diese berücksichtigt werden. In allen Fällen erfasst die telefonische Befragung sehr präzise und mit geringer Unsicherheit die wahren Werte der nachfolgenden Wahl. Der Effekt der Unterschiede *zwischen* den Interviewern ist, selbst durch die teilweise hohe Zahl an Interviews pro Person, erneut begrenzt.

Resümierend kann festgehalten werden: Ist das Ziel der Rückschluss auf die anvisierte Grundgesamtheit auf der Basis einer zufallsbasierten Auswahl an Personen, kann eine Korrektur notwendig sein. Diese ergibt sich durch die Anwendung eines mehrstufigen Erhebungsdesigns. Kommt es zu einer Clusterung auf Grund der regionalen Stratifizierung oder der Aufteilung auf Interviewer, können die befragten Personen innerhalb dieser Gruppen sich zu stark ähneln. Hierdurch sinkt die effektive Stichprobengröße und inferenzstatistische Verfahren treffen zu optimistische Aussagen. Wird dieser Effekt berücksichtigt, können sich daher gefundene sichere „signifikante“ Verbesserungen von Parteien als weiterhin im

Bereich der statistischen Unsicherheit dieser Aussage herausstellen. Ebenso ist die Trennschärfe der ermittelten Zweitstimmanteile der Stichprobe nicht so stark, wie die lehrbuchhafte Vermittlung der Konfidenzintervalle von Anteilswerten suggeriert. Persönliche Befragungen scheinen, auf Basis der hier dargestellten Vergleiche, stärker von diesem Problem, bedingt durch die regionale Stratifizierung, betroffen zu sein.

Um auf der Basis komplexer Auswahldesigns valide Aussagen zu treffen, muss jedoch noch eine weitere Möglichkeit bedacht werden: Das Kapitel 4.2.1 hat bereits Transformationsgewichte vorgestellt. Diese sind notwendig, da Haushaltstichproben durch das ADM-Design erhoben werden und bei Aussagen betreffend der Individualebene in Personenstichproben transformiert werden sollten. Hierdurch wird die Auswahlwahrscheinlichkeit in Abhängigkeit der Haushaltsgröße berücksichtigt. Dieser wirkt jedoch die schwierigere Erreichbarkeit von Einpersonenhaushalten entgegen. Das nachfolgende Kapitel wird diese Problematik nun im Kontext verschiedener Datensätze beleuchten.

6.2. Der neue Erhebungsmodus: Onlinebasierte Befragungen

In den vergangenen Jahren hat sich ein neuer Typus der Datenerhebung herausgebildet: Die onlinebasierte Befragung. Diese hat sich, neben der persönlichen und telefonischen Befragung, als zentraler Erhebungsmodus innerhalb der Umfragelandschaft etabliert (siehe Abbildung 1.1 des einleitenden Teils). Onlineumfragen weisen eine große Bandbreite hinsichtlich ihrer konkreten Durchführung auf (Couper 2000: z.B. zur Übersicht; Couper/Miller 2009). Auch ist die technische Entwicklung keineswegs abgeschlossen, neuere Möglichkeiten der Datenerhebung ergeben sich beispielsweise durch den Einsatz von ‘Mobile-Web-Surveys“ durch die zunehmende Verbreitung von Smartphones (z.B. Raento/Oulasvirta/Eagle 2009; Buskirk/Andres 2013; Mavletova 2013). Diese Form der flexiblen mobilen Befragung zeigt eine steigende Akzeptanz dieser neuen zeit- und ortsunabhängigen Erhebungsformen durch Umfrageteilnehmer (z.B. Toepoel/Lugtig 2014).

Die Beliebtheit onlinebasierter Befragungen ergibt sich durch ihre Vorteile. Die Möglichkeit der Umsetzung von Experimentalsituationen durch die Einbindung multimedialer Formate hat zu einer hohen Verbreitung von Onlineumfragen in der Politikwissenschaft (Clarke et al. 2008: z.B. Faas/Huber 2010) und Psychologie (Skitka/Sargis 2006: z.B.) beigetragen. Generell sind die Kosten

im Vergleich zu den interviewerad ministrierten Erhebungsformaten gering, da keine Interviewer eingesetzt und bezahlt werden müssen. Eine große Anzahl an Personen kann somit schnell und günstig befragt werden (z.B. Bethlehem/Biffignandi 2012: 45 ff. Groves 2011: 866; Evans/Mathur 2005: 196-199; Bethlehem 2009: 276 f. Fricker/Schonlau 2002; Schnell 2012: 290 f. van Selm/Jankowski 2006). Durch die Abwesenheit des Interviewers kann dieser zudem keinen Effekt auf das Antwortverhalten sozial sensibler Fragen verursachen (z.B. Holbrook/Green/Krosnick 2003; Kreuter/Presser/Tourangeau 2008; Chang/Krosnick 2009; Heerwegh/Loosveldt 2009; Chang/Krosnick 2010). Onlineumfragen weisen somit ein geringeres Potential für einen Messfehler auf, als es bei persönlichen und telefonischen Befragungen der Fall ist.

Als Nachteil der Selbstadministration ergibt sich jedoch eine mangelnde Möglichkeit der Kontrolle des Erhebungsprozesses (Groves/Fowler et al. 2009) und eine mögliche geringere kognitiven Involvierung des Befragten. Die Folge ist eine höhere Wahrscheinlichkeit der Nutzung von „Weiß nicht"-Kategorien (Heerwegh 2009). Die Selbstadministration kann ebenfalls eine soziale Enthemmung erzeugen. Personen agieren in ihrem Alltag in einen sie begrenzenden sozialen Kontext. Die Anwesenheit eines Interviewers simuliert diesen Effekt. Durch dessen Abwesenheit werden die Aussagen jedoch nicht mehr sozial normiert, die befragten Personen zeigen ein von in ihrem sozial eingebetteten Alltag abweichendes Verhalten (z.B. Taddicken 2009). Das Ziel sozialwissenschaftlicher Umfragen ist jedoch gerade die Erfassung der gesellschaftlichen und sozialen Wirklichkeit und nicht die enthemmte individuelle Gedankenwelt einer Person. Ein weiterer Nachteil sind die, noch einmal, geringeren Teilnahmeraten an Onlineumfrage gegenüber anderen Erhebungsmodi (z.B. Crawford/Couper/Lamias 2001; Manfreda et al. 2008; Shih/Xitao 2008; Maurer/Jandura 2009: 66). Die Gründe für die erhöhte Wahrscheinlichkeit eines Nonresponse ausgewählter Personen ziehen sich durch den gesamten Prozess der Konzeption und Durchführung von Onlineumfragen (Fan/Yan 2010). Schwierigkeiten im technischen Umgang mit dem Internet erhöhen zudem die Gefahr, dass die Teilnahme an der Umfrage verweigert oder abgebrochen wird (Fricker/Schonlau 2002). Diese aufgeführten Nachteile relativieren die aufgeführten Vorteile, erlauben jedoch auch weiterhin valide Schätzungen auf der Basis dieser, sofern diese keinen systematischen Fehler mit sich bringen.

Es existieren jedoch *drei weitere zentrale Probleme*, welche schnell zu der Einschätzung führen können, dass „[...] die schwerwiegenden methodischen Probleme aller Formen von Internet-Surveys [...] den Einsatz dieser Erhebungs-

form für fast alle ernsthaften wissenschaftliche Zwecke lange Zeit unmöglich machen.“ (Schnell 2012: 291). Erstens führt der Ausschluss der Bevölkerung ohne einen Internetzugang zu einem nicht zu vermeidenden Noncoverage-Fehler qua Design, da sich in der Regel die Offliner systematisch von den Onlinern unterscheiden. Zweitens ist die Teilnahmewahrscheinlichkeit an einer Onlineumfrage nicht unabhängig von der Nutzungsintensität des Internet. Steht diese technische Kompetenz auch in einem Bezug zu inhaltlichen Variablen, kommt es ebenfalls zu einem Nonresponse-Fehler (z.B. Bethlehem 2009: 277). Wie stark diese Probleme auftauchen, variiert sehr stark über verschiedene Länder und Zeiträume hinweg (z.B. Mohorko/Leeuw/Hox 2013: 615 ff.). Drittens lässt sich für onlinebasierte Befragungen nur selten eine zufallsbasierte Auswahl der zu befragenden Personen aus der anvisierten Grundgesamtheit durchführen, welche eine zentrale Voraussetzung für einen Inferenzschluss auf diese ist (Baker/Brick et al. 2013). Onlineumfragen werden meist nur online geschaltet und darauf aufmerksam werdende Personen können an dieser teilnehmen. Eine Person muss daher sowohl Kenntnis von der Umfrage haben und sich zusätzlich auch noch entscheiden an dieser zu partizipieren. Durch diese Selbstrekrutierung sind jedoch die Auswahlwahrscheinlichkeiten nicht bestimmbar und ein Inferenzschluss auf eine nicht definierbare Grundgesamtheit wird unmöglich (z.B. Couper 2000: 477; Bethlehem 2009: 285 ff.). Eine, im Sinne der Inferenzsstatistik, adäquate Durchführung von Onlineumfragen kann daher nur bei kleinen und speziellen Populationen gewährleistet werden. Hierbei sind mit einem gewissen Aufwand die Auswahlwahrscheinlichkeiten bestimmbar (Maurer/Jandura 2009: 71).[21] Onlineumfragen droht daher, dass deren Einsatz „[...] für wissenschaftliche verallgemeinerbare Studien für sehr lange Zeit sehr speziellen, hochmotivierten Teilpopulationen einerseits, begrenzten experimentellen Studien andererseits, vorbehalten [bleibt].“ (Schnell 2012: 305).

Access-Panels sind eine Herangehensweise, um zu mindestens das Problem nicht-zufallsbasierter Auswahldesigns zu umgehen. Registrierte Personen dieser haben im Vorhinein eingewilligt an Umfragen teilzunehmen und können dann im Rahmen einer Zufallsstichprobe zur Befragung ausgewählt werden. Von diesen ist dann der Inferenzschluss auf die Population wiederum möglich. Stellt das Access-Panel ein verkleinertes Abbild der Bevölkerung insgesamt dar, ist somit eine onlinebasierte Befragung der Bevölkerung insgesamt möglich. Die zentrale Herausforderung liegt jedoch darin, diese Übereinstimmung zu realisieren.

[21] Steht die Erfassung intraindividueller Veränderungen ohne eine Verallgemeinerung auf eine zu Grunde liegende Grundgesamtheit im Fokus, können nicht-zufallsbasierte Auswahldesigns ebenfalls eine adäquate Herangehensweise sein (Bergmann 2015: 113 f.).

Die nachfolgenden Kapitel werden nun diese Probleme von Onlineumfragen konkretisieren. Hierzu wird in einem ersten Schritt gezeigt, dass bereits mit der Entscheidung einer onlinebasierten Befragung in der Bundesrepublik ein nicht zu vermeidender Noncoverage-Fehler verbunden ist (Kapitel 6.2.1). Bevor überhaupt die erste Person ausgewählt und befragt wurde, sind die Ergebnisse der nachfolgenden Umfrage bereits mit einem systematischen Fehler belegt. Dies ergibt sich durch den starken Unterschied zwischen den Personen ohne einen Internetzugang und dem Rest der Bevölkerung. Anschließend kann das Kapitel 6.2.2 und 6.2.3 den zweiten und dritten Punkt betrachten: Wie lassen sich die typischerweise mit onlinebasierten Befragungen assoziierten nichtzufallsbasierten Auswahldesigns generell einordnen und wann können Access-Panels eine Lösung hierfür bereitstellen? Letzteres wird zudem durch das Beispiel zweier prominenter Access-Panels in der Bundesrepublik diskutiert und deren Möglichkeit der Abbildung eines bevölkerungsrepräsentativen Querschnitts analysiert.

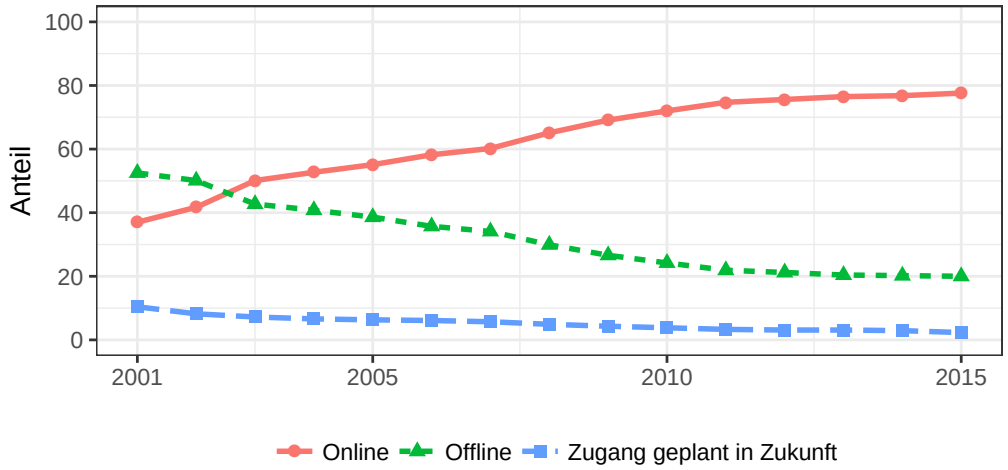
6.2.1. Noncoverage-Probleme

Ein systematischer Noncoverage-Fehler durch das Erhebungsdesign ergibt sich bei onlinebasierten Befragungen durch die von grundlegenden soziodemografischen Merkmalen abhängige Verteilung des Internetzugangs (z.B. Couper/Kapteyn et al. 2007; Bethlehem 2009; Ragnedda/Muschert 2013; Schnell 2012: 278 ff.; Bandilla et al. 2009; Bosnjak et al. 2013; Leenheer/Scherpenzeel 2013; Blom/Gathmann/Krieger 2015: 398 f.), sowie auf weiteren interessierenden politikwissenschaftlichen relevanten Variablen (z.B. Blasius/Brandt 2009) oder auch Gesundheitsindikatoren (z.B. Couper/Kapteyn et al. 2007). Ist das Ziel die Befragung der allgemeinen Bevölkerung, wird somit ein spezifischer Teil dieser Grundgesamtheit schwerer oder gar nicht erreicht. Ein Inferenzschluss auf der Basis einer online befragten Stichprobe kann somit nicht gültig für die allgemeine Bevölkerung sein (z.B. Bethlehem 2010: 43 ff. Mohorko/Leeuw/Hox 2013; Dillman/Smyth/Christian 2009; Couper 2000: 467 ff.). Trotz steigender Nutzungsraten führt daher der weiterhin vorhandene Ausschluss der Bevölkerung ohne einen Internetzugang zu einer hohen Gefahr systematischer Verzerrungen bei Bevölkerungsumfragen (Brick 2011: 885). Onlineumfragen sorgen somit für einen Selektionsmechanismus, welcher ein Noncoverage qua Design kaum vermeiden lässt (z.B. Hoonakker/Carayon 2009).

Dies wird das nachfolgende Kapitel nun für die volljährige und wahlberechtigte Bevölkerung der Bundesrepublik validieren. Grundsätzlich zeichnet die

Ausweitung des Internetzugangs ein positives Bild, seit 2001 steigt der Anteil an Personen innerhalb der Bundesrepublik, welche über einen solchen verfügen. Waren dies in 2001 nur 40 % der Bevölkerung, hat sich der Anteil in 2015 auf ungefähr 80 % verdoppelt, wie die Abbildung 6.4 an Hand der regelmäßig telefonisch und zufallsbasiert erhobenen Daten des (N)Online-Atlas zeigt.[22]

Abbildung 6.4.: Entwicklung Internetzugang in der Bundesrepublik 2001 bis 2015 ((N)Onliner-Atlas)



Quelle: (N)Onliner-Atlas. Erhebungen dieser Daten resultieren aus CATI-Stichproben der deutschsprachigen Bevölkerung ab 14 Jahren. Die Auswahl erfolgt nach dem RLD-Verfahren als ADM-Stichprobe. Befragt werden zwischen 20.000 und 50.000 Personen pro Jahr (Initiative D21 2015: 5).

In den letzten Jahren hat sich dieser Trend jedoch auf einem hohen Niveau stabilisiert: Jede fünfte Person ab 14 Jahren hat auch weiterhin keinen Internetzugang zur Verfügung und nur ein geringer Anteil plant dies auch in der nahen Zukunft zu ändern. Zu einem Noncoverage-Fehler führt dies, da die Zugehörigkeit zu den drei aufgeführten Gruppen nicht unabhängig von soziodemografischen Rahmendaten der Befragten im Rahmen der (N)Onliner-Erhebungen ist (Initiative D21 2015: 58 f.).[23] Auch die Ost-West-Zugehörigkeit erklärt beinahe die Hälfte der Unterschiede des Internetzugangs auf der aggregierten Ebene der

[22] Die „Internet World Stats“ reportiert einen Internetzugang für die Bundesrepublik von beinahe 90 % in 2015 (Quelle: http://tinyurl.com/d746aj4, Abruf: 11. Mai 2018). Als Internetzugang wird dort jedoch auch ein beruflicher Internetzugang definiert. Diese Quelle überschätzt somit das Potential von Onlineumfragen, da diese in der Regel nicht am Arbeitsplatz ausgeführt werden dürfen. Um eine internationale Vergleichbarkeit zu gewährleisten, kann sich diese Quelle jedoch anbieten (Blom/Bosnjak et al. 2016: 16).

[23] Ebenfalls: (N)Onliner-Atlas Homepage, aufrufbar über http://www.initiatived21.de/portfolio/d21-digital-index-2015/ (Abruf: 11. Mai 2018).

Bundesländer und sorgt für eine regionale Spaltung.[24] Stehen diese Variablen in einem Zusammenhang mit interessierenden Merkmalen einer Onlineumfrage, ist ein Noncoverage-Fehler qua Design bereits unvermeidlich.

Zusätzlich bedeutet die reine Verfügbarkeit eines Internetzugangs nicht auch die aktive und intensive Nutzung dieses (Groves/Fowler et al. 2009: 164 f.). Diese Nutzungshäufigkeit ist vielmehr selbst von weiteren Faktoren abhängig (z.B. Brandtzæg/Heim/Karahasanović 2011). Auch führt letztlich eine stärkere Ausweitung des Internetzugangs zu einem immer stärkeren Unterschied zwischen der Gruppe der Onliner und den verbliebenen Offlinern (Callegaro/Baker et al. 2014: 2). Das Ausmaß des qua Design verursachten Noncoverage-Fehlers bei Onlineumfragen ist daher sowohl vom Anteil der Bevölkerung ohne Internetzugang (N_{NI}/N) als auch vom Unterschied zwischen On- und Offlinern hinsichtlich eines interessierenden Merkmals ($\bar{Y}_I - \bar{Y}_{NI}$) abhängig (Bethlehem/Cobben/Schouten 2011: 101):

$$B(\bar{y}) = E(\bar{y}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N}(\bar{Y}_I - \bar{Y}_{NI}) \tag{6.8}$$

Dieser Bias lässt sich mit Hilfe der Daten des Allbus auf der Individualebene zeigen, da die befragten Personen dieser Trendstudie seit 2006 auch hinsichtlich eines vorhandenen Internetzugangs befragt werden.[25] Der Allbus eignet sich, da das Kapitel 6.1.1 persönliche Interviews bereits als wenig anfällig gegenüber einem Noncoverage-Fehler herausgestellt hat, als dies bei telefonischen Umfragen der Fall ist. Ebenfalls ist bezüglich des Internetzugangs kein durch den Interviewer bedingtes Antwortverhalten denkbar, für welches persönliche Interviews anfällig sein können. Tabelle 6.2 zeigt die Entwicklung des arithmetischen Mittels des Alters der Befragten ohne Internetzugang (NI) und die Veränderung dieses durch den Wechsel zur Gruppe mit Internetzugang (I), also den Kontrast ($\bar{Y}_I - \bar{Y}_{NI}$). Ebenso ist der Anteil der Personen ohne (privaten) Internetzugang an der Gesamtstichprobe (NI/N) angegeben und der Bias nach der Formel 6.8 bestimmt ($B(\bar{y})$).

Der im Allbus erfasste Anteil der Offliner sinkt von über 50 % in 2006 auf nur noch 20 % in 2014, was sich mit der vorherigen Entwicklung auf der Basis des (N)Onliner-Atlas deckt (siehe Abbildung 6.4). Die Ausweitung des Internetzugangs sorgt gleichzeitig für einen positiven Effekt durch die Verringerung

[24] Mit den Daten aus Initiative D21 (2015: 56) ergibt sich $\eta^2 = 0.4715$ hinsichtlich der Erklärung des jeweiligen aggregierten Anteils an Internetnutzern durch die „Ost-West-Zugehörigkeit“ des jeweiligen Bundesland.

[25] Abgefragt wurde jeweils: „Nutzen Sie privat das Internet?“ (Allbus Kumulation Variable *V1730*).

Tabelle 6.2.: Verzerrung Alter „Onliner“ und „Offliner“ im Allbus 2006-2014

Jahr	$\bar{Y}_{NI}$	$\bar{Y}_I - \bar{Y}_{NI}$	R^2	N_{NI}/N	$B(\bar{y})$
2006	56.11	-14.64	0.18	0.53	-7.77
2008	60.06	-18.31	0.26	0.45	-8.22
2010	63.60	-20.98	0.31	0.33	-6.90
2012	65.25	-21.36	0.29	0.26	-5.63
2014	66.74	-21.99	0.26	0.20	-4.44

Quelle: Allbus-Kumulation, durch Gewichtung wurden Ost-West-Ungleichgewichte der Erhebung korrigiert. AV: Alter in Jahren, UV: Internetzugang (1: Ja, 0: Nein).

des Bias des Alters auf Basis der Personen. Die steigenden Altersunterschiede zwischen den Onlinern und noch verbliebenen Offlinern relativieren jedoch diese positive Entwicklung. Auch in 2014 verbleibt daher, hinsichtlich des Alters der befragten Personen, ein klarer Unterschied zwischen der volljährigen Bevölkerung mit Internetzugang und denjenigen ohne, obwohl die zweite Gruppe nur noch 20 % der befragten Personen ausmacht.

Trifft dies auch weitere Variablen zu, welche in einer Verbindung zu inhaltlich interessierenden Variablen stehen können? Die Abbildung 6.5 weist hierzu die Ergebnisse logistischer Regressionsanalysen für alle Erhebungen des Allbus seit 2006 aus. Die zu erklärende dichotome Variable erfasst, ob eine Person über einen Internetzugang verfügt (1) oder nicht (0). Als erklärende Variablen der Wahrscheinlichkeit dieses Zugangs wurden, neben häufig verwendeten soziodemografischen Variablen, auch das Interesse an Politik und das interpersonelle Vertrauen einbezogen. Hierdurch ist eine Einschätzung möglicher Verzerrungen durch einen selektiven Internetzugang über die grundlegenden soziodemografischen Charakteristika hinaus möglich.

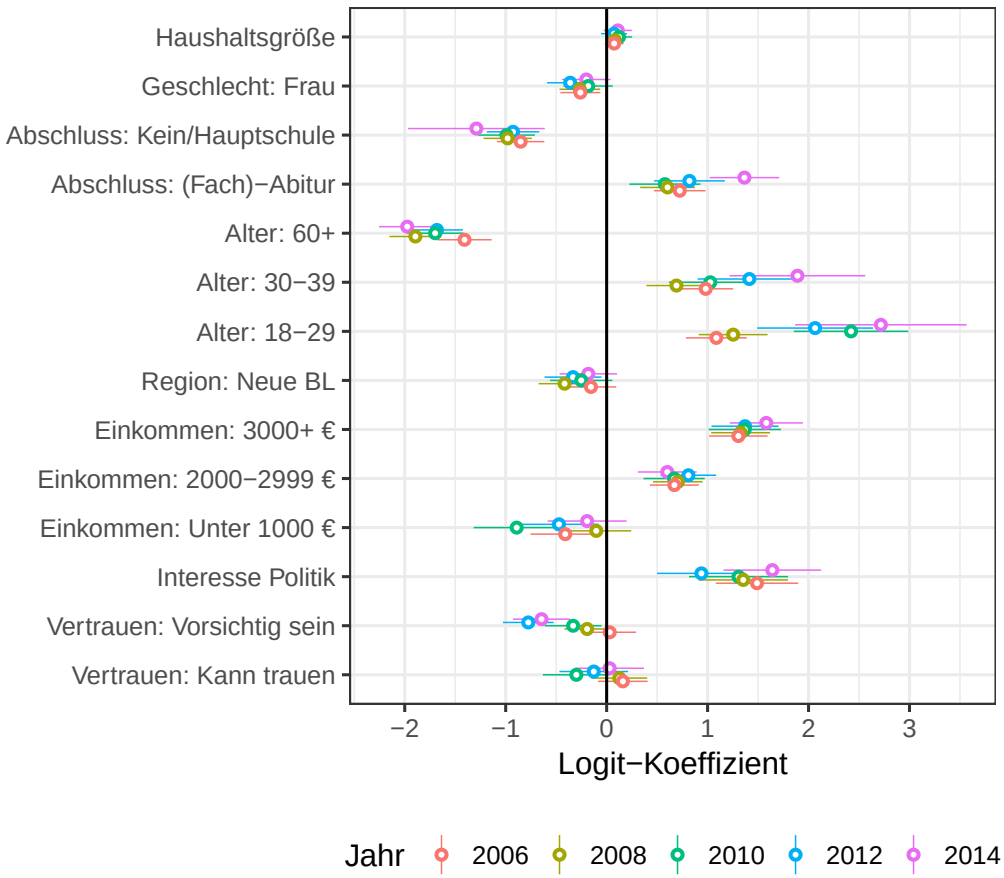
Das zentrale Ergebnis ist: Die Verteilung des Internetzugangs ist sehr selektiv und die Unterschiede zwischen On- und Offlinern nehmen im Zeitverlauf weiter zu. Dies zeigen, bei einer annähernd konstanten Fallzahl, die steigen Werte des zentralen Gütekriteriums der logistischen Regression, dem McFadden Pseudo R^2 in Tabelle 6.3. Die Möglichkeit der Differenzierung zwischen den beiden Gruppen, auf der Grundlage der verwendeten erklärenden Variablen, steigt mit jeder weiteren Erhebung des Allbus.

Tabelle 6.3.: Beobachtung und McFadden Internetzugang Allbus

Jahr	2006	2008	2010	2012	2014
Beobachtungen	2639	2796	2318	2950	3011
McFadden	0.29	0.35	0.42	0.43	0.45

Quelle: Eigene Darstellung.

Abbildung 6.5.: Modelle Internetzugang 2006 bis 2014 (Allbus)



Angegeben sind die Logitkoeffizienten und ihr 95 %-Konfidenzintervall. Quellen: Allbus-Kumulation (ZA4580) und Allbus 2014 (ZA5240), Designkorrekturen für Ost-West-Ungleichgewicht angewandt. Referenzkategorien sind (sofern nicht eindeutig): „Realschule“ (Abschluss), „1000-1999 €“ (Einkommen) und „Kommt darauf an“ (Vertrauen). Ein steigender Wert für „Politisches Interesse“ bedeutet ein verstärktes Interesse an Politik. Dessen fünfstufige Skala wurde auf einen Wertebereich von 0 bis 1 normalisiert (Min-Max-Normalisierung).

Neben der zentralen Erklärungskraft des Alters, des Bildungshintergrundes und der Einkommenssituation des Haushalts lassen sich auch Effekte des politischen Interesse und des interpersonellen Vertrauens finden. Stehen diese ebenfalls in Verbindung zu inhaltlich interessierenden Variablen, existiert auch auf diesen eine Verzerrung. Da es für die Korrektur dieser Verzerrungen die kausale Wirkungsrichtung nicht relevant ist, kann daher die Korrektur der aufgeführten Faktoren die verzerrte Erfassung weiterer inhaltlicher Variablen erreichen.

Ergänzend zum Allbus erlauben auch die Erhebungen der Komponenten 1 und 2 der GLES in 2013 die Abschätzung der Verfügbarkeit eines Internetzugangs. Zusätzlich wurde in beiden Umfragen die Nutzungshäufigkeit dieses Zugangs in

Tagen pro Woche erfasst.[26] Durch diese ist ein Bias genauer abschätzbar, wenn mit einer Onlineumfrage die wahlberechtigte Bevölkerung der Bundesrepublik anvisiert werden soll und ob sich eine zusätzliche Selektivität durch unterschiedliche Nutzungsgrade des Internet ergibt.

Wie die Tabelle 6.4 zeigt, ist ein vorhandener Internetzugang noch keine Garantie, dass dieser auch effektiv genutzt wird. Sowohl innerhalb der persönlichen,

Tabelle 6.4.: Internetnutzung an Tagen pro Woche (Komponente 1 und 2 der GLES)

Tage	0	1	2	3	4	5	6	7	$\bar{x}$
Komponente 1 (%)	20.94	2.64	6.95	9.77	8.97	11.34	6.87	32.52	4.07
Komponente 2 (%)	18.18	3.64	5.26	6.60	4.42	6.46	3.70	51.74	4.68

Quelle: Eigene Darstellung. Anmerkung: Standardabweichung der Wochenstichproben des „Rolling-Cross-Section“ (RCS)-Designs der Komponente 2 (*pre_woche*) beträgt für die durchschnittliche Nutzungshäufigkeit $\bar{x} = 4.68$ lediglich $s_{\bar{x}} = 0.14$ Tage.

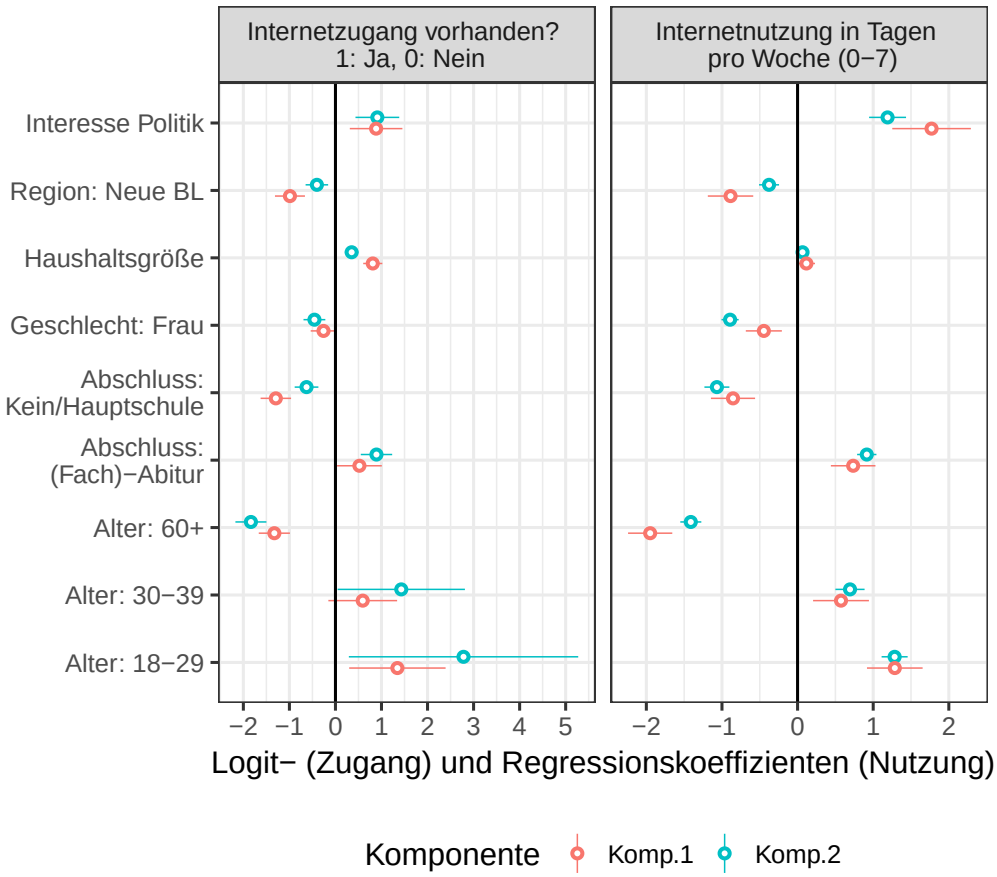
als auch der telefonischen Befragung gibt jede fünfte Person an, diesen weniger als einen Tag pro Woche zu nutzen. Wird eine Umfrage und ihr Auswahldesign nur online durchgeführt, hat dieser Personenkreis nur eine geringe Wahrscheinlichkeit, das Ziel dieser zu werden. Auch kann die seltene Nutzung zu höheren Vorbehalten der Teilnahme an einer Onlineumfrage führen, da die geringeren Erfahrungswerte die persönlichen Kosten der Teilnahme erhöhen. Zu den 20 % der Bevölkerung ohne Internetzugang, gesellt sich somit ein weiterer Teil der Onliner, welcher eine geringere Teilnahmewahrscheinlichkeit an Onlineumfragen gegenüber den Onlinern mit ausgeprägterer Expertise aufweist.

Wird das dichotome Vorliegen eines Internetzugangs *und* die Nutzungshäufigkeit in Tagen pro Woche über regressionsanalytische Verfahren erklärt, lässt sich sowohl der *reine Zugang*, als auch die *tatsächliche Nutzung* analysieren. Damit ergibt sich die Möglichkeit der Analyse einer zweiten Komponente des TSE: Neben der Verteilung des Internetzugangs und die damit einhergehende mögliche Erreichbarkeit durch Onlineumfragen („Coverage“), ist auch die aktive Nutzung entscheidend. Erweist sich die Häufigkeit der Nutzung als nicht unabhängig verteilt über die Auswahlgesamtheit der Internetnutzer, droht ein systematischer Nonresponse-Fehler durch unterschiedliche Teilnahmewahrscheinlichkeiten.

[26] Beide Informationen basieren auf den Variablen *q114* (Komponente 1) und *pre057* (Komponente 2): „An wie vielen Tagen in der Woche nutzen Sie im Durchschnitt das Internet?“. Möglichkeiten der Antworten waren „kein Internetzugang vorhanden“, „nutze nie das Internet“, „seltener als 1 Tag pro Woche“ und weitere Ausprägungen auf einer Skala von „1 Tag pro Woche“ bis „7 Tage pro Woche“.

Die Ergebnisse in Abbildung 6.6 bestätigen dieses Potential für einen kombinierten Noncoverage- *und* Nonresponse-Fehler bei onlinebasierten Befragungen der wahlberechtigten Bevölkerung in der Bundesrepublik.

Abbildung 6.6.: Modelle Internetzugang 2013 (GLES)



Angegeben sind Logit-/Regressions-Koeffizienten und die zugehörigen 95 %-Konfidenzintervalle. Quellen: Vorwahlbefragungen der Komponenten 1 und 2 der GLES 2013. „Oversampling" der Bevölkerung in den neuen Bundesländern ist durch Gewichtung korrigiert und Datensätze in Personenstichproben transformiert (Komponente 1: *w_trow* und Komponente 2: *w_trang*). Referenzkategorien sind identisch zu Abbildung 6.5. Das Interesse an Politik ist ebenfalls auf einen Wertebereich von 0 („kein Interesse") bis 1 („sehr starkes Interesse") normalisiert.

Auf Basis der beiden aufgeführten zufallsbasierten Auswahldesigns der GLES, zeigen sich klare Effekte hinsichtlich des Vorliegens eines Internetzugangs, welche die Ergebnisse auf der Basis des Allbus in Abbildung 6.5 bestätigen. Das zusätzliche Potential eines Nonresponse-Fehlers existiert ebenfalls, die Nutzungshäufigkeit eines vorhandenen Internetanschlusses steht ebenfalls in einer Verbindung zu den aufgeführten Variablen. Insgesamt weisen junge und politisch interessierte westdeutsche Männer mit einem hohen Bildungsstand

die höchste durchschnittliche Nutzungsfrequenz auf.[27] Neben dem Alter kommt dabei dem politischen Interesse eine zentrale Bedeutung der Unterscheidung zu. Je höher dies ausgeprägt ist, umso intensiver ist die Internetnutzung und umso höher auch die Wahrscheinlichkeit, das Ziel einer Onlineumfrage zu werden.

Diese Ergebnisse relativieren die starke Ausweitung des Internetzugangs innerhalb der Bevölkerung der Bundesrepublik. Die Reduktion der Gefahr eines Noncoverage-Fehlers ist nicht so groß, wie die Ausweitung suggeriert. Es muss davon ausgegangen werden, dass eine grundlegende Selektivität des Internetzugangs und der Nutzungshäufigkeit des Internets gegeben ist. Onlineumfragen sind daher qua Erhebungsformat bereits mit einem Noncoverage-Fehler belastet, durch die unterschiedlichen Grade der Nutzungshäufigkeit über die wahlberechtigte Bevölkerung hinweg ergibt sich zusätzlich eine erhöhte Gefahr für einen Nonresponse-Fehler. Der rein quantitative Fokus auf die Ausweitung der Zugangsraten zum Internet stellt daher einen ähnlichen Trugschluss dar, wie der Fokus auf die Teilnahmeraten an Umfragen (Kapitel 3.1). Auch hier verringert eine Ausweitung das *Potential* eines Noncoverage-Fehlers, es handelt sich jedoch nicht um einen Automatismus. Nachfolgend wird nun der zweite angesprochene grundlegende Aspekt diskutiert, welcher die Qualität onlinebasierter Befragungen einschränken kann: Die häufig mangelnde Möglichkeit der Realisierung einer zufallsbasierten Auswahl aus der allgemeinen Bevölkerung.

6.2.2. Das entscheidende Qualitätskriterium: Zufall oder kein Zufall

Die zufallsbasierte Auswahl von Befragungsteilnehmern ist die zentrale Voraussetzung, um einen Rückschluss auf eine anvisierte Grundgesamtheit vollziehen zu können (Baker/Brick et al. 2013). Hierzu muss diese jedoch definitorisch abgrenzbar sein. Ebenso wird ein Rahmen (z.B. in Form einer Liste) benötigt, um eine Zufallsstichprobe mit bekannten Auswahlwahrscheinlichkeiten zu erlangen. Ansonsten handelt es sich um eine willkürliche und nicht verallgemeinerbare Auswahl (Schnell 2012: 291 f.). Die nicht-zufallsbasierte Auswahl der befragten Personen kann eine Verzerrung dadurch erzeugen, dass der Zugang zur Umfrage selber nicht mehr unabhängig von verschiedenen Merkmalen ist. Die Verteilungen stellen keine Realisierung der eigentlich anvisierten Grundgesamtheit dar und Inferenzsschlüsse ohne weitere Korrekturen sind fehlerhaft oder

[27] Die Modellgüte beträgt für das Vorliegen eines Internetzugangs laut McFaddens Pseudo R^2 0.35 (Komp. 1) und 0.29 (Komp. 2) sowie für die Nutzungsintensität laut R^2 0.35 (Komp. 1) und 0.30 (Komp. 2).

sogar ungültig (Baker/Brick et al. 2013: 93 f.). Diese „Willkürlichkeit“ ergibt sich bei Onlineumfragen häufig durch die Selbstselektion der Teilnehmer in eine Umfrage. Um teilnehmen zu können, muss eine Person sowohl Kenntnis über die Existenz der Umfrage haben und sich zusätzlich entscheiden auch zu partizipieren. Aus der Gesamtheit aller Internetnutzer lässt sich jedoch nur dann eine repräsentative Stichprobe erlangen, wenn die Wahrscheinlichkeit der Auswahl bestimmbar ist und die Selektion nicht abhängig von weiteren interessierenden Variablen (Bethlehem 2009: 285 ff.). Solche eine Bestimmung der Auswahlwahrscheinlichkeiten der teilnehmenden Personen sind bei einer Selbstrekrutierung über eine online geschaltete Umfrage jedoch unmöglich. Inferenzstatistische Verfahren verlieren ihre Rechtfertigungsgrundlage (Couper 2000: 477).[28] Wird die Auswahl der zu befragenden Personen nicht über einen Zufallsmechanismus gesteuert, haben Personen daher unterschiedliche Wahrscheinlichkeiten von der Existenz einer Onlineumfrage zu erfahren. Mit Bezug auf das vorherige Kapitel impliziert dies eine weitere Verschärfung des Potentials für einen Nonresponse-Fehler. Dies ergibt sich, da die Häufigkeit der Internet*nutzung* nicht unabhängig von Kernmerkmalen innerhalb der volljährigen und wahlberechtigten Bevölkerung ist (siehe das vorangegangene Kapitel 6.2.1).

Ist hingegen bei Online-Umfragen eine zufallsbasierte Auswahl der Personen und die *anschließende* Übermittlung der Onlineumfrage möglich, ist die Einladung zur Teilnahme an dieser unabhängig vom Nutzungsverhalten des Internet. In einem solchen Fall können onlinebasierte Befragungen sogar qualitativ hochwertigere Ergebnisse erzielen, als dies bei vergleichbaren zufallsbasierten telefonischen Umfragen der Fall ist (Chang/Krosnick 2009; Yeager et al. 2011). Beispielsweise trifft lässt sich dies hinsichtlich der Prognose der Präsidentschaftswahl 2012 beobachten, wo Onlineumfragen teilweise bessere Resultate ergeben haben als die auf reinen Festnetzstichproben basierende Telefonumfragen.[29] Dies ist durch die Selbstadministration einer Onlineumfrage bedingt, welche Personen mit geringeren kognitiven Fähigkeiten durch die Möglichkeit der Bestimmung der eigenen Geschwindigkeit der Beantwortung der Umfrage entgegenkommt (Chang/Krosnick 2010). Ist eine zufallsbasierte Auswahl nicht möglich, dann weisen diese Erhebungen klare Schwächen auf (z.B. Sanders et al. 2006; Malhotra/Krosnick 2007; Chang/Krosnick 2009).

[28] Siehe hierzu z.B. auch die Darstellung des Pew Research Center http://www.people-press.org/methodology/sampling/why-probability-sampling/ (Abruf: 11. Mai 2018).

[29] Zur Aufschlüsselung siehe die Darstellung von Nate Silver: http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/?_r=0 (Abruf: 11. Mai 2018).

Dass nicht-zufallsbasierte Auswahlverfahren bei Onlineumfragen weitverbreitet sind, liegt an der mangelnden Möglichkeit der Erstellung einer adäquaten Liste, aus der eine Auswahl getroffen werden kann. Auch sind E-Mailadressen nicht immer klar einer Person zuordbar oder eine Person betreibt mehrere E-Mailadressen gleichzeitig (Groves/Fowler et al. 2009: 83). Für spezielle Populationen ist dieses Problem mit einem gewissen Aufwand behebbar.[30] Für die Durchführung von Bevölkerungsumfragen auf der Basis einer onlinebasierten Auswahl ist dies hingegen kaum möglich, selbst wenn lediglich der Teil der Bevölkerung mit Internetzugang anvisiert wird (z.B. Bethlehem/Cobben/Schouten 2011: 45 ff. Schnell 2012: 293).

Einen Ausweg stellt die Auswahl der zu befragenden Personen einer Onlineumfrage über eine zufallsbasierte offline stattfindende Stichprobenziehung. So kann beispielsweise das RLD-Verfahren telefonischer Befragungen zur Auswahl und Kontaktierung verwendet werden, die eigentliche Befragung findet dann jedoch im Rahmen einer Onlineumfrage statt. Dies verhindert ungleiche Auswahlwahrscheinlichkeiten in Abhängigkeit der Nutzungsfrequenz des Internets. Wird zusätzlich ein Internetzugang gestellt, sofern dieser nicht vorhanden ist, kann auch der Teil der Bevölkerung ohne einen Zugang zum Internet in die Onlineumfrage einbezogen werden. Ein Noncoverage-Fehler kann vermieden werden.

Dies trifft jedoch nur zu, wenn der Personenkreis ohne einen Internetzugang auch auf dieses Angebot eingeht. Sind zudem Personen ohne einen Zugang auch generell skeptischer gegenüber Onlineumfragen und partizipieren daher seltener an diesen, wird der vorherige Noncoverage-Fehler lediglich durch einen Nonresponse-Fehler eingetauscht (Blom/Herzing et al. 2016: 3). Im Bereich der sozialwissenschaftlichen Umfrageforschung finden sich innerhalb der vergangenen Jahre einige Beispiele der Etablierung solcher Formate der Offlinerekrutierung einer bevölkerungsweiten Stichprobe, welche anschließend onlinebasiert befragt wird. Das „Longitudinal Internet Studies for the Social sciences“ (LISS)-Panel (seit 2007), das „German Internet Panel“ (GIP) (seit 2012), die „Longitudinal Study by Internet for the Social Sciences“ (ELIPSS)-Panel (seit 2012) und das GESIS-Panel (seit 2013) stellen einige Versuche dar, die Repräsentativität von onlinebasierten (Panel-)Umfragen durch die Integration von Personen ohne Internetzugang über die technische Bereitstellung zu verbessern.

Alle Stichprobenziehungen dieser basieren entweder auf Einwohnermelde-

[30] Beispielsweise lassen sich ohne Probleme eine zufallsbasierte Auswahl aller Mitarbeiter eines Unternehmens oder aller Studenten einer Hochschule über deren jeweilige *individualisierte* E-Mailadressen realisieren.

registern oder Flächenstichproben zur Durchführung von persönlichen Befragungen. Zudem werden diese teilweise durch eine telefonische (LISS) und/oder postalische Befragung (ELIPSS) zur Stichprobenerhebung unterstützt. Hat eine ausgewählte und erfolgreich befragte Person keinen Internetzugang, stellt das LISS-Panel, das GIP und das ELIPSS die technische Infrastruktur. Im GESIS-Panel hingegen wird als Mixed-Mode das Angebot der zukünftigen postalischen Befragung offeriert (Blom/Bosnjak et al. 2016). In allen Erhebungen mit einer Bereitstellung dieser technischen Infrastruktur zeigt sich jedoch auch weiterhin ein zu hoher Anteil der jungen und hoch gebildeten Teile der jeweils anvisierten Grundgesamtheit (Blom/Bosnjak et al. 2016: 11). Dies ergibt sich, da längst nicht jede Person auf das Angebot der Bereitstellung der technischen Infrastruktur eingeht, um an künftigen Befragungen partizipieren zu können. Im GIP trifft dies beispielsweise nur auf jede fünfte Person zu (Blom/Herzing et al. 2016: X). Bereits die Berücksichtigung dieser Gruppe sorgt jedoch für eine stärkere Übereinstimmung mit der gesamten Bevölkerung hinsichtlich des Alters und des Bildungsstand, auch wenn dies für weitere Variablen, wie das politische Interesse, nicht klar beantwortbar ist (Blom/Herzing et al. 2016: 16). Auch im LISS-Panel lässt sich nur jede dritte Person ohne Internetzugang durch die Bereitstellung der technischen Infrastruktur zu einer zukünftigen Teilnahme bewegen. Verzerrungen hinsichtlich soziodemografischer Variablen lassen sich ebenfalls reduzieren, auf das reportierte Wahlverhalten zeigt sich hingegen kein relevanter Einfluss (Leenheer/Scherpenzeel 2013: 20 ff.).

Die Integration von „Offlinern“ über die Bereitstellung der technischen Infrastruktur stellt daher prinzipiell den Königsweg des Umgangs eines drohenden Noncoverage-Fehlers onlinebasierter Befragungen dar (Blom/Herzing et al. 2016: 4). Die erhöhte Repräsenativität, konditional der soziodemografischen Rahmendaten, bei geringeren Kosten als bei klassischen persönlichen Umfragen und hohen Wiederbefragungsraten stimmen zuversichtlich (Blom/Gathmann/Krieger 2015: 402 f.), ein endgültiges Fazit zum Potential solcher offline-rekrutierter zufallsbasierter Onlinepanels fällt zum jetzigen Zeitpunkt jedoch auf Grund der wenigen Erfahrungen schwer. Zudem stellt die Form der offlinebasierende Stichprobenziehung die Ausnahme dar. Der überwiegende Teil onlinebasierter Befragungen basiert auf keiner wahrscheinlichkeitstheoretischen Grundlage der Auswahl. Diese nicht-zufallsbasierten Auswahlverfahren stellen eine Bandbreite verschiedenster Auswahlmöglichkeiten dar, welche kein gemeinsames theoretisches Konzept vereint. Ihre Bewertung muss daher auch immer vor ihrem spezifischen Hintergrund erfolgen (Baker/Brick et al. 2013: 99 f.).

Ob Daten aus einer solchen nicht-zufallsbasierten Auswahl nutzbar sind, hängt daher essentiell davon ab, ob der Selektionseffekt in die Umfrage durch adäquate Korrekturen unter Kontrolle gebracht werden kann (Baker/Brick et al. 2013: 94 f.) Die Einschränkungen nicht-zufallsbasierter Stichproben ergeben sich durch eine erhöhte Variabilität und systematischen Abweichungen zu Stichproben auf der Basis echter Zufallsziehungen, wobei auch Gewichtungsverfahren dieses Problem eher selten beheben können (z.B. Faas/Schoen 2006; Loosveldt/Sonck 2008; Yeager et al. 2011; Erens et al. 2014). Die häufig angewandte einfache Korrektur auf soziodemografische Abweichungen ist in jedem Fall nicht ausreichend (z.B. Faas 2004; Couper/Kapteyn et al. 2007; Pasek 2015). Die alleinige Bereitstellung verhindert nicht den durch die nicht-zufallsbasierte Auswahl hervorgerufenen Selektionseffekt in die Umfrage hinein (Blom/Herzing et al. 2016: 17). Kann dies gewährleistet werden, kommt der Erhebungsmodus der Onlinebefragung über den Ruf eines ergänzenden – weniger ersetzenden – Modus der Datenerhebung hinaus (z.B. Fricker/Schonlau 2002).

Die Qualitätsprobleme onlinebasierter Befragungen stellt sich daher nun als ein dreiteiliges dar: *Erstens* existieren Probleme hinsichtlich der Abdeckung der gesamten Bevölkerung durch eine ungleiche Verteilung des Internetzugangs innerhalb dieser. Wie die Analysen der zufallsbasierten Stichproben des Allbus seit 2006 und der ebenfalls zufallsbasierten Stichproben der Komponenten 1 und 2 der GLES gezeigt haben, zeigen sich systematische Unterschiede zwischen Personen mit und ohne Internetzugang. Der sich ausweitende insgesamte Zugang zum Internet innerhalb der Bundesrepublik hat also das Potential der Reduzierung von Verzerrungen. Die sich verstärkenden Unterschiede zu dem Teil der Bevölkerung ohne Zugang konterkarieren diesen Effekt jedoch. Da *zweitens* die Häufigkeit der Internetnutzung von ebenfalls diesen Faktoren abhängt, kombiniert sich zu dem Noncoverage-Fehler ein Nonresponse-Fehler. Dieser, bedingt durch die Selektivität der Entscheidung zur Teilnahme an einer Umfrage, wird *drittens* durch die häufige Kombination mit einer nicht-zufallsbasierten Auswahl verstärkt. Hierdurch kommt es zu einer weiteren Verzerrung durch die selektive Kenntnis der Befragung. Personen mit einer erhöhten Internetnutzung haben auch eine erhöhte Wahrscheinlichkeit der Auswahl auf der Basis einer nicht-zufallsbasierten Stichprobenziehung.

Eine Möglichkeit die mangelnde Möglichkeit einer zufallsbasierten Auswahlgrundlage bei Onlineumfragen zu umgehen, ist auf einen bereits konstruierten Kreis an möglichen zu befragenden und bekannten Personen zurückzugreifen. Sofern es sich bei diesem Kreis selber um ein repräsentatives Abbild der anvi-

sierten Grundgesamtheit handelt, lassen sich über eine Zufallsauswahl aus dieser Gruppe Onlineumfragen mit einem Anspruch der Repräsentativität hinsichtlich der gesamten Bevölkerung realisieren. „Access-Panels“ versuchen solch eine Lösung zu sein. Sie stellen einen Pool an potentiell zu befragenden Personen bereit, aus denen adhoc per Zufallsmechanismus eine Stichprobe ausgewählt werden kann. Da Stichproben aus diesen „Access-Panels“ bereits eine weite Verbreitung in qualitativ hochwertige Journals gefunden haben (Ansolabehere/ Schaffner 2014: 285) und auch die Komponenten 3 und 8 der GLES-Studie aus solchen Access-Panels zu Stande kommen, wird das nachfolgende Kapitel 6.2.3 diese Form der Bereitstellung eines „Pools“ an möglichen zu befragenden Personen vorstellen und an Hand der Zusammensetzung der bereits in Kapitel 5 diskutierten Access-Panels von Respondi und LINK analysieren.

6.2.3. Access-Panels als Lösung?

„Klassische“ Panels als Längsschnittdesign verfolgen das Ziel der Befragung eines identischen Kreises verschiedener Personen zu mehreren Zeitpunkten und sind mit spezifischen Vorteilen, aber auch Nachteilen behaftet (z.B. Bergmann 2015). Ihr Ziel ist die Erfassung von individuellen Veränderungen (z.B. Lynn 2009: 1 ff.). Diese Form der Erhebung blickt seit den ersten Erhebungen in einem Wahlkontext durch Lazarsfeld/Fiske (1938) auf eine lange Tradition zurück.

Die Rolle von Access-Panels weicht von der Wirkungsweise „gewöhnlicher“ Panels der empirischen Sozialforschung klar ab. Hierbei handelt es sich um einen „Pool“ an erfassten Personen. Diese haben ihre Bereitschaft signalisiert, in Zukunft an Umfragen zu unterschiedlichsten Themen teilzunehmen. Die Ziehung einer Stichprobe ist damit auf den Kreis der registrierten Teilnehmer des Access-Panels beschränkt. Bei der Registrierung werden lediglich einige wenige soziodemografische Rahmendaten abgefragt, auf denen eine spätere mögliche Quotierung der Stichprobe basieren kann (Couper 2000: 482). Mit der Evolution dieses Formats Anfang der 2000er Jahre waren Erwartungen innerhalb der kommerziellen Marktforschung verknüpft, andere Befragungsformen durch dieses Prinzip der Adhoc-Befragung von registrierten Panelisten obsolet werden zu lassen (Couper 2013: 149).

Für Access-Panels existiert eine ausgearbeitete ISO-Norm, welche die zentralen Begriffe klärt. Es handelt sich bei ihnen, per Definition der Norm, um eine Datenbank von Personen. Diese haben im Vorfeld ihre Zustimmung zur Einladung an zukünftigen Befragungen gegeben. Der Provider des Panels ist verantwortlich für die Qualitätssicherung und Bereitstellung der Stichproben aus

diesem. Als aktives Panelmitglied wird eine eine Person definiert, wenn sie in den vergangenen 12 Monaten an einer eingeladenen Umfrage teilgenommen, ihr Profil überarbeitet oder sich auf dem Panel registriert hat.[31] Als Kritikpunkt an dieser Norm lässt sich jedoch festhalten, dass diese „[...] so ungenau formuliert [ist], dass daraus keine Handlungsempfehlungen oder gar Sanktionen herleitbar wären." (Schnell 2012: 291). Access-Panels umfassen schnell eine Größe von mehreren 10.000 Personen.[32] Aus diesem Kreis lassen sich dann zufallsbasierte Stichproben realisieren. Aufgrund der niedrigen Teilnahmeraten an eingeladenen Umfragen ist es jedoch erwartbar, dass der überwiegende Teil der registrierten Personen nicht auf die Einladungen reagiert oder mit zu vielen konkurrierenden Umfragen konfrontiert ist. Im Zuge der starken Ausweitung der Nutzung von Access-Panels findet sich daher der generelle Effekt, dass immer mehr Umfragen auf der Basis einer kleinen, auf Access-Panels registrierten Subpopulation, der Bevölkerung durchgeführt werden (Tourangeau/Conrad/Couper 2013: 13).

Access-Panels sind mit zwei zusätzlichen Problemen gegenüber einmaligen Onlinebefragungen konfrontiert. Erstens muss ein hoher Ausfall aus dem Panel und eine damit einhergehende mögliche hohe Fluktation der registrierten Personen vermieden werden („Panel-Pflege"). Zweitens können Befragungen von einzelnen Personen nur gemäßigt eingesetzt werden, da es ansonsten zu Professionalisierungseffekten im Rahmen der Umfragen („Panel-Effekte") kommt (Schnell 2012: 297). Wird dies nicht verfolgt, verstärkt sich insgesamt weiter das Problem einer zunehmenden selektiven Befragung der Bevölkerung durch professionalisierte Panelmitglieder (Couper 2013: 150). Zur „Panel-Pflege" werden häufig Incentivierungen bei Befragungen eingesetzt. Vielfach eingesetzte monetäre Anreize in Umfragen erhöhen relativ zuverlässig die Teilnahmerate in postalischen (Church 1993), intervieweradministrierten (Yu/Cooper 1983; Singer/Groves/Corning 1999) und Online-Umfragen (Göritz 2006). Die bedingungslose Auszahlung monetärer Entlohnung im Vorfeld der Befragung ist dabei die effektivste Variante. Dies trifft jedoch nur mit einem abnehmendem Grenzeffekt je eingesetzten Euro auf die weitere Steigerung der Teilnahmebereitschaft zu. Die Erhöhung der Ausschöpfung geht tendenziell mit keinem oder nur nur einem sehr geringen Effekt auf die Datenqualität und die Zusammensetzung von Verteilungen einher. Um hier eine Reduzierung einer Verzerrung zu erreichen,

[31] Einsicht ISO-Norm 26362 : 2009(*en*) unter https://www.iso.org/obp/ui/#iso:std:iso:26362:ed-1:v1:en, Abruf: 11. Mai 2018.

[32] Beispielsweise enthält das Access-Panel der Respondi AG, als größten Anbieter, eine Auswahlgesamtheit von 100.000 Personen (Quelle: Panelbook über https://www.respondi.com/downloads, Abruf: 11. Mai 2018). Das LINK Internet-Panel enthält in der Bundesrepublik ca. 40.000 Personen zur Befragung (Quelle: https://www.link.ch/, Abruf: 11. Mai 2018).

müssen klar unterrepräsentierte Gruppen vielmehr gezielt durch die Incentivierung angesprochen werden (Singer/Ye 2012).

Ob bestimmte Subgruppen stärker auf Incentivierungen reagieren als andere, lässt sich mit dem momentanen Forschungsstand nicht klar beantworten (Church 1993; Singer/Groves/Corning 1999; Singer/Ye 2012). Ein möglicher Wirkungskanal ist die Abmilderung der höheren Wahrscheinlichkeit der Selbstselektion einer Person in die Stichprobe, wenn auch Personen mit einem verringerten Interesse durch die Incentivierung vermehrt an einer Umfrage partizpieren (Groves 2004). Beispielsweise finden Ryu/Couper/Marans (2005) und Petrolia/Bhattacharjee (2009) stärkere Anreizeffekte für gering gebildete Personen, erstere ebenso für Singlehaushalte und Arbeitslose. Es finden sich jedoch auch genug Studien ohne eine solche Variation der Teilnahmewahrscheinlichkeit in Abhängigkeit von Charakteristika von befragten Personen bei eingesetzter Incentivierung (zum Überblick Toepoel 2016: 112). Es zeigt sich daher, selbst bei vorhandenen steigenden Ausschöpfungsquoten, ein nur sehr verhaltener Einfluss auf die Datenqualität von Umfragen durch den Einsatz von Incentivierungen (z.B. Sánchez-Fernández et al. 2010). Bei onlinebasierten Befragungen zeigen sich Effekte der Incentivierung der Teilnahme unabhängig davon, ob ein zufallsbasiertes Auswahlverfahren angewandt wurde (Göritz 2004). Incentivierungen in Online-Umfragen rufen jedoch einen geringeren Effekt hervor als in anderen Umfrageformen (Göritz 2006), auch wenn sich hier ebenfalls positive Effekte hinsichtlich einer möglichen Unterrepräsentierung ökonomisch schwächerer Gruppen durch die Incentivierung zeigen (Goritz 2008). Dies ist gerade durch die Verwendung von Stichproben aus Access-Panels erklärbar, deren Teilnehmer über die Incentivierung bereits informiert sind und diese erwarten. Der nachträgliche Effekt der Incentivierung schwächt sich ab (Toepoel 2012: 216).

Professionalisierung und Incentivierung stehen oftmals in einer engen Verbindung: Professionelle Umfrageteilnehmer nehmen primär durch die Incentivierung an der Umfrage und weniger auf Grund des Inhaltes. Sie weisen im Kontext politischer Umfragen daher in aller Regel ein geringeres politisches Interesse auf, wodurch ihre Anwesenheit sogar den Bias auf damit assoziierten Variablen reduzieren kann (Hillygus/Jackson/Young 2014: 232 f.). Der gesetzte monetäre Anreiz zur Teilnahme bei Access-Panels wird jedoch dann zu einem Problem, wenn es der hauptsächlich treibende Motivationsfaktor ist, der Wahrheitsgehalt der Antworten sinkt (Stern/Bilgen/Dillman 2014: 289 f.). Es handelt sich um eine Form von Professionalisierung mit dem Ziel die Umfragen mit einem maximalen Erlös durch die Incentivierung zu bearbeiten. Solche trainierte Befragte

unterscheiden sich vor allem bei Wissensfragen von untrainierten Befragten (Toepoel/Das/van Soest 2009). Hierbei tritt der Effekt auf, dass diese Befragten schneller Antworten bereitstellen und eine Umfrage beenden, da sie sowohl erfahrener im Umgang mit dem Internet im Allgemeinen als auch mit Online-Umfragen im Spezifischen sind (Yan/Tourangeau 2008).

Das maßgebliche Problem der Befragungen auf Access-Panels bilden jedoch weiterhin nicht zu vermeidende Selektionseffekte, auch wenn diese sich gegenüber anderen Formen der onlinebasierten Befragung stärker kontrollieren lassen. Auch Stichproben aus Access-Panels sind weiterhin von Noncoverage betroffen, da die Befragung online durchgeführt wird. Ebenso ergeben sich mögliche Selektionseffekte durch die ungleiche Nutzungshäufigkeit des Internets. Zusätzlich ist die Art der Rekrutierung in das Panel essentiell, da hier eine weitere Ebene der Selbstselektion auftritt (Lee 2006: 330). Die Bewertung von Stichproben aus Access-Panels unterscheidet sich daher vor allem nach der Art der Rekrutierung in das zu Grunde liegende Access-Panel. Folgt dies einer nicht-zufallsbasierten Auswahl, kann ein Selbstselektionseffekt in das Panel nicht ausgeschlossen werden. Die Auswahlwahrscheinlichkeiten sind dann nicht bekannt und die Anwendung inferenzsstatistischer Verfahren kann nur erfolgen, wenn eine Zufallsstichprobe aus dem Panel gezogen wird. Der Rückschluss ist dann nur auf die Grundgesamtheit des Access-Panel möglich. Bei Anwendung einer zufallsbasierten Auswahl kann jedoch der Selbstselektionseffekt auf der Grundlage der Kenntnissnahme des Panels verhindert werden (Callegaro/Baker et al. 2014: 6 f.).

Die Qualität eines solchen Access-Panel ist daher immer vor dem Kontext zu bewerten, wie die Auswahl der bereits registrierten Personen als Auswahlgesamtheit der nachfolgend zu ziehenden Stichprobe durchgeführt wurde und ist nur schwer allgemein zu beantworten (Baker/Brick et al. 2013: 91 f.). Der überwiegende Teil der weltweiten Access-Panels folgt einer primär online stattfindenden und passiven Selbstrekrutierung. Die Entscheidung zur Registrierung wird in diesem Fall alleine den zukünftigen Mitgliedern überlassen (Vehre 2012: 38 f.). Hierzu werden Personen mit Hilfe eines Schneeballprinzips über Online-Marketinganzeigen rekrutiert. Die Aussicht auf eine zukünftige monetäre Entlohnung durch die Teilnahme an Umfragen soll als ein Anreiz dienen. Eine anvisierte Grundgesamtheit – abseits einer räumlich und zeitlich nicht klar abgrenzbaren „Internet-Population“ – ist daher nicht möglich. Die willkürliche Auswahl der registrierten Personen ist nicht zufallsbasiert und eine Annahme konstanter Auswahlwahrscheinlichkeiten nicht gegeben (Schnell 2012: 293).

Eine solche Rekrutierung in ein Access-Panel überspringt den kompletten Schritt der Entwicklung eines Auswahldesigns durch die direkte Anwendung einer breiten Online-Rekrutierung. Eine im Vorhinein getroffene Einschränkung auf bestimmte, nach einem Zufallsprinzip ausgewählte, Personen findet nicht statt (Groves/Fowler et al. 2009: 83 f.). Die Bereitstellung einer solchen Auswahlgesamtheit an möglichen Teilnehmern einer hieraus gezogenen Stichprobe ist daher nur sehr bedingt ein Garant zur Erlangung einer als repräsentativ einzustufenden Stichprobe für eine mögliche hierdurch anvisierte Grundgesamtheit (Bethlehem/Biffignandi 2012: 52). Die Nutzung dieser Stichproben ist also fragwürdig, sofern das Ziel ein Inferenzschluss auf eine definierte anivisierte Grundgesamtheit darstellt (Baker/Blumberg et al. 2010: 714).

Eine Möglichkeit dies zu umgehen ist die Nutzung einer aktiven und offline stattfindenden Rekrutierung nach einem Zufallsprinzip. Hierdurch werden sowohl Personen ohne Zugang zum Internet erreicht, als auch Selektionseffekte an Hand der Internetnetzung durch die Möglichkeit einer zufallsbasierten Auswahl in das Panel abgemildert. Der Personenkreis des Panels kann hierbei über eine telefonische Befragung und einem dazu gehörigen zufallsbasierten Auswahlverfahren erfolgen (Vehre 2012: 49 ff.). Wird Personen ohne Internetzugang dieser ebenfalls gestellt und ist dies die einzige Barriere der Teilnahme gewesen, sind theoretisch bevölkerungsrepräsentative Befragungen auf der Basis von Onlineumfragen möglich (Blom/Herzing et al. 2016: 3).

Zusammenfassend gibt es somit eine Reihe an Anforderungen an ein Access-Panel, um bevölkerungsweite Prognosen auf Basis dieser Stichproben zu ermöglichen. Die *Rekrutierung* sollte *aktiv durch den Provider* gesteuert werden, um Effekte einer Selbstselektion zu vermeiden. Eine *offline* stattfindende Rekrutierung ermöglicht dies, da nur hierdurch eine zufallsbasierte Auswahl und die Vermeidung der Selbstselektion in das Access-Panel garantiert werden kann. Dies ist nicht der Fall, wenn die Rekrutierung online und passiv geschieht. Personen ohne Internetzugang, sollte zudem die *technische Infrastruktur* zur Teilnahme *gestellt* werden. Dies verhindert einen Noncoverage-Fehler qua Design. Ebenfalls sollte verhindert werden, dass die letztendliche Entscheidung zur Registrierung auf der Basis möglicher interessierender Merkmale getroffen wird. Dann bildet das vorliegende Access-Panel ein verkleinertes Abbild der interessierenden und anvisierten Grundgesamtheit. Eine zufallsbasierte Ziehung aus dem Panel ermöglicht dann einen Schluss auf die allgemeine Bevölkerung.

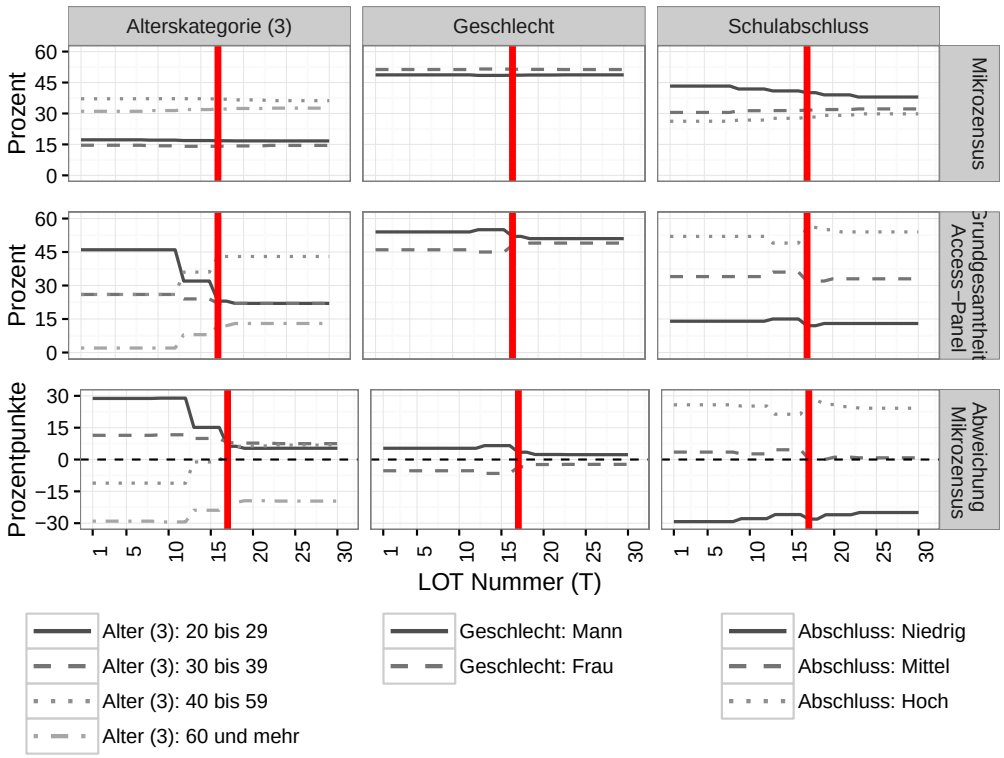
Können die Access-Panel von Respondi und LINK, welche ebenfalls für die Erhebungen der Komponenten 3 und 8 der GLES genutzt werden, dies leisten?

Beide Betreiber wenden verschiedene Maßnahmen eines Qualitätsmanagements und Maßnahmen zur Panelpflege an (siehe zur Darstellung Kapitel 5.1). Ebenfalls werden die angewandten Incentivierungen unkonditional und im Vorfeld angewandt, was die beste Variante laut dem momentanen Stand der Forschung zur Incentivierung von Onlineumfragen entspricht. Ebenfalls gehen beide Betreiber aktiv bezüglich ihrer Rekrutierung vor und überlassen die Entscheidung zur Registrierung nicht den späteren Teilnehmern von Umfragen. Der maßgebliche Unterschied findet sich jedoch in der Art der Rekrutierung. Basiert das Respondi-Panel rein auf online und nicht-zufallsbasierten geschalteten Rekrutierungsmaßnahmen, nutzt das LINK Internet Panel eine offline stattfindende und zufallsbasierte Auswahl über Telefonfestnetzstichproben. Dies reduziert die Selektivität, da theoretisch die Häufigkeit und Kompetenz des Umgangs mit dem Internet keinen Effekt mehr auf die Wahrscheinlichkeit der Rekrutierung hat. Wie die Ergebnisse der Abbildung 6.6 gezeigt haben, dürfte dies bereits auf grundlegenden soziodemografischen Variablen und dem Interesse an Politik bereits positive Auswirkungen haben. Da jedoch auch LINK keinen Internetzugang stellt, besteht weiterhin ein qua Design nicht zu vermeidender Noncoverage-Bias durch den Ausschluss des Teils der Bevölkerung ohne einen Zugang zum Internet. Auch auf der Basis des LINK Internet Panels sind somit bevölkerungsrepräsentative Befragungen ohne weitere Korrekturen nicht möglich.

Relevant für die Arbeit ist zudem, ob es Access-Panels als Form der Bereitstellung von potentiell zu befragenden Personen schaffen, ein adäquates Abbild der volljährigen, bzw. wahlberechtigten, Bevölkerung durch die aus ihnen resultierenden Stichproben zu erhalten. Die Komponente 8 ist für diesen Vergleich besonders geeignet, da diese in einer vergleichsweise hohen Frequenz der Erhebung vorliegt und zusätzlich ein Wechsel von einem primär online zu einem primär offline rekrutierten Acces-Panel im Zeitablauf stattgefunden hat (Kapitel 5.1). Abbildung 6.7 zeigt die Angaben der Verteilung des Bildungsstatus, verschiedener Altersgruppen und des Geschlechts der Access-Panels von Respondi und LINK als Grundgesamtheit daraus zu ziehender Stichproben.[33] Als Referenz dienen die Verteilungen der amtlichen Statistik durch den „Mikrozensus“ (MZ), wozu die Verteilungen innerhalb der realisierten Stichproben in Vergleich gesetzt wurden. Die beiden Access-Panel von Respondi (bis T17) und LINK (ab T17) zeigen klare Abweichungen zur allgemeinen volljährigen Bevölkerung, welcher

[33]Quelle für die Verteilungen sind die Angaben der Betreiber, entnommen aus den Studienberichten der Komponente 8 der GLES (T1 bis T29). Diese lassen sich über die GESIS frei einsehen.

Abbildung 6.7.: Access-Panel Respondi und LINK in Referenz zum Mikrozensus 2009-2015



Quelle: Eigene Darstellung. Daten: Studienbeschreibungen der Komponente 8 zur Angabe der Verteilungen der Access-Panel der jeweiligen Anbieter. Bis Welle 17 Respondi, von da an LINK. Erhebung $T = 1$ ist 04/05-2009, $T = 30$ ist 12-2015.

der MZ erfasst (Abbildung 6.7). Dem Access-Panel von LINK gelingt es jedoch besser, die Altersstruktur abzubilden. Das beinahe ausschließlich onlinerekrutierte Access-Panel von Respondi ist hier sehr starken Verzerrungen ausgesetzt, auch wenn die letzten Erhebungen aus diesem (ab T=13) eine klare Tendenz der Besserung zeigen. Hier zeigt sich ein möglicher Effekt der Offlinerekrutierung. Nutzen Personen vergleichsweise selten das Internet, ist ihre Wahrscheinlichkeit aktiv durch den Panelbetreiber online rekrutiert zu werden geringer. Durch die Kontaktierung per Telefon tritt dieses Problem nicht auf. Ältere Personen, mit einer tendenziell geringeren Frequenz der Nutzung ihres Internetzugangs, erhalten somit dieselbe Auswahlchance über die RDD-Auswahl der Nutzung des Angebots zur Registrierung in das LINK Internet-Panel. Es kann aber an dieser Stelle ein zeitlicher Effekt nicht ausgeschlossen werden. Demnach weist das Internt Panel von LINK eine bessere Abbildung auf, da seit 2012 ein zunehmender Anteil der *älteren* Bevölkerung das Internt nutzt. Auch das LINK Internet Panel weist als Grundgesamtheit jedoch einen eindeutig zu geringen Anteil der

Altersgruppe ab 60 Jahren auf. Hier zeigt sich eine mögliche altersbedingte Abhängigkeit der Entscheidung zur Registrierung im Panel, da diese Gruppe per Festnetzstichproben sehr gut zu erreichen ist. Das Problem der schwierigen Erreichbarkeit dieser Gruppe bei einer online durchgeführten Rekrutierung scheint sich also mit einem Problem der schwierigeren Rekrutierung bei Erreichbarkeit zu kombinieren. In den kommenden Jahren könnte sich die Situation bessern. Dies liegt darin begründet, dass die heute 40- bis 50-jährigen bereits Erfahrungen im Umgang mit der „Online-Welt“ sammeln und als kommende 60- bis 70-jährige der Bevölkerung dann auch mit onlinebasierten Befragungen umgehen können.

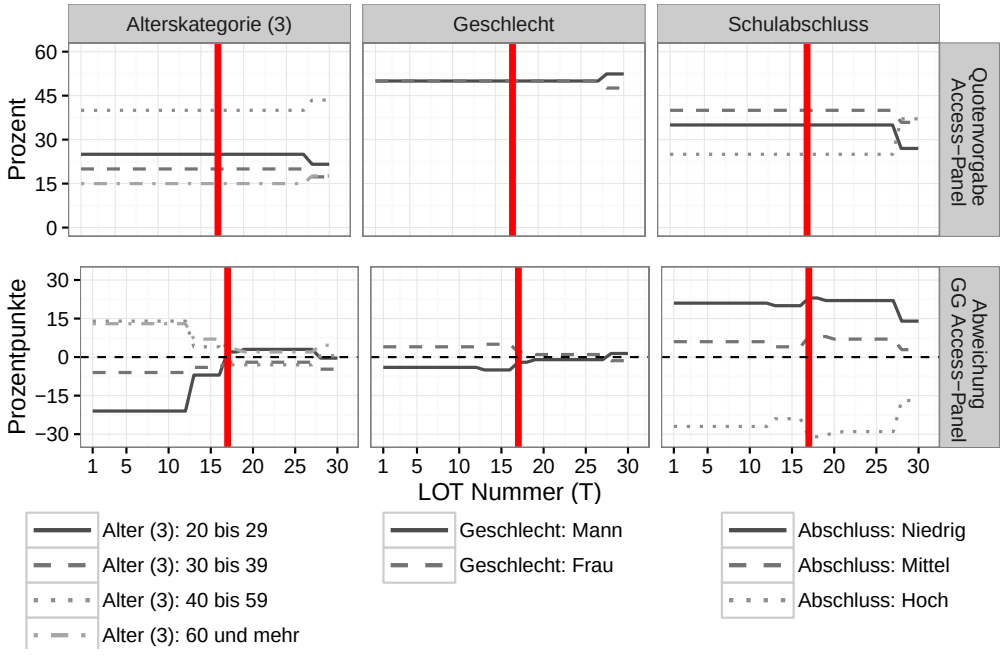
Gravierender wirkt hier die, ausgesprochen konstante, zeitliche Verzerrung des Bildungshintergrunds. Die Abweichungen zum Mikrozensus reduzieren sich auch durch den Wechsel des Panelbetreibers nicht. Es findet im Zeitablauf lediglich eine marginale Angleichung der Grundgesamtheiten der Access-Panels an die Rahmendaten der amtlichen Statistik statt. Diese Angleichung ist jedoch durch Veränderungen innerhalb der allgemeinen volljährigen Bevölkerung bedingt, da der allgemeine Qualifizierungsgrad im betrachteten Zeitraum weiter zugenommen hat. Es nähert sich somit die Verteilung des Bildungshintergrunds des MZ an die Verteilung der beiden Access-Panels an, deren Zusammensetzung jedoch keiner Veränderung unterliegt.

Im Vergleich zur allgemeinen Bevölkerung stellt der Pool der hier betrachteten Access-Panels daher einen zu jungen und zu hoch gebildeten Personenkreis dar. Bereits auf diesen beiden sehr grundlegenden soziodemografischen Variablen finden sich starke Abweichungen zwischen diesen beiden möglichen Grundgesamtheiten der empirischen Wahlforschung in Form von Access-Panels und dem MZ. Da das Alter und der Bildungshintergrund für die Wahlforschung ebenfalls grundlegend relevante Variablen darstellen (siehe hierzu Kapitel 7.2.1), kann bereits apriori von einer Verzerrung der aus Access-Panels resultierenden Stichproben ausgegangen werden. Dies trifft insbesondere auf Aussagen zu, durch welche ein Rückschluss auf die allgemeine volljährige oder wahlberechtigte Bevölkerung gezogen werden soll. Da Access-Panels mit diesem Problem der verzerrten Zusammensetzung zu kämpfen haben, wird zur Erlangung von Stichproben aus nicht-zufallsbasierten Access-Panels üblicherweise gar nicht der Versuch einer zufallsbasierten Auswahl angewandt, da sich diese Verzerrung dann innerhalb der Stichprobe widerspiegeln würde. Stattdessen wird mit einer Quotierung der zu erlangenden Stichprobe eine bewusste Auswahl in den Ziehungsprozess eingebracht (Callegaro/Baker et al. 2014: 11 f.). Quotenstich-

proben weisen verschiedene Schwächen auf und sorgen für eine weitere Selektivität während des Auswahlprozesses der resultierenden Stichprobe (Schnell/Esser/Hill 2013: X). Die Hoffnung der Quotierung ist, dass die hierdurch wiederum hervorgerufenen Verzerrungen die ursprünglich auf diesen Variablen bestandenen und durch die Quotierung beseitigten Verzerrungen aufwiegen.

Abbildung 6.8 zeigt die jeweiligen Quotenvorgaben durch die Betreiber der Access-Panels und deren Abweichungen zu den angegebenen Verteilungen innerhalb dieser.[34] Eine effektiv abweichende Quotierung der resultierenden Stichproben von der Verteilung des Respondi-Panels bezüglich des Alters wird insbesondere in den ersten Erhebungen vorgenommen. Die Gruppe der 20 bis 29-jährigen wird durch die Quotierung um beinahe 20 Prozentpunkte nach unten korrigiert. Die Gruppe der Älteren ab 60 Jahren wird hingegen um ca. 15 Prozentpunkte nach oben quotiert. Innerhalb des LINK-Panels wird hingegen nur eine sehr geringe Korrektur durch eine Quotierung, im Vergleich zum Mikrozensus, angewandt. Das Geschlecht wird durch beide Betreiber kaum korrigiert.

Abbildung 6.8.: Quotenvorgaben und Grundgesamtheit Access-Panel Respondi und LINK



Quelle: Eigene Darstellung. Daten: Studienbeschreibungen der Komponente 8 zur Angabe der Quoten der Access-Panel der jeweiligen Anbieter. Bis Welle 17 Respondi, ab da an LINK.

[34]Quelle der Angaben: Jeweilige Studienbeschreibung der Erhebungen der Komponente 8. Abruf: 11. Mai 2018.

Anders gestaltet sich dies bezüglich des Bildungshintergrunds. Hier findet sich betreiberübergreifend eine sehr konstante Quotierung. Die Gruppe mit einem (Fach-)Abitur wird um ca. 30 Prozentpunkte nach unten korrigiert, maßgeblich zu Gunsten derjenigen mit einem geringen Bildungsabschluss. Lediglich die Durchführung der letzten Erhebungen der Komponente 8 zeigen die Tendenz im LINK-Panel, die Quotierung zu Gunsten des niedrigen Bildungsstatus zu lockern. Ob für die Veränderung der Quotierung jedoch eine Einschätzung der Verbesserung der Qualität des eigenen Panels oder die Aufgabe des (anscheinend) nicht zu erfüllenden Anspruchs der Quotierung maßgeblich ist, kann hier nicht beantwortet werden.

Die eigentliche Frage ist jedoch, wie sich diese bisher dargestellten Formen der Selbstselektion, Noncoverage- und Nonresponse-Probleme und die daraus folgenden Entscheidungen der Anwendung nicht-zufallsbasierter Auswahlverfahren in Form von Quotierungen in den Verteilungen der *resultierenden Stichproben* aus diesen beiden Access-Panels zeigen. Hierzu ist der Vergleich mit den beiden anderen Erhebungsmodi notwendig. Erst dann lässt sich aufklären, ob telefonische Umfragen ebenfalls mit bestimmten selektiven Teilnahmeprozessen auf der Grundlage soziodemografischer Globalvariablen konfrontiert sind. Dies kann die Tendenz erklären, warum die Altersverteilung durch die Offlinrekrutierung in Abbildung 6.7 im Vergleich zum Mikrozensus verbessert wurde, die Erfassung des Bildungshintergrunds jedoch nicht.

6.3. Zwischenfazit: Fehlerquellen und Erhebungsmodi

Mit den gewonnen Erkenntnissen ist nun ein Zwischenfazit möglich. Der Teil der Arbeit war angetreten, um die Erhebungsvielfalt von Umfragen darzustellen. Für die drei als relevant herausgestellten Erhebungsmodi der persönlichen, telefonischen und onlinebasierten Befragung wurden die jeweiligen Vor- und Nachteile beleuchtet. Hierzu wurde die jeweilige Anfälligkeit gegenüber den Fehlerquellen des TSE herausgestellt. Ebenfalls wurde eine starke Verknüpfung zwischen der anvisierten Befragungsmethode und dem damit einhergehenden Auswahldesign aufgezeigt. Wie Personen befragt werden sollen, hat bereits einen sehr starken Einfluss darauf, wie diese zur Befragung ausgewählt werden.

Umfragen lassen sich nach ihrem angewandten Erhebungsmodus unterscheiden. Bei persönlichen Befragungen wird in der Bundesrepublik eine Stichprobe meist über ein mehrstufiges Zufallsverfahren im Rahmen des ADM-Design oder über die Ziehung von Registerstichproben erhoben. Auch telefonische

Befragungen weisen über das RLD-Verfahren eine mehrstufige Zufallsauswahl als Auswahlgrundlage auf. Dieses ist jedoch mit zunehmenden Problemen hinsichtlich der Abdeckung der volljährigen oder wahlberechtigten Bevölkerung, bedingt durch die steigende alleinige Mobilfunknutzung, konfrontiert. Da dieses Noncoverage-Problem zudem nicht unabhängig von soziodemografischen Merkmalen der zu befragenden Personen ist, existiert ein hohes Potential für einen systematischen Fehler. Hier ist zudem eine pessimistische Prognose angebracht: Das Problem erweist sich als zunehmende Herausforderung der Nutzung telefonischer Umfragen. Die Integration der Mobilfunknutzer im Rahmen eines „Dual-Frame“ als mögliche Lösung wirft jedoch das Problem der Vergleichbarkeit der durchgeführten Interviews auf. Zudem dürfte die Kontaktierung über das Mobilfunktelefon zu einer immer weiter absinkenden Teilnahmebereitschaft an telefonischen Umfragen führen.

Für onlinebasierte Befragungen ist es hingegen unmöglich, eine Zufallsauswahl der volljährigen oder wahlberechtigten Bevölkerung zu realisieren. Die Anwendung nicht-zufallsbasierter Auswahldesign bei onlinebasierten Befragungen ergibt sich daher aus der mangelnden Möglichkeit der Erstellung und Nutzung eines Auswahlrahmens. Ebenfalls kombiniert sich dies mit einem qua Erhebungsformat nicht zu vermeidenden Noncoverage-Fehler. Onlineumfragen schließen per se die Bevölkerung ohne Internetzugang aus, diese hat keine Möglichkeit der Teilnahme. Der Anteil dieser „Offliner“ macht seit einigen Jahren ca. 80 % der Bevölkerung aus. Zu einem nicht zu vermeidenden Bias kommt es jedoch, da systematische Unterschiede zwischen den Gruppen bestehen. Der höhere Anteil der „Onliner“ wird daher durch die sich dadurch bedingenden weiter zunehmenden Unterschiede zu den verbliebenen Personen ohne Internetzugang relativiert. Eine Möglichkeit dieses Problem zu umgehen ist die Auswahl von Personen über ein offline stattfindendes Auswahldesign *und* die Bereitstellung der technischen Infrastruktur für Personen ohne Internetzugang. Es finden sich, nicht zuletzt auf Grund der Kosten, jedoch nur sehr wenige Durchführungen dieser Form der onlinebasierten Befragung. Dies bedingt zudem eine Entscheidung zur Teilnahme an der Umfrage, welche nicht abhängig von der Kompetenz des Umgangs mit dem Medium Internet ist.

Neben der Garantie der zufallsbasierten Auswahl aller Personen der anvisierten Grundgesamtheit zur Teilnahme an einer Umfrage ist jedoch ebenfalls die tatsächliche Entscheidung durch die ausgewählte Person entscheidend. Hier haben sich persönliche Befragungen im direkten Vergleich als der Erhebungsmodi mit einer vergleichsweise hohen Ausschöpfungsquote herausgestellt, auch

wenn diese ebenfalls mit ca. 40 % Potential für einen dadurch bedingten Bias ergibt. Telefonische Umfragen hingegen weisen Teilnahmeraten zwischen 10 und 20 % auf. Hier besteht eine sehr hohe Wahrscheinlichkeit, dass es zu einer zufallsbasierten Auswahl, jedoch nicht zu einer zufallsbasierten Entscheidung der Teilnahme kommen wird. Onlineumfragen haben aus einem anderen Grund eine hohe Wahrscheinlichkeit für einen Nonresponse-Fehler: Die Entscheidung zur Teilnahme ist nicht unabhängig von der Kompetenz des Umgangs mit dem Internet. Die Kompetenz, als definierte Folge der Häufigkeit des Umgangs mit dem Medium, ist wiederum nicht gleichmäßig über die Bevölkerung verteilt.

Ein zentraler Vorteil der onlinebasierten Befragung ist die Abwesenheit von Interviewern. Neben dem Kostenaspekt ergibt sich hierdurch kein Reaktivität des Interviewers auf die Umfrageteilnehmer. Daher ist eine bessere Erfassung von Antworten auf Fragen möglich, welche von einer Tendenz der sozialen Erwünschtheit betroffen sind. Im Rahmen von Wahlumfragen können sich hierdurch Vorteile der korrekten Erfassung rechtspopulistischer Ansichten und des sich daraus ergebenden Potentials solcher Parteien ergeben. Ebenfalls kann die abgeschätzte Wahlbeteiligungsrate besser eingeschätzt werden, sofern die Tendenz bestimmter Bevölkerungsgruppen besteht, als sozial empfundende Norm wählen zu gehen.

Nur persönliche und telefonische Befragungen lassen sich mit einer zufallsbasierten Auswahlmöglichkeit einer Stichprobe aus der allgemeinen oder wahlberechtigten Bevölkerung kombinieren. Onlinebasierte Befragungen sind hingegen per se mit einem systematischen Fehler durch ihr Auswahldesign belegt. Dies resultiert in einem hohen Potential für das Vorliegen eines Noncoverage-Fehlers.

Dieses weisen auch telefonische Befragungen auf, da ihr Auswahldesign zunehmende Bevölkerungsgruppen nicht mehr ausreichend erreicht *und* sich diese systematisch von den erfassten Personenkreisen unterscheiden. Die niedrigen Ausschöpfungsquoten telefonischer Befragungen sorgen zudem für eine erhöhte Wahrscheinlichkeit eines Nonresponse-Fehlers im Vergleich zur persönlichen Befragung. Diese wiederum haben sich am anfälligsten gegenüber einem Messfehler bedingt durch den Interviewereinsatz erwiesen. Die Selbstadministration der onlinebasierten Befragung weist hierbei einen klaren, und auch ihren einzigen, Vorteil auf.

Insgesamt kommt es daher zu einem Eindruck der ordinalen Rangfolge: Der Goldstandard der Erhebung ist die persönliche Befragung, hier ist die zufallsbasierte Auswahl größtenteils deckungsgleich mit einer Teilnahme, welche nicht in Kombination mit interessierenden Merkmalen steht. Die telefonische Befragung

weist auch eine zufallsbasierte Auswahl auf, sieht sich jedoch mit zunehmenden Einschränkungen konfrontiert. Die onlinebasierte Befragung schließlich ist, sofern das Ziel eine bevölkerungsweite Aussage ist, ein Erhebungsformat, dessen Wahl sich nur schwer rechtfertigen lässt. Der nachfolgende empirische Teil hat zur Aufgabe, die daraus resultierenden Einschränkungen der Datenqualität in Relation zu den anderen beiden Erhebungsmodi zu betrachten. Hierzu wird im folgenden Schritt die dazu ausgewählte Datengrundlage besprochen, um dieses Ziel des Vergleichs der Qualität persönlicher, telefonischer und onlinebasierter Befragungen auf der Basis von Access-Panels zu realisieren.

7. Inhaltliche Faktoren zur Korrektur

An dieser Stelle der Arbeit ist geklärt, welche Fehlerquellen die Qualität von Umfragen reduzieren können (Kapitel 2). Ebenso wurden Möglichkeiten aufgezeigt, den aus diesen resultierenden Gesamtfehler einer Umfrage messbar zu machen (Kapitel 3). Zudem wurde die Anwendung von Gewichtungsverfahren vorgestellt, welche durch die Bestimmung von Bedeutungsgewichten eine Reduzierung dieses Gesamtfehlers bewirken können (Kapitel 4). Dies ergibt sich durch die Angleichung an Referenzverteilungen im Rahmen einer „Iterative Proportional Fitting“ (IPF)-Gewichtung oder über die Modellierung von „Propensity Scores“ (PS) mittels logistischer Regressionsverfahren. Die PS stellen dann die Wahrscheinlichkeit der Zugehörigkeit einer Person zu einer Referenzumfrage dar. Eine Alternative ist die Nutzung der PS als Grundlage eines weiterführenden „Propensity Score Matching“ (PSM). Der Erfolg dieser Vorgehensweise(n) ist zentral von der Güte des spezifizierten Korrekturmodells zur Bestimmung der PS abhängig.

Auf der Grundlage dieser Erkenntnisse wird das nachfolgenden Kapitel 7.1 definieren, welche generellen Eigenschaften eine gute „Information“ als Grundlage eines solchen Modells mitbringen sollte. Anschließend stellt das Kapitel 7.2 persönliche Eigenschaften von Umfrageteilnehmern als solch eine mögliche Information heraus. Hierzu wird dargelegt, dass diese sowohl den Teilnahmeprozess an einer Umfrage, als auch das Antwortverhalten auf inhaltliche Fragen beeinflussen. Im Rahmen der Arbeit sind diese Individualinformationen die soziodemografischen Rahmendaten der befragten Personen einer Umfrage (Kapitel 7.2.1), deren Persönlichkeitsmerkmale (Kapitel 7.2.2) sowie deren Selbsteinschätzung hinsichtlich der Kompetenz des Verständnis von und ihres Interesses an Politik aufgearbeitet. Dies wird mit der Erfassung ihrer Einstellungen gegenüber dem politischen System und seinen Akteuren kombiniert (Kapitel 7.2.3). Diese Faktoren beeinflussen inhaltliche Zielvariablen einer Wahlumfrage, wie das Wahlverhalten und die Beteiligungsabsicht, welche das Kapitel 7.3 herausstellt. Als Resultat ergibt sich ein Gesamtmodell in Kapitel 7.4 der, im Rahmen der Arbeit verstandenen, Einflussfaktoren auf die Umfrageteilnahme und das daraus resultierende Antwortverhalten.

7.1. Eigenschaften und Quellen von Informationen zur Korrektur

Schätzer auf Basis einer Umfrage weisen eine Verzerrung in Relation zur anvisierten Grundgesamtheit auf, wenn der gedankliche Prozess der Entscheidung der Teilnahme an einer Umfrage in Verbindung zu den inhaltlichen Abfragen dieser steht. In diesem Fall ist der Ausfall als systematisch zu klassifizieren (Bethlehem 2002). Zwei Anforderungen muss ein Korrekturmodell erfüllen, damit die darauf basierende Bestimmung von Bedeutungsgewichten Verzerrungen in Umfragen beseitigen kann: Sowohl der Zugangs- als auch Verzerrungsprozess einer Umfrage muss durch dieses erklärt werden. Informationen zur Korrektur müssen daher sowohl Unterschiede der Wahrscheinlichkeiten der Teilnahme an einer Umfrage zwischen Personen erklären, als auch mit den eigentlich inhaltlich interessierenden Variablen der Umfrage korrelieren (z.B. Bethlehem/Cobben/Schouten 2011: 248 f. Kreuter/Olson et al. 2010; Stoop 2016: 419).

Lässt sich hingegen nur der Teilnahmeprozess einer Umfrage durch ein Korrekturmodell aufdecken, führt die Korrektur lediglich zu einer Erhöhung der Varianz von Schätzern inhaltlich interessierender Variablen einer Umfrage, da deren Verzerrung nicht mit dem modellierten Zugangsprozess korreliert. Die Folge ist eine sinkende Effizienz der Schätzung bei einer konstanten Verzerrung (z.B. Little/Vartivarian 2005; Groves/Heeringa 2006). Erklärt eine für das Korrekturmodell genutzte Variable hingegen nicht den Teilnahmeprozess einer Umfrage, kann sie auch keinen Einfluss auf die resultierenden Bedeutungsgewichte entfalten. Umfasst das Korrekturmodell andererseits nicht alle die Verzerrung von Variablen aufklärenden Variablen, ist eine vollständige Reduktion der Verzerrung nicht möglich (Stuart 2010: 5). Sehr gute Ergebnisse lassen sich mit Informationen in Form solcher Variablen erzielen, welche relativ schwach mit dem Teilnahmeprozess der Umfrage und relativ stark mit den von Verzerrungen betroffenen Zielvariablen korrelieren. Die Korrektur wird beseitigt, die Erhöhung der Varianz von Schätzern durch die Gewichtung steigt nur gering (Ho et al. 2007: 216).

Generell lässt sich die Partizipation an einer Umfrage als abhängig vom jeweiligen gesellschaftlichen und sozialen Klima, verschiedenen Aspekten des Designs der Erhebung und den Eigenschaften der zu befragenden Personen begreifen. Die (soziale) Umwelt prägt hierbei Einstellungen gegenüber Umfragen (z.B. die als allgemein empfundene Einstellung gegenüber Umfragen), welche die Entscheidung zur Teilnahme beeinflussen. Auch die Art der Datenerhebung,

eine mögliche Incentivierung oder die Existenz eines Interviewers wirken auf diese. Neben diesen äußeren Kontextfaktoren ist die Entscheidung zur Teilnahme jedoch vor allem von individuellen Charaktereigenschaften, des individuellen Interesse und der persönlichen Kompetenz des Verständnis der abgefragten Themen abhängig (Groves/Cialdini/Couper 1992; Groves/Couper 1998; Bethlehem/Cobben/Schouten 2011: 148).[1] Dies wird zudem von der Einordnung der die Umfrage durchführenden Institution durch die ausgewählte Person begleitet (Dillman 2002).

Damit die herausgestellten Faktoren zur Korrektur genutzt werden können, müssen diese verfügbar sein. Bei persönlichen Charakteristika ist dies über die Abfragen in Umfragen möglich, sofern die Einladung zur Teilnahme wahrgenommen wird. Neben der direkten Befragung der Personen können aggregierte Verteilungen durch administrative Daten der amtlichen Statistik oder kommerzielle Quellen genutzt werden. Auch bereits erhobene Umfragen können als Referenz dienen. An diese kann die vorliegende Umfrage dann angeglichen werden (Bethlehem/Schouten 2016: 562), was bereits das vorangegange Kapitel herausgearbeitet hat. Neben der Art der Quelle ist zusätzlich relevant, welche Gruppen diese Information umfasst. Aggregierte Referenzverteilungen der amtlichen Statistik stellen die gemeinsamen Ergebnisse von potentiellen Respondenten und Nonrespondenten bereit. Abweichungen von diesen Werten können daher durch eine wirklich vorhandene Verzerrung, als auch durch den sich aus der Nutzung einer Stichprobe resultierenden Stichprobenfehler ergeben (Bethlehem/Schouten 2016: 563). Generell werden Daten der amtlichen Statistik häufig mit einem hohen Qualitätsstandard assoziiert. Elementar für eine Nutzung ist jedoch, dass die bereitgestellten Informationen auch die identische Information der Abfrage in der Umfrage enthalten (Bethlehem/Cobben/Schouten 2011: 257).

Für die Korrektur soziodemografischer Variablen kann auf Referenzdaten der amtlichen Statistik zurückgegriffen werden. Für darüber hinausgehende Variablen ist eine solche Referenz der (annähernd) wahren Werte in der Regel nicht verfügbar. Eine Möglichkeit zur Erlangung dieser besteht in der Nutzung von „follow-up“-Studien von Nonrespondenten. Diese haben den Vorteil der exakten Möglichkeit der Quantifizierung einer Verzerrung (z.B. Roberts/Vandenplas/Stähli 2014). Dies gilt jedoch nur, wenn dies nicht zu einem Wechsel des Erhebungsmodus führt und einen damit möglicherweise auftauchenden Messfehler

[1] Für eine intensive Aufarbeitung des Literaturstands zu diesen Aspekten, insbesondere dem Design der Umfrage, siehe Keusch 2015.

impliziert. Ebenfalls kann ein möglicher zeitlicher Abstand zur ersten Erhebung einen negativen Effekt haben (Vandenplas et al. 2015: 142). Des Weiteren können für den Vergleich von Umfragen auch Referenzverteilungen anderer Umfragen herangezogen werden, sofern diese eine identische Frageformulierung aufweisen und die identische Population beschreiben. Im Falle von Onlineumfragen kann der Ausschluss der Bevölkerung ohne Internetzugang entscheidend sein (Callegaro/Villar et al. 2014: 27).

Die notwendige Identifizierung von Informationen zur Korrektur kann also auf der individuellen Personenebene des Umfrageteilnehmers, des spezifischen Kontextes und der Rahmenfaktoren einer Umfrage geschehen. Erst durch diese Informationen ist überhaupt eine Modellschätzung möglich, welche zwischen Teilnehmergruppen unterschiedlicher Umfragen unterscheiden kann. Die Entscheidungsmechanismen zur Teilnahme an einer Befragung sind jedoch sehr umfrage- und kontextspezifisch. Daher lässt sich bereits früh diagnostizieren, dass diese „[...] complexity that makes global remedies for nonresponse such as a universal weighting system to correct bias error-prone, even though there are various useful techniques for use on a survey-by survey basis.“ (Goyder 1987: 187). Ein Korrekturmodell ist daher sehr kontextspezifisch zu konstruieren, ein Universalmodell der Erklärung jeglicher Verzerrungen in thematisch unterschiedlichsten Umfragen eine Utopie. Wahlumfragen weisen einen solchen spezifischen Kontext auf. Ein solches Modell muss erklären können, warum Personen sich entscheiden an Wahlumfragen teilzunehmen und das deren Teilnahmeverhalten einen Einfluss auf inhaltliche Zielvariablen haben kann. Es sind also die Merkmale der befragten Personen von Relevanz, welche ihr politisches Verhalten kennzeichnen (können). Daher ist es notwendig darzustellen, bei welchen Variablen ein solcher „Wirkungskanal der Korrektur“ denkbar ist.

7.2. Wirkungskanäle der Korrektur von Wahlumfragen

Diese werden nun vorgestellt und deren Einfluss auf den Teilnahmeprozess an einer Wahlumfrage und das durch diese erfasste politische Verhalten skizziert. Gezeigt wird dies für drei verschiedene Gruppen von Variablen, welche im empirischen Teil der Arbeit für die Bestimmung von Korrekturmodellen genutzt werden. Dies sind *soziodemografische Variablen*, *Persönlichkeitseigenschaften* sowie *Einstellungen und Kompetenzeinschätzungen*.

7.2.1. Soziodemografie

Soziodemografische Variablen werden am häufigsten zur Korrektur von Verzerrungen herangezogen. Sie sind die am offensichtlichsten erscheinenden Eigenschaften von befragten Personen und in beinahe allen Umfragen Teil der Abfrage. Ihre Korrektur liegt daher nahe (z.B. Groves/Cialdini/Couper 1992; Toepoel 2016: 202 f.).[2] Entsprechend häufig findet sich ihre Nutzung auch im Bereich der Literatur zum Umgang mit Nonresponse in Umfragen. Matsuo et al. (2010) nutzen beispielsweise in einem „follow-up“ von Nonrespondenten des ESS für Belgien und Norwegen die Variablen Alter, Bildungsgrad, Erwerbsstatus und die Haushaltsgröße zur Korrektur. Nonrespondenten zeigen sich vor allem als älter und weniger gut gebildet (Matsuo et al. 2010: 167). Billiet/Philippens et al. 2007 finden auf der Grundlage der Durchführung des „European Social Survey“ (ESS) für Deutschland ebenfalls Indizien, dass Nonrespondenten älter sind und häufiger in Großstädten leben (Billiet/Philippens et al. 2007: 151). Individualisiertes Verhalten lässt sich jedoch nur selten durch grundlegende soziodemografische Charakteristika einer Person erklären (Schnell 1997: 198 ff.). Gewichtungskorrekturen auf der Basis soziodemografischer Variablen zeigen daher häufig nur eine geringe qualitative Verbesserung der Datenqualität einer Stichprobe. Ihr Einfluss auf das Antwortverhalten eigentlich interessierender Variablen einer Umfrage ist sehr gering (z.B. Schnell 1997; Shadish/Clark/Steiner 2008; Peytcheva/Groves 2009).

Für politikwissenschaftliche Fragestellungen kann sich insbesondere das Alter jedoch als eine relevante Größe äußern, da mit einem zunehmendem Alter eine konservativere Einstellungen weiter verbreitet sind (z.B. Feather 1977; Truett 1993). Ältere Kohorten weisen eine erhöhte Wahrscheinlichkeit eines konservativeren Lebensstils und damit der Wahl einer konservativen Partei auf (Tilley/Evans 2014) und mit einem zunehmenden Alter entwickeln bestimmte Persönlichkeitstypen eher konservativere Wertvorstellungen (van Hiel/Cornelis/Roets 2007). Auch der Bildungsgrad kann entscheidend als Teil eines Korrekturmodells einer Wahlstudie sein. Personen mit einem höheren Bildungsstand weisen häufig eine erhöhte Bereitschaft zur politischen Partizipation auf, auch wenn der dahinterstehende Wirkungskanal in der Literatur nicht abschließend klar ist (z.B. Brady/Verba/Schlozman 1995; Norris 2004; Hillygus 2005). Zudem bedingt die Kombination verschiedener soziodemografischer Variablen auch die soziale Einbettung einer Person, wodurch sich das soziale Umfeld auch auf

[2] Praktisch alle Anwendungen eines IPF-Verfahrens beruhen auf dieser Gruppe von Variablen, da diese Informationen über die amtliche Statistik unproblematisch beziehbar sind.

das ausgeübte Partizipations- und Wahlverhalten auswirkt (Voogt/Saris 2003). Innerhalb der Bundesrepublik äußert sich dies durch eine Ungleichheit der Wahlbeteiligung in Abhängigkeit des Bildungs- und Einkommensgrades (z.B. Roßdeutscher/Schäfer 2015) und verstärkt sich innerhalb der jüngeren Alterskohorten durch eine bildungsabhängige Beteiligungsentscheidung weiter (z.B. Abendschön/Roßteutscher 2016). Das Alter und der Bildungsstand von zu befragenden Personen kombinieren damit als eine erste Gruppe bereits die notwendigen Eigenschaften zur Korrektur: Sie korrelieren mit dem Teilnahmeprozess an einer Umfrage und können einen Einfluss auf politische Ansichten und das Wahlverhalten entwickeln.

Neben diesen soziodemografischen „Globalvariablen“ existieren jedoch weitere individuellen Eigenschaften einer Person, welche einen gleichzeitigen Einfluss auf das Teilnahmeverhalten an Umfragen und das Antwortverhalten auf inhaltliche Zielvariablen einer Wahlumfrage haben: Die Persönlichkeit der zu befragenden Personen.

7.2.2. Persönlichkeit

Die Persönlichkeit von Personen lässt sich innerhalb des dominierenden „trait“-Paradigmas der Psychologie durch die Unterteilung der „BIG Five“ nach dem Fünf-Faktoren-Modell erfassen (z.B. Digman 1990; Goldberg 1990; Ostendorf/Angleitner 1994).[3] Die Persönlichkeit definiert sich demnach an Hand fünf verschiedene Kernmerkmale. *Offenheit* äußert sich durch eine hohe Kreativität und Neugier. Sie geht einher mit Intelligenz und einem höheren Bildungsgrad. *Gewissenhaftigkeit* hingegen kennzeichnet sich durch Ordentlichkeit, Zuverlässigkeit und Beharrlichkeit aus. Ein hoher Grad an *Extraversion* zeigt sich durch ein geselliges, aktives und ungehemmtes Verhalten. Personen mit einem freundlichen, hilfsbereiten und Wärme signalisierenden Umgang mit anderen Personen weisen hingegen eine hohe *Verträglichkeit* auf. *Neurotizismus* wiederum zeigt sich durch Nervosität, Ängstlichkeit und Gefühlsschwankungen. Diese fünf Dimensionen äußern sich in hunderten verschiedener möglicher Charaktertypen von Personen, lassen sich jedoch in Umfragen auf sehr wenige zentrale Fragen verkürzen (z.B. Asendorpf/Neyer 2012: 107 f. Rammstedt/Kemper/Klein et al. 2013: 234). Im Extremfall ist dies mit einer Fragebatterie von nur zehn oder sogar fünf Fragen möglich (Gosling/Rentfrow/Swann 2003). Auch für den deutschsprachigen Raum existieren Vorschläge zur Operationalisierung der „BIG Five“ (Rammstedt 2007; Rammstedt/Kemper/Klein

[3] Für eine breite Aufarbeitung der Genese des Konzept siehe z.B. Saßenroth (2012: 65 ff.).

et al. 2013). Diese Versuche der Standardisierung und psychometrischen Validierung solcher psychologischer „Kurzskalen“ soll dazu dienen, eine stärkere Vergleichbarkeit zwischen Bevölkerungsumfragen und eine Erhöhung der Datenqualität ebendieser zu erreichen (Rammstedt/Kemper/Schupp 2013: 148). Was die Persönlichkeitsmerkmale einer Person zur Korrektur von Verzerrungen eignet, ist ihre hohe Stabilität. Trotz möglicher, sich vor allem in jungen Jahren vollziehender Veränderungen der Persönlichkeitsmerkmale einer Person, stellt die Persönlichkeit eine sehr zeitinvariante Größe dar. Auch die Verteilung von Persönlichkeitstypen innerhalb einer Bevölkerung ist durch eine solche Stabilität gekennzeichnet (Caspi/Roberts/Shiner 2005: 476 f.). Durch diese Eigenschaften lassen sich zur Korrektur einer erhobenen Umfrage die Verteilung der Persönlichkeitseigenschaften einer Referenzumfrage mit einigem zeitlichen Abstand nutzen.

Für Wahlumfragen ist die korrekte Erfassung der Verteilung der Persönlichkeitsmerkmale nicht unerheblich, um inhaltliche Fragen ohne eine Verzerrung zu erfassen. Konzeptionell und theoretisch sind diese, mittlerweile, eine innerhalb der Politikwissenschaft als relevant erachtete Größe zur Erklärung des politischen (Wahl-)Verhaltens und der Einstellung gegenüber politischen Parteien (z.B. Capara/Barbaranelli/Zimbardo 1999; Schumann 2002; Schumann/Schoen 2005; Schoen/Schumann 2007; Vecchione et al. 2011; Schumann 2014). Existierende liberale oder konservative Orientierungen haben, unter anderem, ihren Ursprung in unterschiedlichen Persönlichkeitsprofilen, insbesondere der Offenheit und Gewissenhaftigkeit (z.B. Carney et al. 2008). Dies äußert sich in einer unterschiedlichen Sympathie gegenüber Parteien auf den Enden des klassischen Links-Rechts-Spektrums. Offenheit erhöht die Tendenz eines höheren Sympathiegrades für die Parteien des linken Spektrums, Personen mit einer hohen Gründlichkeit neigen hingegen eher zu Parteien des rechten Spektrums (Schumann 2001: 265, 2002: 70). Auch wirkt sich die Persönlichkeit auf den Grad der politischen Aktivität (Rammstedt 2007; Mondak/Halperin 2008) und die Existenz und Stärke einer Parteiidentifikation aus (Gerber/Huber/Doherty/Dowling 2012). Der Einfluss der Persönlichkeitsmerkmale auf politische Einstellungen ist hinsichtlich der Erklärungskraft mit derjenigen von soziodemografischen Variablen vergleichbar und übernimmt damit eine sehr generelle Funktion des Einflusses (Gerber/Huber/Doherty/Dowling/Ha 2010), welche sich in unterschiedlichen Reaktionen auf Fragestellungen im Rahmen einer Umfrage äußern (Gerber/Huber/Doherty/Dowling 2011: 268 f.).

Auch auf die Teilnahmebereitschaft an Umfragen wirken sich die Persönlichkeitsmerkmale der BIG Five aus (z.B. Rogelberg et al. 2003; Marcus/ Schutz 2005; Brüggen/Dholakia 2010; Saßenroth 2012), ein erhöhter Grad an Neurotizismus oder Offenheit senkt diese. Ersteres lässt sich durch die „Social Isolation Hypothesis“ erklären. Die, von einer objektiven Isolation unabhängige, subjektive Isolation einer Person reduziert ihre Bereitschaft zur Interaktion und damit der Teilnahme an Umfragen. Offenheit hingegen senkt diese durch eine zunehmende „Interviewgewöhnung“. Personen empfinden in dem Fall die Teilnahme nicht mehr als etwas spannendes und aufregendes (Saßenroth 2012: 161 ff.). Dies erklärt Unterschiede zu früheren Studien der Umfrageforschung, welche einen positiven Effekt der Offenheit auf die Teilnahme attestieren (z.B. Dollinger/Leong 1993; Marcus/Schutz 2005: 968 ff.). Extraversion, Verträglichkeit oder auch Gewissenhaftigkeit erhöhen hingegen die Teilnahmewahrscheinlichkeit, da Personen mit diesen Persönlichkeitsmerkmalen eine geringere subjektiv wahrgenommene Isolation aufweisen (Saßenroth 2012: 161 ff.).

Persönlichkeitsmerkmale der BIG Five beeinflussen daher letztlich den Grad der als subjektiv wahrgenommenen Isolation einer ausgewählten Person und somit auch ihr Teilnahmeverhalten an Umfragen. Die Persönlichkeit einer Person und ihr Wahlverhalten lassen sich jedoch konzeptuell nicht direkt in einen Ursache-Wirkungszusammenhang bringen. Das spezifische Wahlverhalten ist vielmehr von den Einstellungen gegenüber den jeweiligen Akteuren eines Parteiensystems abhängig. Damit es zu einer Wirkung kommt, muss die Persönlichkeit daher für die Herausbildung von Einstellungen gegenüber Parteien verantwortlich sein (Schumann 2014: 613 f.). Dies leitet zu einer weiteren Gruppe an möglichen Korrekturfaktoren von Umfragen im Kontext von Wahlumfragen über: Einstellungen gegenüber den Akteuren und Institutionen des politischen Systems, sowie die generelle Kompetenz des Verständnisses und das Interesse an Politik als solches.

7.2.3. Kompetenz- und Einstellungsabfragen

Einstellungen von Personen sind, im Gegensatz zur Persönlichkeit, zielgerichtete Wertungen bestimmter Objekte. Dies lässt sich konzeptuell innerhalb des eindimensionalen Einstellungsmodells nach Fishbein (1963) erfassen. Zur Herausbildung einer Einstellung gegenüber einem Objekt, wie beispielsweise einer Partei, verknüpft eine Person verschiedene Merkmale mit einer unterschiedlichen inhaltlichen Bedeutung mit diesem. Die gewichtete Bewertung der verknüpften Einzelmerkmale bildet dann in ihrer Summe eine generelle Einstellung

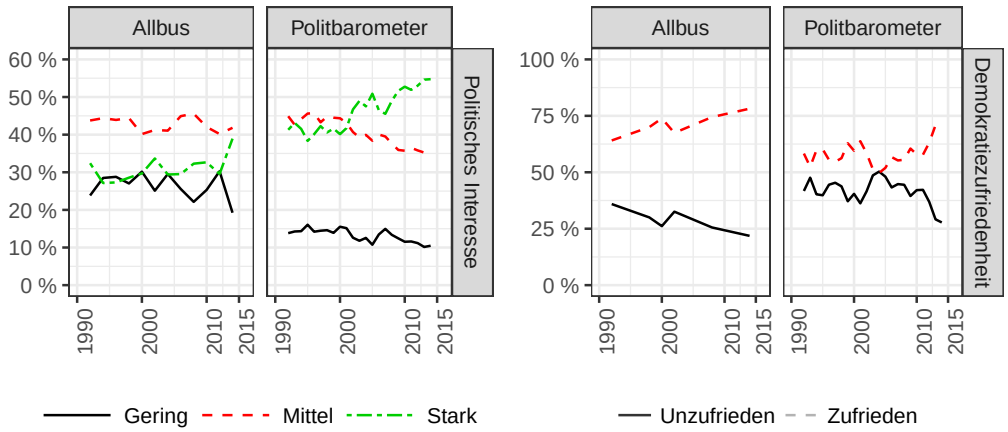
gegenüber dem Objekt (Schumann 2002: 71 f.). Durch die Veränderung der Merkmale eines Objekts kommt es zu einer affektiven Reaktion durch eine Person, welche durch die Persönlichkeit gesteuert wird. Hierdurch verändern sich Einstellungen gegenüber Objekten (Schumann 2014: 598 ff.). Die Persönlichkeit einer Person kann daher auch als eine Art „Proxy-Variable“ zur Erfassung von Einstellungen aufgefasst werden (Bosnjak et al. 2013: 342).

Zeigen sich langfristig Einstellungsänderungen innerhalb einer Bevölkerung durch sich ändernde Sichtweisen in jüngeren Kohorten (Gummer 2015: 163), sind diese kurzfristig nur in einem sehr geringen Ausmaß von Veränderungen betroffen. Durch diese Stabilität sind Bewertungen der Zufriedenheit mit dem Status der Demokratie und des Vertrauens in Institutionen daher zur Korrektur von weiteren inhaltlichen Verzerrungen geeignet (Vandenplas et al. 2015: 146 ff.). Auf der Grundlage des ESS können Billiet/Philippens et al. 2007 beispielsweise zeigen, dass Nonrespondenten geringere Grade an politischen und sozialen Vertrauens, ein geringeres Partizipationsverhalten, Interesse an Politik und stärkere Vorbehalte gegenüber Immigranten innerhalb der Bundesrepublik aufweisen (Billiet/Philippens et al. 2007: 153). Auch Matsuo et al. 2010 finden einen geringeren Grad sozialen Vertrauens, politischen Interesses und eine geringere Zufriedenheit mit dem Status der Demokratie, welche effektiv für eine Korrektur weiterer inhaltlicher Abfragen genutzt werden können (Matsuo et al. 2010: 167 ff.). Damit stellen sowohl die Einstellungen gegenüber Objekten, wie der „Demokratie“ im Allgemeinen, als auch die Kompetenz zum Verständnis und das Interesse am Umfragethema mögliche stabile Informationen bereit und eignen sich daher auch zur Nutzung einer zeitversetzten Korrektur durch herangezogene früher erhobene Referenzumfragen.

Dies belegt auch die Abbildung 7.1. Dargestellt sind die prozentualen Verteilungen des Interesse an Politik und des Grades der Zufriedenheit mit der Situation der Demokratie in der Bundesrepublik. Die Datengrundlage stellen die seit 1980 erhobenen Umfragen des Allbus und des Politbarometers der Forschungsgruppe Wahlen dar.[4] Beide Umfrageprogramme weisen keine sprunghaften Veränderungen der erfassten Verteilungen des politischen Interesse und der Zufriedenheit mit der Situation der Demokratie auf, welches die hohe Stabilität dieser unterstreicht. Sehr wohl weist das Politbarometer jedoch klare Trends auf, welche sich innerhalb des Allbus nicht in der Stärke finden. Der unterschiedliche angewandte Erhebungsmodus und eine damit verbundene

[4]Genutzt wurden für das Politbarometer die kumulierten Datensätze der GESIS. Für jedes Jahr wurden Gewichte berechnet und angewandt, sofern Ost und West getrennt bereitgestellt wurden. Siehe zu einer näheren Beschreibung das nachfolgende Kapitel 5.2.

Abbildung 7.1.: Stabilität zentraler Variablen (Allbus und Politbarometer)



Quelle: Eigene Darstellung. Allbus hinsichtlich Ost-West-Ungleichgewichte korrigiert. Sofern notwendig wurde die Stichprobe in eine Personenstichprobe transformiert. Die Unterbrechung der Zeitreihe des Politbarometers zeigt den Wechsel von der persönlichen zur telefonischen Befragung an. Siehe hierzu http://www.gesis.org/wahlen/politbarometer (Stand: 11. Mai 2018).

zunehmende Selektivität kann hierfür jedoch ein Grund sein. Die nachfolgenden Kapitel werden dies noch weiter beleuchten.

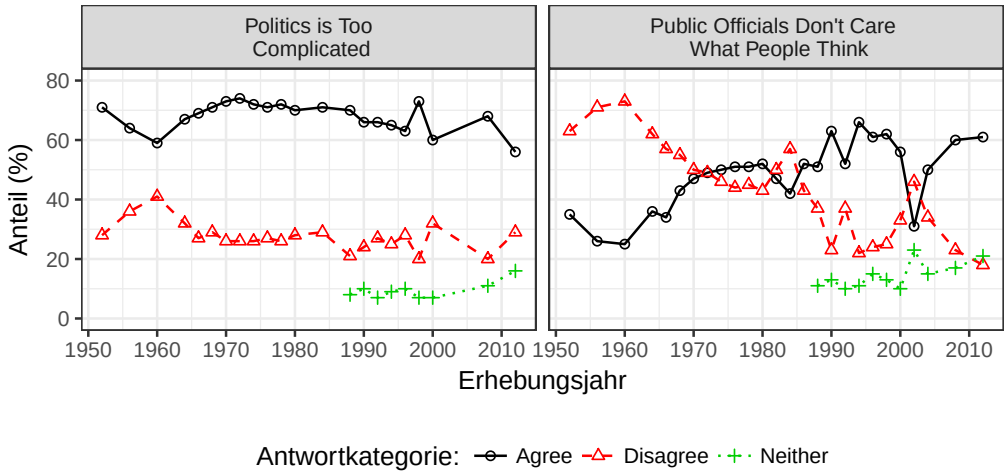
Eine weitere Gruppe möglicher Korrekturfaktoren von Wahlumfragen betrifft die Kompetenz einer Person zum Verständnis verschiedener Politikfelder und der Struktur des politischen Systems. Solche Kompetenzabfragen haben innerhalb der Politikwissenschaft seit Anfang der 1990er Jahre (wieder) eine zunehmende Bedeutung gefunden, „[...] both as a causal or intermediary variable and as a phenomenon to be explained in its own [...]“ (Carpini/Keeter 1993: 1180). Die Kompetenz zum Verständnis und das Interesse an Politik bedingen sich. Je höher das Verständnis eines Themas ist, umso stärker ist auch das Interesse ausgeprägt. Da das Interesse am Thema einer Umfrage auch generell die Teilnahmebereitschaft beeinflusst (z.B. Groves/Presser/Dipko 2004), haben diese beiden Variablen somit auch eine zentrale Bedeutung als Korrekturfaktor. Sie beeinflussen die Teilnahmewahrscheinlichkeit *und* das Antwortverhalten an einer Umfrage.

Einstellungen gegenüber (politischen) Objekten und das Verständnis politischer Sachverhalte lassen sich durch das Konzept der „Political Efficacy“ erfassen. Dieses geht auf Campbell (1954) zurück und wird bereits seit 1952 innerhalb „American National Election Studies“ (ANES) erhoben. Eine geeignete Operationalisierung ist die Unterteilung in einen Bereich der Selbsteinschätzung des Verständnis von Politik („internal political efficacy“) und in die wahrgenommene Offenheit des politischen Systems („external political efficacy“) (Vetter 1997b;

Arzheimer 2016: 107). Durch die beiden Dimensionen soll die Einschätzung von Umfrageteilnehmern bezüglich ihrer Kompetenz des Verständnisses politischer Prozesse und die Möglichkeit der aktiven Beeinflussung dieser durch sie erfasst werden (Vetter 1997a: 53). Die „Internal Efficacy“ kann daher auch als ein Persönlichkeitsmerkmal, die „External Efficacy“ als Überzeugungen gegenüber dem politischen System aufgefasst werden (Beierlein et al. 2012: 15).

Um zur Korrektur von Umfragen mit einer selektiven Teilnahme hinsichtlich der beiden Dimensionen der Efficacy geeignet zu sein, ist erneut eine Stabilität dieser eine wünschenswerte Eigenschaft. Dann können zeitlich weiter entfernt durchgeführte Referenzumfragen zur Korrektur herangezogen werden. Die Abbildung 7.2 zeigt die Entwicklung der Zustimmungsraten der umgesetzten Formulierungen der beiden Dimensionen innerhalb der ANES-Erhebungen seit 1952.[5]

Abbildung 7.2.: Entwicklung „Efficacy“ in der „American National Election Study“ (ANES) von 1950 bis 2012



Quelle: Eigene Darstellung. Daten: Table 5B.1 und Table 5.B der Kategorie 5.B.Efficacy, URL: http://www.electionstudies.org/nesguide/gd-index.htm (Stand: 11. Mai 2018).

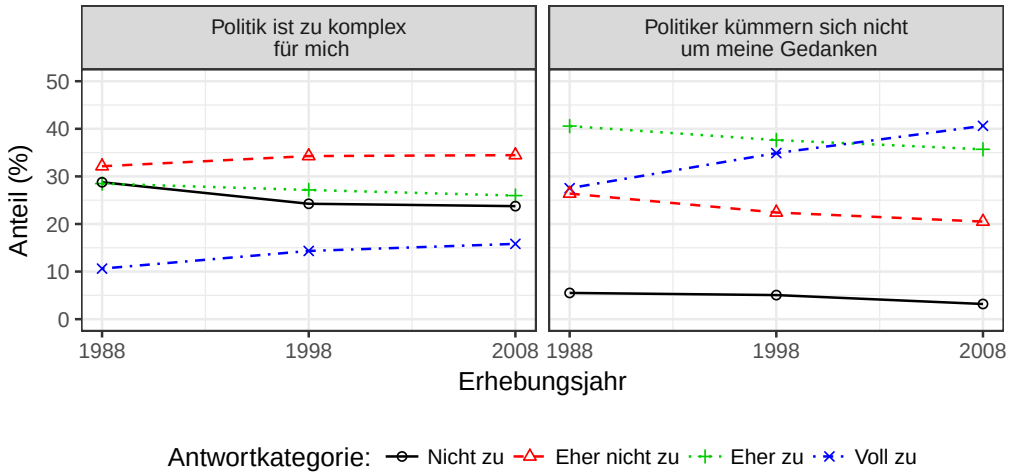
Durch den langen Zeitraum von 60 Jahren lassen sich Aussagen hinsichtlich der gesamtgesellschaftlichen Stabilität der Zu- und Ablehnungsraten treffen. Die einzige methodische Änderung der Abfrage ist die zusätzliche Aufnahme einer Mittelkategorie für die Kompetenzabfrage. Die Abfrage dieser durch „Politics is Too Complicated“ weist einen hohen Grad gesamtgesellschaftlicher Stabilität auf, ein konstant hoher Teil von über 60 % der amerikanischen Bevölkerung

[5]Es wurden zwei der vier aufgeführten Indikatoren ausgewählt. Das zentrale Kriterium ist die Vergleichbarkeit zu den nachfolgend dargestellten Formulierungen. Zur Übersicht der Indikatoren innerhalb der ANES siehe http://www.electionstudies.org/nesguide/gd-index.htm (Stand: 11. Mai 2018) Punkt 5.B.Efficacy.

stimmt dieser Aussage zu. „Public Officials Don't Care What People Think" als Operationalisierung der External Efficacy zeigt sich hingegen bereits volatiler. Lehnt anfänglich eine klare Mehrheit diese Aussage ab und attestiert den Repräsentanten des politischen Systems somit eine hohe Responsivität, findet sich seit Anfang der 1970er Jahre eine Mehrheit der Zustimmung zu dieser Aussage. Obwohl es vereinzelt zu starken Veränderungen zwischen zwei Erhebungen kommen kann,[6] zeichnet sich auch hier die Tendenz einer starken Stabilität über längerfristige Zeiträume ab.

Innerhalb der Erhebungen des Allbus finden sich seit 1988 in einem Abstand von 10 Jahren zur ANES vergleichbare Abfragen der beiden „Efficacy"-Dimensionen. Innerhalb der Bundesrepublik zeigt sich ebenfalls eine hohe Stabilität der Verteilung der Kompetenz des Verständnis von Politik, die Einstellung gegenüber den politischen Akteuren ist jedoch ebenfalls volatiler und durch eine zunehmende Skepsis geprägt. Die Zustimmung zu der Aussage „Politiker

Abbildung 7.3.: Entwicklung „Efficacy" im Allbus von 1988 bis 2008



Quelle: Eigene Darstellung. Daten: Kumulierter Allbus (ZA-Nr.: 4584, Gewicht: V1741). Variablen: V131 („Politiker kümmern sich nicht um meine Gedanken") und V134 („Politik ist zu komplex für mich"). Skalierung des Allbus hinsichtlich der Richtung angepasst an Skalierung ANES (vorher: 1: „Voll zu"). Angaben in Prozent (%).

kümmern sich nicht" zeigt einen klaren Trend des Anstiegs. Insgesamt lassen sich die jeweiligen Verteilungen und das Jahr der Erhebung jedoch nur in einen sehr schwachen Zusammenhang bringen.[7]

[6] Die größte Schwankung findet sich zwischen den Erhebungen 2000 und 2002 und dürfte als Ursache die Reaktion der Regierung auf den 11.September 2001 haben.

[7] Für beide Kontingenztabellen ergibt sich ein χ^2, welches statistisch signifikant von Null abweicht. Mit $V = 0.06$ („Politik ist zu komplex für mich") und $V = 0.08$ („Politiker kümmern sich nicht um meine Gedanken").

Die adäquate Messung der Efficacy steht immer wieder im Fokus der wissenschaftlichen Diskussion (Arzheimer 2016: 107) und die durch die ANES und den Allbus genutzten Operationalisierungen sind daher nicht die einzig denkbaren. Eine weitere Möglichkeit der Messung der beiden Dimensionen des Begriffs bieten die Skalen der „Political Efficacy Kurzskala“ (PEKS), welche zu einer „[...] reliablen und validen Erfassung von subjektiven politischen Kompetenz- und Einflusserwartungen in der psychologischen und politikwissenschaftlichen Forschung einsetzbar [ist].“ (Beierlein et al. 2012: 17). Die beiden Dimensionen des „Efficacy“-Begriffs werden durch jeweils zwei Aussagen erfasst.[8] Die Internal Efficacy wird durch „Wichtige politische Fragen kann ich gut verstehen und einschätzen.“ und „Ich traue mir zu, mich an einem Gespräch über politische Fragen aktiv zu beteiligen.“ operationalisiert. Die Messung der External Efficacy durch die beiden Aussagen „Die Politiker kümmern sich darum, was einfache Leute denken.“ und „Die Politiker bemühen sich um einen engen Kontakt zur Bevölkerung.“ (Beierlein et al. 2012: 8).

Diese Formulierungen der beiden Dimensionen der Efficacy unterstreichen die enge Verknüpfung zu weiteren prominenten Abfragen in Wahlumfragen. Eine empfundene Komplexität von Politik sollte mit einem geringeren Interesse an Politik und einer geringeren Wahrscheinlichkeit der Partizipation an politischen Abstimmungen auftreten. Wird die Responsivität der politischen Akteure angezweifelt, dürfte dies ebenfalls Auswirkungen auf die Bewertung des politischen Systems insgesamt haben und eine generell empfundene Unzufriedenheit mit der Situation der Demokratie ist wahrscheinlich.

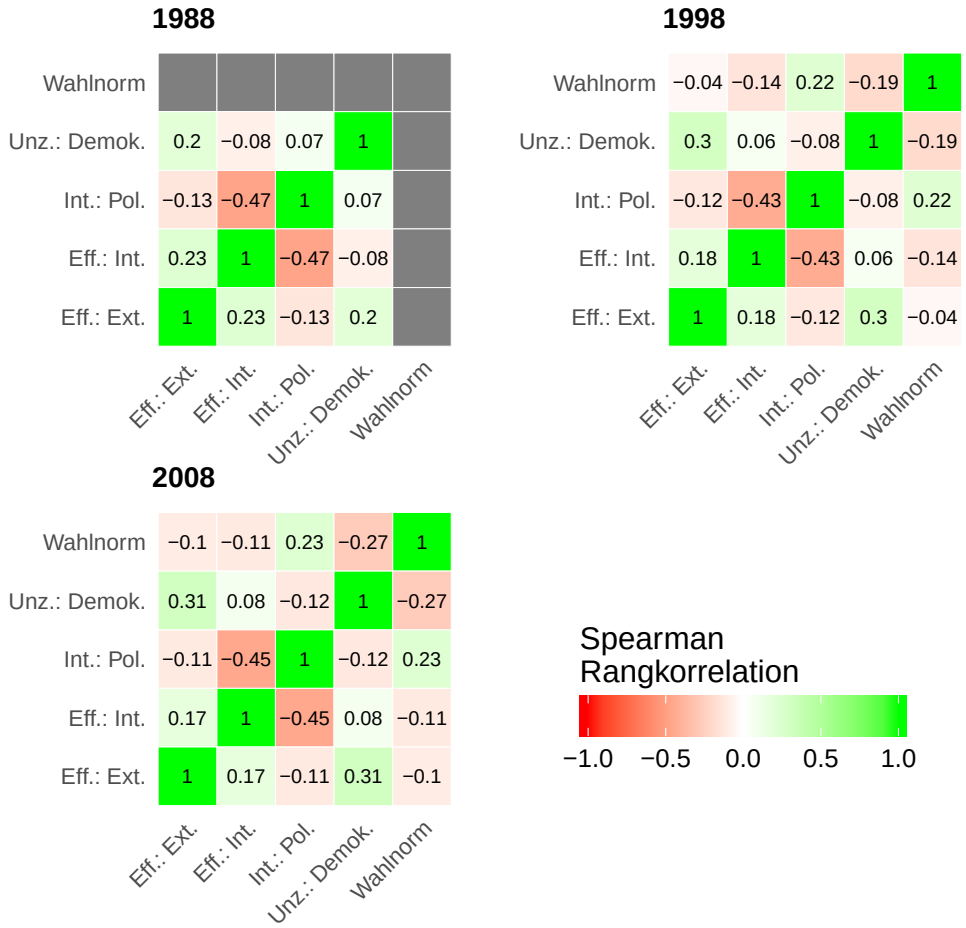
Die Abbildung 7.4 bestätigt dies. Ergänzt man die Variablen aus Abbildung 7.3 um das Interesse an Politik, die empfundene Wahlnorm und die Demokratiezufriedenheit, zeigen sich die vermuteten Zusammenhänge.[9]

Wird einer zu starken Komplexität von Politik zugestimmt, findet sich gleichzeitig ein eindeutig geringes politisches Interesse. Auch die Zustimmung des Wählens als Bürgerpflicht ist niedriger ausgeprägt. Eine Zustimmung zur man-

[8] Siehe hierzu auch die Beschreibung auf der Seite der „Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen“ (GESIS): http://www.gesis.org/kurzskalen-psychologischer-merkmale/kurzskalen/political-efficacy (Stand: 11. Mai 2018).

[9] Innerhalb des Allbus wurden die Personen zu ihrer Zustimmung zu den Aussagen „Kommen wir nun zu der Demokratie in Deutschland (1988: in der Bundesrepublik Deutschland): Wie zufrieden oder unzufrieden sind Sie - alles in allem - mit der Demokratie, so wie sie in Deutschland (1988: in der Bundesrepublik Deutschland) besteht?“ (1: „Sehr zufrieden“ und 6: „Sehr unzufrieden“) und „Wie stark interessieren Sie sich für Politik?“. Die letztere Aussage basierte 1988 auf einer zehnstufigen Likert-Skala (1: „Überhaupt nicht“ und 10: „Sehr stark“) und 1998 und 2008 auf einer fünfstufigen (1: „Sehr stark“ und 5: „Überhaupt nicht“). Für die Berechnung der Zusammenhänge wurde die Skala von 1988 hinsichtlich der Richtung (nicht der Kodierung) an die anderen beiden Durchführungen angepasst.

Abbildung 7.4.: Entwicklung Zusammenhänge Wahlnorm, Politisches Interesse, Efficacy und Demokratieunzufriedenheit 1988 bis 2008 (Allbus)



Quelle: Eigene Darstellung. Datenquelle: Kumulierter Allbus (ZA-Nr.: 4584, Gewicht: V1741). Efficacy-Variablen: Siehe Abbildung 7.3. Wahlnorm: V139, Skalierung angepasst (1: „Stimme gar nicht zu“). Demokratieunzufriedenheit: V22. Interesse Politik: V25 (1998,2008) und V26 (1988, Skalierung an V25 angepasst).

gelnden Responsivität der Akteure hingegen erhöht vor allem die generelle Unzufriedenheit mit der Situation der Demokratie. Zudem zeigt sich eine positive Verbindung zwischen den beiden Dimensionen der Efficacy. Verändern sich zwischen 1998 und 2008 diese Zusammenhänge nur geringfügig, lässt sich zwischen 1988 und 1998 ein stärkerer Unterschied festmachen. Eine mögliche Erklärung, gegenüber einer wirklichen Verschiebung der Zusammenhänge, könnte der zwischenzeitliche Beitritt der „Deutschen Demokratischen Republik“ (DDR) zum Geltungsbereich des Grundgesetzes sein.

Im Rahmen der Arbeit ist jedoch zentral, dass die hier aufgezeigten Verteilungen und Zusammenhänge zwischen den Kompetenzabfragen des Verständnisses und damit einhergehenden Interesse an Politik und die Einstellungsfragen gegenüber dem politischen System und seinen Akteuren ein hohes Maß an Sta-

bilität aufweisen. Zudem findet sich ein Bezug der beiden Dimensionen der Efficacy zu den weiteren aufgeführten und häufig verwendeten Abfragen in Wahlumfragen. Hieraus ergeben sich zwei weitere Blöcke möglicher Korrekturfaktoren, welche sich durch eine zeitliche Stabilität auszeichnen. Einerseits die intern empfundene Kompetenz von und das Interesse an Politik sowie die Beteiligungsabsicht. Andererseits die sich extern äußernden Einstellungen gegenüber den politischen Akteuren oder dem politischen System insgesamt. Sind diese Variablen mit einem systematischen Fehler in einer Umfrage erfasst *und* wirken sich auf das Antwortverhalten weiterer inhaltlich interessierender Variablen aus, findet sich auch dort ein Bias.

Ein Beispiel für eine solche weitere Zielvariable einer Umfrage sind die empfundenen Sympathien für verschiedene Parteien durch eine befragte Person. Die Persönlichkeit einer Person bestimmt dabei, wie externe Ereignisse in eine Veränderung der empfundenen Sympathie für eine Partei umgewandelt werden. Stehen diese ebenfalls in einer Verbindung zu den Kompetenz- und Einstellungsabfragen, beispielsweise durch einen unterschiedlichen Grad eine Protestpartei gegen „die etablierte Politik“ zu sein, äußert sich eine Verzerrung letzterer auch in einer fehlerhaften Erfassung der wahrgenommenen. Die Korrektur der Kompetenz- und Einstellungsabfragen über Gewichtungsverfahren kann diese Verzerrungen dann jedoch beseitigen. Da Referenzverteilungen für diese Variablen nicht über die amtliche Statistik zur Verfügung stehen, ist die Korrektur durch die Angleichung einer Umfrage an eine Referenzumfrage über ein PSM prädestiniert. Das nachfolgende Kapitel wird nun die angegebenen Sympathiegrade für verschiedene politische Parteien der Bundesrepublik ebenfalls hinsichtlich ihrer Stabilität analysieren. Sind diese zu volatil, deutet dies auf eine klare Dominanz kurzfristiger Faktoren hin und eine zentrale Bedeutung dieser für die Korrektur. In diesem Fall wären die aufgeführten Variablen nicht geeignet, um eine konditionale Repräsentativität für eine vorliegende Umfrage herzustellen.

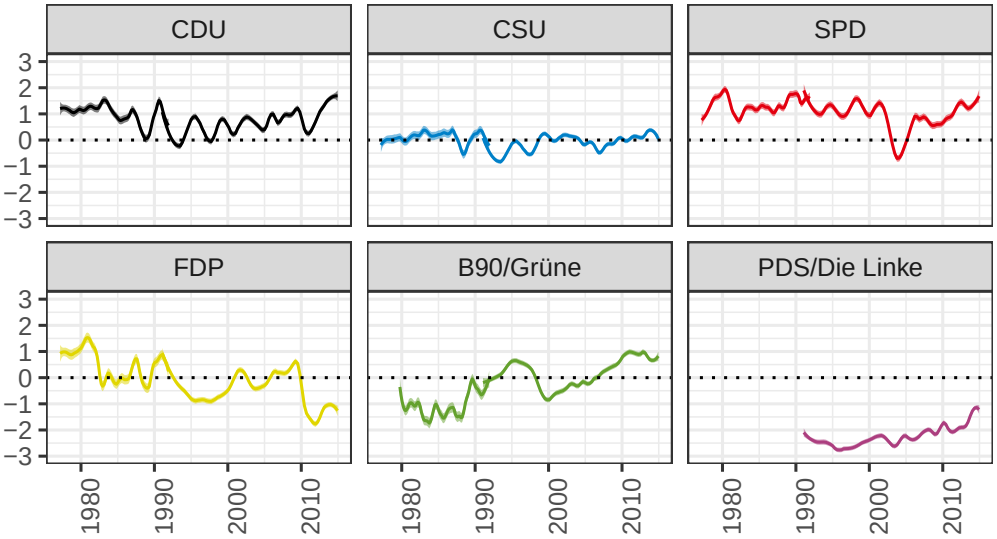
7.3. Inhaltliche Zielvariablen von Wahlumfragen

Nach dem Ansatz von Fishbein (1963) führt die Verknüpfung von gewichteten Merkmalseigenschaften eines Objektes zu einer insgesamten Bewertung dieses Objekts. Veränderungen der Merkmalseigenschaften führen wiederum zu Veränderungen der Bewertung. Für eine Partei sind beispielsweise Merkmale die mit ihr assoziierten Politiker oder ihre Rolle innerhalb des politischen Systems

als Oppositions- oder Regierungspartei. Kommt der Persönlichkeit eine moderierende Funktion zu, haben sich wandelnde Einstellungen gegenüber Politikern und dem politischen System einen direkteren Einfluss auf die Sympathiebewertungen von Parteien.

Wie die Abbildung 7.5 darstellt, sind die Parteibewertungen bereits weniger konstant als die bisher aufgeführten Variablen. Insbesondere für die FDP und Die Grünen lassen sich Phasen unterschiedlicher gesamtgesellschaftlicher Zustimmungsraten herausstellen.

Abbildung 7.5.: Parteiensympathie Skalometer 1977 bis 2014 (Politbarometer)



Quelle: Eigene Darstellung. Datengrundlage: Kumulierte Jahresstichproben Politbarometer. Ab 1991 die neuen Bundesländer berücksichtigt und hinsichtlich der Ungleichgewichte auf Basis der nachfolgenden Darstellung in Kapitel 5.2 korrigiert. Dargestellt ist die LOESS-Schätzung des Verlaufs der arithmetischen Mittel aller verfügbaren Stichproben seit 1977 ($n = 521$, Bandbreite 0.15).Skala: -5 („Unsympathisch“) bis +5 („Sympathisch“)

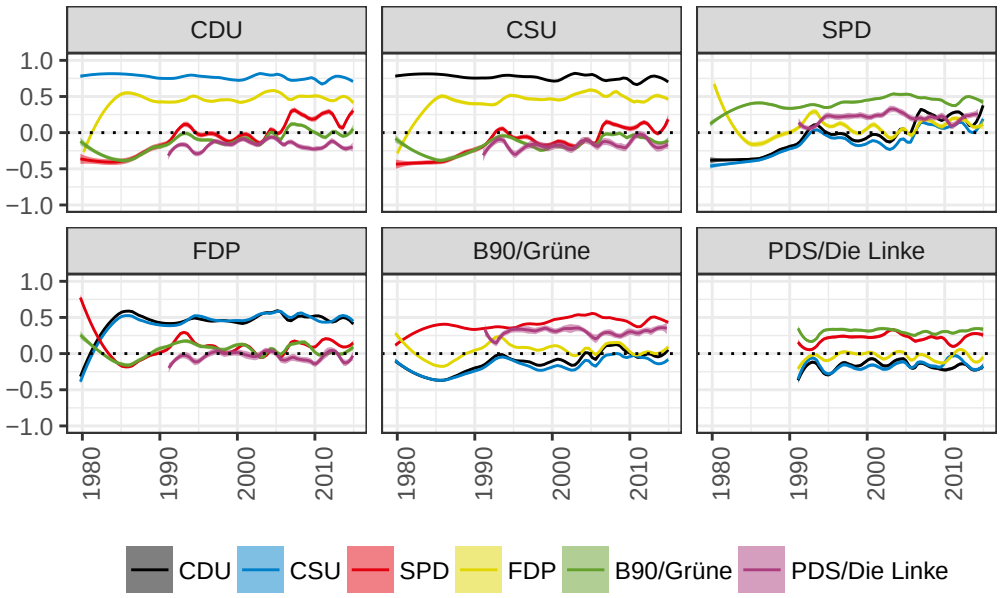
Zudem zeigt sich ein Einfluss zeitlich abgrenzbarer politischer Ereignisse: Das Misstrauensvotum gegen Helmut Schmidt am 01.10.1982 und das erstmalige Unterschreiten der 5 %-Hürde durch die FDP bei der Wahl 2009 führten zu einem jeweiligen langfristigen Sympathieeinbruch für diese. Für „Die Grünen“ findet sich hingegen, nach einem vorher stattfindenden Einbruch, ein seit den 2000er Jahren einsetzender Sympathiegewinn. Auch die SPD konnte die durch den Beschluss der „Agenda 2010“ bedingten Einbruch Anfang der 2000er-Jahre entgegnen. Für die CSU, CDU und „Die Linke“ zeichnet sich hingegen kein langfristiger Trend ab, auch wenn die letzten beiden eine zunehmende

Verbesserung in der Legislaturperiode zwischen 2009 und 2013 aufweisen.[10]

Eine noch viel stärkere Stabilität weist die politische Lagerbildung auf. Die Abbildung 7.6 führt die paarweisen Korrelationen der angegebenen Sympathie

Sympathien und damit verbundene Einstellungen gegenüber Parteien als Objekte sind, nicht zuletzt wegen ihres letztendlichen Einflusses auf die Wahlentscheidung bei einer Wahl, eine zentrale Zielvariable von Wahlumfragen dar. Sie können eine langfristige Stabilität des Niveaus aufweisen, auch wenn sie kurzfristig starken Schwankungen unterworfen sind. Da sie eine Folge der bisher aufgeführten anderen Variablen sind, wird der empirische Teil diese zur Evaluation des Erfolgs einer Korrektur durch Gewichtungsverfahren und nicht als Teil eines Korrekturmodells nutzen.

Abbildung 7.6.: Zusammenhänge (r_{xy}) Parteiensympathie Skalometer 1977 bis 2014 (Politbarometer)



Quelle: Eigene Darstellung. Dargestellt ist der Verlauf der LOESS-Schätzung (Bandbreite 0.15) der paarweisen Zusammenhänge gemäß Pearsons r zwischen den Parteien, ausgehend von der jeweiligen Partei. Daten/Beschreibung/Skala: Siehe Abbildung 7.5.

Es zeigen sich Phasen konstanter Zusammenhänge, als auch Phasen der Verschiebung dieser. CDU und CSU als Unionsparteien vereint eine sehr stabile Beziehung, Pearsons r weist einen konstanten Wert von ca. 0.8 auf. Seit dem Ende der sozialliberalen Koalition empfinden Unions-Sympathisanten zudem auch

[10] Dies bestätigt ebenfalls ein durchgeführter „Augmented Dickey-Fuller"-Test auf Stationarität der Zeitreihen mit einem Verzögerung von $k = (n-1)^{\frac{1}{3}}$, welche über alle Parteien hinweg $k = 7$ entspricht. Die Alternativhypothese der Stationarität lässt sich für die CDU ($p = 0.01$), CSU ($p = 0.01$) und Die Linke ($p = 0.016$) für $\alpha = 0.05$ annehmen.

eine starke Sympathie für die FDP als Partei. Präferierten die Anhänger der SPD über einen langen Zeitraum bis Mitte der 1990er Jahre ebenfalls nur „B90 / Die Grünen“, dreht sich dieses Bild seitdem. Nach einem bis dahin negativen Bild der Unionsparteien und der FDP, kommt es hier zu einem Zeitraum ohne einen relevanten Zusammenhang. Anschließend beginnt eine hohe Sympathie für die SPD mit einer ebenfalls erhöhten Sympathie für diese Parteien einherzugehen.

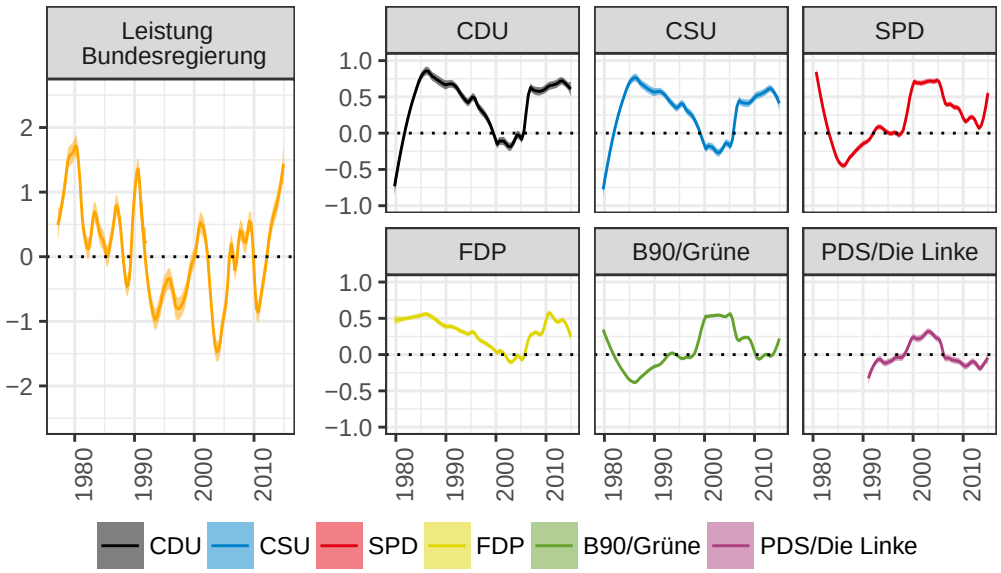
Der Vorsprung von „B90/Die Grünen“ als Sympathieträger derjenigen Personen, welche auch die SPD favorisieren, ist seit Mitte der 2000er Jahre (und der ersten Legislaturperiode der Großen Koaliation) geringer geworden. Hinsichtlich der „PDS/Die Linke“ zeigt sich keine relevante Verschiebung der Zusammenhänge der Sympathie zu dieser Partei mit allen weiteren aufgeführten. Personen mit einer Sympathie für diese weisen ebenfalls eine Sympathie für die SPD und „B90/Die Grünen“ auf. Gegenüber den Unionsparteien und der FDP kommt es hingegen eher zu einer geringeren Sympathie.

Auch die Bewertung der Leistung der Bundesregierung in Abbildung 7.7 unterliegt starken Veränderungen der im arithmetischen Mittel gemessenen Bewertung durch das Elektorat. Es finden sich sowohl Phasen starker Euphorie, als auch klarer negativer Bewertungen. Insgesamt zeigt sich jedoch keine Niveauverschiebung über den gesamten betrachteten Zeitraum, welche eine dauerhaft positivere oder negativere Haltung gegenüber der Bundesregierung als Bezugsobjekt von Einstellungen erlaubt.[11] Der zweite Teil der Abbildung 7.7 zeigt für jede Erhebung des Politbarometers seit 1978 die Entwicklung der Zusammenhänge zwischen den angegebenen Sympathien der jeweiligen Parteien und der Bewertung der Leistung der Bundesregierung. Dies lässt einen Schluss zu auf die Entwicklung der langfristigen Zusammenhänge zwischen diesen beiden möglichen Bezugsobjekten, auf welche Einstellungen einen Einfluss haben können. Diese Zusammenhänge sind äußerst stabil, sofern es nicht zu einerm Regierungswechsel nach einer Wahl kommt. In diesem Fall lässt sich eine abrupte Veränderung der Bewertung feststellen.[12] Dies zeigt: Die Bewertung der Regierungsleistung geht in einem sehr hohen Maß mit der Sympathie für die jeweilige Regierungspartei(en) einher. Es lässt sich ein klarer Effekt der Regierungswechsel 1982, 1998 und 2005 ausmachen. Der jeweilige Moment führt zu

[11] Auf der Grundlage eines „Augmented Dickey-Fuller Test“ kann mit einer Verzögerung von $k = 7$ die Alternativhypothese einer stationären Zeitreihe mit $p = 0.01$ für $\alpha = 0.05$ angenommen werden.

[12] Beispielsweise hat weder der Einsatz im Kosovo noch der Beschluss der Agenda 2010 innerhalb der Regierunszeit der rot-grünen Koalition unter Gerhard Schröder einen Effekt auf die Stärke des Zusammenhangs der Sympathie für die SPD und „B90/Die Grünen“ und die Bewertung der Regierungsleistung hat.

Abbildung 7.7.: Regierungsleistung ($\bar{x}$) und Zusammenhänge (r_{xy}) Parteiensympathie Skalometer 1977 bis 2014 (Politbarometer)



Quelle: Eigene Darstellung. Dargestellt ist der Verlauf der LOESS-Schätzung (Bandbreite 0.15) des arithmetischen Mittel der Leistung Bundesregierung und die paarweisen Zusammenhänge gemäß Pearsons *r* der jeweiligen Parteiensympathie zu dieser. Daten/Beschreibung/Skala Sympathie: Siehe Abbildung 7.5. Skala Leistung Bundesregierung: -5 („Zufrieden“) bis +5 („Unzufrieden“)

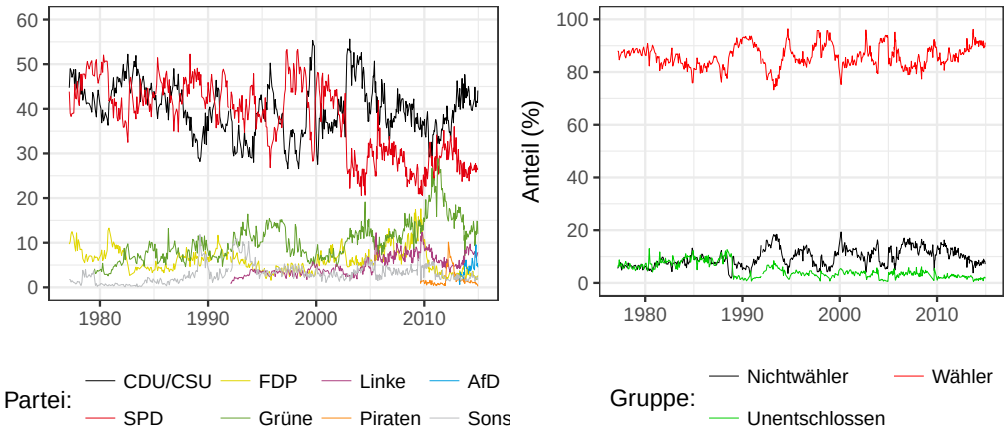
einer sofort auftretenden positiven Korrelation der Sympathie für die jeweilige neue Regierungspartei und der Bewertung der Leistung der Regierung insgesamt. Gleichzeitig ist der positive Zusammenhang zwischen der Sympathie für die abgelöste Regierungspartei und die Sympathie für diese nicht mehr existent, oder führt nun sogar zu einer negativen Assozioation, wie nach der Ablösung des Kabinett Schmidt III am 01.10.1982. Der Wechsel des Kabinett Schröder II auf die von Angela Merkel geführte Große Koalition am 22.11.2005 sorgt auch für diesen Effekt, jedoch durch die weiterhin vorhandene Regierungsbeteiligung in einer nur abgeschwächten Form. Auch hinsichtlich der Sympathisanten von „B90/Die Grünen“ verbleibt ein positiver Effekt in dieser Zeit.

Im Rahmen der Arbeit führt dies zu einem wichtigen Schluss: Die Bewertung der Leistung der Bundesregierung variiert kurzfristig. Dies berührt jedoch nicht die starke Verknüpfung mit der Sympathie für die jeweiligen Regierungs- und Oppositionsparteien. Nach dem Ansatz von Fishbein (1963) stellt die Bundesregierung ein Bezugsobjekt dar. Die jeweiligen Regierungsparteien sind als Merkmalsausprägungen mit dieser verknüpft. Ändert sich die Bewertung einer Partei, ändert sich daher auch die Bewertung der Regierungsleistung insgesamt. Verzerrungen, welche die Sympathie einer Regierungspartei betreffen, wirken

sich daher auch auf die Zufriedenheit der Leistung der Bundesregierung auf. Jede Regierungsphase weist daher Phasen höherer Zustimmung oder Ablehnung innerhalb des *gesamten* Elektorats auf. Das Niveau der Bewertung ist jedoch in einem sehr hohen Ausmaß durch die Sympathie für die jeweiligen Regierungsparteien bestimmt.

Ebenfalls, neben weiteren Faktoren, ist die Wahlentscheidung eine Folge der Parteiensympathien. Die Abbildung 7.8 zeigt die Entwicklung der Zweitstimmenanteile der Politbarometer-Stichproben.

Abbildung 7.8.: Zweitstimmenpräferenz und Wahlbeteiligung 1977 bis 2014 (Politbarometer)



Quelle: Eigene Darstellung. atengrundlage: Kumulierte Jahresstichproben Politbarometer. Ab 1991 die neuen Bundesländer berücksichtigt und hinsichtlich der Ungleichgewichte auf Basis der nachfolgenden Darstellung in Kapitel 5.2 korrigiert.

Im Gegensatz zur sich erholenden durchschnittlichen Sympathie für die SPD nach dem Abfall durch die Einführung der Agenda 2010 (Abbildung 7.5), zeigt sich keine Erholung der Zweitstimmenanteile. Diese verharren auf einem niedrigeren Niveau. Ebenfalls sind, bereits in kürzeren Phasen zwischen zwei Bundestagswahlen, erkennbare Entwicklungen der präferierten Zweistimmenanteile bemerkbar. Entgegen dem Wahl*verhalten*, äußert sich dies nicht hinsichtlich der angegebenen Wahl*beteiligungsabsicht*. Es zeigen sich im Zeitverlauf kaum relevante Veränderungen des Anteils der befragten Personen, welche sich an der kommenden Wahl beteiligen würden. Einzig der Sprung ab Mitte 1988 ist erkennbar. Der Anteil der „Unentschlossenen“ nimmt stark ab. Dies kann jedoch eine zu der Zeit stattfindende Folge des Wechsel des Erhebungsmodus von persönlichen zu telefonischen Befragungen innerhalb des Politbarometers

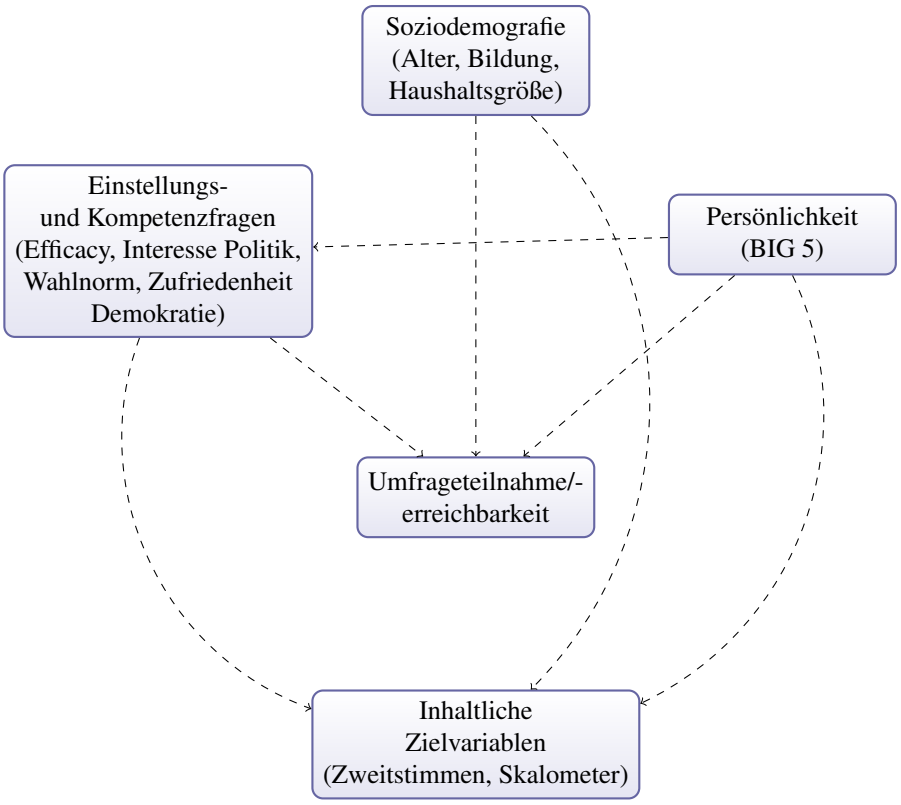
darstellen.[13] Seit 2010 ist ebenfalls ein Trend der steigenden Anteile an Wählern erkennbar, was vor dem Hintergrund der sinkenden Wahlbeteiligung bei Bundestagswahlen 2009 bemerkenswert ist. Das Kapitel 9 wird dies jedoch auf eine zunehmende selektive Teilnahme an telefonischen Wahlumfragen zurückführen können.

7.4. Ein Gesamtmodell der Korrektur

An dieser Stelle hat die Arbeit die relevanten „Informationen“ in Form von möglichen Variablen für den kommenden empirischen Teil vorbereitet, welche zur Untersuchung von Unterschieden zwischen Umfragen und darauf aufbauenden Korrekturmodellen genutzt werden können. Dies resultiert in einem *Gesamtmodell der Korrektur*, welches in Abbildung 7.9 dargestellt ist. Auf dieser Basis lassen sich die skizzierten und verstandenen Prozess der Umfrageteilnahme und aufzulösender Verzerrungen in Wahlumfragen auflösen. Dies betrifft sowohl das Wahl*verhalten* als auch die Wahl*beteiligungsabsicht*. Die *Soziodemografie* wurde in Form des Alters und des Bildungsstands herausgearbeitet. Diese beiden Variablen erklären sowohl den Zugangsprozess zu einer Umfrage, als auch Unterschiede hinsichtlich möglicher Zielvariablen empirischer Wahlstudien . Ihr Potential als „Globalvariablen“ zur Erklärung individuellen Verhaltens lässt sich jedoch als gering einschätzen. Ebenfalls wurde die *Persönlichkeit* einer Person als möglicher Korrekturfaktor herausgestellt. Die gängige Operationalisierung dieser in Form der „BIG Five“ bietet eine langfristig stabile Information über verschiedene Personentypen, welche einen Einfluss auf die Entscheidung zur Teilnahme an Umfragen hat. Auch hat die Ausprägung der Persönlichkeit einen Einfluss auf die empfundene Parteiensympathie, wodurch ebenfalls das Wahlverhalten differieren kann. Dass es zu diesem Einfluss kommt, lässt sich auf den Einfluss der Kombination der Persönlichkeitsmerkmale auf die Steuerung der Verarbeitung externer Informationen zur Umwandlung von Einstellungen zurückführen. Hierdurch kommt es zur Veränderung von *Einstellungen gegenüber Akteuren des politischen Systems*. Neben den sich kurzfristig ändernden affektiven Einstellungen gegenüber spezifischen politischen Parteien als Objekte, betrifft dies vor allem auch die Einstellung gegenüber dem politischen System insgesamt in Form der *Bewertung der Zufriedenheit mit der Demokratie* zu. Daher lässt sich die Zufriedenheit mit der Situation der Demokratie und die

[13] Zur Beschreibung des Zeitpunkts des Wechsels siehe https://dbk.gesis.org/dbksearch/SDESC2.asp?no=2391&search=Politbarometer&search2=&DB=D (Stand: 11. Mai 2018).

Abbildung 7.9.: Visualisierung Gesamtmodell Korrektur



Quelle: Eigene Darstellung

„External Efficacy“ ebenfalls als eine mögliche Information zur Korrektur von Verzerrungen in Wahlumfragen nutzen. Neben diesen gerichteten Einstellungen wurden ebenfalls *Kompetenzeinschätzungen* als Faktor herausgearbeitet. Das Verständnis von und das Interesse an Politik erhöht die Wahrscheinlichkeit der Teilnahme an einer Wahlumfrage. Da es ebenso auch einen Einfluss auf die Entscheidung zur Partizipation am politischen Prozess in Form der Wahlbeteiligung hat, ist die Berücksichtigung des „Interesse an Politik“ und der „Internal Efficacy“ als langfrisitig und gesamtgesellschaftlich stabile Größen naheliegend. Dies gilt ebenfalls für die Empfindung der Beteiligung an Wahlen als eine soziale Norm, welche sich mit der Kompetenzeinschätzung des Verständnis und des Interesse an Politik verbunden gezeigt hat.

Nachdem nun geklärt ist, welche Art von Informationen in Wahlumfragen generell für eine Korrektur prädestiniert ist, wird der nachfolgende Teil der Arbeit nun die Erhebungslandschaft der Umfrageforschung darstellen. Dies dient dazu die maßgeblich verwendeten Formen der persönlichen, telefonischen und onlinebasierten Befragung hinsichtlich ihres angewandten Erhebungsdesigns zu beleuchten und hierdurch ihre qualitative Einordnung vorzunehmen.

Teil II.

Empirie: Die Qualität von Wahlumfragen

8. Der Vergleich zur amtlichen Statistik

Um die Qualität onlinebasierter Befragungen zu steigern, ist ihre Angleichung an eine als qualitativ hochwertiger angesehene Referenzerhebung notwendig. Das Kapitel 4 hat hierzu bereits herausgestellt, dass der Erfolg statistischer Matchingverfahren im Rahmen der „Propensity Scores“ (PS)-Modellierung zentral von der Qualität der gewählten Referenzquelle abhängt. Sowohl persönliche als auch telefonische Umfragen lassen eine zufallsbasierte Auswahl zu, welche sich bei Onlineumfragen nur in einem sehr begrenzten Rahmen realisieren lässt. Access-Panels können dieses Problem theoretisch lösen, die Realisierung dieser wird diesen Ansprüchen jedoch nur selten gerecht (siehe hierzu das Kapitel 6).

Ein zufallsgesteuertes Stichprobenverfahren bezieht sich jedoch nur auf die Garantie einer identischen (oder bekannten und korrigierbaren) Wahrscheinlichkeit der *Auswahl* einer Person, die Entscheidung zur *Teilnahme* liegt hingegen vollständig außerhalb der Kontrolle durch das Erhebungsverfahren. Ist die Verweigerungsquote einer relevanten Gruppe systematisch höher, oder das Erhebungsdesign erschwert die Teilnahme signifikant, ist auch die Wahrscheinlichkeit einer Verzerrung der erfassten Umfrageergebnisse hoch.

Das Kapitel 6.1 hat einen komparativen Vorteil der persönlichen gegenüber der telefonischen Befragung herausgearbeitet, trotz der stärkeren Anfälligkeit für einen Design-Effekt (Kapitel 6.1.3). Diese Befragungsform weist in der Regel höhere Ausschöpfungsquoten und somit eine geringere Anfälligkeit für einen Nonresponsefehler auf. Teilnahmeraten von bestenfalls 20 % lassen in telefonischen Befragungen hingegen Zweifel an einer zufallsbasierten *Teilnahme* aufkommen, welches durch zunehmende Probleme hinsichtlich des Coverage der jüngeren Bevölkerung verstärkt wird.

Onlinebasierte Befragungen hingegen haben trotz ihrer Vorteile (Kapitel 6.2), auch weiterhin mit starken komparativen Nachteilen gegenüber der persönlichen *und* telefonischen Befragung zu kämpfen. Das Kapitel 6.2.1 hat bereits den kaum zu vermeidenden Noncoveragefehler durch die ungleiche Verteilung des Internetzugangs herausgestellt. Auch ist die Häufigkeit der Nutzung nicht unabhängig von zentralen Merkmalen der ausgewählten Personen. Ein qua Design kaum zu vermeidender kombinierter Nonresponse- und Noncoveragebias ist das Resultat.

Die Verwendung einer zufallsbasierten Auswahlgrundlage kann das Ausmaß dieser beiden Fehlerquellen abmildern (Kapitel 6.2.2). Die Zusammensetzung der für die Arbeit zentralen Access-Panel von Respondi und LINK zeigen jedoch bereits, dass diese lediglich eine selektive Abbildung der Bevölkerung darstellen (Kapitels 5.4).

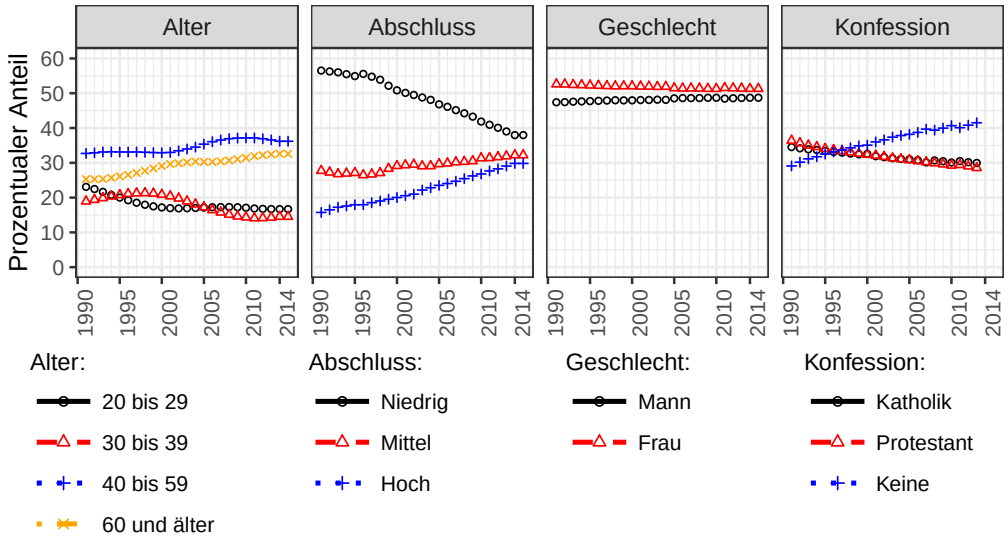
Das nachfolgende Kapitel wird die Frage geeigneter Referenzumfragen beantworten: Durch welche Befragungsform ist eine adäquate Steigerung der Qualität onlinebasierter Verfahren möglich? Die Bedeutung soziodemografischer Rahmendaten sollte hinsichtlich ihrer Bedeutung für diese Korrektur nicht überschätzt werden (Kapitel 7.2.1). Ihre Analyse liefert jedoch einen guten ersten Ansatzpunkt, ob eine Befragungsform einen zunehmend selektiven Kreis der Bevölkerung befragt. Ebenfalls sind für diese amtliche Referenzdaten verfügbar, welche einen guten Vergleich durch ihren hohen Qualitätsstandard ermöglichen.

Nachfolgend werden die in Kapitel 5 vorgestellten Umfragen hinsichtlich ihrer Abweichungen von den Referenzdaten der amtlichen Statistik in Form des „Mikrozensus“ (MZ) (Kapitel 5.4) untersucht. Hierdurch wird die qualitativ aufgestellte Rangfolge der Qualität der Erhebungsmethoden in Kapitel 6.3 empirisch belegt. In einem ersten Schritt wird hierzu die Entwicklung der zufallsbasierten Erhebungen des Allbus, ESS, Politbarometers und Forsa seit 1991 in Kapitel 8.1 vorgenommen. Anschließend werden die seit 2009 durchgeführten onlinebasierten Erhebungen der Komponente 8 der „German Longitudinal Election Study“ (GLES) in Kapitel 8.2 untersucht. Abschließend kann dann in Kapitel 8.3 die Entscheidung eines „Soziodemografischen Goldstandards“ der Datenerhebung getroffen werden. Diese Befragungsform kann dann als Referenz zur Nutzung von PS-Korrekturen genutzt werden (siehe hierzu auch das Kapitel 4.3). Dies impliziert die zu treffende Annahme, dass diejenigen Umfragen, welche soziodemografische Informationen besser erfassen können, dies auch tendenziell für weitere inhaltliche Variablen *eher* gewährleisten können.

8.1. Zufallsbasierte Bevölkerungs- und Wahlumfragen seit 1991

Bevölkerungsumfragen versuchen auf der Basis einer ausgewählten Stichprobe ein repräsentatives Abbild der Bevölkerung zu erlangen. Durch einen Vergleich mit der nachgezeichneten Bevölkerungsentwicklung der Bundesrepublik durch die amtliche Statistik lässt sich dieser Anspruch validieren. Die Abbildung 8.1 zeigt die Entwicklung der volljährigen Bevölkerung der Bundesrepublik seit dem Jahr 1991 auf der Basis der Verteilungen des Mikrozensus und des durchgeführten Zensus in 2011.[1] Wenig überraschende gesellschaftliche Trends lassen sich identifizieren. Die (volljährige) Bevölkerung der Bundesrepublik altert: Insbesondere der Anteil der Personen ab 60 Jahren hat innerhalb von ca. 20 Jahren von einem Viertel auf ein Drittel der Bevölkerung zugenommen, wohingegen die jüngeren Altersgruppen an relativer Größe verlieren.[2]

Abbildung 8.1.: Entwicklung Alter, Abschluss, Geschlecht und Konfessionszugehörigkeit 1990 bis 2014 (Mikrozensus)



Quelle: Eigene Darstellung. Daten: GENESIS Online-Datenbank von Destatis, Auswahlgrundlage volljährige Bevölkerung. Für die Jahre einer „Bundestagswahl“ (BTW) wurden die Daten der Repräsentativstatistik des Bundeswahlleiters genutzt, Auswahlgrundlage wahlberechtigte Bevölkerung.

[1] Das Anfangsjahr der Darstellung wurde durch die Verfügbarkeit der zu vergleichenden Umfragen gewählt. Da ab 1991 erst die Erhebungen des Forsa-Bus beginnen, wurde dieses Jahr gewählt. Der kumulierte Politbarometer für Westdeutschland wurde somit, ebenso wie die Erhebungen des ALLBUS vor 1991, nicht berücksichtigt.

[2] Für alle nachfolgenden Analysen wird diese Einteilung der Altersklassen die Grundlage bilden, um eine Vergleichbarkeit zur amtlichen Statistik zu garantieren.

Gleichzeitig machen sich die Effekte einer Bildungsexpansion bemerkbar: Der Anteil der erwachsenen Bevölkerung mit einem erlangten (Fach-)Abitur (Bildung: Hoch) hat sich im angegebenen Zeitraum von knapp unter 20 % auf ca. 30 % gesteigert. Der Anteil der Personen mit keinem Abschluss oder einem Hauptschulabschluss (Bildung: Niedrig) ist hingegen von 55 % auf unter 40 % gesunken. Einzig die Gruppe der Personen mit einem Realschulabschluss (Bildung: Mittel) erweist sich als sehr konstant mit einem Anteil von ca. 30 %. Diese Klassifizierung des Bildungsniveaus wird den gesamten empirischen Teil über verfolgt.[3] Gleichzeitig nimmt innerhalb der Bevölkerung der Anteil von Katholiken und Protestanten ab: Von einem ausgeglichenen Verhältnis in 1994 überwiegt der Anteil der Personen mit keiner oder einer anderen Konfession mit über 40 % eindeutig.[4] Das Verhältnis von Männern zu Frauen hat sich hingegen kaum verändert. Frauen überwiegen prozentual gesehen geringfügig, was auf ihre erhöhte Lebenserwartung zurückzuführen sein dürfte. Diese skizzierten Entwicklungen stellen durch ihre Basis der jährlichen und qualitativ hochwertigen Mikrozensus-Erhebungen, sowie ihre Aktualisierungen durch Hochrechnungen auf Basis des Zensus 2011, die bestmögliche Annäherung an die wahren Anteile dieser Gruppen in der Bundesrepublik dar (zur Beschreibung siehe Kapitel 5.4). Sie können somit als Referenzdaten für die innerhalb der Arbeit verwendeten Umfragen herangezogen werden.

Wie gut schaffen es nun die in Kapitel 5 vorgestellten Datensätze, diese Entwicklungen widerzuspiegeln? Abbildung 8.2 zeigt die in Kapitel 3.4 vorgestellte Verzerrungskennzahl A_i von Arzheimer/Evans (2014) für die Stichproben des „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (Allbus) seit 1991 und des „European Social Survey“ (ESS) seit 2002.[5] Die gewählten Referenzverteilungen stellen die Datensätze der amtlichen Statistik des jeweiligen Jahres dar.

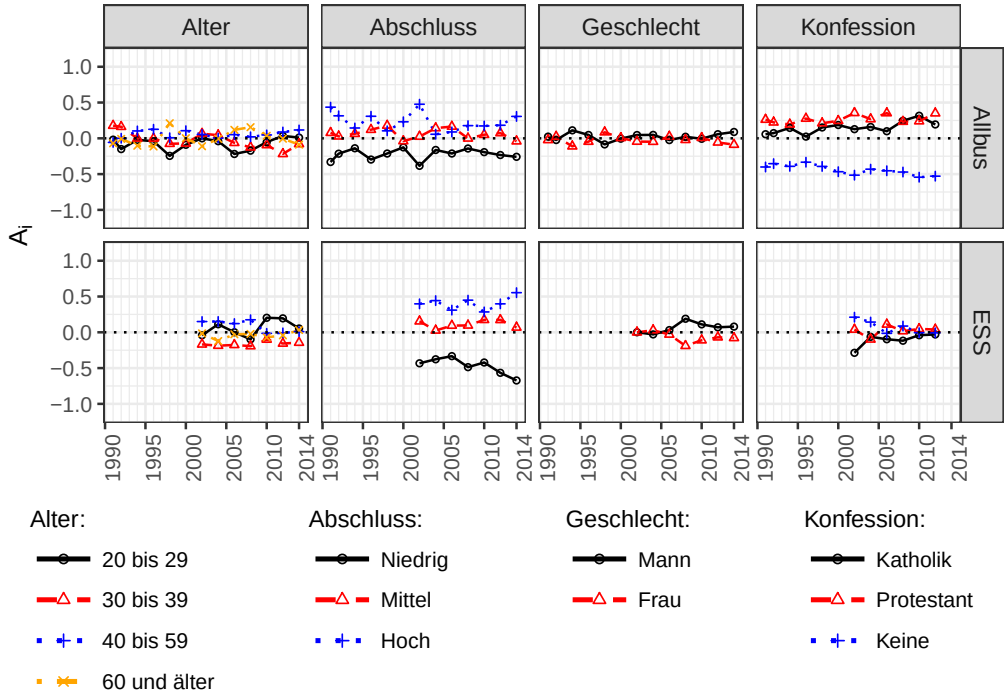
Der Verlauf der zweijährlichen Erhebung des Allbus zeigt eine insgesamt hohe Qualität und Stabilität der Erfassung des Niveaus und der Entwicklungen des Bildungsstands, der Anteile der Altersgruppen und des Geschlechts der volljährigen Bevölkerung. Es finden sich kaum relevante Abweichungen von den amtlichen Referenzverteilungen des jeweiligen Jahres ($A_i = 0$), auch wenn auf einem

[3] An diese Einteilung wurden nachfolgend alle Datensätze angepasst. Ein „anderer Schulabschluss“ und „bin noch Schüler“ führten zu einem Ausschluss dieser Person.

[4] Ursächlich hierfür ist eine starke Zunahme der Kategorie „Keine“. Die Kategorie „Sonstige“ Konfession steigert sich hingegen nur gering.

[5] Als Alternative hatte das Kapitel 3.4 auch den „Relative Bias“ (RB) herausgestellt. Dieser wurde ebenfalls berechnet. Da für A_i und RB eine Korrelation von $r = 0.97$ gilt, wird aus Platzgründen an dieser Stelle nur eine der beiden Kennzahlen ausgewiesen.

Abbildung 8.2.: Verzerrung Allbus und ESS in Referenz zum Mikrozensus



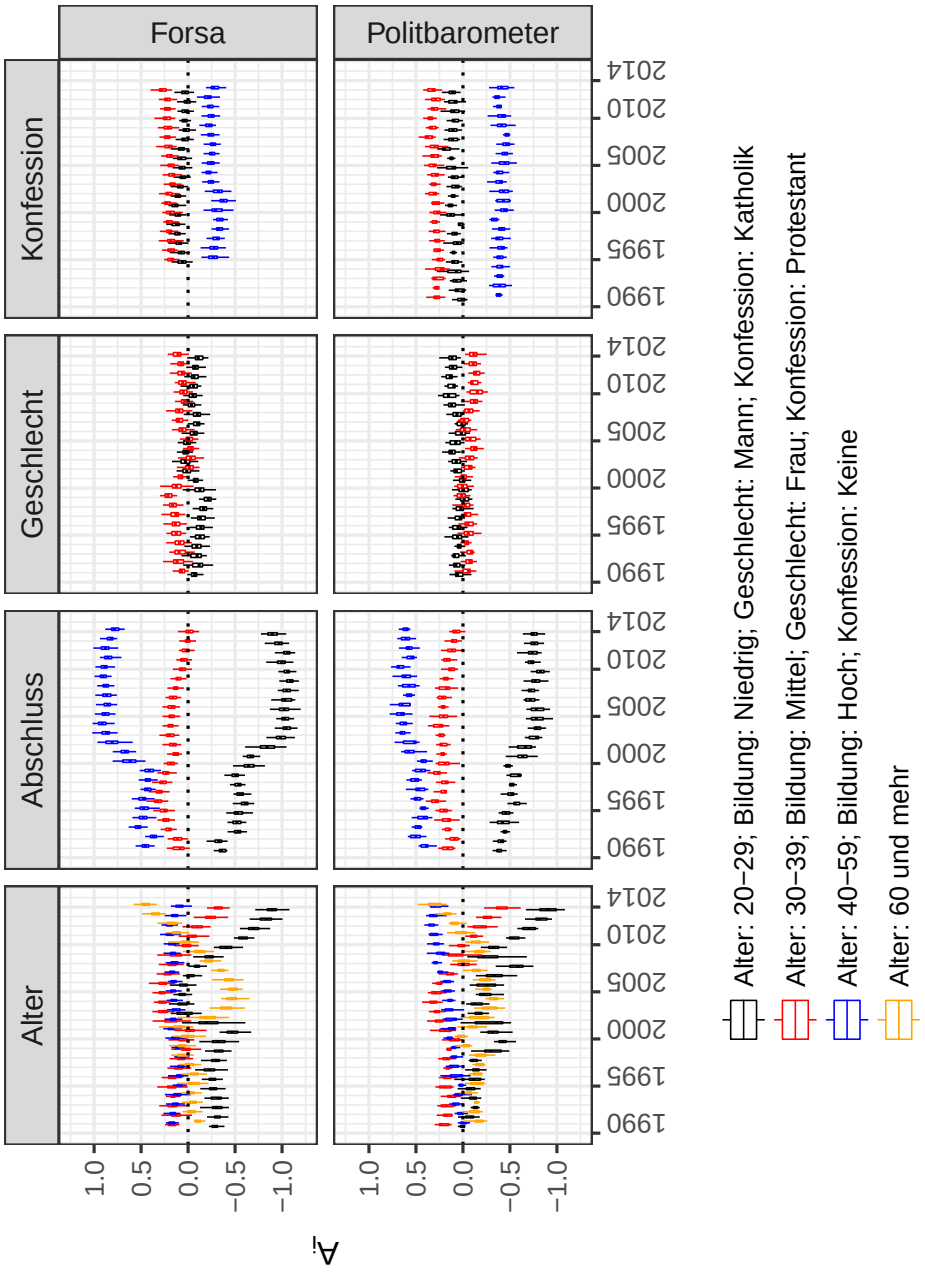
Quelle: Eigene Darstellung. Beide Datensätze wurden hinsichtlich der Ost-West-Ungleichgewichte durch die bereitgestellten Design-Gewichte korrigiert.

geringen Niveau die Gruppe der Personen mit einem niedrigen Bildungsstand („Abschluss: Niedrig“) unterrepräsentiert ist. Lediglich die korrekte Erfassung des konfessionellen Hintergrunds bereitet seit Beginn der genutzten Erhebungen Probleme, welche eine systematische Verzerrung erkennen lassen. Auch der ESS schafft die Abbildung der Altersstruktur der volljährigen Bevölkerung ohne erkennbare Verzerrungen. Stärkere Probleme zeigen sich hingegen bzgl. des Bildungsniveaus: Die Gruppe der befragten Personen mit höchstens einem Hauptschulabschluss („Abschluss: Niedrig“) erweist sich in den letzten beiden Erhebungen 2012 und 2014 als zunehmend schwieriger zu berücksichtigen. Hinsichtlich des Ausmaß der Erfassung des konfessionellen Hintergrunds zeigen sich jedoch, im Gegensatz zum Allbus, keine Probleme.

Wie haben sich in dem identischen Zeitraum von 1990 bis 2014 die Zusammensetzung der telefonischen Befragungen verändert? Können auch diese ein vergleichbares Abbild ergeben? Abbildung 8.3 stellt für diesen Zeitraum die Bemessung der Qualität der Stichproben von Forsa und Politbarometer ebenfalls gemäß A_i dar. Da durch ihren komparativen Kostenvorteil gegenüber persönlichen Befragungen eine höhere Befragungsfrequenz möglich ist, wurde

eine etwas andere Form der Darstellung gewählt.

Abbildung 8.3.: Verzerrung Forsa und Politbarometer in Referenz zum Mikrozensus



Quelle: Eigene Darstellung. Für jedes Jahr sind Boxplots der berechneten Kennzahl A_i über alle erhobenen Stichproben des jeweiligen Jahres dargestellt. Für eine Darstellung der Anzahl an Umfragen und Jahr durch Forsa und Politbarometer siehe Abbildung 5.2 in Kapitel 5.2. Die Politbarometer-Stichproben sind hinsichtlich des Oversamplings der ostdeutschen Bevölkerung durch die getrennten Kumulationen eines Jahres korrigiert.

Für jede erhobene Umfrage eines Jahres wurde die Kennzahl A_i berechnet und anschließend über alle Werte eines Jahres die Verteilung in Form von Boxplots dargestellt. Hierdurch wird deutlich, dass Erhebungsdesigns telefonischer

Befragungen mit zunehmenden Problemen zu kämpfen haben. Lediglich das Geschlecht der befragten Personen wird auf einem ähnlichen Niveau erfasst wie durch die Erhebungsdesigns persönlicher Befragungen. Insbesondere zeigen sich bereits seit Mitte der 1990er Jahre Probleme, den Teil der Bevölkerung mit einem niedrigen Bildungsstand adäquat zu berücksichtigen, welche zu Gunsten der Personen mit einem (Fach-)Abitur unterrepräsentiert sind.

Seit dem Beginn der 2000er Jahre hat sich dieses Problem noch einmal intensiviert, wobei die Stichproben von Forsa stärkere Probleme aufweisen, als es die Erhebungen des Politbarometers tun.[6] Der Verlauf der Erfassung der Altersstruktur der volljährigen Bevölkerung zeigt sehr stark variierende Episoden: Beginnt seit Anfang der 2000er Jahre der Anteil der Personen ab 60 Jahren systematisch geringer zu werden, relativiert sich diese Entwicklung bis 2010 wieder. Seitdem wird der Personenkreis unter 40 Jahren, relativ gesehen zur Gesamtbevölkerung, zunehmend nicht mehr durch Telefonumfragen erreicht. Hier dürfte das Phänomen der steigenden „Coverage"-Probleme durch telefonische Befragungen einen Teil dieser Entwicklung erklären (siehe hierzu das Kapitel 6.1.2). Insbesondere seit 2010 intensiviert sich dieses Problem weiter, weshalb eine positive Prognose für die kommenden Jahre kaum möglich ist.[7] Die simultane Entwicklung und geringe Variation der Stärke der Verzerrungen beider Forsa- und Politbarometer-Stichproben zeigt, dass es sich um ein zunehmendes Problem telefonischer Befragungen im Allgemeinen handelt. Diese sind mit einem systematischen Alters- und Bildungsbias belegt. Dies kann auch mit ein Grund sein, warum die Offlinerekrutierung über Telefonstichproben in ein Access-Panel (wie das von LINK) kein Allheilmittel zur Erlangung qualitativ hochwertiger Onlinestichproben sein kann (siehe Kapitel 5.4).

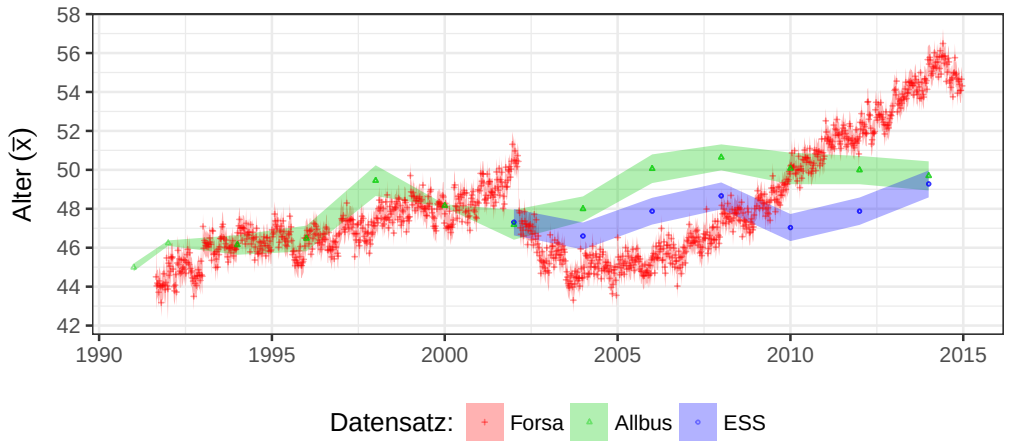
Das Problem der korrekten Abbildung des Durchschnittsalters der Bevölkerung durch telefonische Befragungen zeigt ebenfalls die Abbildung 8.4 durch einen Vergleich der Entwicklungen der arithmetischen Mittel des Alters der Forsa-, Allbus- und ESS-Stichproben.[8] Auch hier findet sich eine

[6] Folgende Zahlen illustrieren die Dimensionen des Problems: Für 2010 weist der Mikrozensus einen Anteil von ca. 42 % der Bevölkerung mit einem „niedrigen Schulabschluss" aus. Die Stichproben des Politbarometers enthalten hingegen im Durchschnitt für 2010 nur ca. 24 %, die Forsa-Daten sogar nur ca. 20 % Personen mit einer solchen Angabe. Hieraus resultieren die starken Verzerrungen von $A_{i=\text{Politbarometer}} = -0.60$ und $A_{i=\text{Forsa}} = -0.79$ im Durchschnitt.

[7] Auch hier verdeutlichen wenige Zahlen bereits die Dimension: 2014 betrug der Anteil der Altersgruppe 20 bis 29 Jahre an der erwachsenen Bevölkerung 16.70 %. Der Durchschnitt der Forsa-Stichproben für dieses Jahr liegt jedoch nur bei 7.64 % ($A_i = -0.81$), derjenige der Politbarometer-Daten nur bei 7.34 % ($A_i = -0.85$). Der junge Teil der Bevölkerung ist damit in den beiden groß angelegten Telefonumfrageprogrammen kaum noch existent.

[8] Die Stichproben des Politbarometers enthalten lediglich eine kategorisierte Abfrage auf einer zehnstufigen Skala, weshalb diese hier nicht berücksichtigt wurden.

Abbildung 8.4.: Entwicklung Alter Forsa, Allbus und ESS 1990 bis 2015



Quelle: Eigene Darstellung. Dargestellt sind die arithmetischen Mittel des Alters in Jahren der 1080 wöchentlichen Forsa-Stichproben und der jeweiligen Durchführungen des Allbus und des ESS. Die eingezeichneten Linien kennzeichnen den Verlauf der 95 %-Konfidenzintervalle der arithmetischen Mittel des jeweiligen Datensatzes. Für die persönlichen Befragungen wurde das komplexe Stichprobendesign hinsichtlich der Bestimmung der Standardfehler berücksichtigt.

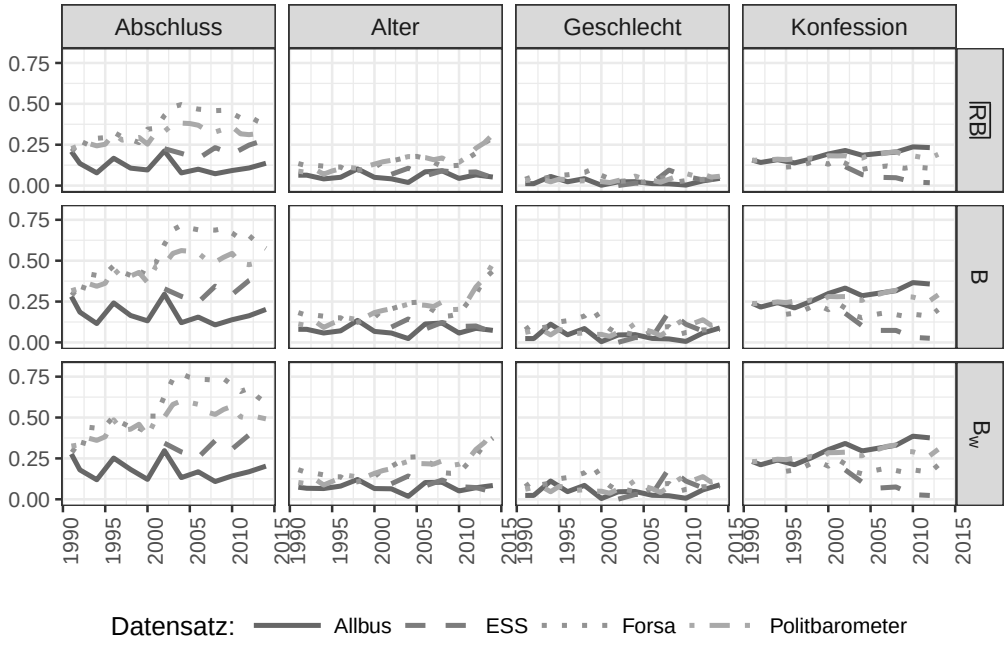
klare Tendenz der zunehmenden „Überalterung“ der Forsa-Stichproben im arithmetischen Mittel. Ist bis Anfang der 2000er Jahre eine nahezu konstante Entwicklung zwischen dem Allbus und Forsa erkennbar, kommt es zwischen 2002 und 2003 zu einem abrupten Absinken des Durchschnittsalters der Forsa-Stichproben. Nach einer Phase der konstanten Erfassung von Werten um ca. 45 Jahre zwischen 2003 und 2007, kommt es seitdem zu einer durchgehenden Steigerung. Seit 2011 übersteigen diese Werte klar diejenigen des Allbus und ESS, welche weiterhin eine sehr konstante Entwicklung skizzieren. Dies ist auf die bereits in Abbildung 8.3 gezeigten zunehmenden Probleme der Erfassung der Bevölkerung unter 40 Jahren in Telefonumfragen zurückzuführen.[9]

Wie lassen sich diese Entwicklungen im Aggregat beantworten? Die Abbildung 8.5 zeigt die Entwicklung der insgesamten Verzerrungen der *jeweiligen* Variablen über alle Gruppen hinweg gemäß B, B_w und RB (siehe zur Konzeption der Kennzahlen Kapitel 3.4). Dies bestätigt die bisherige Einschätzung: Telefonische Erhebungen und die zugehörigen Auswahldesigns lassen sich, bezüglich ihrer Qualität, nur hinsichtlich des Geschlechts und der konfessionellen Zugehörigkeit mit persönlichen Interviews vergleichen.

Die Ausgaben der Regressionsanalysen in Tabelle 8.1 fassen dies zusammen.

[9]Folgende Entwicklung verdeutlicht dies: Findet sich von 1991 bis einschließlich 2001 ein Zusammenhang von $r = 0.31$ zwischen dem Erhebungsjahr und den arithmetischen Mitteln des Alters, sinkt dies auf $r = -0.05$ für den Zeitraum 2002 bis 2006. Für den Zeitraum 2007 bis 2014 ergibt sich schließlich ein $r = 0.61$, mit jedem Jahr steigen die gemessenen arithmetischen Mittel des Alters der Stichproben stark an.

Abbildung 8.5.: Allbus, ESS, Politbarometer und Forsa im Vergleich zum Mikrozensus (Variablen)



Quelle: Eigene Darstellung. Für den Allbus und ESS sind die einzelnen Erhebungen dargestellt (Umfrageebene). Für Forsa und Politbarometer wurde über die Verteilung der Kennzahlen aller Umfragen eines Jahres hinweg das arithmetische Mittel zur Darstellung berechnet (Jahresebene). Korrelation zwischen B und B_w ist quasi linear perfekt ($r \approx 1$, sowohl für Jahres- als auch Umfrageebene), da die Gruppengröße keinen Einfluss auf die Höhe der Verzerrung hat.

Seit 1991 haben Verzerrungen hinsichtlich soziodemografischer Rahmendaten in Befragungen *nicht* per se zugenommen. Für die Stichproben des Allbus und ESS („Umfragen: Persönlich“) findet sich nur eine marginale Zunahme (*Jahre seit 1991*). Der Bildungshintergrund und die konfessionelle Zugehörigkeit lassen sich durch diese schlechter erfassen, als das Geschlecht oder das Alter. Der ESS weist hierbei etwas stärkere Probleme auf, welche jedoch nicht das problematische Niveau der telefonischen Befragungen („Umfragen: Telefonisch“) erreicht. Seit 1991 hat sich die Kennzahl B der telefonischen Befragungen um 0.01 pro Jahr gesteigert, für den betrachteten Zeithorizont von 20 Jahren ist dies ein Sprung um 0.2, bei einem möglichen Wertebereich von 0 bis 1. Die größte Herausforderung stellt der Bildungshintergrund dar, wobei sich keine der beiden betrachteten telefonischen Erhebungen als qualitativ besser erweist. Daher findet sich im kombinierten Modell („Umfragen: Alle“) für Forsa und Politbarometer ein höheres Verzerrungsniveau auf, als es der Referenzdatensatz des Allbus zeigt.

Tabelle 8.1.: Modelle: Verzerrungen Allbus, ESS, Forsa und Politbarometer

Umfragen:	Persönlich			Telefonisch			Alle		
Referenz:	Allbus			Forsa			Allbus		
Abhängige Variable:	$\overline{\lvert RB \rvert}$	B	B_w	$\overline{\lvert RB \rvert}$	B	B_w	$\overline{\lvert RB \rvert}$	B	B_w
Umfrage: ESS	0.002	0.001	-0.001				-0.02	-0.03	-0.03
	(0.01)	(0.02)	(0.02)				(0.01)	(0.02)	(0.02)
Umfrage: Forsa							0.07***	0.11***	0.12***
							(0.01)	(0.01)	(0.01)
Umfrage: Politbarometer				-0.002	-0.002	-0.003	0.07***	0.11***	0.11***
				(0.002)	(0.003)	(0.003)	(0.01)	(0.01)	(0.01)
Variable: Abschluss	0.13***	0.18***	0.18***	0.32***	0.44***	0.47***	0.32***	0.44***	0.47***
	(0.02)	(0.02)	(0.03)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Variable: Alter	0.04**	0.03	0.03	0.10***	0.12***	0.11***	0.10***	0.12***	0.10***
	(0.02)	(0.02)	(0.03)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Variable: Konfession	0.11***	0.16***	0.17***	0.09***	0.11***	0.12***	0.09***	0.11***	0.12***
	(0.02)	(0.03)	(0.03)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Jahre seit 1991	0.0003	0.001	0.001	0.003***	0.01***	0.01***	0.003***	0.01***	0.01***
	(0.001)	(0.001)	(0.001)	(0.0001)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0002)
Konstante	0.02	0.04*	0.04*	0.01***	0.02***	0.03***	-0.07***	-0.08***	-0.09***
	(0.02)	(0.02)	(0.03)	(0.002)	(0.003)	(0.003)	(0.01)	(0.01)	(0.01)
Beobachtungen	78	78	78	6,000	6,000	6,000	6,078	6,078	6,078
R^2	0.54	0.52	0.52	0.84	0.81	0.82	0.83	0.80	0.82
Korrigiertes R^2	0.51	0.48	0.49	0.84	0.81	0.82	0.83	0.80	0.82

Note: *p<0.1; **p<0.05; ***p<0.01

Quelle: Eigene Darstellung. Zur Definition der Kennzahlen siehe Formel 3.13 ($\overline{\lvert RB \rvert}$) und Formel 3.19 ($B_w$) in Kapitel 3.3

Diese Ergebnisse zeigen sich robust bezüglich der Nutzung der verschiedenen Kennziffern. Die durchschnittliche Stärke und Richtung der Erfassung der Verzerrung der Datensätze ist vergleichbar; die Nutzung des durchschnittlichen RB weist lediglich eine geringere Bandbreite auf. Es zeigt sich somit eine variablen- und eine modusspezifische Verzerrung in Umfragen. Den Bildungshintergrund korrekt zu erfassen ist schwieriger als die Abbildung der Altersstruktur der volljährigen Bevölkerung. Für telefonische Umfragen trifft dies zudem in einem höheren Ausmaß zu – konditional der hier aufgeführten soziodemografischen Variablen.

Dies bestätigt sich auch im Rahmen der Durchführung der wissenschaftlichen Wahlumfragen der GLES in Tabelle 8.2. Über beide Erhebungszeiträume ergibt sich auf der Basis der mehrstufigen räumlichen Zufallsauswahl der persönlich befragten Personen der Komponente 1 jeweils eine bessere Abbildung der (volljährigen) Bevölkerung.

Tabelle 8.2.: Verzerrung Komponenten 1 und 2 der GLES in Referenz zum Mikrozensus 2009 und 2013

Jahr	**Komponente**	**Gesamt ($\bar{B}$)**	**Alter**	**Bildung**	**Geschlecht**	**Konfession**
2009	1 (VW)	0.17	0.07	0.16	0.01	0.43
2009	2 (RCS)	0.30	0.14	0.70	0.10	0.25
2013	1 (VW)	0.23	0.25	0.13	0.09	0.45
2013	2 (RCS)	0.31	0.27	0.59	0.12	0.27

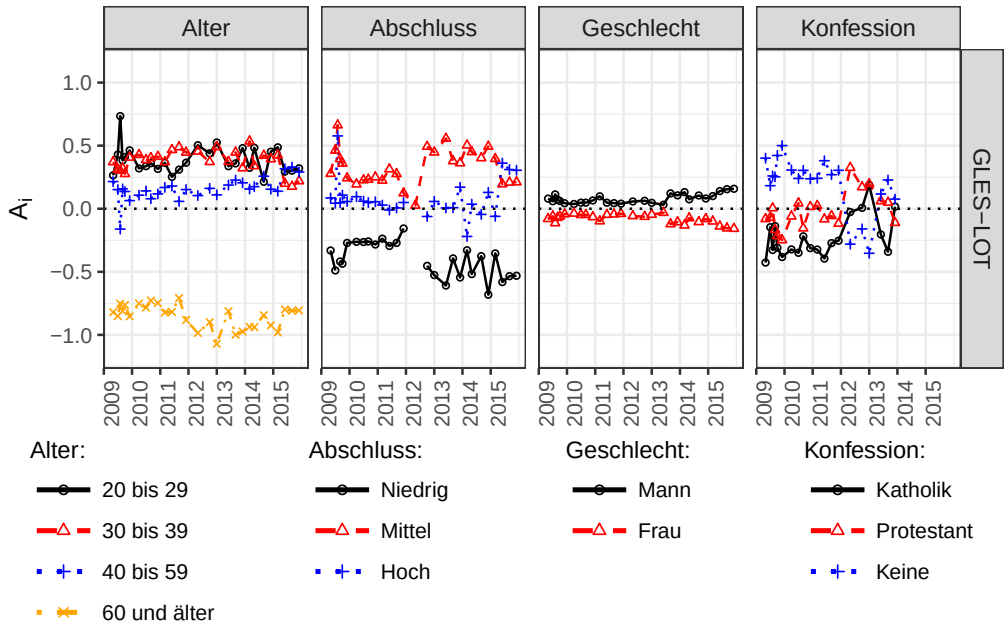
Quelle: Eigene Darstellung. Genutzt wurde der jeweilige Vorwahl-Querschnitt. Angewandte Gewichte: *w_trow* (Komponente 1) und *w_trang* (Komponente 2).

Die telefonische Befragung weist starke Probleme der Erfassung des Bildungsniveaus in beiden Erhebungen auf, in 2013 zusätzliche Probleme der Abbildung der Altersstruktur. Für die persönliche Befragung trifft dies vor allem auf den konfessionellen Hintergrund zu. 2013 findet sich dort ebenfalls, im Gegensatz zu den Durchführungen des Allbus, Probleme hinsichtlich der korrekten Erfassung des Alters. Auch persönliche Befragungen sind nicht per se geeignet soziodemografische Rahmendaten perfekt zu erfassen. Das jeweilige Umfrageprogramm scheint einen Teil der Qualität erklären zu können. Insgesamt zeigt sich jedoch auch hier eine klar bessere Erfassung der soziodemografischen Rahmendaten der befragten Personen durch die Komponente 1 der persönlichen Befragung gegenüber der telefonischen Variante. Ist das Ziel die Korrektur vorhandener soziodemografischer Verzerrungen in onlinebasierten Befragungen, sind telefonische Befragungen als Referenzumfragen nicht geeignet.

8.2. Onlinebasierte Wahlumfragen seit 2009

Der Verlauf der Zusammensetzung der Stichproben der Komponente 8 der GLES in Abbildung 8.6 bestätigt die aufgestellte Vermutung in Kapitel 5.4. Die beiden genutzten Access-Panel von Respondi (2009 bis 2011) und LINK (2012 bis 2016) sind nicht in der Lage, die (volljährige) Bevölkerung der Bundesrepublik annähernd zu replizieren. Weicht bereits die angewandte Quotierung von den (annähernd) wahren Werten der Bevölkerung ab (siehe Abbildung 6.8 in Kapitel 5.4), zeigt sich dies ebenfalls in den realisierten Stichproben.

Abbildung 8.6.: Verzerrung Komponente 8 der GLES in Referenz zum Mikrozensus 2009 bis 2016



Quelle: Eigene Darstellung. Die Erhebungen bis 2012 basieren auf dem Respondi-Panel, die anschließenden Erhebungen auf dem LINK-Panel. Die Erhebung $T3$ und $T17$ wurde für einen geringen Bildungsstand ($A_{i=3} = -1.44$ und $A_{i=17} = -1.92$) und die $T17$ für einen hohen Bildungsstand ($A_{i=17} = 1.29$) zur verbesserten grafischen Übersicht herausgenommen, da es sich hier um klare einmalige Ereignisse handelt.

Die Gruppe der Personen ab einem Alter von 60 Jahren ist eindeutig in den Stichproben beider Access-Panels zu gering vertreten. Dies trifft ebenso auf die Gruppe der Personen mit einem niedrigen Bildungsstand zu. Hierbei finden sich zudem stärkere Abweichungen der Stichproben des LINK Internet-Panel. Eine Ursache hierfür kann die telefonbasierte Rekrutierung in das Panel sein, ist diese Befragungsform (wie das vorangegangene Kapitel gezeigt hat) ebenfalls durch einen starken Bildungsbias gekennzeichnet. Einzig die Erfassung des Geschlechts erfolgt in den Stichproben der Komponente 8 nahezu unverzerrt.

Nur hier kann anscheinend die Aufgabe einer nicht-zufallsbasierten Auswahl durch die Anwendung einer (realisierbare) Quotierung so erfolgen, dass diese auch die volljährige Bevölkerung insgesamt widerspiegelt.

Bereits die Betrachtung dieser soziodemografischen Rahmendaten zeigt die Probleme onlinebasierter Befragungen. Zudem tauscht die offlinebasierte Rekrutierung die Möglichkeit einer zufallsbasierten Auswahl gegen eine Verstärkung des Bildungsbias ein. Wie lassen sich diese bisherigen Erkenntnisse zusammenfassen?

8.3. Zwischenfazit: Existiert ein Goldstandard?

Das vergangen Kapitel hatte zum Ziel, eine ordinale Rangfolge der Qualität der verschiedenen Befragungsformen aufzustellen. Dies ist möglich durch die enge Verknüpfung zwischen der gewählten Befragungsform und des sich dadurch ergebenden Auswahldesigns. Für den Zeitraum 1991 bis 2014 haben sich die persönlichen Befragungen des Allbus und ESS als den Telefonumfragen von Politbarometer und Forsa qualitativ überlegen hinsichtlich der Erfassung soziodemografischer Merkmale gezeigt (Tabelle 8.1). Dieses Ergebnis bestätigte sich auch durch einen Vergleich der Komponenten 1 und 2 der GLES (Tabelle 8.2). Zudem konnten die Schwächen der onlinebasierten Befragung validiert werden. Die Nutzung von quotierten Stichproben aus konstruierten Access-Panels führt zu Stichproben, welche eine sehr klare Abweichung zur amtlichen Statistik hinsichtlich der Altersverteilung und Bildungsstands aufweisen (Abbildung 8.6).

Dies ergibt ein erstes Zwischenfazit: Konditional soziodemografischer Rahmenvariablen ist die Durchführung persönlicher Befragungen, inklusive der damit verbundenen registergestützten Auswahl *oder* der Nutzung des „Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.“ (ADM)-Stichprobendesigns, den Erhebungen telefonischer Befragungen, auf der Basis einer „Randomized Last Digit“ (RLD)-Auswahl, überlegen. Persönliche Interviews stellen einen *soziodemografischen Goldstandard* als Referenzumfragen dar, wenn zur Korrektur von Verzerrungen keine Vergleichsdaten der amtlichen Statistik vorliegen. Werden hingegen die telefonischen Befragungsprogramme zur Korrektur genutzt, werden die vorhandenen Verzerrungen dieser mit übernommen.

Fraglich bleibt jedoch, ob sich dieser Vorteil hinsichtlich der Erfassung soziodemografischer Rahmendaten auch auf weitere inhaltliche Zielvariablen von Wahlumfragen auswirkt. Ist dies nicht der Fall, ist eine Korrektur soziodemo-

grafischer Verzerrungen nicht notwendig, sofern beispielsweise das Wahlverhalten im Fokus steht. Um dies zu überprüfen, werden nun die dargestellten telefonischen Befragungen an die persönlichen Befragungen angeglichen. Wirken sich die daraus ergebenden Korrekturgewichte auf inhaltliche Zielvariablen aus, stehen diese in einem Zusammenhang mit soziodemografischen Merkmalen und eine Korrektur dieser kann sinnvoll sein.

9. Telefonumfragen: Unterschiede zu persönlichen Befragungen

Nachdem das Kapitel 8.1 die Probleme telefonischer Befragungen hinsichtlich der korrekten Erfassung soziodemografischer Merkmale dargestellt hat, ist nun ein weiterer Vergleich dieser mit den persönlichen Interviews des Allbus und ESS als Referenzumfragen notwendig. Erst hierdurch lässt sich die verzerrte Erfassung von Merkmalen untersuchen, für die keine externen Referenzverteilungen durch die amtliche Statistik bereitstehen. Zudem ergibt sich der Vorteil der Betrachtung der multivariaten Verteilungen durch die Nutzung der Individualebene, an Stelle externer aggregierter Referenzverteilungen. Nachfolgend wird das Kapitel 9.1 erneut die Stichproben von Politbarometer und Forsa als langfristige Entwicklung telefonischer Befragungen betrachten: Diese werden durch „Propensity Scores“ (PS)-Modellierungen und verschiedene Algorithmen eines „Propensity Score Matching“ (PSM) (siehe Kapitel 4.3) an die Referenzdatensätze der jeweiligen Allbus- und ESS-Erhebungen angeglichen. Anschließend wird das Kapitel 9.2 Unterschiede und Möglichkeiten der Korrektur im Rahmen der Bundestagswahlen 2009 und 2013 aufzeigen. Hierzu wird die Komponente 2 der GLES an die simultan erhobene Komponente 1 und die durchgeführten Erhebungen des Allbus 2010 und 2014 angeglichen.

9.1. Politbarometer und Forsa

Der nachfolgende Teil wird in einem ersten Schritt die Unterschiede der Stichproben von Forsa und Politbarometer in Referenz zum ESS und dem Allbus in Kapitel 9.1.1 aufzeigen und darauf aufbauend Bedeutungsgewichte zur Korrektur dieser Unterschiede bestimmen. Anschließend wird das Kapitel 9.1.2 dazu übergehen, den Einfluss der Anwendung dieser an Hand des Wahlverhaltens und der Beteiligungsabsicht zu evaluieren.

9.1.1. Unterschiede und Korrekturmodelle – Soziodemografie und das Interesse an Politik

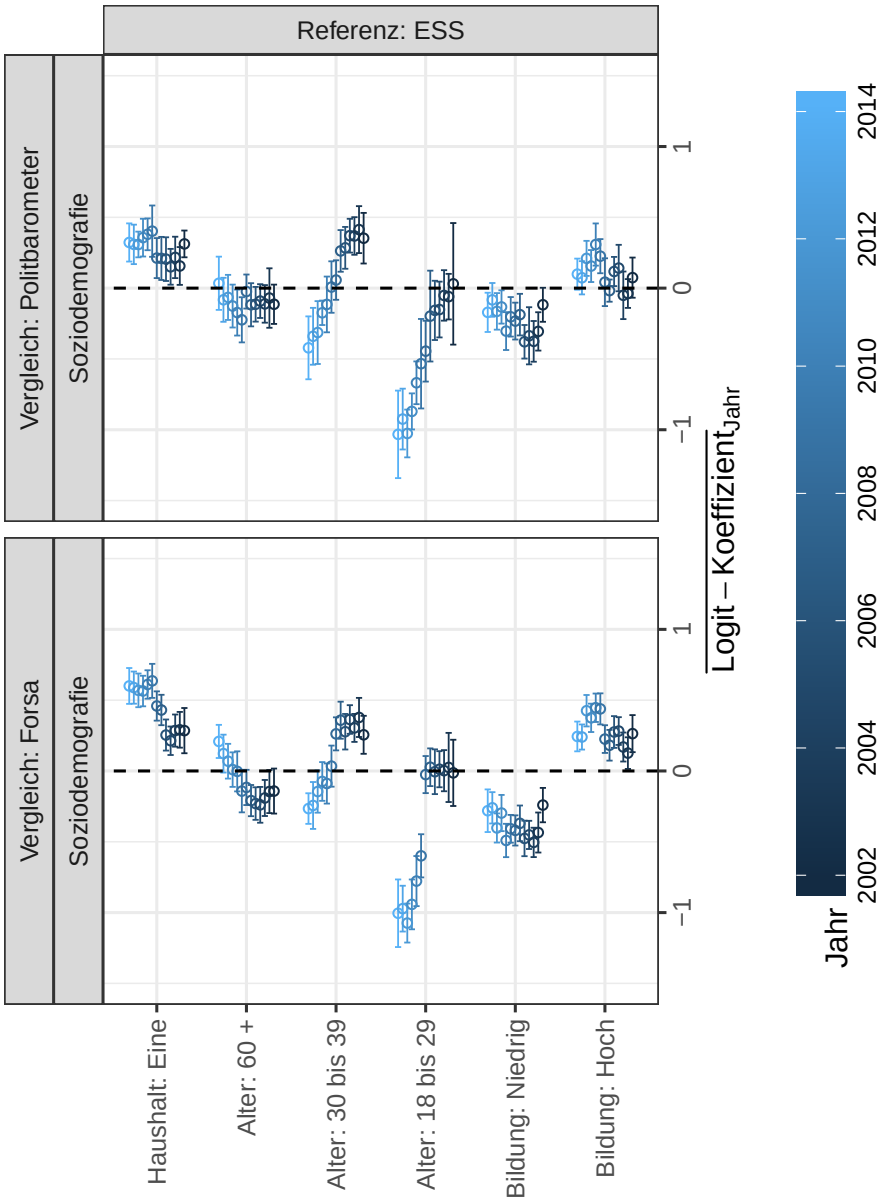
In welcher Form eine Gewichtungskorrektur über die Nutzung von PS möglich ist, ist von den Unterschieden zwischen der jeweiligen Stichprobe und ihres Referenzdatensatzes abhängig. Sowohl für die Erhebungen des Politbarometers und Forsa (Gruppe 1: Treatment) wurde für jede durchgeführte Umfrage deren Unterschiede zur zeitlich nächstgelegenen Erhebung des „European Social Survey“ (ESS) (Gruppe 0: Kontrolle) mit Hilfe logistischer Regressionsanalysen festgestellt. Die Abbildung 9.1 zeigt die Ergebnisse dieser Modelle. Die resultierenden Logit-Koeffizienten eines Jahres wurden zusammengefasst, der Ausgangspunkt der dargestellten Konfidenzintervalle stellt daher das arithmetische Mittel dieser Koeffizienten eines Jahres dar. Das Konfidenzintervall hingegen das $\pm$ 1.96-fache der Standardabweichung der berechneten Koeffizienten des jeweiligen Jahres und Umfrageprogramms. Enthält das jeweilige Konfidenzintervall nicht den Wert Null, kann somit mit einer unter 0.05 liegenden Irrtumswahrscheinlichkeit davon ausgegangen werden, dass alle denkbaren telefonischen Erhebungen dieses Jahres systematisch von den Werten des ESS abweichen.

Da das vorangegangene Kapitel bereits gezeigt hat, dass persönliche Umfragen Probleme hinsichtlich der korrekten Erfassung des konfessionellen Hintergrunds haben, wurde auf diese Aufnahme verzichtet. Die Aufnahme der Haushaltsgröße ergibt sich auf Grundlage der Überlegungen der Probleme der Abdeckung von Einpersonenhaushalten in telefonischen Befragungen (siehe hierzu auch Kapitel 4.2.1).

Auf der Grundlage dieser Modelle zeigen sich bereits eindeutige Entwicklungen der Unterschiede von Politbarometer und Forsa zum ESS. Beide telefonbasierten Umfrageprogramme weisen systematische und simultane Entwicklungen der Abweichungen hinsichtlich der Altersstruktur auf, welche die Ergebnisse des vorherigen Kapitels bestätigen. Insbesondere die Gruppe der 18 bis 29-jährigen Personen wird mit einer zunehmend geringeren Wahrscheinlichkeit in telefonischen Befragungen erfasst. Mit einer steigenden Tendenz trifft dies auch auf die Gruppe der 30- bis 39-jährigen Personen zu. Eine telefonische Stichprobe in 2014 zu erlangen, welche diesbezüglich keine systematischen Unterschiede zum ESS aufweist, ist ein sehr unwahrscheinliches Ereignis. Ebenfalls ergibt sich auf der Basis telefonischer Befragungen nur eine äußerst geringe Wahrscheinlichkeit der Realisierung einer Stichprobe, welche *nicht* einen systematisch höheren Anteil an Einpersonenhaushalten gegenüber dem ESS realisiert.

Auch hinsichtlich des Bildungshintergrunds zeigen sich sehr homogene Ent-

Abbildung 9.1.: Unterschiede Forsa und Politbarometer in Referenz zum ESS von 2002 bis 2014 (Soziodemografie)



Quelle: Eigene Darstellung. Für den ESS und die Politbarometer-Stichproben wurde eine Design-Korrektur auf Grund der Ost-West-Ungleichgewichte angewandt. Alle Variablen sind 0/1-Indikatoren. Referenzdatensatz ist der jeweilige ESS. Punkte: Arithmetisches Mittel der Logit-Koeffizienten *aller* Erhebungen eines Jahres. Linien: 1.96-fache der Standardabweichung der Logit-Koeffizienten eines Jahres. Ein positiver Koeffizient impliziert eine erhöhte Wahrscheinlichkeit der Zugehörigkeit zu den Stichproben von Forsa/Politbarometer. Wurde in einem Jahr der ESS nicht erhoben, wurde das nächste *folgende* Jahr gewählt. Anzahl Umfragen: 201 Politbarometer und 553 Forsa.

wicklungen der Stichproben von Forsa und Politbarometer. Kommt es Anfang 2000 zu einer geringeren Wahrscheinlichkeit der Berücksichtigung von Personen mit einem geringen Bildungshintergrund, zeigt sich seit 2010 eine Tendenz der

erneuten Angleichung. Wie das vorangegangene Kapitel jedoch gezeigt hat, ist die Durchführung des ESS seit 2010 ebenfalls von zunehmenden Problemen der korrekten Abbildung des Bildungshintergrunds konfrontiert, wenn die Referenzverteilungen des Mikrozensus als Referenz herangezogen werden (siehe Abbildung 8.2 in Kapitel 8.1).

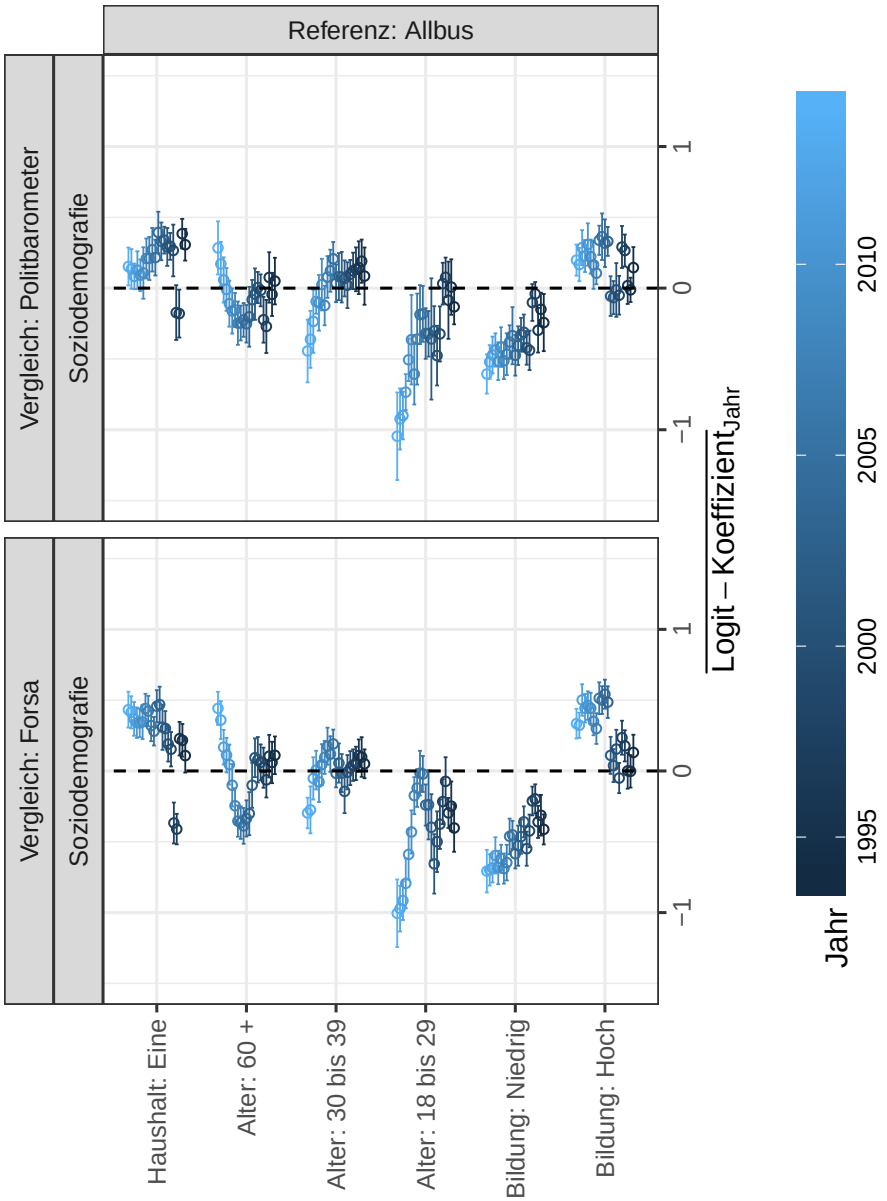
Die Abbildung 9.2 überträgt diese Logik der Berechnung aus Abbildung 9.1 auf den Vergleich zum „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (Allbus), welcher nun als Kontrollgruppe herangezogen wird (Gruppe 0: Kontrolle). Durch die bereits länger vorhandene Erhebungsreihe ist daher nun ein Vergleich seit 1994 möglich. 1994 wurde zudem als Ausgangsjahr gewählt, da das Politbarometer die Haushaltsgröße erst seitdem enthält, in den vorherigen Erhebungen wurde lediglich die reduzierte Haushaltsgröße abgefragt. Der Vergleich mit dem Allbus bestätigt vor allem eine zunehmende verzerrte Erfassung des Bildungshintergrund durch den ESS. Wird der Allbus als Referenz genutzt, zeigt sich keine zeitliche Tendenz der Angleichung auf der Basis der jeweiligen Kategorien des Bildungsstands.

Da der Allbus keine systematischen Abweichungen zum Mikrozensus in diesen Jahren aufweist, ist dies auf die nicht korrekte Erfassung der Telefonumfragen zurückzuführen. Im Vergleich zur Referenzkategorie der Personen mit einem Realschulabschluss haben diejenigen mit keinem oder dem Hauptschulabschluss eine eindeutig geringere Wahrscheinlichkeit, Teil einer telefonischen Befragung von Forsa oder Politbarometer zu werden. Bezüglich der Unterschiede hinsichtlich der Alterskategorien bestätigt sich, in einem vergleichbaren Ausmaß zur Nutzung des ESS als Referenz, die Zunahme der altersbedingten Verzerrungen in telefonischen Befragungen.

Im Gegensatz zum ESS enthalten die Erhebungen des Allbus zudem eine zu den Erhebungen des Politbarometers vergleichbare Abfrage des **Interesse an Politik**. Innerhalb des Allbus wurden die befragten Personen auf einer Skala von 1 („Sehr stark“) bis 5 („Überhaupt nicht“) „Wie stark interessieren Sie sich für Politik?“ befragt (Variable *V25* Allbus-Kumulation ZA4578_v1-0-0). Im Politbarometer unterscheidet sich lediglich die Formulierung für das Skalenende 5 („gar nicht“). Auch der ESS fragt das politische Interesse in einer vierstufigen Skala ab. Durch die nicht vorhandene mittlere Antwortkategorie ist sie jedoch nicht vergleichbar zur fünfstufigen Abfrage innerhalb des Politbarometers und der Vergleich daher nicht berücksichtigt.

Die Abbildung 9.3 zeigt das resultierende multinomiale Modell, welches diese Variable berücksichtigt. Um einen direkten visuellen Vergleich zu ermöglichen

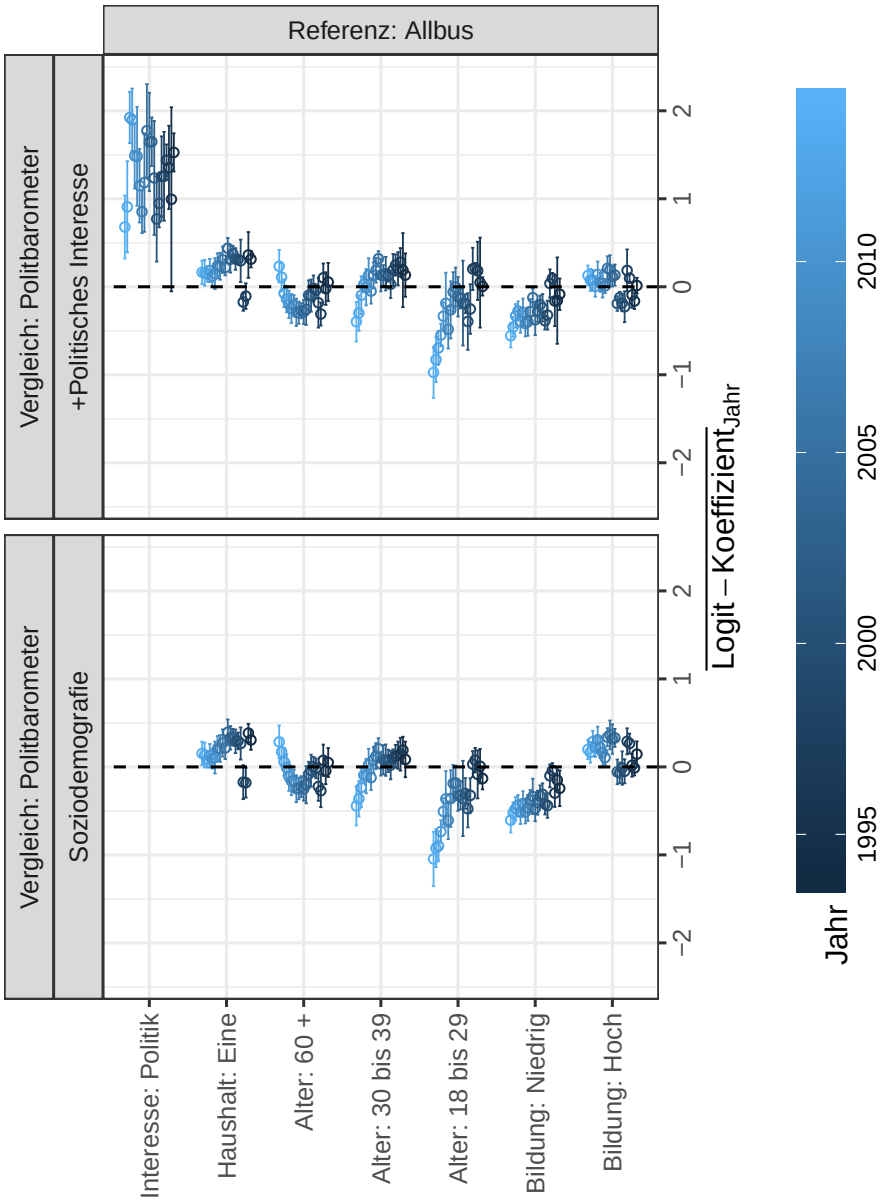
Abbildung 9.2.: Unterschiede Forsa und Politbarometer in Referenz zum Allbus 1994 bis 2014 (Soziodemografie)



Quelle: Eigene Darstellung. Für den Allbus und die Politbarometer-Stichproben wurde eine Design-Korrektur auf Grund der Ost-West-Ungleichgewichte angewandt. Alle Variablen sind 0/1-Indikatoren. Referenzdatensatz ist der jeweilige Allbus. Wurde in einem Jahr der Allbus nicht erhoben, wurde das nächste *folgende* Jahr gewählt. Anzahl Umfragen: 278 Politbarometer und 894 Forsa. Punkte/Linien: Siehe Beschreibung Abbildung 9.1.

ist zudem das vorherige Modell aus Abbildung 9.2 ebenfalls aufgeführt. Da nicht jede Erhebung des Politbarometers das Interesse an Politik berücksichtigt hat, reduziert sich die Gesamtanzahl der Umfragen pro Jahr. Die Breite der jeweilig eingezeichneten Linien des 1.96-fachen der Standardabweichung der Logit-Koeffizienten eines Jahres lässt jedoch einen ausreichend eindeutigen Schluss

Abbildung 9.3.: Unterschiede Politbarometer in Referenz zum Allbus 1994 bis 2014 (Soziodemografie und Interesse Politik)



Quelle: Eigene Darstellung. Für den Allbus und die Politbarometer-Stichproben wurde eine Design-Korrektur auf Grund der Ost-West-Ungleichgewichte angewandt. Das Interesse an Politik wurde über eine Min-Max-Normalisierung auf einen Wertebereich von 0 bis 1 reskaliert, alle anderen Variablen sind 0/1-Indikatoren. Referenzdatensatz ist der jeweilige Allbus. Ein positiver Koeffizient impliziert eine erhöhte Wahrscheinlichkeit der Zugehörigkeit zu den Stichproben von Forsa/Politbarometer. Wurde in einem Jahr der Allbus nicht erhoben, wurde das nächste *folgende* Jahr gewählt. Anzahl Umfragen Politbarometer: 278 für „Soziodemografie“, 200 für „+Politisches Interesse“. Punkte/Linien: Siehe Beschreibung Abbildung 9.1.

zu. Die Unterschiede der bereits vorher aufgenommenen Variablen bleiben weitestgehend konstant. Einzig die Bedeutung des Bildungshintergrund reduziert

sich leicht, was auf eine Verbindung mit dem Interesse an Politik hindeutet.

Hinsichtlich der Stärke der Erklärung der Unterschiede zwischen den Umfragen zeigt sich das politische Interesse jedoch als die maßgebliche Variable, um die Zugehörigkeit einer wahllos ausgewählten Person zwischen den Stichproben des Politbarometers und derjenigen des Allbus vorhersagen zu können. Der Sprung von „kein Interesse“ auf ein „sehr starkes Interesse“ lässt die Wahrscheinlichkeit der Teilnahme an einer Politbarometer-Befragung, im Vergleich zum Allbus, stark ansteigen. Dieser Effekt zeigt sich zudem konstant über alle Erhebungsjahre hinweg. Eine Stichprobe auf der Basis einer telefonischen Befragung zu erlangen, welche kein systematisch höheres politisches Interesse im Vergleich zum Allbus aufweist, ist ein sehr unwahrscheinliches Ereignis.[1]

Somit zeigt sich bereits eine zusätzliche Ursache von Verzerrungen, welche die bisherigen soziodemografischen Faktoren überwiegt: Das Interesse an Politik. Für dieses ist es wenig erstaunlich, wenn es in einem hohen Maße mit weiteren inhaltlichen Zielvariablen einer Umfrage korreliert. Durch die höhere Wahrscheinlichkeit politisch interessierter Personen an telefonischen Befragungen teilzunehmen, kann es daher auch zu einer zu optimistischen Einschätzung der Zustimmung zu Fragen kommen, welche positiv mit dem Interesse an Politik korrelieren. Ein typisches Beispiel für dies ist die Teilnahmebereitschaft an Wahlen und Abstimmungen.

Die aufgezeigten Unterschiede lassen sich als Korrekturmodelle zur Vorhersage der PS nutzen. Sie stellen in dem aufgezeigten Kontext die Wahrscheinlichkeit einer Person dar, Teil der jeweiligen telefonischen Befragung zu sein. Auf der Basis dieser bestimmten PS können dann PS-Gewichte als die Inverse dieser bestimmt. Ebenfalls lassen sich verschiedene Algorithmen eines PSM anwenden, wobei nachfolgend die Subklassifikation für das nachfolgende Beispiel eingesetzt wird und auf der Einteilung in sechs Klassen beruht. Für das PS-Gewicht wurde die Inverse der PS genutzt und auf 1 normalisiert, um die Gesamtzahl konstant zu halten. Damit die resultierenden Bedeutungsgewichte ebenfalls die Zusammensetzung der kumulierten Politbarometer-Stichproben hinsichtlich der Ost-West-Zugehörigkeit korrekt widerspiegeln, wurden diese zusätzlich mit der bereits vorhanden Design-Korrektur multipliziert (siehe hierzu das Kapitel 5.2). Als Folge ergibt sich ein kombiniertes Bedeutungs- und Designgewicht.

Die Bestimmung von Gewichtungsfaktoren kann jedoch den Nachteil der Zuweisung hoher individueller Bedeutungsgewichte haben, über alle Varianten

[1] Einzig in 1995 umschließt das 1.96-fache der Standardabweichung der Logit-Koeffizienten den Wert Null. Dort wurde jedoch nur in sehr wenigen Umfragen diese Abfrage berücksichtigt.

haben sich jedoch Werte der Maximalgewichte von unter 5 ergeben (siehe hierzu auch die Darstellung in Kapitel 4.2). Diese werden jedoch insbesondere größer, wenn zusätzlich das Interesse an Politik mit berücksichtigt wird. Ein korrigierender Eingriff ist also stärker notwendig, um diese Verzerrung ebenfalls zu negieren (siehe Abbildung B3 des Anhang B). Ein negativer Aspekt der Anwendung von Bedeutungsgewichten kann ebenfalls im Ausschluss bestimmter Gruppen von Beobachtungen liegen. Durch das angewandte Korrekturmodell ist dieser Ausschluss, bedingt durch fehlende Angaben auf einer der beteiligten Variablen, jedoch sehr gering. Der Anteil von befragten Personen ohne gültige Angabe auf einer der Kovariaten liegt bei bei unter 6 %. Dieser Ausfall ist zudem (beinahe) ausschließlich auf die Angaben des Bildungshintergrunds (z.B. Schüler) zurückzuführen (siehe Abbildung B2 des Anhang B). Auch die resultierende Balancierung der Stichproben durch das angewandte PSM zeigt eine Angleichung, welche zufriedenstellend ist. Wird die standardisierte Differenz zu Grunde gelegt, ergibt sich eine konstante Reduktion der Unterschiede der Verteilungen der PS beider beteiligten Umfragen von nahe 100 % (siehe Abbildung B1 des Anhang B).

9.1.2. Evaluation: Beteiligungsabsicht

Welcher Einfluss lässt sich durch die berechneten Bedeutungsgewichte feststellen? Die Nutzung zufallsbasierter Erhebungen auf Basis von Festnetzstichproben hat für die Wahlforschung eine besondere Bedeutung. Nahezu alle relevanten Umfrageinstitute in diesem Bereich nutzen dieses Medium.[2] Eigentlich dürfte keine Korrektur notwendig sein: Forsa und Politbarometer wählen ihre zu befragenden Personen nach einem zufallsbasierten Schlüssel aus. Abgesehen von einem Stichprobenfehler sollten die aggregierten Verteilungen daher auch das Elektorat bereits ohne weitere Korrekturen abbilden können. Ist dies nicht der Fall, hat die Selektivität der Teilnahme an den Umfragen auch einen Einfluss auf das Beteiligungsverhalten.

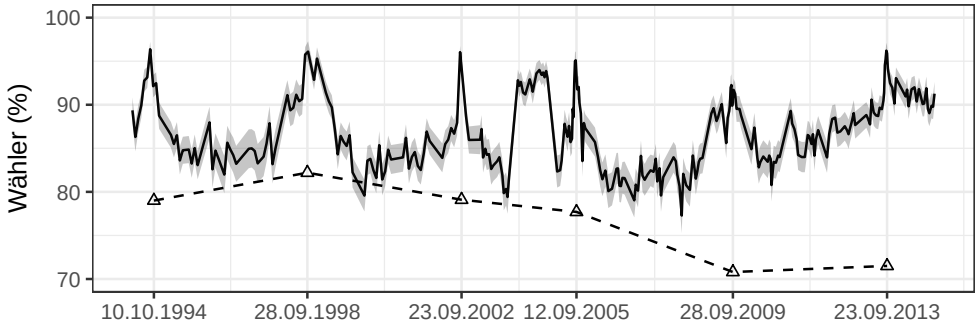
Das vorangegangene Kapitel hat nun aber bereits die zunehmenden Probleme der korrekten Erfassung des Bildungshintergrunds und des Interesses an Politik durch Telefonumfragen aufgezeigt. Stehen diese Variablen auch in einem Zusammenhang mit dem Beteiligungsverhalten, eignen sie sich für eine Korrektur. Die Anwendung darauf basierender Bedeutungsgewichte kann dann ebenfalls

[2] Ausnahme sind die Erhebungen von Allensbach auf der Basis persönlich-mündlicher Befragungen und die (relativ neuen) Erhebungen von INSA auf Basis von Erhebungen aus einem Access-Panel. Quelle: wahlrecht.de.

die Selektivität bezüglich des Teilnahmeverhaltens an Bundestagswahlen verringern. Innerhalb des Politbarometers wurden die befragten Personen durchgehend gefragt: „Wenn am nächsten Sonntag Bundestagswahl wäre, würden Sie dann zur Wahl gehen?“. Die Antwortmöglichkeiten sind konstant gehalten über alle Jahre der Erhebung mit „Ja“, „Nein“ und „Weiß nicht“. Es handelt sich um die identischen Anteile der Abbildung 7.8 in Kapitel 7.3. Innerhalb der Forsa-Stichproben findet sich keine Abfrage der Wahlbeteiligungsabsicht und des politischen Interesses, weshalb deren Umfragen an dieser Stelle nicht berücksichtigt werden.

Die Abbildung 9.4 zeigt den Verlauf des Anteils der Personen mit einer Beteiligungsabsicht an der jeweils nachfolgenden Bundestagswahl auf der Basis der Umfragen des Politbarometer. Es finden sich generell nur sehr wenige Erhe-

Abbildung 9.4.: Entwicklung Wahlbeteiligungsabsicht Politbarometer 1994 bis 2014



Quelle: Eigene Darstellung. Daten hinsichtlich der Ost-West-Ungleichgewichte korrigiert. Angegeben ist zudem das 95 %-Konfidenzintervall. Auffassung als einfache Zufallsziehung nach Formel 6.5 in Kapitel 6.1.3. Eingezeichnete Referenzlinie sind die offiziellen Wahlbeteiligungsraten des Bundeswahlleiters der jeweiligen Bundestagswahl.

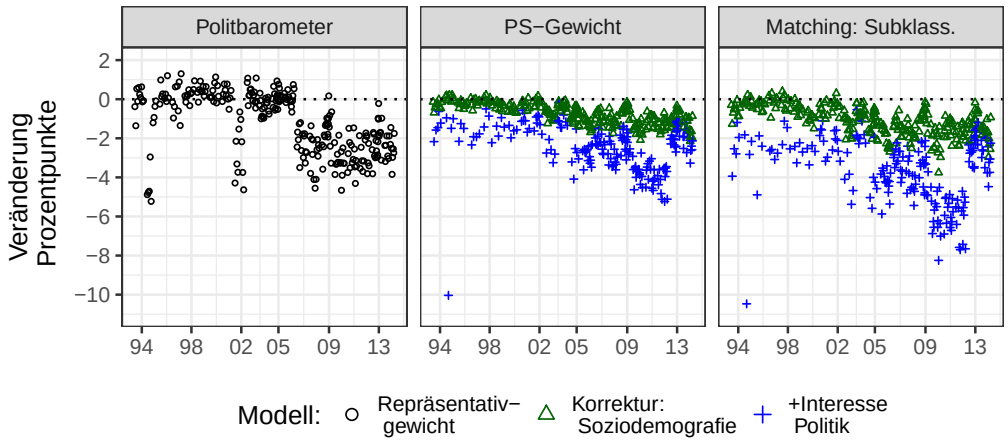
bungen mit einer Wahlabsicht, welche ein zu den offiziellen Beteiligungsraten der jeweiligen Bundestagswahlen vergleichbares Niveau aufzeigen, auch wenn in den ersten Jahren seit 1994 hohe Abweichungen vor allem nur im Vorfeld von Bundestagswahlen entstehen. Markanter ist eine zunehmende Abweichung seit 2007 zwischen den Umfragen des Politbarometers und den offiziellen Beteiligungsraten der Bundestagswahlen 2009 und 2013. Im Gegensatz zu einem offiziellen Niveau von nur noch knapp über 70 % seit der Bundestagswahl 2009, zeigen sich Anteilswerte auf der Basis des Politbarometers von über 90 % der Personen mit einer Beteiligungsabsicht, wenn an diesem Punkt die Bundestagswahl stattfinden würde.

Zwei prominente Erklärungen lassen sich hierfür anführen: Personen geben

an zu wählen, obwohl sie dies nicht tun werden. Dies wäre das „overreporting“ auf Grund sozialer Erwünschtheit durch die Anwesenheit des Interviewers, also ein Messfehler im Sinne des „Total Survey Error“ (TSE) (siehe hierzu das Kapitel 2.3). Die zweite Möglichkeit ist eine Selektivität der Teilnahme an den Befragungen des Politbarometers durch einen Nonresponsefehler (siehe hierzu auch das Kapitel 2.2.1), Personen ohne eine Beteiligungsabsicht nehmen tendenziell auch seltener an diesen teil. Die Anteilswerte der Wähler der Politbarometer-Stichproben weichen als Folge systematisch von ihrem Populationswert ab, da die Erhebungen verzerrt sind. Korreliert diese Entscheidung zur Verweigerung der Teilnahme an den Umfragen mit einem niedrigem Interesse an Politik und einem niedrigem Bildungsniveau, ist dies ein systematischer Fehler konditional dieser beiden Variablen. Eine Korrektur dieser Unterschiede sollte daher die Wahlbeteiligungsabsicht der Politbarometer-Stichproben absenken.

Wie die Abbildung 9.5 aufzeigt, ist dies auch der Fall. Um eine

Abbildung 9.5.: Veränderung Wahlbeteiligungsabsicht durch Gewichtung Politbarometer 1994 bis 2014



Quelle: Eigene Darstellung.

übersichtlichere visuelle Darstellung zu ermöglichen, wurden zunächst erst zwei mögliche Varianten der Korrektur aufgezeigt (einfache PS-Korrektur und ein Matchingverfahren basierend auf der Subklassifikation, siehe auch das Kapitel 4.3 zur Übersicht). Für diese wurde für jede Umfrage der jeweils nächste verfügbare Allbus-Datensatz als Referenzumfrage verwendet. Zudem wurde das durch die Forschungsgruppe Wahlen berechnete „Repräsentativgewicht“ angewandt, um eine bessere relative Bewertung der erzielten Ergebnisse zu erreichen.

Führt die Anwendung des letzteren bis zur Bundestagswahl 2005 größtenteils zu einer weiteren Erhöhung der Wahlbeteiligungsrate, setzt seit 2007 eine kla-

re Verbesserung ein. Es zeigen sich durchgehend positive Auswirkungen auf die Wahlbeteiligungsabsicht, eine Reduktion von ca. 2 bis 4 Prozentpunkten ist möglich. Dies trifft ebenso auf die berechneten Bedeutungsgewichte im Rahmen der Arbeit zu. Im Gegensatz zu den Institutsgewichten findet zudem keine Erhöhung der Beteiligungsrate vor 2005 statt. Die zunehmende Selektivität in den Politbarometer-Stichproben lässt sich also bereits durch einfachere Korrekturen begegnen. Erst die Berücksichtigung des Interesse an Politik bei der Konstruktion der Bedeutungsgewichte führt jedoch zu einer signifikanteren Reduktion der Beteiligungsabsicht. Die Subklassifikation erreicht hierbei Werte von bis zu acht Prozentpunkten im Nachgang der Wahl 2009. Zudem zeigt sich seit 2005 keine Umfrage, welche nicht von der Angleichung an den Allbus profitiert.

Da der beginnende Erfolg der Korrektur mit einem offiziellen Rückgang der Wahlbeteiligung bei Bundestagswahlen koinzidiert, lässt dies folgende Feststellung zu: Im Zuge des Niedergangs der Wahlbeteiligung bei Bundestagswahlen tritt ebenfalls eine verstärkte Selektivität der Teilnahme an Telefonumfragen innerhalb der Politbarometer-Daten auf. Das sich innerhalb dieser *steigende* Beteiligungsraten beobachten lassen, ist auch ein Resultat der zunehmend niedrigeren Teilnahmeraten an Umfragen durch die Bevölkerung mit einem geringeren Interesse an Politik. Aus diesem Grund steigen die Raten in Umfragen konträr zur realen Entwicklung. Wird jedoch eine Korrektur auf der Basis des Allbus durchgeführt, reduziert sich dieser Effekt. Dies funktioniert, da der Allbus nicht von diesen Verzerrungen betroffen ist.

Zwei Beobachtungen werden hierzu näher betrachtet: Das erste ist ein technischer Aspekt. Die Nutzung einer Subklassifikation im Rahmen eines PSM ist gegenüber der einfachen inversen PS-Gewichtung besser in der Lage, eine Korrektur der Wahlbeteiligungsabsicht zu erreichen. Zweitens lässt sich beobachten, dass der Effekt der Korrektur auf der Basis des Interesse an Politik seit dem Vorfeld der Bundestagswahl 2013 nicht mehr möglich ist. Dass die technische Art der Korrektur relevant ist, zeigen die Ergebnisse der Regressionsanalysen der nachfolgenden Tabelle 9.1. Die zeitliche Einteilung sind hierbei die stattfindenden Bundestagswahlen.

Als abhängige Variable ist durchgehend die Reduktion des Wähleranteils der Politbarometer-Stichproben in Prozent*punkten* definiert. Als Ausgangsbasis des Modells dient der Erfolg der Korrektur durch soziodemografische Variablen, welche auf der Basis einer PS-Gewichtung in einem direkten Kontext einer Bundestagswahl („KW bis Wahl“ gleich Null) durchgeführt wird. Dieser direkte

Tabelle 9.1.: Modell: Veränderung Wahlbeteiligungsabsicht durch Gewichtung Politbarometer 1994 bis 2014

	Abhängige Variable: Reduktion Prozentpunkte Wähleranteil					
Anfang:	16.10.1994	28.09.1998	23.09.2002	19.09.2005	28.09.2009	23.09.2013
Ende:	27.09.1998	22.09.2002	18.09.2005	27.09.2009	22.09.2013	Heute
Konstante	0.33	0.01	-0.79^{***}	-0.75^{***}	-0.62^{***}	-4.22^{***}
	(0.26)	(0.14)	(0.14)	(0.12)	(0.16)	(0.50)
KW bis Wahl	-0.004^{***}	-0.001^{*}	0.003^{**}	-0.002^{***}	-0.004^{***}	0.02^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)
Korrektur:	Referenzkategorie					
Soziodem. + Int. Pol.	-2.13^{***}	-1.55^{***}	-1.39^{***}	-1.63^{***}	-2.57^{***}	-1.05^{***}
	(0.20)	(0.10)	(0.12)	(0.08)	(0.11)	(0.12)
Gewichtung:	Referenzkategorie					
PS-Gewicht						
Match: Subklass.	-0.34	-0.50^{***}	-0.61^{***}	-0.83^{***}	-1.08^{***}	-0.47^{**}
	(0.29)	(0.15)	(0.16)	(0.13)	(0.17)	(0.18)
Match: Nearest	0.51^{*}	0.64^{***}	0.94^{***}	1.69^{***}	2.23^{***}	1.54^{***}
	(0.29)	(0.15)	(0.16)	(0.13)	(0.17)	(0.18)
Match: CEM	-0.46	-0.67^{***}	-0.55^{***}	-0.83^{***}	-1.30^{***}	-0.79^{***}
	(0.29)	(0.15)	(0.16)	(0.13)	(0.17)	(0.18)
Match: Full	-0.24	-0.89^{***}	-0.61^{***}	-0.93^{***}	-1.47^{***}	-0.94^{***}
	(0.29)	(0.15)	(0.16)	(0.13)	(0.17)	(0.18)
Beobachtungen	255	305	335	605	620	210
Korrigiertes R^2	0.34	0.54	0.45	0.63	0.67	0.63

*p<0.1; **p<0.05; ***p<0.01

Quelle: Eigene Darstellung.

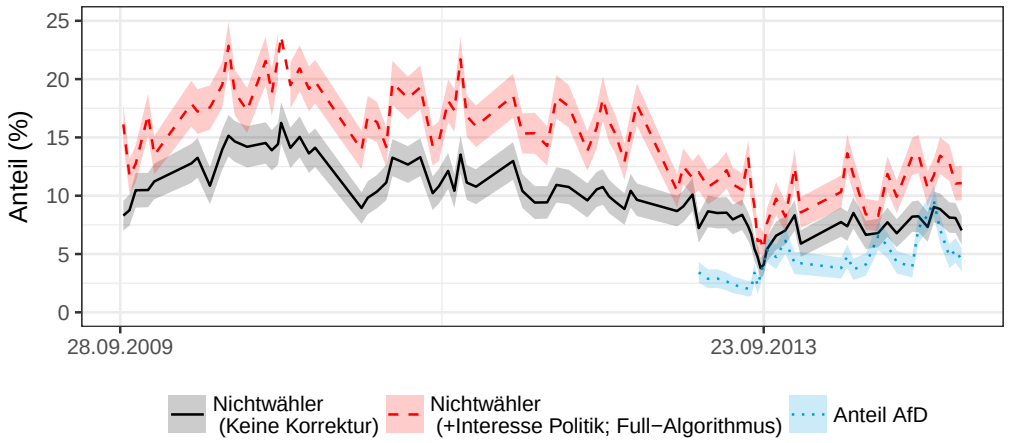
Kontext ergibt sich durch die Aufnahme der zeitlichen Nähe zur kommenden Bundestagswahl in kalendarischen Wochen, da sich im Vorfeld eine Intensivierung der Beteiligungsabsicht in Umfragen ergeben kann. Dies belegt auch der bereits dargestellte Verlauf der (beabsichtigten) Beteiligungsraten in Abbildung 9.4. Zusätzlich ist der Effekt der Aufnahme des Interesses an Politik in das PS-Modell, sowie die Wahl verschiedener Matchingalgorithmen an Stelle der einfachen PS-Gewichtung berücksichtigt worden.

Es bestätigt sich hierdurch der visuelle Eindruck der Abbildung 9.5. Die zusätzliche Berücksichtigung des Interesse an Politik im Rahmen eines Korrekturmodells führt über alle Zeiträume zu einer zusätzlichen Reduzierung des prozentualen Anteils von Personen mit einer Wahlabsicht. Der Effekt tritt am stärksten zwischen den Bundestagswahlen 2009 und 2013 zu Tage. Zudem zeigen sich, mit Ausnahme des Nearest-Algorithmus, alle angewandten Varianten eines PSM überlegen gegenüber der einfachen Nutzung der PS zur Bestimmung der Bedeutungsgewichte. Die durchgehend stärkste Reduktion erreicht

der Full-Algorithmus in Kombination mit einer Korrektur auf der Basis des Interesses an Politik. Zwischen den Wahlen 2009 und 2013 ist hierdurch eine im Durchschnitt um vier Prozentpunkte bessere Reduktion der prognostizierten Wahlbeteiligungsrate gegenüber der inversen PS-Nutzung auf der Basis der Soziodemografie möglich. Neben der Wahl des Korrekturmodells ist somit auch die Wahl der technischen Variante entscheidend für eine erfolgreiche Korrektur.

Dass sich die Möglichkeit der Korrektur auf der Basis des Interesse an Politik seit dem Kontext der Wahl 2013 stark reduziert, obwohl die Politbarometer-Stichproben seit dieser ein anhaltend hohes Niveau der Beteiligung in der Abbildung 9.5 verzeichnen, koinzidiert mit einer anderen Entwicklung innerhalb des politischen Systems: Dem Auftreten der „Alternative für Deutschland“ (AfD). Die Abbildung 9.6 zeigt hierzu den Verlauf der Anteilswerte der *Nicht*wähler in den Umfragen und derjenigen Personen, welche sich hinsichtlich einer (gültigen) Angabe der präferierten Zweitstimme für die AfD entscheiden würden. Auch ist

Abbildung 9.6.: Wahlbeteiligungsabsicht, Gewichtung und Zweitstimmenanteile AfD 2009 bis 2014



Quelle: Eigene Darstellung. Daten hinsichtlich der Ost-West-Ungleichgewichte korrigiert. Angegeben ist zudem das 95 %-Konfidenzintervall. Auffassung als einfache Zufallsziehung nach Formel 6.5 in Kapitel 6.1.3.

der Anteil der Nichtwähler ausgewiesen, welcher sich durch die Anwendung des gerade aufgeführten Bedeutungsgewichts ergibt. Hierbei wurde also ebenfalls das Interesse an Politik einbezogen und die Unterschiede zum Allbus der jeweiligen Jahre (genutzt wurden die Versionen 2010, 2012 und 2014) über ein PSM auf der Basis des Full-Algorithmus balanciert.

Der Abstand hinsichtlich der Entwicklung des Anteils der Nichtwähler ohne und mit Korrektur reduziert sich sichtbar im Kontext der Bundestagswahl 2013. Von diesem Zeitraum an kommt es ebenfalls zu einer expliziten

Berücksichtigung der AfD als Kategorie innerhalb der Abfragen der Zweitstimmenpräferenz durch Politbarometer. Führt die Anwendung der Bedeutungsgewichte bis zur erstmaligen Berücksichtigung der AfD[3] im arithmetischen Mittel zu einer Erhöhung des Nichtwähleranteils um ca. fünf Prozentpunkte, reduziert sich der Erfolg nach dem Aufkommen dieser auf nur noch 3.3 Prozentpunkte.[4] Es deutet sich daher an: Die Gewichtungskorrektur des politischen Interesses wirkt geringer, da die AfD Personen aus dem vorherigen Nichtwählerlager mobilisieren konnte. Der Anteil der Nichtwähler sinkt daher auf ein Niveau seit der Wahl 2013 ab, welches konstant unter 10 % liegt. Diese Gruppe der vorherigen Nichtwähler gibt ebenfalls an, ein geringes Interesse an Politik zu haben. Als Folge ist die Differenzierungs- und Korrekturmöglichkeit auf der Grundlage dieses Interesse nicht weiter möglich, da dieses nicht mehr über eine Partizipationsbereitschaft an Wahlen entscheidet.

Resümierend zeigt sich somit bereits auf lange Sicht eine Möglichkeit der Verbesserung der Datenqualität telefonischer Befragungen. Hierzu muss jedoch über die Korrektur grundlegender soziodemografischer Variablen hinausgegangen werden und weitere Faktoren, wie das hier aufgezeigte Interesse an Politik, berücksichtigt werden. Finden sich diese Unterschiede auch im Rahmen der (wissenschaftlichen) Wahlumfragen der „German Longitudinal Election Study“ (GLES)? Das nachfolgende Unterkapitel 9.2 wird nun diese Erkenntnisse über telefonische Befragungen auf den Kontext der Bundestagswahlen 2009 und 2013 übertragen. Hierdurch ist eine bessere Evaluation möglich, da diese Umfragen explizit im Vorfeld einer Bundestagswahl entstanden sind. Eine Bezugnahme auf die amtlichen Endergebnisse ergibt daher, im Gegensatz zu den langen Reihen der Politbarometer-Daten, einen tieferen Sinn.

9.2. Der Kontext der Bundestagswahlen 2009 und 2013

Im Kontext der beiden Bundestagswahlen 2009 und 2013 ergibt sich durch die Verfügbarkeit der Komponenten 1 und 2 der GLES eine zusätzliche Möglichkeit der Analyse. Durch eine Betrachtung dieser lassen sich ebenfalls Unterschiede zwischen einer persönlichen (Komponente 1) und telefonischen Befragung

[3] Dies fand Ende April 2013 statt. Siehe hierzu die Darstellung auf http://www.wahlrecht.de/umfragen/politbarometer/politbarometer-2013.htm (Abruf: 11. Mai 2018).

[4] Grundlegende deskriptive Statistiken betragen ohne Erfassung der AfD: $\tilde{x} = 5.58$; $\bar{x} = 5.4$, $s_x = 1.69$ und mit Erfassung der AfD: $\tilde{x} = 3.16$; $\bar{x} = 3.3$, $s_x = 1.34$.

(Komponente 2) herausstellen. Zudem ist durch das angewandte „Rolling-Cross-Section“ (RCS)-Design die Unterteilung der Komponente 2 in unabhängige wöchentliche Stichproben möglich. Die Tabelle 9.2 zeigt die Anzahl der befragten Personen dieser Unterteilung. Mit den sich daraus ergebenden Stichprobe-

Tabelle 9.2.: Umfang Wochenstichproben der Komponente 2 der GLES

(a) 2009

KW	31	32	33	34	35	36	37	38	39
n	(138)	616	719	669	697	706	750	842	871

(b) 2013

KW	28	29	30	31	32	33	34	35	36	37	38
n	634	657	755	777	755	639	698	734	726	809	698

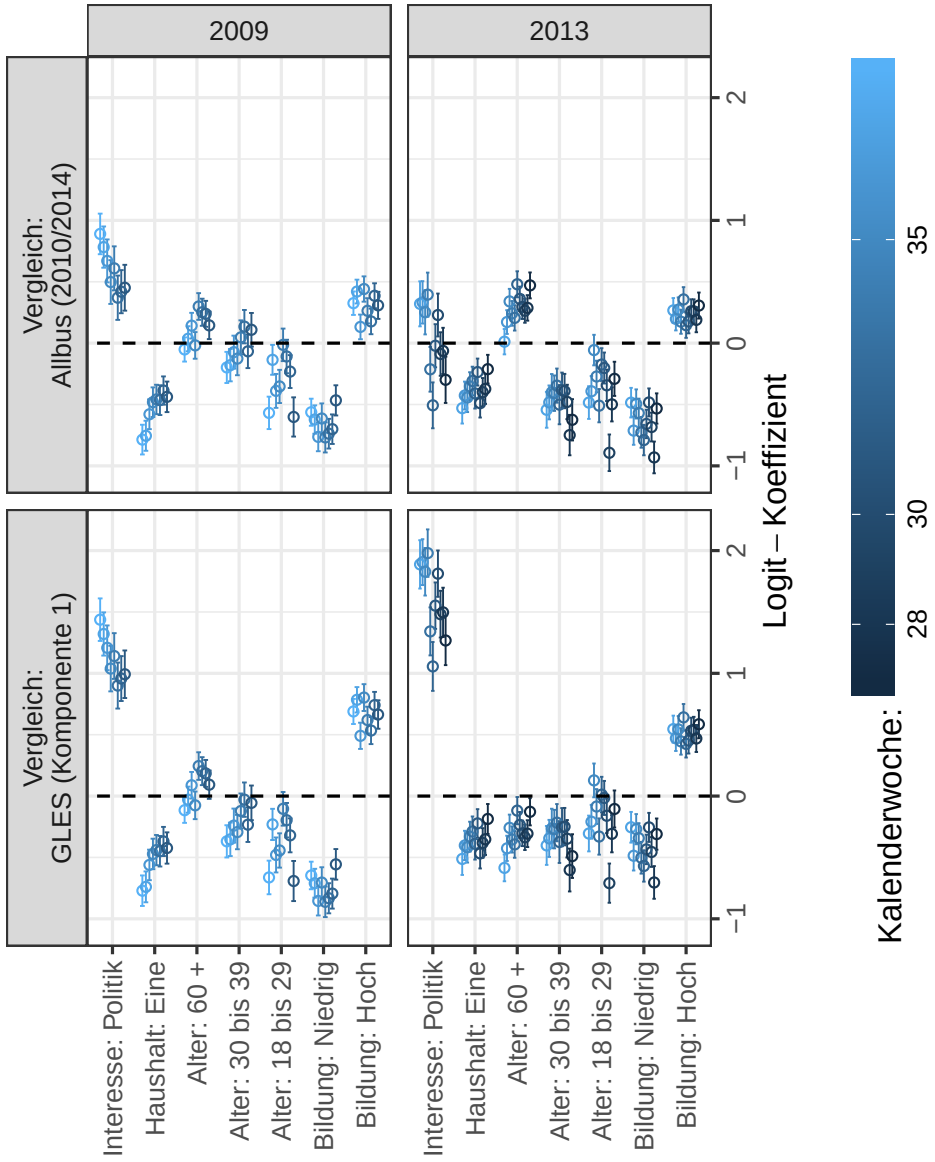
Quelle: Eigene Darstellung. Informationen sind Auszählungen der Variable *pre_woche* des jeweiligen Datensatz (2009: ZA5303 und 2013: ZA5703).

numfängen von ca. 600 bis ca. 800 Personen weisen diese eine ausreichende Größe für einen separaten Infererenzschluss auf das Elektorat im Vorfeld der Wahlen 2009 und 2013 auf.[5] Durch die Unterteilung nach den kalendarischen Wochen lässt sich zudem herausstellen, ob der zeitliche Abstand zur nachfolgenden Bundestagswahl einen Einfluss auf die Verteilungen inhaltlich interessierender Variablen und den Erfolg einer möglichen Korrektur hat.

In Abbildung 9.7 sind die Ergebnisse von vier multinomialen logistischen Modellen aufgeführt, welche die Unterschiede zwischen den beiden Erhebungsmodi der persönlichen und telefonischen Befragung herausstellen. Als Referenz für die telefonischen Wochenstichproben der Komponente 2 wurde sowohl die Erhebungen der Komponente 1, als auch die bereits genutzten qualitativ hochwertigen Erhebungen des Allbus als eine weitere Referenz genutzt. Ein positiver Logit-Koeffizient impliziert daher eine erhöhte Wahrscheinlichkeit der Zugehörigkeit zur jeweiligen Wochenstichprobe der Komponente 2. Die Ergebnisse validieren die Analysen auf der Basis des vorherigen Kapitel 9.1.1. Telefonische Befragungen ergeben Stichproben, welche ein erhöhtes Bildungsniveau aufweisen. Insbesondere die Gruppe der Personen mit keinem oder einem Hauptschulabschluss („Niedrig“) lässt sich im Vergleich zur Referenzgruppe mit einem Realschulabschlus („Mittel“) erneut schlecht erfassen. Personen mit einem (Fach-)Abitur („Hoch“) neigen hingegen zu einer erhöhten Teilnahmewahrscheinlichkeit. Zudem ist, im Vergleich zur Gruppe der 40- bis 59-jährigen, der Personenkreis ab 60 Jahren mit einem erhöhten Anteil vertreten. Die jüngeren Bevölkerungsgruppen

[5] Die KW 31 in 2009 wurde nachfolgend daher nicht berücksichtigt, da in dieser nur 138 Personen befragt wurden.

Abbildung 9.7.: Unterschiede Komponente 2 in Referenz zur Komponente 1 und zum Allbus 2009 und 2013 (Soziodemografie und Interesse Politik)



Quelle: Eigene Darstellung. Dargestellt sind die Logit-Koeffizienten der multinomialen Modelle inklusive ihrer 95 %-Konfidenzintervalle. Gewichte GLES: *w_trow* (Komponente 1) und *w_tranw* (Komponente 2). Für den Allbus wurde jeweils das „Ost-West Transformationsgewicht Person“ des jeweiligen Datensatz angewandt.

und Personen in Einpersonenhaushalten zeigen hingegen eine verringerte Wahrscheinlichkeit der Zugehörigkeit zu den Wochenstichproben auf. Die stärkste Differenzierungsmöglichkeit zwischen den Personenkreisen der beiden Befragungsmodi ergeben sich jedoch erneut auf der Grundlage des angegebenen Interesse an Politik. Alle telefonischen Wochenstichproben weisen ein, im Vergleich zum jeweiligen Referenzdatensatz, erhöhtes politisches Interesse auf. Einzig

die Nutzung des befragten Personenkreis des Allbus 2014 als Referenzgruppe führt nicht zwingend zu diesem Ergebnis. Da diese Durchführung, im Vergleich zu denjenigen der vorherigen Jahre, ebenfalls ein erhöhtes Niveau aufweist, kann dies eine mögliche Erklärung sein. Daher eignet sich dieser Datensatz, hinsichtlich des Interesse an Politik, nicht als eine gute Vergleichsbasis.[6]

Auf Basis der vier Korrekturmodelle wurden Bedeutungsgewichte zur Korrektur der Komponente 2 bestimmt. Sowohl die inverse Nutzung der berechneten PS, die Subklassifikation, als auch der Nearest-, CEM- und Full- Algorithmus wurden angewandt.[7] Die Balancierung der Angleichungen der Verteilungen an Hand der PS zeigt ein gutes Niveau. Für die telefonischen Wochenstichproben ist eine vollständige Reduktion der Unterschiede möglich, wenn die standardisierte Differenz betrachtet wird. Hinsichtlich der weiteren möglichen Kennzahlen zur Evaluation weist der Full-Algorithmus den besten Erfolg auf (zur Darstellung siehe Abbildung B5 in Anhang B). Auch die erzeugten Maximalwerte der Bedeutungsgewichte lassen nicht den Schluss zu, dass es zu einer einseitigen Betrachtung weniger Personen mit einer hohen beigemessenen Bedeutung kommt (Abbildung B4 in Anhang B). Ebenfalls begrenzt sich der Ausfall durch Personen, welche das Korrekturmodell durch fehlende Werte beteiligter Variablen nicht berücksichtigten konnte, auf unter 5 %.[8]

9.2.1. Erfassung der Zweitstimmenanteile

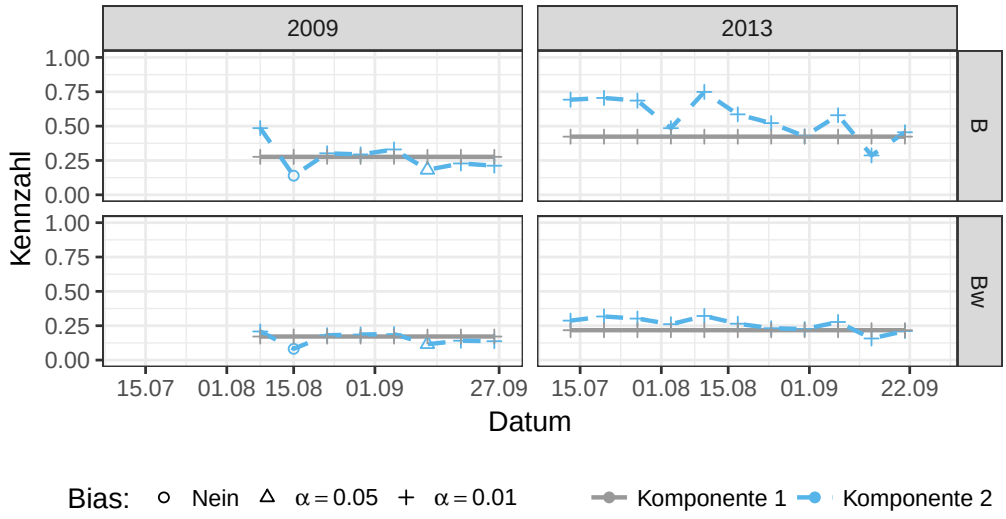
Zeitlich weisen die Wochenstichproben der Komponente 2 der GLES nur geringe Abstände von wenigen Wochen zu den 2009 und 2013 stattgefundenen Bundestagswahlen auf. Es ist trotzdem möglich, dass sich mit einer zunehmenden zeitlichen Nähe zur Wahl auch die Abbildung des Ergebnisses der jeweiligen Bundestagswahl verbessert. Die Abbildung 9.8 weist jedoch ein (beinahe) durchgehend vergleichbares Niveau der Erhebungen eines Jahres auf. In 2009 ist die Stärker der berechneten Abweichungen zum Wahlergebnis an Hand von B und B_w unabhängig vom Zeitpunkt der Erhebung. Für das Jahr 2013 zeigt

[6] Siehe hierzu auch die Darstellung der Abbildung 7.1 in Kapitel 7.2.3.

[7] Für das PS-Gewicht wurde die Inverse dieser genutzt und auf 1 normalisiert. Für das CEM wurde das fünfstufige politische Interesse in drei Kategorien unterteilt („(sehr) gering“, „mittel“, „(sehr) stark“). Die Subklassifikation an Hand der PS basiert auf sechs Klassen. Die Definition des „caliper“ für den Nearest-Algorithmus ist das 0.25-fache der Standardabweichung der PS. Der Full-Algorithmus weist ein Minimum von 0.1 und ein Maximum von zwei Kontrolleinheiten pro Treatmenteinheit auf. Die Gewichtungsvariablen über das MatchIt-Paket in R sind automatisch auf ein arithmetisches Mittel von 1 normiert.

[8] Für 2009 ergibt sich ein Ausfall mit $\bar{x} = 3.32$ % und $\tilde{x} = 3.44$ %, für 2013 $\bar{x} = 3.02$ % und $\tilde{x} = 3.00$ %. Dieser ist bedingt durch den Ausschluss der Kategorien „anderer Schulabschluss“ und „noch Schüler“ des Bildungshintergrunds.

Abbildung 9.8.: Verzerrung Zweitstimmenanteile (B, B_w) Komponenten 1 und 2 (GLES) für 2009 und 2013

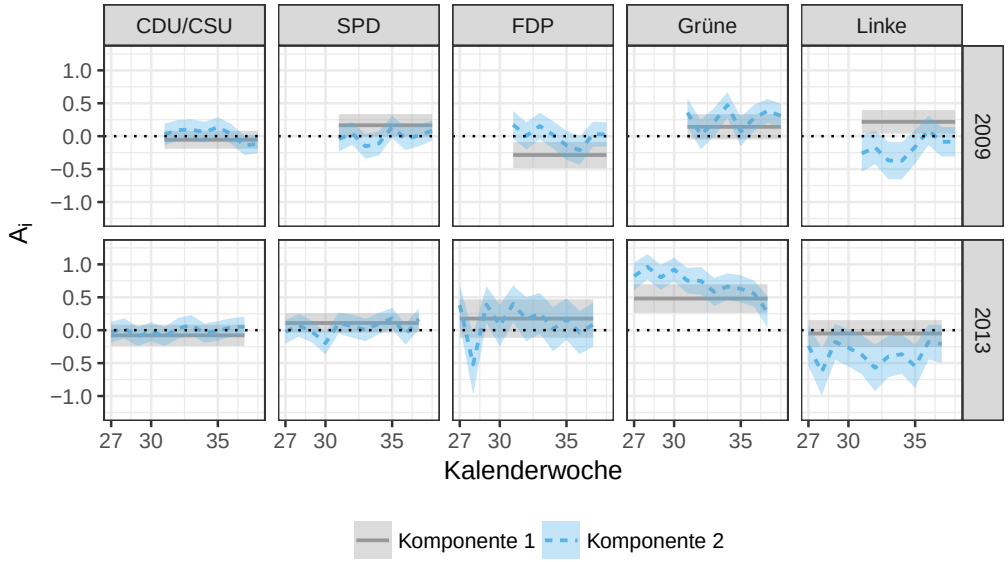


Quelle: Eigene Darstellung. Dargestellt sind die Verzerrungen der Erfassung der Zweitstimmenanteile der wöchentlichen RCS-Stichproben und der Komponente 1 in Referenz zu den Ergebnissen der Bundestagswahl gemäß B und B_w (Gewichte: *w_tranw* (RCS) und *w_trow* (Komponente 1)). Die Komponente 1 wurde als gesamte Stichprobe genutzt und das komplexe Erhebungsdesign über *surveybias* berücksichtigt. Die Gruppe der Wochenstichproben über *surveybiasi* berechnet (siehe hierzu Arzheimer/Evans 2016: 143 f.)).

sich hingegen eine leichte Tendenz der Verbesserung hinsichtlich B, wenn es zu einer zunehmenden Nähe zur kommenden Bundestagswahl kommt. Dies ist auf eine steigende Präzision der Erfassung der, relativ gesehen, kleineren Parteien zurückzuführen, da es hinsichtlich B_w nicht zu diesen Veränderungen kommt. Komparativ gesehen weisen zudem die telefonisch erhobenen Stichproben keinen erkennbaren Nachteil zu den beiden persönlichen Befragungen der Komponente 1 auf. Die, im Vergleich zu dieser, gezeigten selektiven Entscheidungen der Teilnahme an der Befragung der Komponente 2 in Abbildung 9.7 hat somit keine Auswirkung auf Unterschiede des Wahlverhaltens hinsichtlich der *gesamten* Qualität.

Die Abbildung 9.9 zeigt zudem, dass die Genauigkeit der Erfassung der Zweitstimmenanteile der jeweiligen Bundestagswahl sich auch auf der Parteienebene nicht unterscheidet. Hierzu ist der Verlauf von A_i, als Basis der Bestimmung von B und B_w, und der zugehörigen 95 %-Konfidenzintervalle, separiert nach den jeweiligen aufgeführten Parteien, ausgewiesen. Über beide Komponenten hinweg zeigt sich ein zu hoher erfasster Anteil für „B90 / Die Grünen“ in 2013. Für die Anhänger von „Die Linke“ ist es hingegen schwieriger adäquat in Telefonumfragen berücksichtigt zu werden. Die Komponente 1 hingegen schätzt

Abbildung 9.9.: Verzerrung Zweitstimmenanteile (A_i) Komponenten 1 und 2 (GLES) für 2009 und 2013 – I

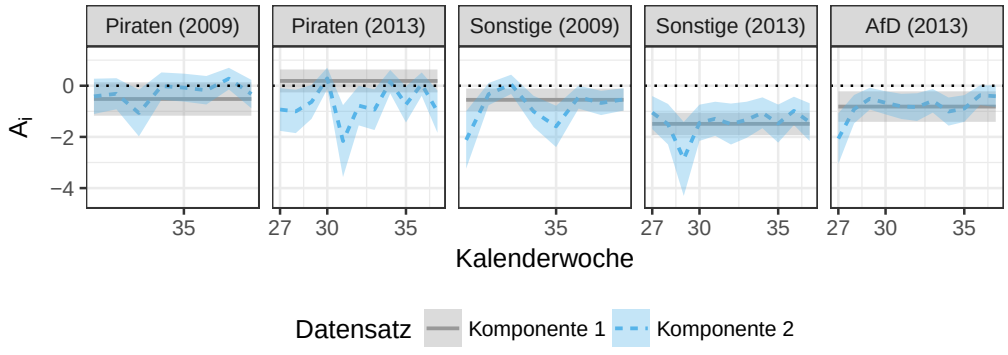


Quelle: Eigene Darstellung. Für die Komponente 1 wurde das komplexe Erhebungsdesign über *surveybias* berücksichtigt. Die Gruppe der Wochenstichproben über *surveybiasi* berechnet (siehe hierzu Arzheimer/Evans 2016: 143 f.)).

deren Anteil 2009 zu hoch ein, denjenigen der FDP als zu gering. Insgesamt sind diese Abweichungen gemäß A_i mit ca. ± 1 gering, es finden sich kaum statistisch signifikante Abweichungen für $\alpha = 0.05$ von Null. Die Anteile von Union und SPD werden zudem gleich gut eingeschätzt. Dies erklärt die niedrige Kennzahl B_w über alle Erhebungen hinweg.

Warum es zu einem starken Unterschied zwischen B und B_w in Abbildung 9.9 für das Jahr 2013 kommt, klärt die Abbildung 9.10 auf. Die dort dargestellte

Abbildung 9.10.: Verzerrung Zweitstimmenanteile (A_i) Komponenten 1 und 2 (GLES) für 2009 und 2013 – II



Quelle: Eigene Darstellung. Vorgehen analog zu Abbildung 9.9, Skalierung Y-Achse jedoch variiert.

Entwicklung von A_i für dieses Jahr zeigt zu geringe Anteile für die AfD und die „sonstigen Parteien“. Die Stärke der verzerrten Erfassung liegt, an Hand der angepassten Skalierung der Abbildung, viel höher. Einzig auf der Basis der letzten Wochenstichprobe der Komponente 2 läge der wahre Anteil der AfD bei der Bundestagswahl 2013 noch im Bereich des Wahrscheinlichen. Dies deutet auf starke Probleme hin, diese beiden Kategorien adäquat auf der Grundlage zufallsbasierter und intervieweradministrierter persönlicher und telefonischer Befragungen zu erfassen.

Lassen sich diese Fehleinschätzungen durch die Korrektur hinsichtlich der verzerrten Erfassung grundlegender soziodemografischer Rahmenfaktoren korrigieren? Die Komponenten der GLES werden mit einem „Iterative Proportional Fitting“ (IPF)-Korrekturgewicht bereitgestellt, welches eine Angleichung an die Randverteilungen des Mikrozensus 2012 hinsichtlich der verwendeten soziodemografischen Variablen erlaubt.[9] Ebenfalls lässt sich das zu hoch erfasste politische Interesse der telefonischen Befragungen über die Korrekturmodelle durch ein PSM negieren. Die Zweitstimmenanteile wurden daher, unter Anwendung der jeweiligen Bedeutungsgewichte, neu berechnet und darauf aufbauend ebenfalls erneut B und B_w bestimmt.

In Tabelle 9.3 sind die Ergebnisse von vier Regressionsanalysen dargestellt. Die abhängige Variable bilden die, über alle möglichen Bedeutungsgewichte hinweg berechneten, B und B_w des jeweiligen Jahres. Da die Genauigkeit einer Schätzung zudem von der Anzahl der befragten Personen und der kalendarischen Nähe zu einer Wahl abhängt, wurden diese beiden Variablen berücksichtigt. Anschließend wurde der Erfolg der Gewichtungsverfahren über die möglichen Modelle und die Art der technischen Bestimmung dieser bestimmt.

Je mehr Personen befragt werden, umso besser lassen sich die Zweitstimmenanteile auch erfassen. Dies sollte lediglich nicht der Fall sein, wenn eine systematische Verzerrung vorliegt. In diesem Fall führt eine Steigerung der befragten Personen zu einer Intensivierung der Abweichung von den wahren Populationsparametern. In 2013 erreichen zudem die Stichproben mit einem stärkeren zeitlichen Abstand zur Wahl ein schlechteres Ergebnis. Dies gilt für 2009 nicht, dort liegt jedoch bereits ein insgesamt sehr niedriges Niveau der Verzerrungen vor (Abbildung 9.8).

Die Anwendung des von der GLES bereitgestellten IPF-Gewichts erhöht hingegen über alle Erhebungen hinweg die Fehlprognose, B und B_w steigt.

[9] Für die Komponente 2 sind dies das Geschlecht, das Alter, der Bildungshintergrund, die BIK-Regionsgrößenklassen und die Gebietszugehörigkeit (Rattinger et al. 2014b: 11).

Tabelle 9.3.: Modell: Veränderung B und B_w Komponente 2 durch Gewichtung

Abhängige Variable:	B	B	B_w	B_w
Jahr:	2009	2013	2009	2013
Konstante	1.048^{***}	0.353^{***}	0.674^{***}	0.192^{***}
	(0.182)	(0.129)	(0.100)	(0.060)
Befragte Personen (n)	-0.001^{***}	-0.00001	-0.001^{***}	-0.00003
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Wochen bis Wahl	-0.008	0.030^{***}	-0.013^{***}	0.012^{***}
	(0.007)	(0.003)	(0.004)	(0.002)
IPF-Gewicht (GLES)	0.021	0.010	0.006	0.024
	(0.046)	(0.052)	(0.025)	(0.024)
Angleichung: Allbus		Referenzkategorie		
Angleichung: GLES	0.002	0.034^{**}	0.008	0.015^{*}
	(0.015)	(0.016)	(0.008)	(0.008)
Korrektur: Soziodemografie		Referenzkategorie		
+Interesse: Politik	-0.006	-0.001	0.004	-0.001
	(0.031)	(0.034)	(0.017)	(0.016)
Interaktion: Interesse Politik	0.002	0.002	0.001	0.001
und Wochen bis Wahl	(0.006)	(0.005)	(0.003)	(0.002)
Transformationsgewicht		Referenzkategorie		
PS-Gewicht	-0.033	-0.035	-0.020	0.006
	(0.038)	(0.042)	(0.021)	(0.020)
Matching: Subklass.	-0.033	-0.017	-0.020	0.011
	(0.038)	(0.043)	(0.021)	(0.020)
Matching: Nearest	-0.034	-0.026	-0.026	0.003
	(0.039)	(0.043)	(0.021)	(0.020)
Matching: CEM	-0.042	-0.008	-0.012	0.019
	(0.039)	(0.043)	(0.021)	(0.020)
Matching: Full	-0.018	-0.041	0.013	0.001
	(0.038)	(0.043)	(0.021)	(0.020)
Beobachtungen	176	242	176	242
Korrigiertes R^2	0.291	0.395	0.211	0.349

Note: *p<0.1; **p<0.05; ***p<0.01

Quelle: Eigene Darstellung. Angegeben sind die Regressionskoeffizienten und ihre Standardfehler in Klammern. Da es sich um eine Evaluation des Erfolg der Anwendung von Bedeutungsgewichten handelt, haben die darauf basierenden Testaussagen lediglich die Interpretation der Stärke eines Abweichens von Null.

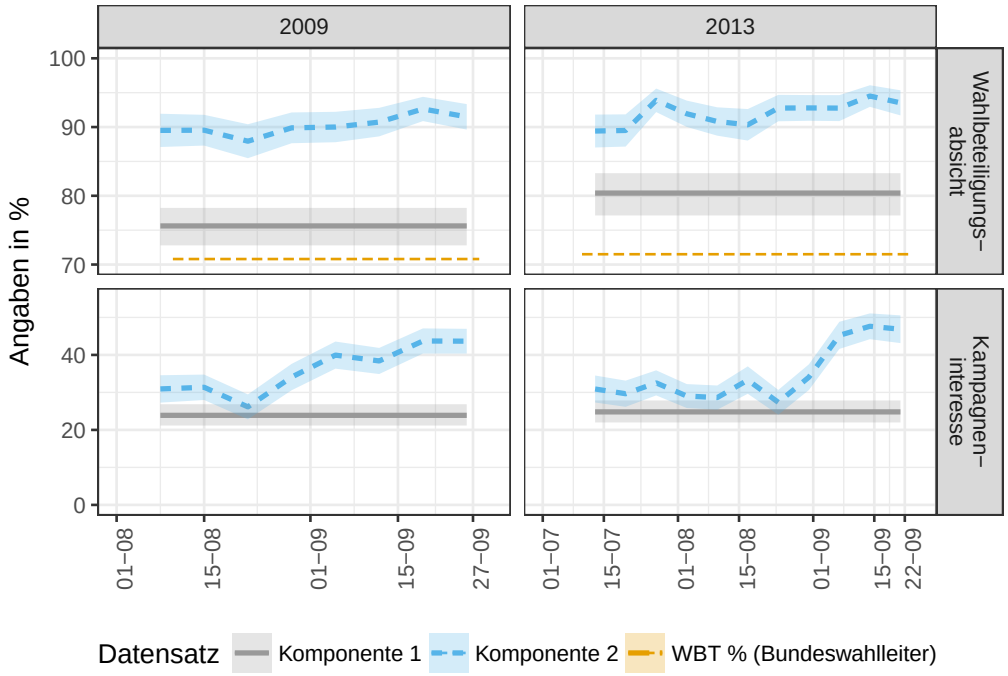
Hinsichtlich der angewandten Korrekturmodelle ist die Bestimmung der Bedeutungsgewichte auf der Basis der Komponente 1 als Referenz, gegenüber dem Allbus, etwas schlechter. Wird in die Korrektur, neben den soziodemografischen Merkmalen, zusätzlich das Interesse an Politik berücksichtigt, kommt es nur zu einer marginalen Verbesserung, oder sogar zu einer Verschlechterung. Daher ist es auch wenig erstaunlich, dass die Berücksichtigung des zeitlichen Abstands zur Wahl keinen Interaktionseffekt zu dieser Korrektur aufzeigt. Zusätzlich ist der Erfolg der jeweiligen Korrektur ausgegeben. Die aufgeführten Effekte treffen auf eine Kombination der Korrektur hinsichtlich der Soziodemografie mit der Nutzung des Allbus als Referenzdatensatz zu. Es lässt sich kein wesentlicher Unterschied feststellen. 2009 erreichen die Bedeutungsgewichte auf der Basis der verschiedenen Algorithmen eine Verbesserung durch eine Reduktion von B um -0.018 bei Anwendung des Full-Algorithmus und -0.042 durch das CEM-Matching. Hinsichtlich B_w dreht sich diese Betrachtung in vergleichbarer Stärke zu Gunsten des Full-Algorithmus. Hinsichtlich des Ausgangsniveau 2009 von ca. 0.25 in Abbildung 9.8, lässt sich dies als Erfolg bezeichnen. Da sich dies 2013 jedoch bereits abschwächt (B) oder die Verzerrung durch die Anwendung der Gewichte erhöht (B_w), lässt sich als Fazit ziehen: Im Vergleich zu persönlichen weisen telefonische Befragungen Verzerrungen hinsichtlich der Erfassung soziodemografischer Merkmale und des politischen Interesse auf. Diese Verzerrungen äußern sich jedoch nicht in einer reduzierten Qualität der Erfassung der Wahlergebnisse im Kontext der Bundestagswahlen 2009 und 2013. Daher führt die Angleichung im Rahmen der Bestimmung von Bedeutungsgewichten auch nicht zu einer Verbesserung. Die Probleme der Erfassung der AfD, der Kategorie „Sonstige“ und der Piraten kann die Korrektur daher nicht lösen.

9.2.2. Wahlbeteiligung und das Interesse am Wahlkampf

Dass die Korrektur durch die erstellten Bedeutungsgewichte keinesfalls obsolet ist, zeigen die Änderungen der Höhe der Wahlbeteiligungsabsicht und das angegebene Interesse am Wahlkampf.[10] Die Abbildung 9.11 zeigt den Verlauf der Anteilswerte dieser, inklusive der berechneten Konfidenzintervalle, für beide Jahre der Erhebungen der Wochenstichproben der Komponente 2.

[10] Diese wurden jeweils auf einer fünfstufigen Likert-Skala abgefragt. Die Angaben „bestimmt zur Wahl gehen“, „wahrscheinlich zur Wahl gehen“ und „habe bereits per Briefwahl meine Stimme abgegeben“ der Wahlbeteiligungsabsicht (*pre003* für 2009 und 2013) wurden als „Wähler“ interpretiert. Als ein Interesse am Wahlkampf (*pre002* für 2009 und 2013) wurden die Angaben „stark“ und „sehr stark“ kodiert. Selbiges gilt für die Erhebungen der Komponente 1 (2009: *q3* und 2013: *q4*).

Abbildung 9.11.: Wahlbeteiligungsabsicht und Wahlkampfinteresse Komponenten 1 und 2 (GLES) für 2009 und 2013



Quelle: Eigene Darstellung. Gewichte: w_trow (Komponente 1) und w_tranw (Komponente 2). Flächen stellen den Bereich der 95 %-Konfidenzintervalle dar. Standarfehler Komponente 2 nach Formel 6.5, Komponente 1 unter Berücksichtigung des komplexen regionalen Erhebungsdesigns nach Formel 6.7. $deft$ für die Wahlbeteiligungsabsicht ist 1.48 (2009) und 1.73 (2013). Für das Kampagneninteresse 1.58 (2009) und 1.54 (2013).

Ebenfalls sind die Anteilswerte und Konfidenzintervalle für die regional stratifizierten Stichproben der Komponente 1 angegeben. Für die Wahlbeteiligungsabsicht wurde zudem die Wahlbeteiligungsrate der jeweils nachfolgenden Bundestagswahl als Referenz eingezeichnet. Das telefonische Befragungen einen zu hohen Anteil an Wählern erfassen, (siehe die Abbildung 7.8 in Kapitel 7.3 oder die Darstellung im vorangegangenen Kapitel) bestätigt sich auch durch den Verlauf der Wochenstichproben der Komponente 2. 2009 finden sich Beteiligungsraten von ca. 90 %, welche in 2013 noch einmal höher ausfallen. Dort liegen die Beteiligungsraten um bis zu 15 Prozentpunkte über dem amtlichen Endergebnis. Die beiden Stichproben der Komponente 1 ergeben klar niedrigere Werte mit ca. 75 % in 2009 und ca. 80 % in 2013 und liegen somit näher an dem amtlichen Referenzergebnissen. Das Interesse am Wahlkampf zeigt an Hand der Erhebungen der Komponente 2 ein anfangs vergleichbares Niveau von ca. 25 % bis 30 % interessierten Personen. Mit zunehmender Nähe zur Bundestagswahl intensiviert sich dies auf über 40 % in 2009 und knapp unter 50 % in 2013.

Zwei Erklärungen sind denkbar: Die Nähe einer Bundestagswahl lässt

tatsächlich das Interesse am Wahlkampf stark zunehmen. In diesem Fall erfassen die RCS-Stichproben eine reale Dynamik im Vorfeld der Wahl. Es ist aber auch möglich, dass mit zunehmender Nähe zur Wahl politisch interessierte Personen vermehrt an Wahlumfragen aus geweckter Neugier partizipieren. In diesem Fall sind die telefonischen Wochenstichproben von einem zunehmendem systematischen Nonresponse betroffen, da nun eine Gruppe systematisch häufiger beginnt an Wahlumfragen teilzunehmen. Die Ergebnisse der Regressionsanalysen in Tabelle 9.4 bestätigen diese Vermutung. Das Prinzip des Vorgehens der Modellierung ist analog zur vorherigen Tabelle 9.3. An der Stelle von B und B_w werden nun die Anteilswerte der Wahlbeteiligung und des Wahlkampfinteresses als abhängige Variablen vorhergesagt. Die aufgeführten „Wochen bis zur Wahl“ unterstreichen den visuellen Eindruck: Je näher eine Bundestagswahl kommt, umso mehr Personen geben in Umfragen an, auch an dieser partizipieren zu wollen. Das Interesse am Bundestagswahlkampf intensiviert sich ebenfalls. Dieser Eindruck entsteht jedoch nur durch eine Selektivität der Teilnahme. Bereits die Angleichung an die Rahmendaten der amtlichen Statistik über das IPF-Gewicht der GLES führt zu einer Reduktion um 2 (2009) und 2.8 (2013) Prozentpunkte. Auch der Anteil der interessierten Personen am stattfindenden Wahlkampf reduziert sich um 3.5 (2009) und 4.5 (2013) Prozentpunkte.

Die Korrektur über eine Angleichung an einen geeigneten Referenzdatensatz entwickelt einen noch stärkeren Effekt. Grundsätzlich ist die Angleichung an die Komponente 1 der GLES erfolgreicher als die Angleichung an den Allbus. Ebenso führt die zusätzliche Berücksichtigung des politischen Interesse zu einer klaren Verbesserung. Diese Verbesserung ist zudem, auf der Grundlage des angegebenen Interaktionseffekts, umso erfolgreicher, je *geringer* der Abstand zur kommenden Bundestagswahl ist. Hinsichtlich der möglichen technischen Varianten der Angleichung erzielt der Full-Algorithmus durchgehend den höchsten Grad der Reduzierung der Anteile. Eine Reduktion von ca. 4.5 Prozentpunkten hinsichtlich der Wahlbeteiligung und von ca. acht bis neun Prozentpunkten bei Betrachtung des Interesse am Wahlkampf ist durch die Anwendung der Bedeutungsgewichte möglich. Insbesondere die Steigerung des Wahlkampfinteresse ist somit teilweise ein Artefakt. Die Intensivierung kommt vor allem zu Stande, da Personen mit einem vorhandenen Interesse mit zunehmender Nähe zu einer Bundestagswahl auch mit einer höheren Wahrscheinlichkeit an einer Wahlumfrage partizipieren.

Tabelle 9.4.: Modell: Veränderung Wahlbeteiligungsabsicht und Wahlkampfinteresse Komponente 2 durch Gewichtung

Abhängige Variable:	Wahlbeteiligung		Wahlkampfinteresse	
Jahr:	2009	2013	2009	2013
Konstante	85.645***	82.128***	47.500***	50.414***
	(2.077)	(1.677)	(5.341)	(4.727)
Befragte Personen (n)	0.008***	0.016***	−0.002	−0.006
	(0.002)	(0.002)	(0.006)	(0.006)
Wochen bis Wahl	−0.330***	−0.264***	−2.274***	−1.814***
	(0.083)	(0.046)	(0.214)	(0.130)
IPF-Gewicht (GLES)	−2.038***	−2.808***	−3.553**	−4.454**
	(0.533)	(0.683)	(1.372)	(1.924)
Angleichung: Allbus	Referenzkategorie			
Angleichung: GLES	−1.281***	−1.288***	−2.134***	−2.294***
	(0.170)	(0.216)	(0.436)	(0.610)
Korrektur: Soziodemografie	Referenzkategorie			
+Interesse: Politik	−1.383***	−1.078***	−3.107***	−2.675***
	(0.169)	(0.216)	(0.434)	(0.608)
Interaktion: Wochen bis	0.165**	0.081	0.333*	0.263
Wahl und +Interesse Politik	(0.070)	(0.065)	(0.181)	(0.184)
Transformationsgewicht:	Referenzkategorie			
PS-Gewicht	−1.363***	−0.505	−1.870*	−0.326
	(0.439)	(0.562)	(1.130)	(1.582)
Matching: Subklass.	−1.954***	−1.140**	−3.161***	−1.827
	(0.441)	(0.562)	(1.134)	(1.585)
Matching: Nearest	2.083***	2.037***	3.065***	3.065*
	(0.444)	(0.565)	(1.143)	(1.592)
Matching: CEM	−1.215***	−0.564	−2.629**	−0.456
	(0.441)	(0.563)	(1.135)	(1.585)
Matching: Full	−2.195***	−2.054***	−3.724***	−3.395**
	(0.441)	(0.562)	(1.134)	(1.585)
Beobachtungen:	176	242	176	242
Korrigiertes R^2:	0.800	0.614	0.809	0.628

Note: *p<0.1; **p<0.05; ***p<0.01

Quelle: Eigene Darstellung. Erläuterungen identisch zu Tabelle 9.3.

9.3. Telefonische Wahlumfragen und Verzerrungen: Zwischenfazit

Als Zwischenfazit lässt sich an dieser Stelle festhalten: Gegenüber persönlichen Befragungen sind telefonische Wahlumfragen mit einer stärkeren selektiven Teilnahme konfrontiert. Bereits durch die Altersangabe, den Bildungshintergrund, die Wohnsituation und dem politischen Interesse finden sich ausreichende Unterschiede der Stichproben beider Befragungsformen. Diese Effekte treffen zudem nicht nur im direkten Kontext einer Bundestagswahl zu, sondern zeigen eine starke Stabilität und klare Entwicklungsmuster im Zeitverlauf seit 1994, wie das Kapitel 9.1.1 gezeigt hat.

Ob dies die Verteilungen inhaltlich interessierender Variablen betrifft, ist ambivalent. Im Rahmen der Arbeit konnte im Kontext der Bundestagswahlen 2009 und 2013 gezeigt werden: Telefonische Befragungen erreichen ohne tiefgreifende Korrekturen ein vergleichbares Qualitätsniveau zu persönlichen Befragungen hinsichtlich des Wahlverhaltens. Die Beteiligungsabsicht an Wahlen und das Interesse am Wahlkampf im Vorfeld ergibt jedoch ein anderes Bild. Die Selektivität der Teilnahme sorgt bei telefonischen Befragungen für eine systematische Verzerrung, durch welche die Beteiligungsabsicht durchgängig zu hoch ausfällt. Die Verzerrungen lassen sich zudem nicht durch grundlegende soziodemografische Faktoren erklären, ihre Ursachen liegen vielschichtiger. Im Rahmen der Arbeit konnte dies für das Interesse an Politik herausgestellt werden. Politisch desinteressierte Personen verabschieden sich zunehmend aus Telefonumfragen, was Auswirkungen wiederum auf die gemessene Wahlbeteiligungsabsicht hat (siehe Abbildung 9.4). Dieser Wirkungskanal der Verzerrung hat sich erst seit dem Kontext der Bundestagswahl 2013 relativiert, welches mit dem Auftreten der AfD koinzidiert. Sofern diese politisch weniger interessierte Personen zur Beteiligung motiviert *und* das politische Interesse mit der Wahlbeteiligungsabsicht korreliert, verschwindet der Kanal der Korrektur. Auch weniger interessierte Personen wählen nun häufiger, die Korrektur funktioniert nicht mehr.

Telefonische Befragungen, auf der Basis einer einer zufallsbasierten Auswahl, können daher insgesamt von einer Angleichung an zufallsbasierte Erhebungen auf der Grundlage persönlicher Interviews profitieren. Dies funktioniert, da der zufallsbasierten Auswahl bei telefonischen Befragungen eine nicht-zufällige Entscheidung zur Teilnahme folgt. Betreffen die resultierenden Verzerrungen weniger die Qualität der Erfassung der Wahlentscheidung, so kommt es doch bereits zu starken Unterschieden hinsichtlich der Wahlbeteiligungsabsicht. An

dieser Stelle können Bedeutungsgewichte systematisch erfolgreich ansetzen, ohne weitere Verzerrungen des Wahlverhaltens hervorzurufen. Mit Bezug auf den Vergleich der Komponente 1 und 2 der GLES lässt sich daher als Resümee festhalten: Beide Erhebungen visieren eine zufallsbasierte Auswahl der wahlberechtigten Bevölkerung der Bundesrepublik an. Im Vergleich zur Komponente 1 erreicht die Komponente 2 jedoch eine Grundgesamtheit, welche politisch interessierter ist. Daher funktionierte die hier präsentierte Korrektur besser.

Nachdem nun umfassend das Verhältnis zwischen persönlichen und telefonischen Befragungen dargestellt wurde, wird sich das nächste Kapitel den onlinebasierten Befragungen widmen. Mit den bis hierhin dargestellten Ergebnissen steht für diese fest: Eine Referenz der Angleichung kann nur eine persönliche Befragung sein.

10. Qualität von Wahlumfragen auf der Basis von Access-Panels

Innerhalb der Umfrageforschung haben onlinebasierte Befragungen hinsichtlich ihrer Reputation einen schweren Stand. Durch einen qua Design nicht zu vermeidenden Noncoverage-Fehler (Kapitel 6.2.1) und einer mangelnden Möglichkeit der Realisierung einer zufallsbasierten Auswahl bei Bevölkerungsumfragen, (siehe Kapitel 6.2.2) ist dieses Erhebungsformat bereits apriori mit einer vorhandenen Selektivität belastet. Die Auswahl von zu befragenden Personen durch ein vorab konstruiertes Access-Panel kann diese Probleme effektiv lösen, wenn die Auswahl in dieses über einen zufallsbasierten Mechanismus, also in der Regel offline, möglich ist. Zudem muss Personen ohne Internetzugang die technische Infrastruktur bereitgestellt werden, um eine weitere Fehlerquelle hinsichtlich des Noncoverage auszuschließen. Sofern schließlich auch die Wahrscheinlichkeit der Teilnahme nicht mit der Kompetenz und Erfahrung des technischen Umgangs mit dem Internet korreliert, ist eine qualitativ hochwertige Befragung auch onlinebasiert möglich (siehe Kapitel 6.2.3).

Das nachfolgende Kapitel wird zeigen, dass dies für prominente onlinebasierte Erhebungen der Wahlforschung ohne weitere Korrekturen durch qualitativ hochwertigere Referenzumfragen in Form persönlicher Befragungen nicht zutrifft. Stehen solche Referenzumfragen jedoch zur Verfügung, ist eine Steigerung der Qualität onlinebasierter Befragungen möglich. Gelingt die Korrektur zudem durch die Nutzung zeitlich konstanter Faktoren, ist eine Nutzung von Referenzumfragen mit abweichender Feldzeit zur onlinebasierten Befragung möglich.

Die Erhebungen der Komponente 8 der „German Longitudinal Election Study“ (GLES) basieren bis einschließlich der Erhebung T16, ebenso wie diejenigen der Komponente 3, auf dem (beinahe ausschließlich) onlinerekrutiertem Access-Panel von Respondi. Seit der Erhebung T17 wird das LINK Internet Panel genutzt, welches eine offline stattfindende zufallsbasierte Auswahl auf der Basis von Telefonumfragen trifft (siehe Kapitel 5.1). Da jedoch keine technische Infrastruktur bereitgestellt wird, kommt es auch bei diesen weiterhin zu einem Ausschluss der Bevölkerung, welche über keinen Internetzugang verfügt. Beide

Access-Panel sind daher mit einem Noncoverage-Problem konfrontiert. Treffen die ausgewählten Personen zudem die Entscheidung zur Registrierung nicht unabhängig von ihrer Kompetenz mit dem Umgang des Internets, lässt sich ein zusätzlicher Nonresponsefehler nicht vermeiden. Auch sind Telefonumfragen selber mit zunehmenden Problemen konfrontiert (siehe hierzu das vorangegangene Kapitel 9), was sich in der Qualität des resultierenden LINK Internet Panel widerspiegelt.

Das nachfolgende Kapitel wird die Qualität der Befragungen dieser beiden Access-Panel in analysieren. In einem ersten Schritt werden in Kapitel 10.1 diejenigen Umfragen der Komponente 3 und 8 der GLES in den Fokus gerückt, welche im zeitlichen Kontext der Bundestagswahl 2013 erhoben wurden. Diese erste Einschränkung hat folgenden Grund: Durch die zeitliche Nähe zu einer Bundestagswahl können die wahren Werte der Zweitstimmenanteile der Wahlen als eine Referenz der Evaluierung der Auswirkung der Angleichung genutzt werden. Die Entscheidung der Betrachtung des Kontext der Bundestagswahl 2013, gegenüber derjenigen von 2009, hat zwei weitere Vorteile: Im Kontext dieser Wahl stehen *erstens* Umfragen auf der Basis des onlinrekrutierten Access-Panel von Respondi und des offlinerekrutierten von LINK zur Verfügung.[1] Damit ist eine Betrachtung des Einflusses der Art der Rekrutierung in das jeweilige Access-Panel auf die Qualität der Ergebnisse möglich. *Zweitens* wurden im Vor- und Nachgang der Wahl 2013 Kontrollquerschnitte der Wellen drei, fünf und sieben des laufenden „Wahlkampfpanel“ (WKP) der Komponente 3 durchgeführt. Neben der Nutzung der Rekrutierungswelle des WKP als einem eigenständigen Querschnitt, ergeben sich daher drei weitere Stichproben (siehe zur Beschreibung dieser auch Kapitel 5.1).

Die Erkenntnisse aus dem Kontext der Bundestagswahl 2013 wird dann das Kapitel 10.2 auf den gesamten Erhebungszeitraum der Komponente 8 übertragen. Durch die zur Verfügung stehenden 31 Erhebungen dieser Komponente[2] bietet sich die Möglichkeit der Nutzung des Zeitraums von Juli 2009 bis September 2016. Hierdurch lassen sich die spezifischen erfassten Ergebnisse des Kontext der Bundestagswahl 2013 über einen längeren Zeithorizont validieren und als systematische Möglichkeit der Korrektur für die jeweiligen Stichprobendesigns beider genutzter Access-Panel herausstellen. Ein Fazit über den Effekt

[1] Im Kontext der Wahl 2009 wurden ebenfalls Umfragen der Komponenten 3 und 8 erhoben. Da die Komponente 8 bis zur Umfrage T16 auf dem Respondi-Panel beruht, ergibt sich diese Möglichkeit nicht.

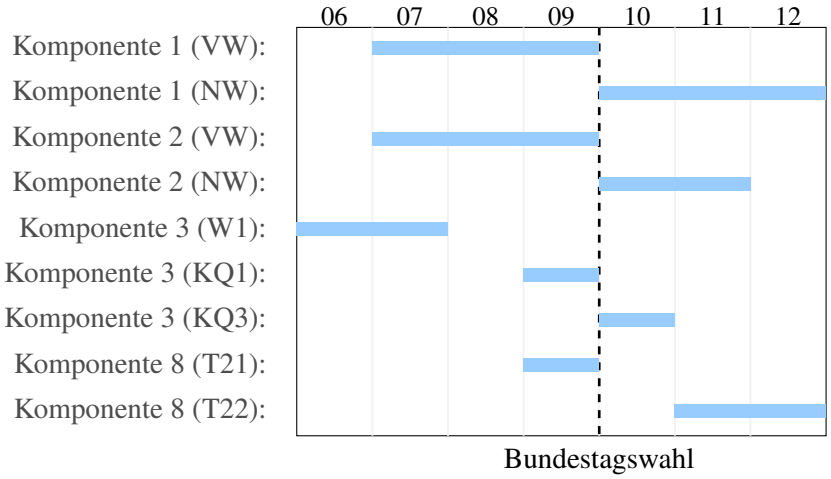
[2] Die ersten beiden Erhebungen (T1, T2) werden nicht berücksichtigt, da elementare Variablen nicht abgefragt wurden. Stand: 10.01.2017.

des Wechsels von einem onlinerekrutierten zu einem offlinerekrutierten Access-Panel ist hierdurch ebenfalls möglich. Im Gegensatz zu den zufallsbasierten Erhebungen der Komponenten 1 und 2 visieren die Stichproben der Komponenten 3 und 8 mit einem Inferenzschluss lediglich die Grundgesamtheit des jeweiligen Access-Panels an. Da dieser Kreis, bereits auf grundlegenden soziodemografischen Variablen, eine systematische Abweichung zur volljährigen Bevölkerung des Mikrozensus aufweist (Kapitel 5.4). Die Ausweisung von Standardfehlern und die Bestimmung von Konfidenzintervallen in Kapitel 10.2 trifft daher keinen Inferenzschluss auf die gesamte Bevölkerung, sondern eine Aussage über die Unterschiede der Grundgesamtheiten der aufgeführten persönlichen und onlinebasierten Befragungen. Stellen erstere ein Abbild der allgemeinen (wahlberechtigten) Bevölkerung dar, wird somit diese mit der Population der genutzten Access-Panel verglichen.

10.1. Der Kontext der Bundestagswahl 2013

Im Vorfeld der Bundestagswahl 2013 wurde eine Reihe von Umfragen im Rahmen der GLES durchgeführt. Dies trifft u.a. auf die Komponenten 1, 2, 3 und 8 des Kapitels 5 zu. Die Abbildung 10.1 skizziert die Feldzeit dieser und ihren (vergleichsweise) geringen Abstand zu dieser Bundestagswahl.

Abbildung 10.1.: Feldzeit Komponenten 1, 3 und 8 (Bundestagswahl 2013)



Quelle: Eigene Darstellung. Markierte Bereiche sind die Erhebungsmonate.

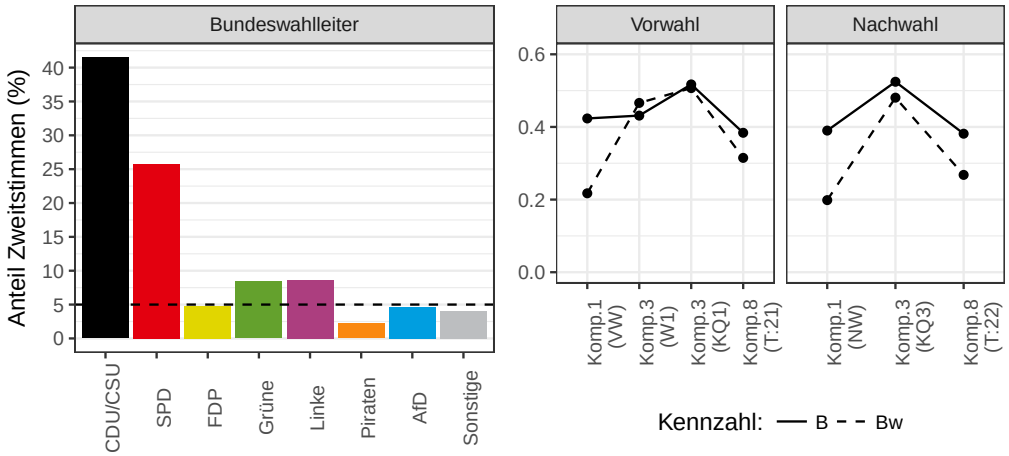
Der längste Abstand der Beendigung der Feldzeit und der durchgeführten Wahl ergibt sich für die Rekrutierungswelle des Wahlkampfpanels der Komponente 3, was für diese Erhebung einen Nachteil des Vergleichs zu den amtli-

chen Endergebnissen der Wahl 2013 darstellt. Die letzten Dynamiken und sich verändernde Wahlpräferenzen können durch diese Erhebung nicht korrekt erfasst werden.

Auf der Grundlage des vorangegangen Kapitels 9.2 hat sich hinsichtlich der Qualität der Erfassung der Zweitstimmenanteile kein wesentlicher Unterschied zwischen der Komponente 1 und 2 ergeben (Abbildung 9.8 in Kapitel 9.2). Telefonische Befragungen sind jedoch mit einem Selektionseffekt hinsichtlich des Bildungsstand, Alter und des politischen Interesse der befragten Personen konfrontiert (siehe hierzu die Darstellungen in Kapitel 8). Als Referenz für die nicht-zufallsbasierten Stichproben der Onlineumfragen wird daher nur die zufallsbasierte Erhebung der Komponente 1 berücksichtigt. Um einen besseren Vergleich der Nachwahlerhebungen T22 (Komponente 8) und „Kontrollquerschnitt“ (KQ) 7 (Komponente 3) zu garantieren, wird an wenigen Stellen die nach der Wahl durchgeführte Auswahl und Befragung der Komponente 1 genutzt. Eine Angleichung an die persönlichen Befragungen der Komponente 1 wird hingegen nur auf der Basis der vor der Wahl ausgewählten und befragten Personen der Komponente 1 durchgeführt. Hierdurch wird gezeigt, dass eine Korrektur auch trotz eines zwischen Referenz- und interessierender Umfrage stattfindenden zentralen Ereignis möglich ist.

Wie gut die in Abbildung 10.1 dargestellten Umfragen das amtlich ermittelte Ergebnis der Bundestagswahl 2013 erfassen können, zeigt die Abbildung 10.2.

Abbildung 10.2.: B und B_w für Komponenten 1, 3 und 8 (GLES) im Kontext der Bundestagswahl 2013



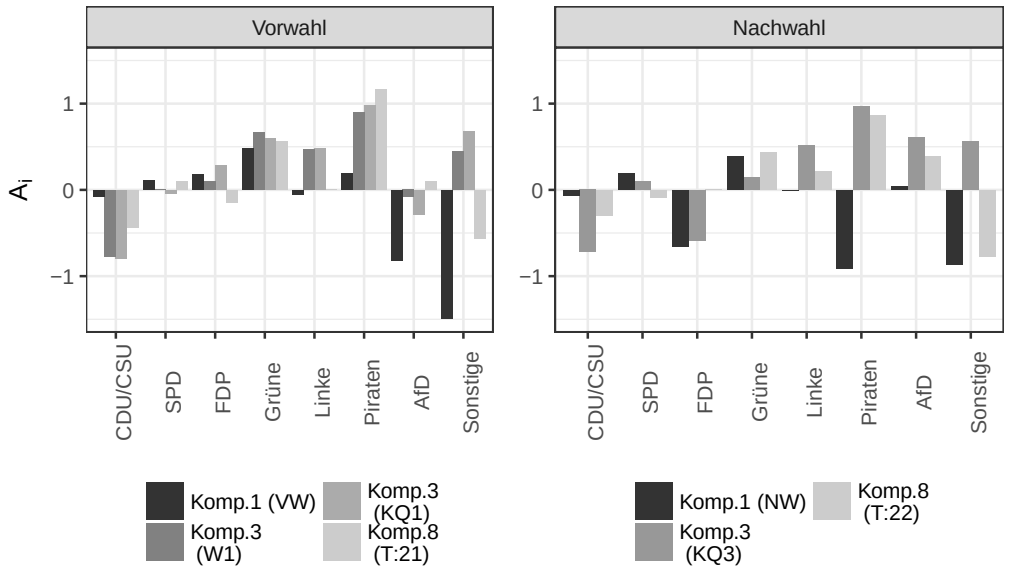
Quelle: Eigene Darstellung. Referenzdaten von B und B_w der Komponenten sind die Daten des Bundeswahlleiters.

Der zeitliche Abstand zur Wahl zeigt keinen Einfluss auf die Genauigkeit der Erhebungen der Komponente 3. Die Rekrutierungswelle des WKP weist, trotz

des längsten Abstands zur Wahl, die beste Erfassung an Hand von B und B_w auf. Im Vergleich zur Komponente 1 zeigen jedoch alle onlinebasierten Erhebungen stärkere Schwierigkeiten der Erfassung der prozentual größeren Parteien, B_w ist stärker erhöht als B. Dies weist auf Schwierigkeiten der Erfassung der Parteien mit einem vergleichsweise großen Anteil der Zweitstimmen hin (Arzheimer/ Evans 2016: 141 f.).

Welche Partei oder Parteien dies genau sind, stellt die Abbildung 10.3 dar. Ausgewiesen ist die Kennzahl A_i der jeweiligen Parteien, auf welchen die Berechnung von B und B_w basiert.

Abbildung 10.3.: A_i für Komponenten 1, 3 und 8 (GLES) im Kontext der Bundestagswahl 2013



Quelle: Eigene Darstellung. Komponente 1 wurde gewichtet (w_trow).

Die erhöhten Werte von B_w der Komponenten 3 und 8 sind auf die zu geringen Anteile der Unionsparteien zurückzuführen. Die Befragungen der Komponente 1 zeigen hingegen nur geringfügige Abweichungen zu den amtlichen Endergebnissen. Zeigt sich hinsichtlich der SPD kein Einfluss der Befragungsform, weist die FDP in den Befragungen nach der Wahl innerhalb der Komponenten 1 und 8 deutlich niedrigere Werte auf. Für die Komponente 3 trifft dies nicht zu, der „Sympathieeinbruch“ als Folge des verpassten Einzugs in den Bundestag wird nicht korrekt erfasst.

Das Potential der Partei „B90 / Die Grünen“ wird von allen Erhebungen, in Referenz zu den Ergebnissen des Bundeswahlleiters, durchgängig als zu hoch eingeschätzt. Die onlinebasierten Erhebungen zeigen zudem ein zu optimistisches Bild hinsichtlich des Potentials für „Die Linke“ und die Piraten. Ein

Vorteil dieser ist jedoch die konstante Erfassung des Zweitstimmenanteils der „Alternative für Deutschland“ (AfD) im Vorfeld der Wahl, welche mit dem amtlichen Anteil nahezu übereinstimmen. Waren die telefonischen Erhebungen der Komponente 2 ebenfalls nicht in der Lage diese Wählergruppe adäquat zu repräsentieren (siehe hierzu Abbildung 9.10 des vorangegangenen Kapitel 9.2), kann eine Erklärung für diesen Qualitätsvorteil onlinebasierter Befragungen ein kombinierter Nonresponse- und Messfehler sein durch den Einsatz von Interviewern sein. Demnach nehmen Wähler der AfD tendenziell seltener an intervieweradminstrierten Erhebungsformen teil. Sofern sie doch zu einer Teilnahme bewegt werden können, offenbaren sie sich dem Interviewer jedoch nicht als solche. Innerhalb der Nachwahlbefragungen kommt es durchweg zu einer Erhöhung des Anteils der AfD in *allen* Komponenten. Eine mögliche Erklärung ist der nur knapp verpasste Einzug in den Bundestag, welcher eine nachträgliche Mobilisierung verursacht. Hinsichtlich der Kategorie „Sonstige“ kommen alle Umfragen zu Fehleinschätzungen. Die Komponenten 1 und 8 unterschätzen deren Anteile systematisch, die Komponente 3 überschätzt diese.

Auch die angegebene Wahlbeteiligungsabsicht unterscheidet sich zwischen den Befragungen (Tabelle 10.1). Auf der Basis des Access-Panel von LINK (T21 und T22) zeigen sich die höchsten Beteiligungsraten, welches durch die telefonische Rekrutierung in das LINK Internet Panel und die Anfälligkeit dieses Erhebungsmodus für diese Verzerrung erklärt werden kann.

Tabelle 10.1.: Wahlbeteiligungsabsicht für Komponenten 1, 3 und 8 der GLES im Kontext der Bundestagswahl 2013

Komponente:	1	3	3	8	1	3	8
Beschreibung:	VW	W1	KQ3	T21	NW	KQ7	T22
Anteil:	77	84	83	90	86	87	93

Quelle: Eigene Darstellung. Daten: Jeweilige Datensätze. Wählerklassifikation: „Sehr wahrscheinlich“, „wahrscheinlich“ teilnehmen oder „bereits Briefwahl gemacht“ (VW, W1, KQ3, T21, T22) oder „ja, habe gewählt“ (NW, KQ7). Gewichtung: Komp.1 (w_trow). Referenzdaten: Bundeswahlleiter. Anteile Umfragen auf ganze Zahlen gerundet.

Durch die stattgefundene Bundestagswahl lässt sich eine Steigerung der Beteiligungsabsicht in allen Erhebungsformaten feststellen, welche in der Komponente 1 am stärksten auftritt. Dies kann, neben einer wirklichen Steigerung, jedoch auf nun zunehmende Probleme persönlicher Befragungen hindeuten. Die Erfahrung der gerade erst stattgefundenen Bundestagswahl 2013 führt zu einer stärkeren Selektion der auch tatsächlich an der Wahl teilgenommenen Personen in die Befragung. Die Nachwahlbefragung ist daher mit stärkeren

Problemen gegenüber der Vorwahlbefragung behaftet. Bereits hinsichtlich des Wahlverhaltens und der Beteiligungsabsicht lassen sich also im Kontext der Bundestagswahl 2013 Unterschiede zwischen den Befragungsformen feststellen. Die zufallsbasierte persönliche Befragung der Komponente 1, die nicht-zufallsbasierten Onlinebefragungen des onlinerekrutiertem Access-Panel von Respondi und die ebenfalls nicht-zufallsbasierten Befragungen auf Basis des offlinerekrutiertem Access-Panel von LINK zeigen abweichende Resultate. Onlinebasierte Befragungen haben Probleme, den wahren Anteil der Unionswähler korrekt widerzuspiegeln. Dies betrifft die Befragungen der Komponente 3 in einem stärkeren Ausmaß, als diejenigen der Komponente 8.

Drei Faktoren sind hierfür nach Kapitel 6.2.3 verantwortlich: Die rein onlinebasierte Rekrutierung in das Respondi-Panel gewährt Personen ohne einen Internetzugang keine Chance der Registrierung. Zudem ist die Wahrscheinlichkeit der Registrierung nicht unabhängig von Kernmerkmalen später zu befragender Personen. Stehen diese Faktoren, wie die Kompetenz des Umgangs mit dem Internet, aber auch beispielsweise die Einstellung gegenüber aktuellen politischen Themen, in einer Verbindung zu dem Wahl- und Beteiligungsverhalten, tritt eine Verzerrung der anschließend stattfindenden Befragungen auf. Stichproben auf der Basis offlinerekrutierter Access-Panel sind nur in einer abgeschwächten Form von dieser Selektivität betroffen. Die Kenntnis über die Möglichkeit zur Registrierung in das Access-Panel ist durch die Auswahl über ein Erhebungsdesign telefonischer Befragungen unabhängig von der Nutzungskompetenz und Häufigkeit der Nutzung des Internet. Dieser Vorteil wird jedoch durch eine erhöhte reportierte Wahlabsicht begleitet, also durch die Anfälligkeit telefonischer Befragungen gegenüber bestimmten Verzerrungen.

Sind die Folgen der selektiven Teilnahme („Nonresponse“) und der mangelnden Abdeckung bestimmter Teile der Bevölkerung („Noncoverage“) der onlinebasierten Befragungen korrigierbar, lassen sich jedoch deren Vorteile nutzen (siehe zur Beschreibung dieser Kapitel 6.2). Hierzu müssen Unterschiede zwischen den onlinebasierten Erhebungen und der zufallsbasierten persönlichen Befragung als Referenzumfrage existieren. Diese Unterschiede können dann über Angleichungen der Befragungen der Komponenten 3 und 8 an die Komponente 1 im Rahmen eines „Propensity Score Matching“ (PSM) korrigiert werden. Das nachfolgende Kapitel 10.1.1 wird daher nun evaluieren, inwiefern sich die Verteilungen dieser Umfragen hinsichtlich derjenigen Variablen unterscheiden, welche in Kapitel 7.2 als „Wirkungskanäle der Korrektur“ präsentiert wurden. Dies ist möglich, da innerhalb der Komponenten der GLES-Studie eine

Vielzahl vergleichbarer Frageformulierungen existieren. Existieren Variablen, welche Unterschiede zwischen den Stichproben der Onlineumfragen und der persönlichen Referenzumfrage erklären *und* zudem mit inhaltlichen Zielvariablen dieser Befragung korrelieren, eignen sich diese zur Korrektur. Das nachfolgende Kapitel wird nun Variablen vorstellen, welche sich zu einem Korrekturmodell verbinden lassen.

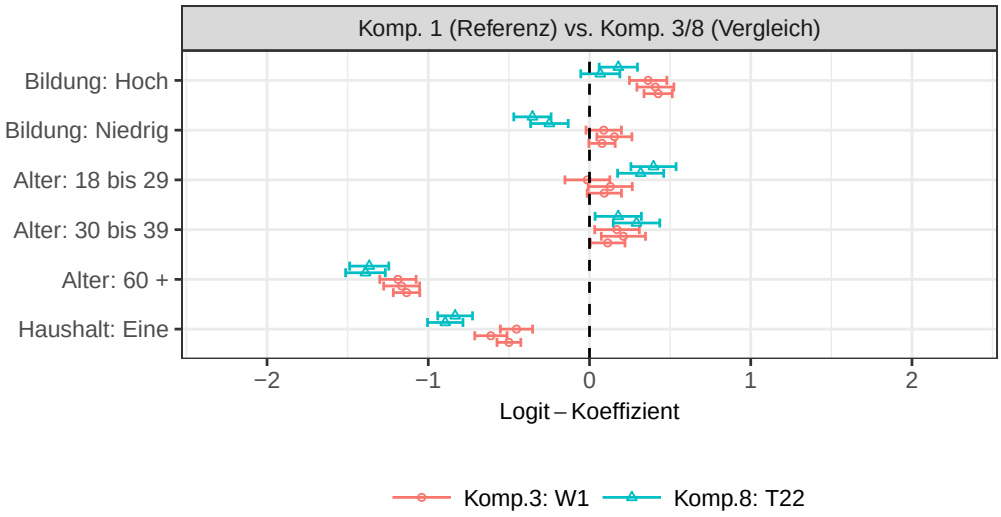
10.1.1. Korrekturmodelle

Die Abbildung 10.4 zeigt die Ergebnisse eines „Minimal"-Modells der Korrektur, welches grundlegende soziodemografische Unterschiede herausstellt. Hierzu wurden durch ein multinomiales Modell alle onlinebasierten Erhebungen in Referenz zur persönlichen Befragung vor der Bundestagswahl 2013 gesetzt. Die sich hierdurch ergebenden Unterschiede sind vergleichsweise gering und zudem maßgeblich durch die in den Befragungen der Komponente 3 und 8. Lediglich ein Einpersonenhaushalt und die Zugehörigkeit zur Altersgruppe ab 60 Jahren, im Vergleich zu den 40- bis 59-jährigen als deren Referenzgruppe, senkt klar die Wahrscheinlichkeit, Teil der Stichprobe der jeweiligen Befragung der Komponente 3 oder 8 zu sein. Zudem lässt sich eine leichte Begünstigung der Personen mit einem hohen Bildungsstand innerhalb der Komponente 3 und einer schwierigeren Berücksichtigung der Personen mit einem niedrigen Bildungsstand innerhalb der Komponente 8 feststellen. Die Effekte sind, an Hand der Abweichungen der Logit-Koeffizienten von Null, jedoch sehr gering. Diese Unterschiede können ebenfalls eine Folge der Quotierung der Stichproben sein (siehe Kapitel 5.1).

Auch wenn diese Unterschiede aktiv über die Auswahl erzwungen sind (siehe zur Erläuterung auch das Kapitel 5.1), erzeugt die angewandte Quotierung der Komponenten 3 und 8 Stichproben, welche Teile der Bevölkerung systematisch benachteiligen. Durch diese bewusste Auswahl der Personen wird eine zufallsbasierte Auswahl aufgegeben und trotzdem kein adäquates Abbild der Bevölkerung erlangt (siehe hierzu auch den vorherigen Vergleich zum Mikrozensus in Kapitel 5.4).

Die mangelnde zufallsbasierte Auswahl erhöht zudem die Gefahr einer weiteren selektiven Teilnahme, welche sich auf weiteren interessierenden Variablen äußern kann. Dies tritt beispielsweise auf, wenn es sich bei den durch die Onlineumfragen befragten älteren Personen um eine sehr spezielle Gruppe dieser handelt. Ihr Denken und ihre Einstellungen können sich daher systematisch von ihrer Altersgruppe insgesamt unterscheiden.

Abbildung 10.4.: Unterschiede Komponenten 3 und 8 in Referenz zur Komponente 1 für 2013 (Minimalmodell)



Quelle: Eigene Darstellung. Gewicht: w_trow (Komponente 1). Negativer Logit-Koeffizient erhöht die Wahrscheinlichkeit, Teil der Referenzgruppe/-umfrage der Komponente 1 zu sein.

Um dies zu überprüfen, werden nun nachfolgend weitere Faktoren in die Betrachtung eingebracht. Die Auswhal hierzu basiert auf dem Kapitel 7.2.3. Diese weiteren Variablen sind, in der Reihenfolge ihrer Berücksichtigung, die Kompetenz des Verständnisses und das Interesse an Politik sowie das Empfinden des Wählens als staatsbürgerliche Pflicht (Kapitel 7.2.3), die Einstellungen gegenüber den Akteuren des politischen Systems und die Diagnose seines Gesamtzustands (ebenfalls Kapitel 7.2.3), sowie die Persönlichkeit einer Person an Hand der Dimensionen der „Big Five“ Persönlichkeitsmerkmale (Kapitel 7.2.2).

Handelt es sich bei soziodemografischen Merkmalen um sehr standardisierte Skalen der Umfrageforschung, muss dies auf Einstellungsfragen nicht zwingend zutreffen. Ein Problem der Vergleichbarkeit der Frageformulierung entsteht dann, wenn die Konzeptspezifikation zwischen Umfragen variiert. Die exakten Frageformulierungen und Antwortmöglichkeiten sind daher für diese Fragen essentiell, um eine Vergleichbarkeit der Korrekturen zu garantieren. Die Tabelle 10.2 führt diese auf.[3]

[3]Eine Möglichkeit der weiteren Berücksichtigung wäre der KQ 2 der fünften Welle des WKP gewesen. Da innerhalb dieser jedoch keine Abfragen hinsichtlich der Dimensionen der „Efficacy“ und dem Grad der Unzufriedenheit mit der Situation der Demokratie enthält, wurde diese nicht berücksichtigt.

Tabelle 10.2.: Frageformulierungen der Efficacy, Wahlnorm und Demokratieunzufriedenheit der Komponenten 1, 3 und 8 der GLES (BTW 2013)

Datensatz	Internal Efficacy	External Efficacy	Wahlnorm	Demokratieunzufriedenheit
Komp. 1 (2013)	„Politische Fragen sind für mich oft schwer zu verstehen.“ (Variable: q73a, Fünfstufig: 5 „trifft voll und ganz zu“)	„Die Parteien wollen nur die Stimmen der Wähler, ihre Ansichten interessieren sie nicht.“ (Variable: q73b, Fünfstufig: 5 „trifft voll und ganz zu“)	„In der Demokratie ist es die Pflicht jedes Bürgers, sich regelmäßig an Wahlen zu beteiligen.“ (Variable: q73d, Fünfstufig: 5 „trifft voll und ganz zu“)	„Sind Sie mit der Art und Weise, wie die Demokratie in der Bundesrepublik Deutschland funktioniert, alles in allem [...][a]?“ (Variable: q6, Fünfstufig: 5 „sehr unzufrieden“)
Komp. 3 (W1, KQ1, KQ3)	„Wichtige politische Fragen kann ich gut verstehen und einschätzen.“ (Variable: kp1_050k, kp3_050k, kp7_050k, Fünfstufig: 5 „trifft voll und ganz zu“)	„Die Politiker kuemmern sich darum, was einfache Leute denken.“ (Variable: kp1_050a, kp3_050a, kp7_050a, Fünfstufig: 5 „Stimme voll und ganz zu“)	„In der Demokratie ist es die Pflicht jedes Bürgers, sich regelmäßig an Wahlen zu beteiligen.“ (Variable: kp1_050l, kp3_050l, kp7_050l, Fünfstufig: 5 „Stimme voll und ganz zu“)	„Wie zufrieden oder unzufrieden sind Sie alles in allem mit der Demokratie, so wie sie in Deutschland besteht?“ (Variable: kp1_020, kp7_020, Fünfstufig: 5 „Sehr unzufrieden“)
Komp. 8 (T21, T22)	„Politische Fragen sind für mich oft schwer zu verstehen.“ (Variable: t156a, Fünfstufig: 5 „trifft voll und ganz zu“)	„Die Parteien wollen nur die Stimmen der Wähler, ihre Ansichten interessieren sie nicht.“ (Variable: t156b, Fünfstufig: 5 „trifft voll und ganz zu“)	„In der Demokratie ist es die Pflicht jedes Bürgers, sich regelmäßig an Wahlen zu beteiligen.“ (Variable: t156d, Fünfstufig: 5 „trifft voll und ganz zu“)	„Und wie zufrieden oder unzufrieden sind Sie - alles in allem - mit der Demokratie, so wie sie in Deutschland besteht?“ (Variable: t6, Fünfstufig: 5 „Sehr unzufrieden“)

Quelle: Eigene Darstellung. Beschreibungen den jeweiligen Fragebögen der Datensätze entnommen. Komp1: ZA5700_fb_v2-0-0; W1: ZA5704_fb_v3-0-0; KQ1: ZA5753_fb_v2-0-0; KQ3: ZA5755_fb_v2-0-0, T21: ZA5721_fb_v3-1-0; T22: ZA5722_fb_v3-0-0. Alle Dokumente frei über die GESIS beziehbar.

Die „Internal Efficacy“ wurde hinsichtlich der Richtung der Skalierung unterschiedlich abgefragt. Beide angewandten Varianten beziehen sich jedoch auf den Kompetenzaspekt dieser Dimension, welchen auch Beierlein et al. (2012: 8) zur Operationalisierung vorschlagen. Die Formulierungen der „External Efficacy“ führen jedoch zu einer Änderung des Bezugsobjektes zwischen den verschiedenen aufgeführten Umfragen. Die Komponente 3 folgt dem Vorschlag von Beierlein et al. (2012: 8), die Responsivität von Politikern zu operationalisieren. Die Komponenten 1 und 8 beziehen sich hingegen auf Parteien als Akteur. Eine Vergleichbarkeit impliziert daher, dass eine Differenzierung zwischen „Partei“ und „Politiker“ keine Auswirkung auf das Antwortverhalten hat. Manifestieren sich Politiker als Merkmale von Parteien im Sinne von Fishbein (1963), ist diese Annahme gegeben. Der Wechsel des Bezugsobjekts der Frage ändert nicht das Zustimmungsverhalten. Gilt dies auch für die Parteienlandschaft insgesamt, sind „die Politiker“ und „die Parteien“ letztendlich ein identischer Bezugspunkt. Diese Annahme wird nachfolgend getroffen, um eine Vergleichbarkeit der unterschiedlichen Frageformulierungen zu garantieren.

Innerhalb der Komponente 3 weisen zudem Zustimmungen zu den beiden „Efficacy“-Dimensionen auf eine hohe Selbsteinschätzung der Kompetenz des Verständnis von Politik und eine positive Einschätzung der Responsivität der politischen Akteure hin. Innerhalb der Befragungen Komponenten 1 und 8 signalisiert dies jedoch einen geringeren Kompetenzgrad und eine niedrigere Einschätzung der Offenheit des politischen Systems. Als Entscheidung der Richtung der Skalierung dient im Rahmen der Arbeit die langjährige Verwendung der Formulierung innerhalb der „American National Election Studies“ (ANES) (siehe hierzu Abbildung 7.2 in Kapitel 7.2.3). Ein *hoher Wert* auf beiden Dimensionen der „Efficacy“ bedeutet somit eine *niedrige* Selbsteinschätzung der Kompetenz des Verständnis von Politik und einen *geringen* Glauben an die Offenheit und Responsivität des politischen Systems. Die Skalierung der Komponente 3 wird daher an die beiden anderen Komponenten angepasst.

Die empfundene Wahlnorm und die Unzufriedenheit mit der Situation der Demokratie in der Bundesrepublik wurde hingegen in allen Umfragen mit einer identischen Formulierung und Skalierung verwendet. Die Komponente 1 zeigt lediglich eine leichte Abweichung der Formulierung. Die Verwendung dieser Variablen impliziert daher die Annahme einer Äquivalenz des Aspekts des „Funktionierens“ und des „Bestehens“, damit sich der Bezug der Bewertung nicht verändert.

Das Kapitel 7.2.3 hat die *zeitliche* Stabilität der Zusammenhänge zwischen der

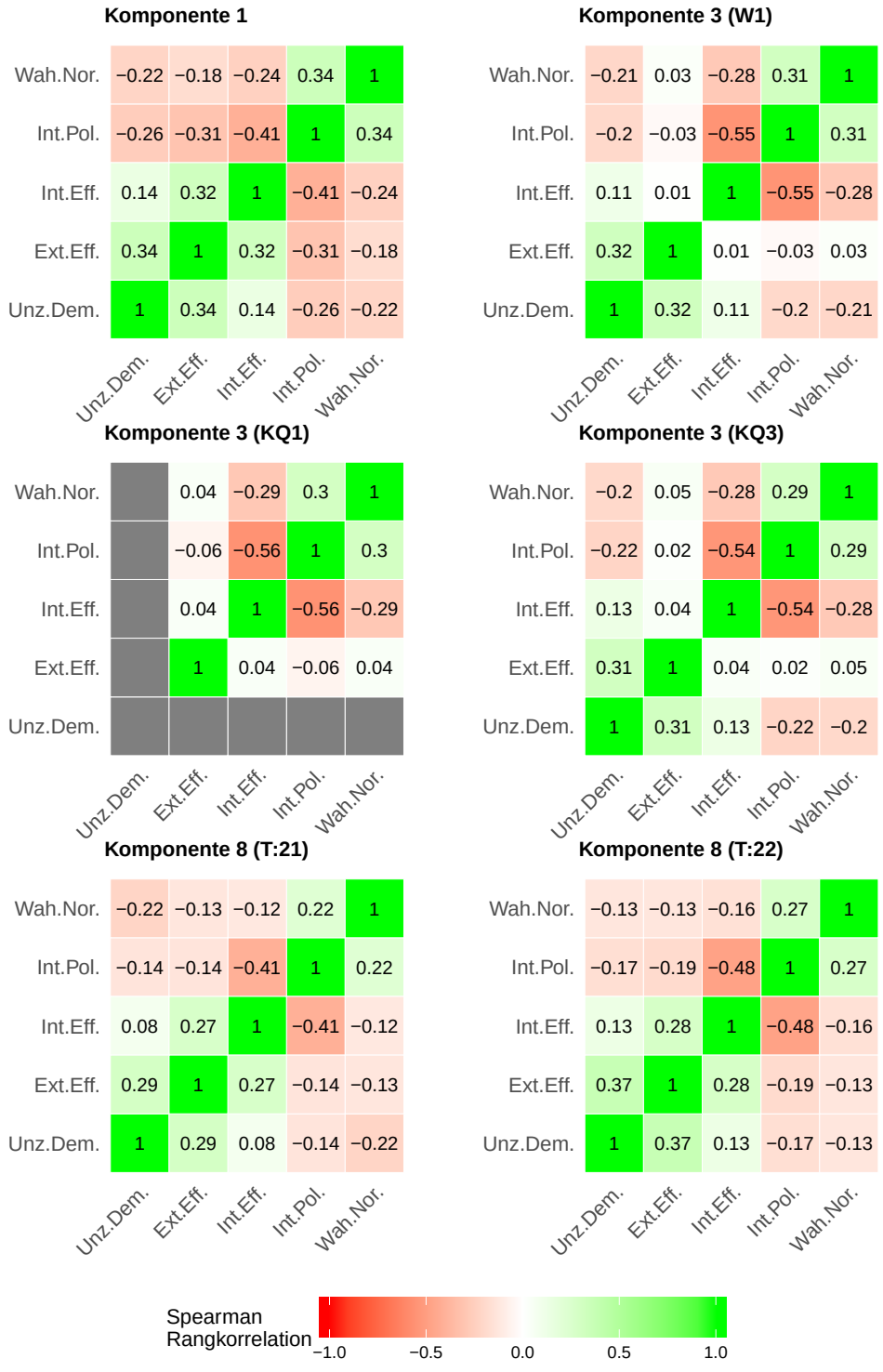
Internal und External Efficacy, sowie dem Interesse an Politik aufgezeigt. Hierdurch sind diese Faktoren für eine Korrektur prädestiniert, wenn ein zeitlicher Abstand der Referenzumfrage zur korrigierenden Umfrage existiert und eine Verzerrung auf Grundlage dieser Variablen vermutet wird. Die Abbildung 10.5 bestätigt die Stabilität diese Zusammenhänge *zwischen* den verschiedenen Erhebungen der GLES, welche ähnliche Strukturen erfassen. Über alle Erhebungen hinweg äußert sich ein starker Zusammenhang zwischen der Internal Efficacy und dem politischen Interesse. Hohe Schwierigkeiten des Verständnis von politischen Fragen gehen mit einem verringerten Interesse an Politik einher.[4] Ebenfalls erweist sich die Richtung und Stärke des Zusammenhangs zwischen der External Efficacy und der Demokratieunzufriedenheit als konstant. Ein geringer Glaube an die Responsivität der Akteure des politischen Systems trifft häufig auf eine ebenfalls vorhandene Unzufriedenheit mit der Situation der Demokratie in der Bundesrepublik.[5] Im Unterschied zu den anderen Erhebungen weisen die Befragungen der Komponente 3 jedoch keine Verbindung der External zur Internal Efficacy, der Wahlnorm und dem Interesse an Politik auf. In allen Fällen weist Spearmans ρ Werte nahe Null aus, Unterschiede der External Efficacy lassen dort keine Prognose über die Ausprägungen der anderen drei Variablen zu. Die durch die External Efficacy verursachten Verzerrungen könnten daher so stark sein, dass sie selbst die Zusammenhänge zu anderen Variablen verschieben.

Um dies zu überprüfen werden nun zwei Modelle präsentiert, welche das „Minimal"-Modell aus Abbildung 10.4 erweitern. Das erste Modell integriert das Interesse an Politik und die Kompetenz des Verständnis ebendieses. Dies wird verbunden mit der empfundenen Norm zur Beteiligung an Wahlen als eine staatsbürgerliche Pflicht. Dieses *„Kompetenz/Norm"-Modell* vereint somit das Interesse am Thema einer Wahlumfrage, die Fähigkeit den Inhalt auch nachvollziehen zu können und die als soziale Norm empfundene Pflicht der Partizipation. Diese drei Faktoren werden als ursächlich für die Teilnahme *und* das Antwortverhalten auf inhaltlichen Fragen einer Wahlumfrage angenommen. Das zweite Modell integriert zusätzlich die Unterschiede der Einstellungsfragen der External Efficacy und der Demokratieunzufriedenheit. Durch dieses *„Einstellungs"-Modell* stehen somit externalisierte Bewertungen des politischen Systems und seiner Akteure im Fokus.

[4] Die Skalierung des politischen Interesse ist über alle Umfragen hinweg auf einer fünfstufigen Skala mit 5 „überhaupt nicht" abgefragt worden. Um die Skalierung konstant zu den vorangegangenen Kapiteln zu halten, wurde die Skala rekodiert. Der Wert 5 entspricht daher der inhaltlichen Bedeutung „sehr stark".

[5] Innerhalb des KQ1 der Komponenten 3 wurde die Demokratieunzufriedenheit nicht abgefragt, weshalb für diese Variable dort keine Zusammenhänge aufgeführt sind.

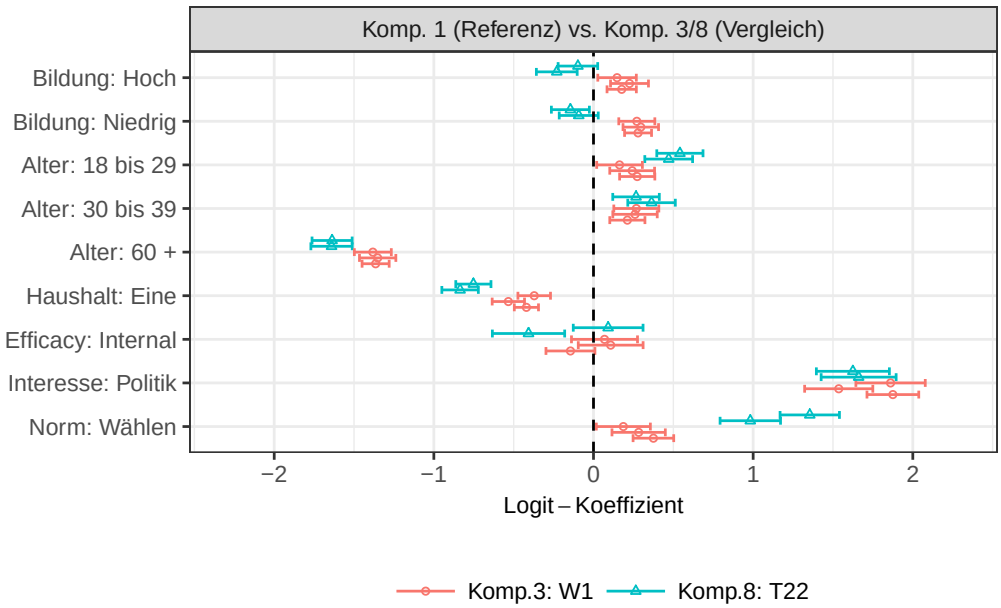
Abbildung 10.5.: Zusammenhänge Wahlnorm, Politisches Interesse, Efficacy und Demokratieunzufriedenheit für Komponenten 1, 3 und 8 der GLES (BTW 2013)



Quelle: Eigene Darstellung. Komponente 1 gewichtet (w_trow).

Abbildung 10.6 zeigt die Ergebnisse der Erweiterung des „Minimal“-Modells zu einem „Kompetenz/Norm“-Modell.

Abbildung 10.6.: Unterschiede Komponenten 3 und 8 in Referenz zur Komponente 1 für 2013 („Kompetenz/Norm“-Modell)



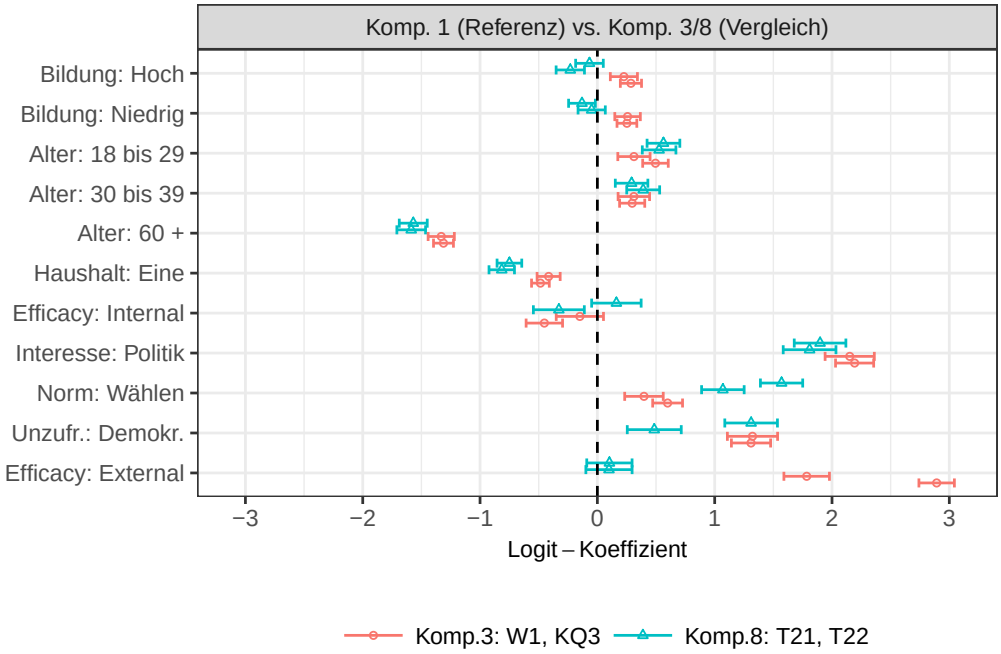
Quelle: Eigene Darstellung. Komponente 1 ist gewichtet (w_trow). Sofern notwendig, wurde über eine Normalisierung an Hand der Minimal- und Maximalwerte ein Wertebereich von 0 bis 1 erzeugt. Die jeweiligen Logit-Koeffizienten stellen somit die Veränderung des vorhergesagten Logit-Werts durch einen Wechsel zwischen diesen beiden Werten dar.

Die grundlegenden soziodemografischen Merkmale zeigen nur geringfügigere Veränderungen durch die zusätzliche Berücksichtigung auf. Der Effekt des Bildungshintergrunds der befragten Personen ist innerhalb der Komponente 8 nicht mehr nachweisbar, diejenigen der Zugehörigkeit zur Altersgruppe ab 60 Jahren und das Führen eines Einpersonenhaushalts verstärken sich. Die Kompetenz des Verständnis von Politik (Efficacy: Internal) lässt (nahezu) keine Differenzierungsmöglichkeit zwischen den onlinebasierten Befragungen und der persönlichen Befragung zu. Dass Interesse an Politik ist jedoch über alle Erhebungen der Komponente 3 und 8 klar erhöht und deutet auf eine Verzerrung hin. Für die Komponente 8 trifft dies ebenfalls auf die angegebene Zustimmung des Verständnis des Wählens als Bürgerpflicht zu. Die befragten Personen dieser Komponente stimmen dieser Aussage eher zu als diejenigen der beiden anderen Komponenten. Eine Korrektur des Interesse an Politik in beiden Komponenten und der empfundenen Wahlnorm innerhalb der Komponente 8, hat somit das Potential einer Korrektur weiterer inhaltlicher Variablen.

Lässt sich auf der Basis der Einstellungen gegenüber dem politischen System

insgesamt („Demokratieunzufriedenheit“) und seinen Repräsentanten („External Efficacy“) ebenfalls eine Vorhersage der jeweiligen Umfragezugehörigkeit treffen? Abbildung 10.7 präsentiert die Ergebnisse des „Einstellungs“-Modells. Die

Abbildung 10.7.: Unterschiede Komponenten 3 und 8 in Referenz zur Komponente 1 für 2013 („Einstellungs“-Modell)



Quelle: Eigene Darstellung. Komponente 1 ist gewichtet (w_trow). „Unzufr.: Demokratie“ und „Efficacy: External“ ebenfalls auf einen Bereich von 0 bis 1 normalisiert (siehe Beschreibung Abbildung 10.6). Ein positiver Logit-Koeffizient impliziert eine erhöhte Wahrscheinlichkeit der Zugehörigkeit zur Komponente 3 oder 8.

Richtung und Stärke der Faktoren des „Kompetenz/Norm“-Modell verändern sich durch die Aufnahme der „External Efficacy“ und der Demokratieunzufriedenheit nicht. Da sich zwischen erster und den Variablen des „Kompetenz/Norm“-Modells bereits kein Zusammenhang gezeigt hat (Abbildung 10.5), ist dies nicht überraschend. Beide Modelle versuchen unterschiedliche Ursachen der Verzerrung in Wahlumfragen zu erfassen.

Die Nutzung des Access-Panel von Respondi und LINK resultiert in Stichproben, welche eine erhöhte Demokratieunzufriedenheit aufweisen. Die beiden Erhebungen der Komponente 3[6] kombinieren diese generelle Unzufriedenheit zusätzlich mit einer negativen Einschätzung der Responsivität der Akteure. Die Stichproben dieser Komponente sind ebenso politisch interessiert wie diejenigen der Komponente 8, die befragten Personen haben jedoch keinen Glauben an

[6]Innerhalb dieses Modell konnte der KQ 2 der Komponente 3 nicht mehr berücksichtigt werden, da dort die Abfrage der Demokratieunzufriedenheit nicht erfolgt ist.

eine ausgeprägte Responsivität der politischen Akteure. Die Auswirkung der möglichen Korrektur der befragten Personen auf Basis dieser angegebenen Einstellung ist vergleichbar mit derjenigen des generellen Interesse an Politik.

Das Kapitel 7.2.2 hat ebenfalls die Bedeutung der Persönlichkeit in Form der „BIG Five“ für das politische (Wahl-)Verhalten identifiziert. In Kapitel 7.2.3 wurde herausgestellt, dass diese die Verarbeitung externer Informationen zur Veränderung von Einstellungen steuern. Hierdurch kann sich die Bewertung eines Bezugsobjekts der Einstellung, wie einer Partei oder einem Politiker, ändern. Vergleichbare Operationalisierungen der Persönlichkeitsmerkmale der „Big Five“ finden sich innerhalb des Referenzdatensatzes der Komponente 1 und den Befragungen T22 (Komponente 8) und der Rekrutierungswelle W1 des WKP (Komponente 3).

Vier der fünf Dimensionen lassen sich durch übereinstimmende Fragestellungen zwischen den Komponenten berücksichtigen. Da das Wahlkampfpanel der Komponente 3 keine übereinstimmende Operationalisierung der „Verträglichkeit“ („Ich gehe aus mir heraus, bin gesellig.“) mit den Formulierungen der anderen beiden Komponenten („Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.“) hat, wurde diese Dimension nicht berücksichtigt. Die „BIG Five“ werden daher durch die Dimensionen der Offenheit, Extraversion, Gründlichkeit und des Neurotizismus als ein „Big Four“ abgedeckt.

Die Tabelle 10.3 präsentiert die verwendeten Aussagen, zu welchen die befragten Personen den Grad ihrer Zustimmung angeben sollten.

Tabelle 10.3.: Big Five Frageformulierungen der Komponenten 1, 3 und 8 der GLES (BTW 2013)

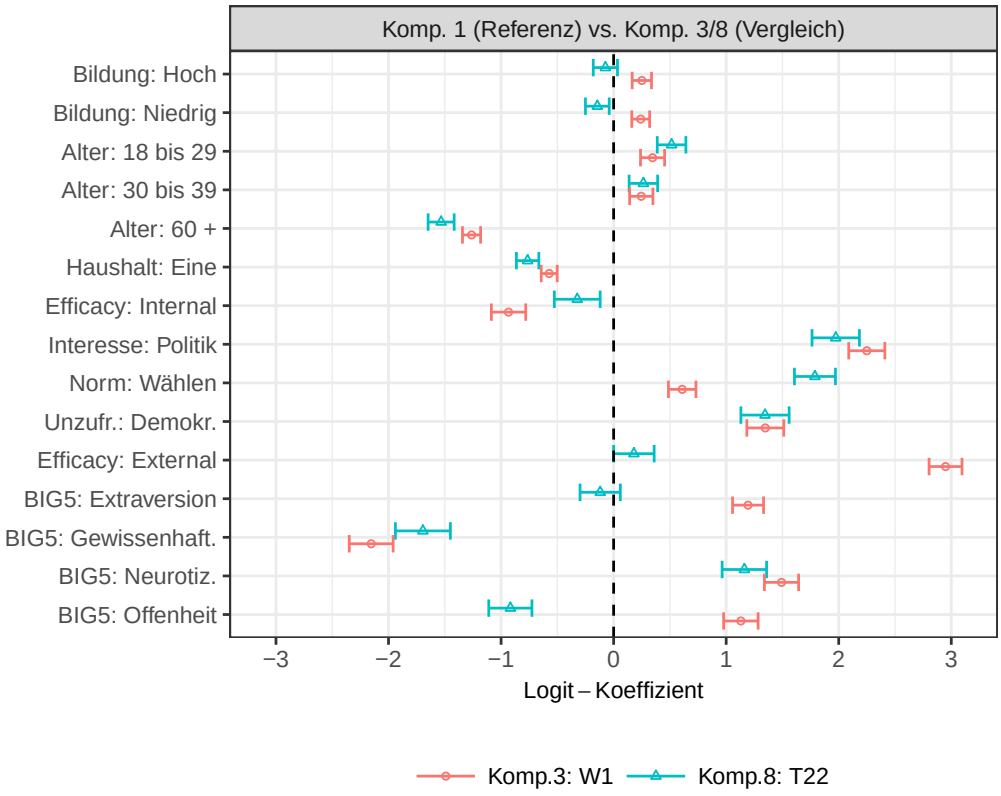
Offenheit:	„Ich habe eine aktive Vorstellungskraft, bin phantasievoll.“
Extraversion:	„Ich bin eher zurückhaltend, reserviert.“
Gründlichkeit:	„Ich erledige Aufgaben gründlich.“
Neurotizismus:	„Ich werde leicht nervös und unsicher“
Geselligkeit:	Keine vergleichbaren Fragen vorhanden

Quelle: Eigene Darstellung. Alle Abfragen wurden auf einer fünfstufigen Likert-Skala mit 1 „trifft überhaupt nicht zu“ und 5 „trifft voll und ganz zu“ durchgeführt. Variablen sind: t163* ($T22$), q124* (Komponente 1) und kpX_2180* (W1 des WKP).

Inwieweit kann die Persönlichkeit einen Beitrag zur Erklärung der Unterschiede zwischen den beiden onlinebasierten und der persönlichen Befragung leisten? Die Abbildung 10.8 weist die Ergebnisse auf der Grundlage des „Persönlichkeits“-Modell aus. Es finden sich Effekte aller vier berücksichtigten Dimensionen. Personen der Erhebung T22 des offlinerekrutierten Access-Panel von LINK

(Komponente 8) weisen im Gegensatz zum verwendeten Referenzdatensatz der Komponente 1 generell niedrigere Zustimmungsraten bezüglich der Offenheit und der Gewissenhaftigkeit als Persönlichkeitsmerkmale auf. Ein hoher Grad an Neurotizismus erhöht hingegen die Wahrscheinlichkeit der Zugehörigkeit zu dieser Umfrage.

Abbildung 10.8.: Unterschiede Komponenten 3 und 8 in Referenz zur Komponente 1 für 2013 („Persönlichkeits“-Modell)



Quelle: Eigene Darstellung. Komponente 1 ist gewichtet (w_trow). Alle „BIG5“-Abfragen sind auf einen Bereich von 0 bis 1 normalisiert worden (siehe Beschreibung Abbildung 10.6).

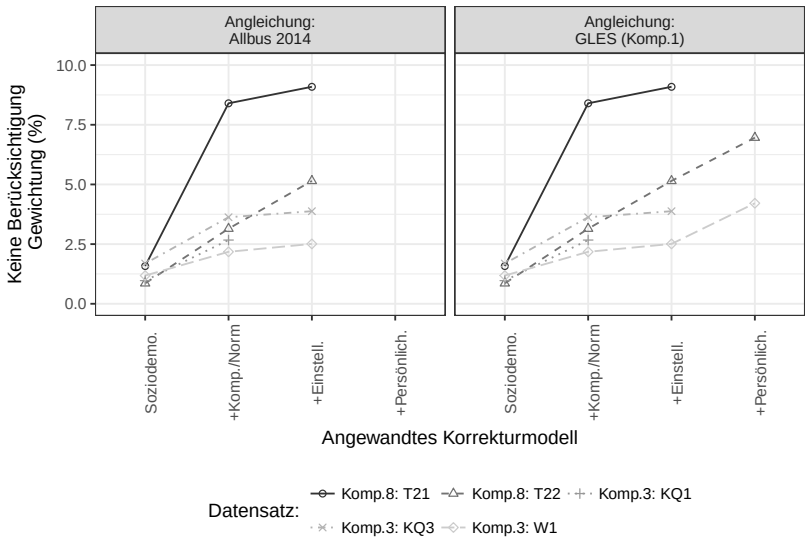
„Neurotische Züge“ teilen auch die befragten Personen der Rekrutierungswelle des WKP aus dem aktiv onlinerekrutierten Access-Panel von Respondi (Komponente 3), kombinieren diese jedoch mit einem erhöhten Grad an Offenheit und der Dimension der Extraversion. Hier kann die Rekrutierungsform des Access-Panel von Respondi ursächlich sein, „offene“ Personen mit einem expressiven Charakter werden mit einer erhöhten Wahrscheinlichkeit durch eine rein onlinebasierte Rekrutierung erreicht und stimmen dieser eher zu.

Wie wirkt sich nun eine Korrektur dieser Unterschiede auf das Wahl- und Beteiligungsverhalten im Kontext der Bundestagswahl 2013 aus? Basierend auf

den Ergebnissen aller vier aufgeführten Modelle („Minimal“-Modell, „Kompetenz/Norm“-Modell, „Einstellungs“-Modell und „Persönlichkeits“-Modell) wurden Korrekturgewichte auf der Basis einer einfachen Nutzung der inversen „Propensity Scores“ (PS) und ein PSM mit einer Subklassifikation und dem Nearest-, CEM- und Full-Algorithmus angewandt.[7]

Ein Problem der Anwendung von Gewichtungsverfahren ist der Ausfall von Personen durch fehlende Angaben auf Variablen des Korrekturmodells. Dieses Problem zeigt sich jedoch nur in einem geringen Ausmaß, wie die Abbildung 10.9 zeigt.

Abbildung 10.9.: Fehlende Werte Korrekturmodell: BTW 2013 (Kapitel 10.1)



Quelle: Eigene Darstellung.

Die Korrektur der Soziodemografie kann lediglich für ca. 2.5 % der Personen kein gültiges Bedeutungsgewicht berechnen. Auch der zusätzliche Einbezug der „Internal Efficacy“, des Interesse an Politik und der Wahlnorm erhöht diese Rate nur für die Erhebung T21 auf über 5 %. Dies trifft ebenso auf die External Efficacy und die Demokratieunzufriedenheit zu. Der Einbezug der Persönlichkeit innerhalb der Erhebungen T22 und W1 sorgt für weitere Steigerungen des Ausfalls, auch hier zeigen sich aber Raten von nur 3 und 7 % der Personen, welche durch die Gewichtung nicht berücksichtigt werden können.

[7] Für das PS-Gewicht wurde die Inverse dieser genutzt und auf 1 normalisiert. Für das CEM wurde alle abgefragten fünfstufigen Likert-Skalen in die Kategorien „Ablehnung“, „Neutral“ und „Zustimmung“ „coarsened“. Die Subklassifikation an Hand der PS basiert auf sechs Klassen. Die Definition des „caliper“ für den Nearest-Algorithmus ist das 0.25-fache der Standardabweichung der PS. Der Full-Algorithmus weist ein Minimum von 0.1 und ein Maximum von zwei Kontrolleinheiten pro Treatmenteinheit auf. Die Gewichtungsvariablen über das MatchIt-Paket in R sind automatisch auf ein arithmetisches Mittel von 1 normiert.

Ein weiteres wichtiges Kriterium ist die erreichte Balancierung zwischen der anzugleichenden Umfrage und des Referenzdatensatzes, welche die Abbildung 10.10 aufführt. Wird diese technisch nicht erreicht, können die vorhandenen und bekannten Unterschiede zwischen den beiden Umfragen nicht ausreichend korrigiert werden, ein Restfehler ist weiter vorhanden. Für die Subklassifikation und den Nearest- und Full-Algorithmus ergeben sich Balancierungen von nahe 100 % der standardisieren Differenz der PS. Bezüglich der restlichen Verteilungen ergeben sich insbesondere für die Subklassifikation Probleme, eine ausreichende Reduzierung der Unterschiede zu erreichen (siehe Abbildung 10.10 des Anhangs ??).

Auch ein zu hohes Maximalgewicht für Gruppen von Person kann nachteilig werden. Dies beseitigt zwar die diagnostizierten Unterschiede, macht die Ergebnisse inhaltlich interessierender Variablen jedoch zunehmend von den Antworten dieser Personen abhängig. Hierbei erweist sich in Abbildung 10.11 der CEM-Algorithmus am nachteiligsten, insbesondere je komplexer das Modell zur Bestimmung der PS wird. Werden alle Korrekturfaktoren mit einbezogen, erhalten Personen dort ein Bedeutungsgewicht von bis zu 60. Dies ist auf die dortige Nutzung eines „Exact Matching“ (EM) zurückzuführen. Auch die Subklassifikation erreicht Werte zwischen 15 und 30. Die direkte Nutzung der PS erzeugt hingegen bereits konstante Bedeutungsgewichte im Bereich von 10 bis zu 15, das PSM mit Hilfe des Full-Algorithmus sogar konstante Maximalgewichte mit wünschenswert niedrigen Werten von unter 5. Das Nearest-Matching zeigt keine Variation, da dieses ohne Wiederholung durchgeführt wurde.

Insgesamt sind die angewandten Korrekturen somit größtenteils in der Lage vorhandene Unterschiede zwischen den jeweiligen onlinebasierten Befragungen und der persönlichen Referenzumfrage der Komponente 1 zu beheben, ohne einen größeren Personenkreis ausschließen zu müssen oder sehr hohe individuelle Bedeutunsgewichte zu zuweisen. Das nachfolgende Kapitel wird nun dazu übergehen die Auswirkungen der Anwendung der Bedeutungsgewichte auf inhaltliche Zielvariablen der Umfragen, namentlich der Wahlentscheidung und der Beteiligungsabsicht, darzustellen. Hierdurch wird die Frage beantwortet: Wirken sich die aufgefundenen Unterschiede der onlinebasierten Befragungen der Komponenten 3 und 8 auch auf das Wahl- und Beteiligungsverhalten aus? Führt somit eine Anwendung der präsentierten Korrekturmodelle auch zu einer Verbesserung inhaltlich interessierender Variablen?

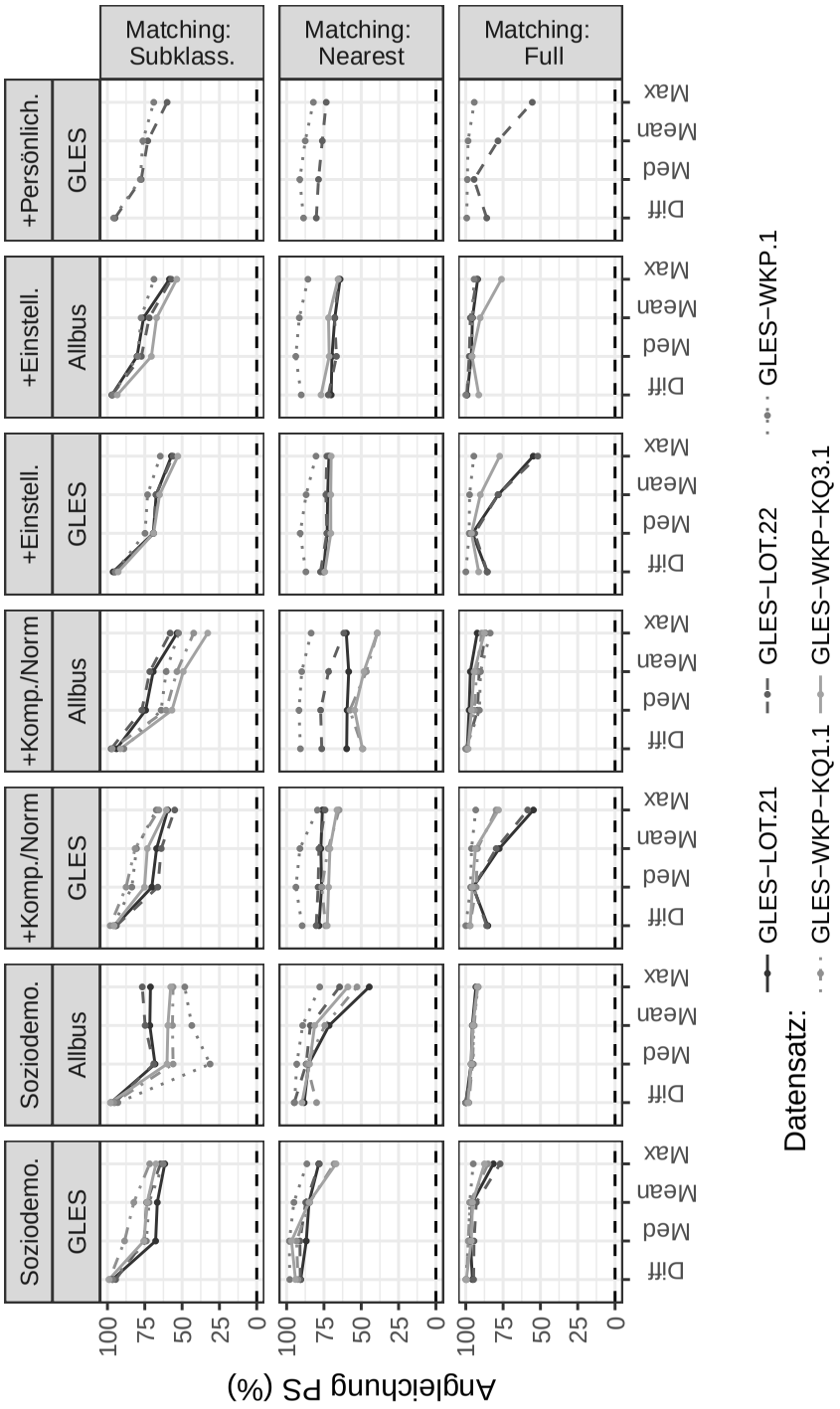


Abbildung 10.10.: Balancierung: BTW 2013 (Kapitel 10.1)

Quelle: Eigene Darstellung.

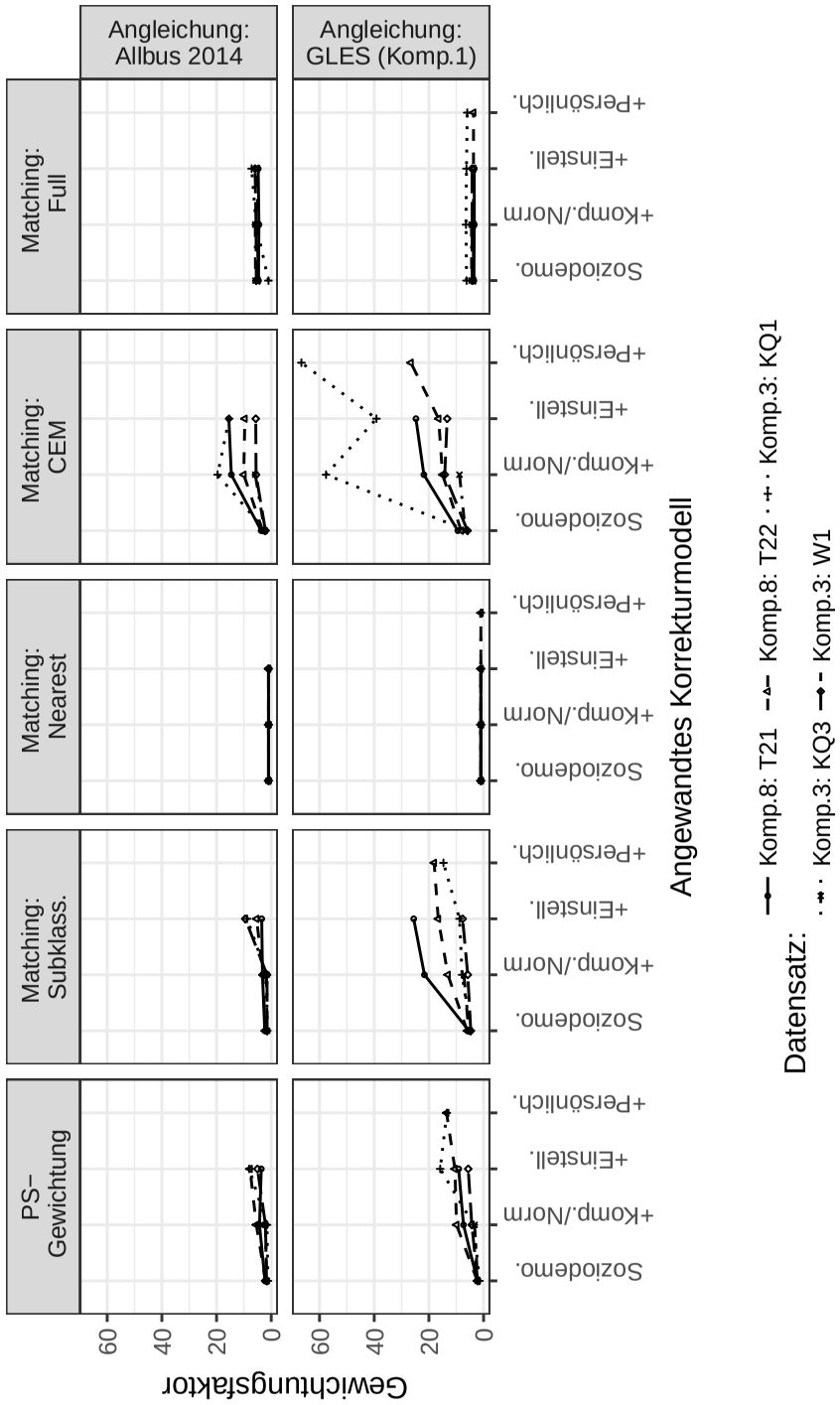
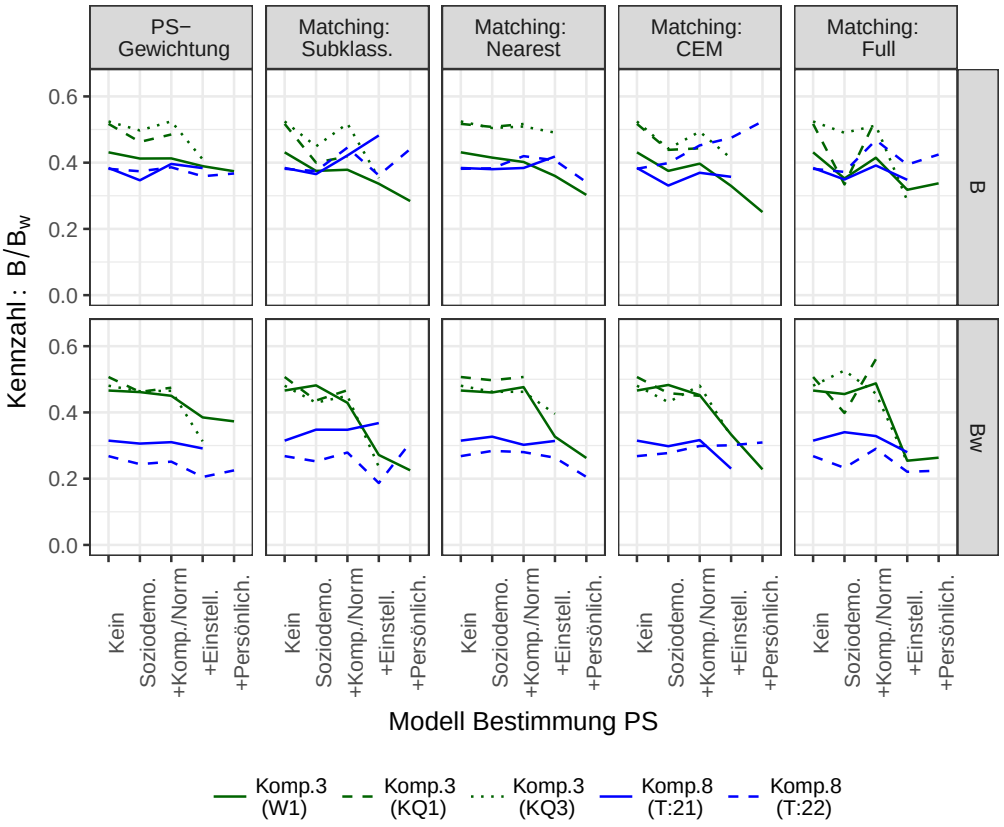


Abbildung 10.11.: Maximalgewichte: BTW 2013 (Kapitel 10.1)

10.1.2. Evaluation: Wahl- und Beteiligungsverhalten

In Abbildung 10.3 wurde bereits die verzerrte Erfassung der Zweitstimmenanteile der verschiedenen Parteien der Bundestagswahl 2013 auf Basis der onlinebasierten Befragungen aufgedeckt. Lässt sich durch die Anwendung der bestimmten Bedeutungsgewichte eine Reduktion der Verzerrungen gemäß B und B_w für die Komponenten 3 und 8 erreichen? Die Abbildung 10.12 zeigt die Entwicklung beider Kennzahlen durch die Anwendung der jeweiligen Korrekturmodelle und technischen Varianten der Erstellung der Bedeutungsgewichte. Die Referenz einer möglichen Verbesserung sind die bestimmten Werte von B und B_w der Abbildung 10.2 ohne eine angewandte Korrektur („Kein“).

Abbildung 10.12.: Entwicklung B und B_w der Komponenten 3 und 8 der GLES durch Gewichtung (BTW 2013)



Quelle: Eigene Darstellung.

Es zeigt sich eine klare Verbesserung der Abbildung der Zweitstimmenanteile der Bundestagswahl 2013 durch die Komponente 3, sofern ein komplexeres Korrekturmodell angewandt wird. Durch das „Minimal“-Modell lässt sich für den KQ der Welle 3 bereits eine geringfügige Verbesserung von B und B_w erreichen. Der Übergang zum „Komp./Norm.“-Modell verursacht hingegen nur

geringe Veränderungen, erst der Einbezug der External Efficacy und der Demokratieunzufriedenheit im Rahmen des „Einstellungs“-Modell sorgt für eine weitere qualitative Verbesserung der Komponente 3.

Werden zudem noch die Unterschiede der Persönlichkeit durch vier der fünf möglichen Dimensionen der BIG Five in das Korrekturmodell integriert, kommt es zu einer weiteren Steigerung der Qualität der Stichprobe der Komponente 3. Diese positiven Effekte der Korrektur zeigt sich über alle Varianten des PSM hinweg, die Nutzung der Inversen der PS als Gewichtungskorrektur ist hingegen nicht in der Lage diesen Erfolg zu erreichen. Da sich durch die Korrektur B_w stärker als B verbessert, ist die genauere Abbildung auf eine verbesserte Repräsentation der Wähler der prozentual größeren Parteien zurückzuführen. Für die Komponente 3 liegt daher nun durch die Gewichtung ein höherer Anteil der vorher unterrepräsentierten Unions-Wähler vor (siehe Abbildung 10.2).

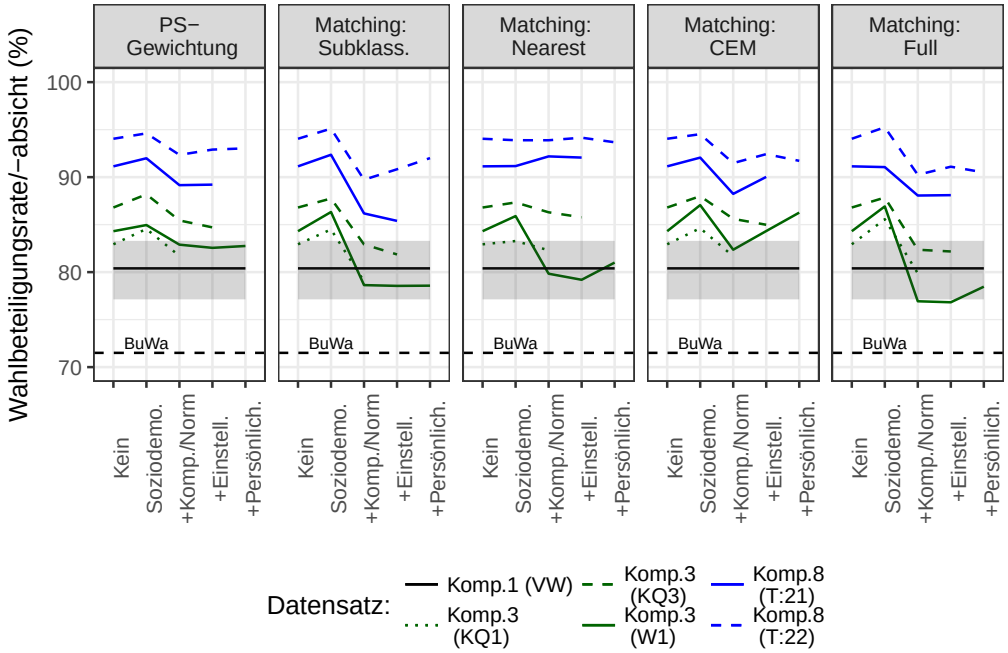
Für die Komponente 8 zeigt sich hingegen nur eine geringfügige bis keine Veränderung durch die angewandten Korrekturen. Werden zudem zur Berechnung der Verzerrung alle Parteien mit einer gleichen Gewichtung berücksichtigt (Kennzahl B), verschlechtert sich das Abbild durch die Stichprobe sogar. Aus dem Verlauf der Entwicklungen von B und B_w bestätigen sich daher die Aussagen in Kapitel 4.3.3: Zur Beseitigung von Verzerrungen in Stichproben ist das angewandte Korrekturmodell, also die Auswahl der inhaltlich zu korrigierenden Faktoren, entscheidender als der angewandte technische Matchingalgorithmus. Die einfache Nutzung der PS-Gewichte erreicht jedoch generell nicht den Erfolg dieser Algorithmen. Ebenfalls ist die Korrektur nur innerhalb der Stichproben auf der Basis des onlinerekrutierten Access-Panels erfolgreich. Dies ist wenig verwunderlich, weisen die Stichproben der Komponente 3 ein viel höheres anfängliches Niveau der Verzerrung auf. Die Probleme der korrekten Abbildung des Wahlverhaltens sind bei diesen Stichproben daher viel stärker ausgeprägt, und damit lassen sich auch schneller und offensichtlicher Erfolge durch Korrekturen erreichen.

Auch die abgefragte Wahlbeteiligungsabsicht vor der Bundestagswahl und die tatsächlich angegebene erfolgte Teilnahme nach der Bundestagswahl hat sich zwischen den Erhebungen unterschieden (Tabelle 10.1). Alle Komponenten weisen erhöhte Werte gegenüber der offiziellen Beteiligungsrate von 71.5 % bei der Bundestagswahl 2013 auf. Der Referenzdatensatz der Komponente 1 weicht hierbei am geringsten von diesen Werten ab. Wie das Kapitel 9.1.2 an Hand der Politbarometer-Daten bereits gezeigt hat, weisen telefonische Befragungen zu hohe Raten politisch interessierter Personen auf. Auch im direkten Kontext

der Bundestagswahl zeigen die Erhebungen des telefonisch offline rekrutierten Access-Panels von LINK die stärksten Abweichungen von den amtlichen Ergebnissen, was somit durch die angewandte Rekrutierungsform bedingt sein kann.

Lässt sich durch die Angleichung an die persönliche Befragung der Komponente 1 eine Verringerung der Wahlbeteiligunsraten der onlinebasierten Befragungen erreichen? Die Abbildung 10.13 stellt den Verlauf der Anteilswerte der klassifizierten Wähler, in Abhängigkeit des angewandten Korrekturmodells und der jeweiligen technischen Variante der Angleichung, dar.

Abbildung 10.13.: Entwicklung Wahlbeteiligungsabsicht der Komponenten 3 und 8 der GLES durch Gewichtung (BTW 2013)



Quelle: Eigene Darstellung. Für die Komp.1 (Gewicht: w_trow) ist zusätzlich das 95 %-Konfidenzintervall für einen Anteilswert angegeben. Hierzu wurde die regionale Stratifizierung nach der Formel 6.7 in Kapitel 6.1.3 berücksichtigt. Der Design-Effekt (Formel 6.7 in Kapitel 6.1.3) beträgt $deft = 1.74$.

Die Anwendung von Bedeutungsgewichten führt in allen Umfragen zu einer Verbesserung, die zu hohen Wahlbeteiligungsraten reduzieren sich von ca. 90 % (Komponente 8) und 85 % (Komponente 3) auf bis zu 85 % (Komponente 8) und unterhalb von 80 % (Komponente 3). In der Tendenz ist die Korrektur erneut erfolgreicher innerhalb der Komponente 3, aber auch die Komponente 8 kann nun von den Bedeutungsgewichten profitieren. Entscheidend ist nun, neben der Wahl des Modells zur Schätzung der PS, ebenfalls die Wahl der technischen Korrektur. Der größte durchschnittliche Erfolg basiert auf der Klassifikation in Gruppen auf der Basis der ermittelten PS („Subklassifikation“) oder

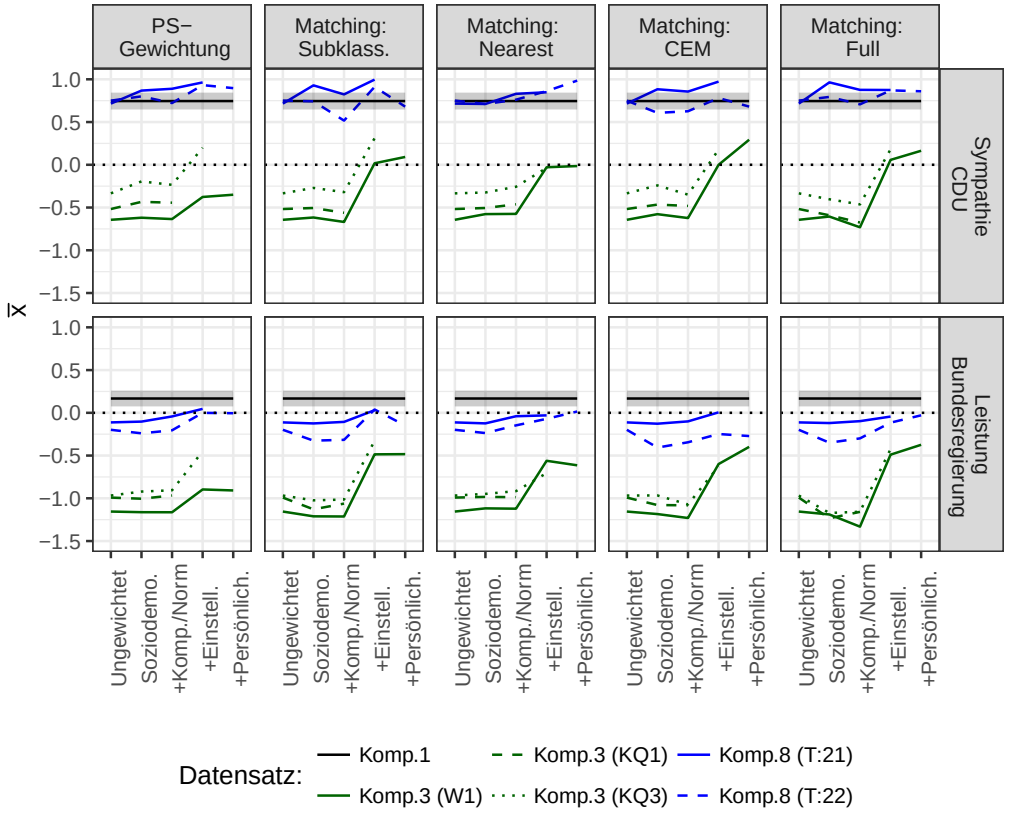
der Anwendung des Full-Algorithmus. Die erfolgreiche Korrektur durch eine Annäherung an die reale Beteiligungsrate der Bundestagswahl 2013 ist zudem (beinahe) vollständig auf die „Internal Efficacy“, das Interesse an Politik und die empfundene Wahlnorm zurückzuführen, welche das „Komp./Norm“-Modell in die Korrektur integriert. Das hierzu ausgewiesene Modell zur Bestimmung der PS in Abbildung 10.6 hatte bereits gezeigt, dass sich keine Unterschiede der Komponenten 3 und 8 zur Komponente 1 hinsichtlich der „Internal Efficacy“ auffinden lassen. Ebenso hat die Wahlnorm nur eine Differenzierung der befragten Personen der Komponente 8 zugelassen. Der Erfolg der Korrektur der Komponente 3 lässt sich somit durch die Reduzierung des Anteils der politisch interessierten Personen über die Anwendung der Bedeutungsgewichte erklären. Die gewichteten onlinebasierten Befragungen erreichen somit ein vergleichbares Niveau zur zufallsbasierten Erhebung der Komponente 1. Sie schaffen es eine, hinsichtlich der Wahlbeteiligung, vergleichbare Grundgesamtheit anzuvisieren.

Auch auf weitere inhaltliche Variablen wirken sich die aufgezeigten Verzerrungen der Stichproben der Onlineumfragen aus, wie die Abbildung 10.14 zeigt.

Die Sympathie für die CDU und die Zufriedenheit mit der Leistung der Bundestregierung wird innerhalb der Erhebungen der Komponente 3 systematisch niedriger bewertet. Beträgt der Unterschied der arithmetischen Mittel zu den Erhebungen der Komponenten 1 und 8 einen Skalenpunkt auf der elfstufigen Skala hinsichtlich Zufriedenheit der damaligen schwarz-gelben Bundesregierung, ist dies bereits 1.5 Skalenpunkte für die Sympathie der CDU als maßgebliche Regierungspartei. Diese Verzerrungen der Komponente 3 führen bereits zu qualitativen Unterschieden: Sind die befragten Personen der Komponenten 1 und 8 im Mittel positiv gegenüber der CDU und indifferent gegenüber der Bundesregierung eingestellt, zeigen diejenigen der Komponente 3 eine kritische Haltung gegenüber der Regierung und eine leichte Abneigung gegenüber der CDU.

Diese Unterschiede lassen sich durch die Korrekturmodelle verringern. Deren Anwendung führt übereinstimmend über alle Algorithmen eines PSM zu ähnlichen Verbesserungen, die PS-Gewichtung weist erneut nur ein geringeres Potential auf. Auch hier ist es jedoch entscheidend, die Einstellungen der Befragten in Form der External Efficacy und der Demokratieunzufriedenheit mit einzubeziehen („Einstellungs“-Modell). Durch dieses Modell wandeln sich die Stichproben der Komponente 3 zu einer nur noch leicht negativen Grundhaltung gegenüber der Bundesregierung und einem im arithmetischen Mittel

Abbildung 10.14.: Entwicklung Sympathie CDU und Bewertung Regierungsleistung (Komponenten 3, 8) durch Gewichtung (BTW 2013)



Quelle: Eigene Darstellung. Für die Komp.1 (Gewicht: w_trow) ist das jeweilige 95 %-Konfidenzintervall angegeben. Hierzu wurde das komplexe regionale Erhebungsdesign („Sample Points“ (SP): vpoint; Strata: ostwest) berücksichtigt. Die Design-Effekte (Formel 6.1 in Kapitel 6.1.3) betragen $deft_{\text{Leist.Reg.}} = 1.59$ und $deft_{\text{Symp.CDU}} = 1.53$. Skala rangiert von -5 (Negative Bewertung) bis +5 (Positive Bewertung).

indifferenten Sympathiegrad gegenüber der CDU. Der zusätzliche Einbezug der Big Five im Rahmen des „Persönlichkeits“-Modell sorgt dann für eine weitere Verbesserung.

Das sich über beide hier betrachteten Variablen hinweg vergleichbare Erfolge hinsichtlich der Richtung und Stärke erzielen lassen, ist auf der Basis der Darstellungen in Kapitel 7.3 wenig überraschend. Die Bewertung der Leistung der Bundesregierung steht in einem engen Zusammenhang zur Sympathie für die jeweilige(n) Regierungspartei(en). Die Selektivität des Zugangs und der Teilnahme in das Respondi-Panel hat den Personen die Teilnahme erschwert, welche tendenziell an Hand ihrer Zweitstimme sich für die Union bei der darauf folgenden Bundestagswahl entscheiden würden. Hierdurch weisen die Erhebungen der Komponente 3 einen zu niedrigen Anteil dieser Gruppe auf. Da die Abgabe der Zweitstimme für eine Partei in einem Zusammenhang mit

der Sympathie für diese steht, kommt es ebenfalls zu einer zu geringen durchschnittlichen Sympathie für die CDU. Und dba diese zudem die maßgebliche Regierungspartei der schwarz-gelben Koalition im Vorfeld der Bundestagswahl 2013 darstellt, wird die Leistung der Bundesregierung ebenfalls als niedriger bewertet.

Der Kontext der Bundestagswahl 2013 hat somit gezeigt: Insbesondere die Stichproben des onlinerekrutierten Access-Panel von Respondi haben mit einer Selektivät der Teilnahme und der Rekrutierung zu kämpfen, welche ein korrektes Abbild des Elektorats nur schwer ermöglicht. Dies wurde am Beispiel der Wahlentscheidung und der Beteiligungsabsicht gezeigt und wirkt sich auf weitere Zielvariablen, wie die Bewertung der Regierungsleistung, aus. Werden die aufgezeigten Verzerrungen korrigiert, verbessert sich das Bild. Der funktionierende Wirkungskanal der Korrektur ist für die Wahlentscheidung und der Bewertung politischer Akteure identisch. Durch die Beachtung möglicher Unterschiede der Demokratie(un)zufriedenheit und der Einschätzung der Responsivität der politischen Akteure ist eine Verbesserung möglich. Hinsichtlich der Entscheidung zur Partizipation sind hingegen die Variablen des „Kompetenz-/Norm“-Modells entscheidend. Die zentrale Bedeutung der Korrektur übernimmt hier das Interesse an Politik, welches systematisch innerhalb der onlinebasierten Befragungen erhöht ist.

Das nachfolgende letzte Kapitel der Arbeit wird das Funktionierens dieses Wirkungskanals nun für einen längeren Zeitraum belegen. Hierdurch wird gezeigt, dass diese Form der Korrektur systematisch und unabhängig von dem Kontext einer Bundestagswahl möglich ist. Hierzu wird nun mit den weiteren zur Verfügung stehenden Stichproben der Komponente 8 der GLES der Analysezeitraum ausgedehnt.

10.2. Die Qualität onlinebasierter Wahlumfragen seit 2009

Das Fazit des vorangegangenen Kapitels ist eindeutig: Onlinebasierte Wahlumfragen haben im Kontext der Bundestagswahl 2013 starke Schwächen hinsichtlich der repräsentativen Erfassung der Beteiligungsabsicht und Wahlentscheidung der wahlberechtigten Bevölkerung der Bundesrepublik. Insbesondere eine online stattfindende Rekrutierung wirkt sich negativ auf die korrekte Abbildung der Grundgesamtheit aus, die Qualität kann mit einer auf den Kontext einer Wahlumfrage passenden Bestimmung von Korrekturgewichten jedoch verbessert werden.

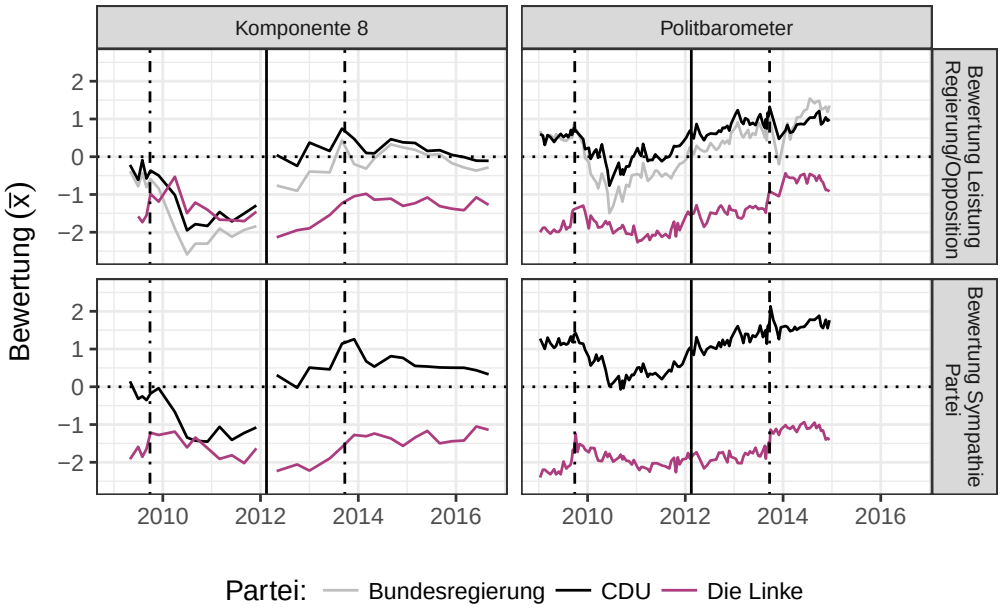
Lässt sich dies auch über einen langfristigen Zeitraum validieren? Dies wird das nachfolgende Kapitel nun an Hand der Erhebungen der Komponente 8 der GLES aufzeigen. Bis einschließlich Dezember 2011 (Erhebung T16) wurden für diese Stichproben das onlinerekrutierte Access-Panel von Respondi genutzt. Seitdem wird das Access-Panel von LINK zur quotierten Auswahl verwendet.[8] Das vorangegangene Kapitel hat dies auf die systematisch zu geringen Zweitstimmenanteile der Union durch die Erhebungen der Komponente 3 aufgezeigt und auf die zu stark erfasste Demokratieunzufriedenheit, eine zu geringe Einschätzung der Responsivität der politischen Akteure und die Verteilung der Persönlichkeitsmerkmale zurückgeführt. Bedeutungsgewichte, welche im Vergleich zur persönlichen Befragung der Komponente 1 diese Verteilungen korrigieren, waren daher in der Lage die Verteilung der Zweitstimmenanteile dieser Erhebungen stärker an die Realität der Bundestagswahl 2013 anzugleichen. Auch den zu hoch erfassten Wahlbeteiligungsraten, sowohl offline- als auch onlinerekrutierter Onlineumfragen, lässt sich durch diese Korrekturgewichte effektiv begegnen. Diese Verzerrung und die Möglichkeit der Korrektur zeigt sich auch langfristig.

Die Abbildung 10.15 stellt den Verlauf der arithmetischen Mittel von fünf erfassten Zufriedenheitsskalen innerhalb der Komponente 8 dar. Der erste Teil konzentriert sich auf die separate Bewertung der gesamten Arbeit der Regierungskoalition und der spezifischen Zufriedenheit mit derjenigen der CDU, welche unter der Kanzlerschaft von Angela Merkel das zentrale Bezugsobjekt der Regierung ist. Ebenfalls ausgewiesen ist die Bewertung der Leistung der Op-

[8] Die Erhebungen T1 und T2 werden nicht berücksichtigt, da elementare Variablen (z.B. die Haushaltsgröße) nicht abgefragt wurden. Ebenfalls wird mit der T17 die erste Erhebung auf der Basis des Access-Panel von LINK nicht berücksichtigt, da sich hier sehr starke Abweichungen von allen anderen Stichproben gezeigt haben. Siehe hierzu auch Abbildung 8.6 in Kapitel 8.2.

positionsarbeit der Linken, welche innerhalb der onlinerekrutierten Erhebungen der Komponenten 3 und 8 im vorangegangenen Kapitel überrepräsentiert war (Abbildung 10.3).

Abbildung 10.15.: Entwicklung Bewertungen Leistung und Parteiensympathie Bundesregierung, CDU und Die Linke für Komponente 8 und Politbarometer 2009 bis 2016



Quelle: Eigene Darstellung. Politbarometer hinsichtlich Ost-West-Ungleichgewichte durch die Nutzung der Jahreskumulationen korrigiert (siehe Kapitel 5.2). Gestrichelte Linien stellen die Termine der Bundestagswahlen 2009 und 2013 dar. Durchgezogene Linie markiert den Wechsel des Betreibers des Access-Panels. Anzahl Umfragen sind: 14 (Respondi), 19 (LINK) und 118 (Politbarometer). Für sieben Politbarometer-Umfragen wurde direkt nach einem Regierungswechsel nicht die Bewertung der Leistung abgefragt.

Zusätzlich sind die Entwicklungen der Sympathiebewertungen der beiden Parteien aufgeführt. Bereits der grafische Verlauf zeigt die starke zeitliche Verbindung der Sympathie gegenüber einer Partei und die Bewertung ihrer Arbeit durch das Elektorat. Die identisch skalierten Fragen der Erhebungen des Politbarometers sind als Referenz einer zufallsbasierten telefonischen Befragung aufgeführt.[9] Auch wenn diese ebenfalls mit starken Problem der korrekten Erfassung soziodemografischer Merkmale konfrontiert sind, erfassen sie jedoch weiterhin auf einem zu persönlichen Befragungen vergleichbaren Niveau das Wahlverhalten. Da ihr Kostenvorteil gegenüber persönlichen Befragungen eine stärkere Erhebungsfrequenz erlaubt, eignen sie sich auch zur kurzfristigen Erfassung gesellschaftlicher Stimmungsschwankungen.

[9] Alle elfstufigen Skalen wurden lediglich mit einer monoten Transformation von +1 (negative Bewertung) bis +11 (positive Bewertung) auf den Bereich -5 bis +5 überführt.

Der Wechsel des Betreibers des Access-Panels (und somit der Rekrutierungsart) innerhalb der Komponente 8 der GLES führt zu einer klaren Veränderung der arithmetischen Mittel aller fünf abgefragten Skalen. Mit der Nutzung des offlinerekrutierten Access-Panel von LINK erlangt die CDU einen nachhaltigen Sympathieschub, ihre Arbeit wird positiver bewertet und auch die Arbeit der gesamten Bundesregierung wird wohlwollender gesehen. Diese Beobachtung unterstützt die Vermutung einer starken Verknüpfung der Sympathie und Bewertung der Arbeitsleistung von Parteien.[10] Gemäß des eindimensionalen Ansatzes von Fishbein (1963) ist folgende Verknüpfung der beiden Variablen denkbar: Die Leistung innerhalb der Bundesregierung stellt ein verknüpftes (gewichtetes) Merkmal der maßgeblichen Regierungspartei CDU dar. Kommt es zu einer stärkeren Unzufriedenheit mit der CDU als Partei, wirkt sich dies daher auch auf die gesamte Bewertung der Regierungsleistung aus.

Ein Effekt des Wechsel des Panel-Betreibers und seiner damit einhergehenden Rekrutierungsart findet sich auch auf die angegebene Sympathie und Bewertung der Oppositionsarbeit der Linken. Dieser Effekt ist nicht so prägnant ausgeprägt wie derjenige auf die CDU, reicht jedoch aus um ein visuell erkennbares Absenken der angegebenen Sympathie und der Bewertung der Oppositionsarbeit dieser Partei zu erreichen. Ein Einfluss der Rekrutierungsart in ein Onlinepanel zeigt somit auch langfristig stabile Auswirkungen auf die Entwicklung aggregierter Zeitreihen.

Die generell erfassten Trends sind größtenteils identisch zwischen den verwendeten Stichproben des jeweiligen Access-Panels und der jeweiligen Zeiträume des Politbarometers. Die Abweichungen zu diesem sind jedoch unterschiedlich zwischen diesen ausgeprägt. Zeigen die offlinerekrutierten Stichproben eine vergleichbare Entwicklung *und* ein vergleichbares Niveau zum Politbarometer auf, trifft dies für die onlinerekrutierten Stichproben nur auf die gezeichneten Veränderungen der Bewertungen zu. Insbesondere eine teilweise bessere Bewertung der Leistung der Oppositionsarbeit der Linken, als diejenige der CDU und der Bundesregierung, lässt starke Zweifel an einer unverzerrten Erfassung der wahren Bewertungen durch das Elektorat aufkommen.

Lassen sich als Erklärung dieser stärkeren Verzerrungen erneut dieselben Merkmale und Bewertungen der befragten Personen anführen, welche bereits das Kapitel 10.1 im Kontext der Bundestagswahl 2013 aufgezeigt hat? Das nachfolgende Unterkapitel wird die vorgestellten Korrekturmodelle des voran-

[10] Für die Politbarometer-Stichproben ergibt sich ein linearer Zusammenhang von $r = 0.95$ zwischen der jeweiligen Sympathie und der Bewertung der Arbeitsleistung, ebenso für die Umfragen $T3$ bis $T16$ der Komponente 8.

gegangenen Kapitels nun auf die gesamte Anzahl der vorliegenden Befragungen der Komponente 8 anwenden. Hierdurch wird die langfristige Stabilität dieser Korrekturmodelle, insbesondere für onlinerekrutierte Befragungen, aufgezeigt.

10.2.1. Korrektur: Unterschiede zu persönlichen Befragungen

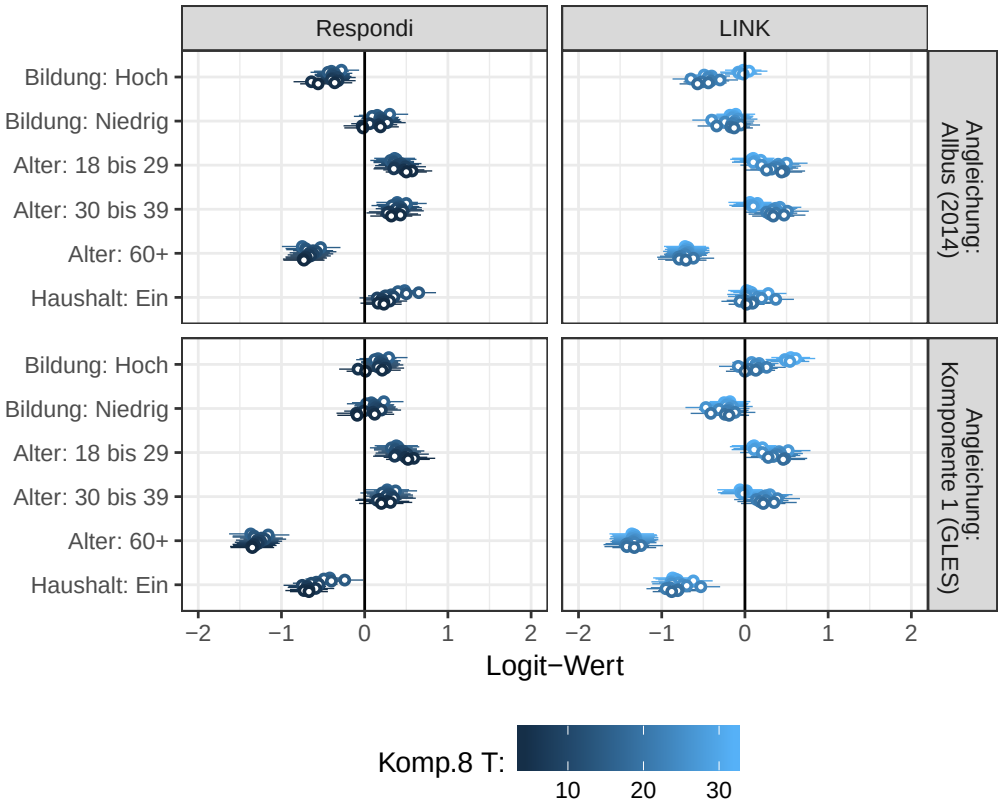
Um eine Vergleichbarkeit zu den Ergebnissen des Kapitels 10.1 zu garantieren, wird auch für die langfristige Analyse die *Vorwahl*befragung 2013 der Komponente 1 genutzt. Ist eine Korrektur durch diese möglich, sind die Verzerrungen der onlinebasierten Befragungen stabil und eine Korrektur durch zeitlich entfernt erhobene persönliche Referenzbefragungen möglich. Ein gesondert erhobener Referenzdatensatz in geringen zeitlichen Abständen ist dann *nicht* zwingend notwendig, die hohe zeitliche Stabilität der in Kapitel 7.2 aufgezeigten Wirkungskanäle der Korrektur sind daher zur Verbesserung der Qualität onlinebasierter Befragungen geeignet.

Eine zusätzliche Validierung erfolgt durch den Einbezug einer zweiten externen persönlichen Referenzbefragung: Der Allbus 2014. Im Rahmen der Erhebung wurden vergleichbare Variablen erhoben, welche das vorangegangene Kapitel auch zur Korrektur genutzt hat. Der Allbus stellt zudem keine explizite Wahlbefragung dar, wodurch die Möglichkeit der Angleichung über eine solche hinaus gezeigt werden kann. Sofern notwendig, werden abweichende Frageformulierungen zur GLES diskutiert und vorgenommene Anpassungen der Skalierungen erläutert. Alle nachfolgenden Korrekturmodelle werden zudem nach dem jeweiligen Access-Panel separat aufgeführt.

Die Abbildung 10.16 präsentiert die Ergebnisse des „Minimal“-Modells, welches für den Kontext der Bundestagswahl 2013 in Abbildung 10.4 durchgeführt wurde. Die Stichproben der Komponente 8 zeigen ein durchweg homogenes Bild der ausgewiesenen Effekte, welche sich zudem robust gegenüber der Wahl des Referenzdatensatz zeigen. Lediglich das Führen eines Einpersonenhaushalts zeigt gegenüber der Komponente 1 einen Unterschied zu den onlinebasierten Befragungen, welcher sich auf der Basis des Allbus nicht ergibt. Hat eine Person das sechzigste Lebensjahr bereits überschritten, ist ihre Teilnahmewahrscheinlichkeit an den onlinebasierten Befragungen, im Vergleich zu den 40- bis 59-jährigen Personen, eindeutig geringer. Die Wahrscheinlichkeit ist somit sehr hoch, dass diese Person innerhalb der beiden persönlichen Referenzumfragen befragt wurde.

Neben diesen grundlegenden soziodemografischen Merkmalen werden nun zusätzlich das Interesse an Politik, die empfundene Wahlnorm, die Demokratie-

Abbildung 10.16.: Unterschiede Komponente 8 2009 bis 2016 in Referenz zur Komponente 1 2013 und Allbus 2014 (Minimalmodell)



Quelle: Eigene Darstellung. Allbus mit V870 („Personenbezogenes Ost-West-Gewicht“) gewichtet, Komponente 1 mit w_trow. $T3$ bis $T32$ der Komponente 8 berücksichtigt. Alle aufgeführten Variablen stellen 0/1-Indikatoren dar. Ein positiver Logit-Koeffizient impliziert eine höhere Wahrscheinlichkeit der Zugehörigkeit zu den Befragungen der Komponente 8.

unzufriedenheit und die beiden Dimensionen der „Efficacy“ berücksichtigt. Das „Kompetenz/Norm“- und „Einstellungs“-Modell wird somit, abweichend zum Vorgehen des vorherigen Kapitels, direkt zusammengeführt. Diese Variablen wurden in einer überwiegenden Anzahl der Erhebungen der Komponente 8 berücksichtigt, erstmalig jedoch erst mit der Befragung T12. Die Stichproben der Erhebungen T3 bis T11 lassen sich daher nicht verwenden. Die Formulierung der Frage der empfundenen Wahlnorm ist durchweg konstant: „In der Demokratie ist es die Pflicht jedes Bürgers, sich regelmäßig an Wahlen zu beteiligen.“. In den Erhebungen 25, 29 und 33 wurde jedoch eine sieben- statt fünfstufige Skala verwendet. Da alle Variablen jedoch auf einen Wertebereich von 0 bis 1 reskaliert werden, ergibt sich kein Einfluss dieses Unterschieds.

Hinsichtlich der Dimensionen der „Efficacy“ wurde mit der Umfrage T20 zudem ein Wechsel der Formulierung und Skalierung der „External Efficacy“ durchgeführt, welche zudem das Bezugsobjekt von „Politikern“ zu „Parteien“

ändert.[11] Da die vorherigen ersten drei Erhebungen auf der Basis des LINK-Panels diese Frage nicht berücksichtigen, koinzidiert die Anpassung der Formulierung zudem mit einem Wechsel des Access-Panel. Diese Unstimmigkeit des Bezugsobjektes traf jedoch auch bereits auf die Erhebungen der Komponente 3 im vorherigen Kapitel zu, da die persönliche Befragung der Komponente 1 ebenfalls das Objekt „Partei“ als Bezug aufweist. Die dort erfolgreich durchgeführte Korrektur rechtfertigt daher die weitere Verwendung, trotz Abweichungen der Frageformulierungen zwischen den Onlineumfragen und der Referenzumfrage. Hinsichtlich der Dimensionen der „Internal Efficacy“ handelt es sich nur um eine leichte Modifikation der Formulierung, ohne jedoch das Bezugsobjekt („Politik als Ganzes“) und die Skalierung (fünfstufige Likert-Skalen) zu ändern.[12]

Der Allbus 2014 berücksichtigt ebenfalls die diskutierten Variablen. Das politische Interesse und der Grad der Demokratieunzufriedenheit wurden in vergleichbaren Formulierungen zu denjenigen der Komponente 8 abgefragt.[13] Im Gegensatz zu den fünfstufigen Likert-Skalen der GLES, wurde die Demokratieunzufriedenheit jedoch durch eine sechsstufige Skala operationalisiert. Die Verwendung dieser Variable trifft somit die implizite Annahme, dass die mangelnde Möglichkeit der Wahl einer neutral erscheinenden Mittelkategorie nicht das grundlegende Antwortverhalten der befragten Personen ändert.

Die beiden Dimensionen der „Efficacy“ wurden ebenfalls durch den Allbus 2014 berücksichtigt, auch wenn die „Internal Efficacy“ die Kompetenz relativ zu anderen Personen einschätzen lässt.[14] Hinsichtlich der „External Efficacy“ wird der Bezug der Responsivität auf der Ebene der Politiker angesiedelt.[15] Auch die empfundene soziale Norm des Wählens als Staatsbürgerpflicht wurde berücksichtigt. Entgegen der (meisten) Abfragen innerhalb der GLES, wurde eine siebenstufige Skala gewählt. Diese lässt sich über eine Normalisierung der Skala auf einen Wertebereich von 0 bis 1 jedoch vergleichen.[16] Zudem findet

[11] Von ‘Politiker kümmern sich darum, was einfache Leute denken.„ bis T20 zu „Die Parteien wollen nur die Stimmen der Wähler, ihre Ansichten interessieren sie nicht.“ ab T21.

[12] Von „Die ganze Politik ist so kompliziert, dass jemand wie ich nicht versteht, was vorgeht.“ bis T20 zu “Politische Fragen sind für mich oft schwer zu verstehen.“ ab T21.

[13] „Wie stark interessieren Sie sich für Politik?“ (V209) und „Wie zufrieden oder unzufrieden sind Sie alles in allem mit der Demokratie, so wie sie in Deutschland besteht?“ (V216).

[14] Abgefragt wurde „Die meisten Menschen in Deutschland sind über Politik und Regierung besser informiert als ich.“ (Variable: V770, Fünfstufig: 5 „Stimme überhaupt nicht zu“, ISSP-Befragung)

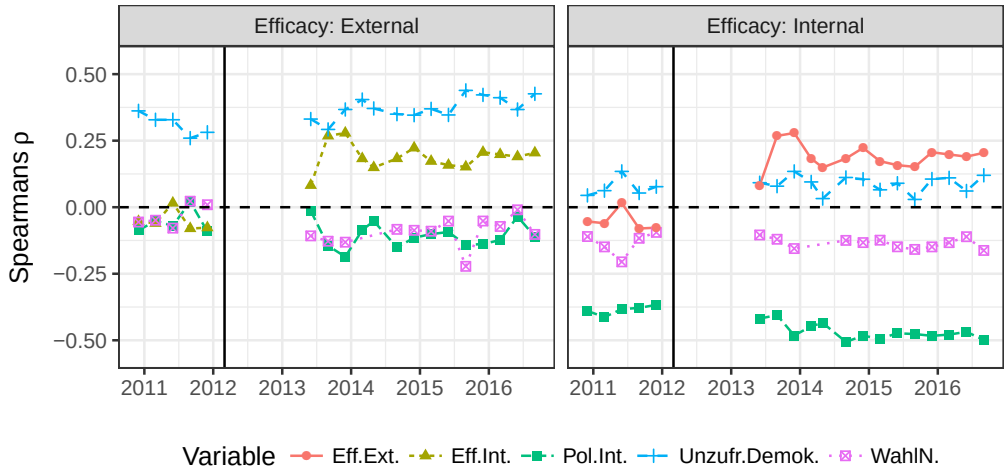
[15] Die Frageformulierung ist: „Die meisten Politiker sind nur wegen ihres persönlichen Vorteiles in der Politik.“ (Variable: V777, Fünfstufig: 5 „Stimme gar nicht zu“, ISSP-Befragung).

[16] Die Formulierung ist: „Was meinen Sie: Inwieweit sind folgende Dinge wichtig, um ein guter Bürger zu sein? Dass jemand...a. immer wählen geht.“ (Variable: V732, Siebenstufig: 7 „Sehr wichtig“, ISSP-Befragung).

die Abfrage der Dimensionen der „Efficacy“ und die Wahlnorm innerhalb der in den Allbus integrierten ISSP-Befragung statt. Da diese Fragen nur 1502 zufällig ausgewählten der insgesamt 3471 befragten Personen vorgelegt wurden, reduziert sich die maximal nutzbare Fallzahl des Allbus 2014 als Vergleichsbasis auf diesen Personenkreis.

Die in Abbildung 10.5 bereits gezeigten Zusammenhänge zwischen den zusätzlich berücksichtigten Variablen des „Kompetenz/Norm“- und „Einstellungs“-Modell zeigt ebenfalls die Abbildung 10.17. Auch langfristig sind diese durch eine hohe Stabilität gekennzeichnet.

Abbildung 10.17.: Zusammenhänge Wahlnorm, Politisches Interesse, Efficacy und Demokratieunzufriedenheit für Komponente 8 der GLES 2010 bis 2016



Quelle: Eigene Darstellung. Dargestellte Punkte sind die genutzten LOT-Erhebungen mit allen relevanten Variablen (T: 12 bis 16, 20 bis 22 und 25 bis 33). Eingezeichnet ist der Wechsel des Access-Panel.

Werden die Akteure des politischen Systems als nicht responsiv wahrgenommen (geringe „Efficacy: External“), geht dies mit einer erhöhten Demokratieunzufriedenheit einher. Eine Zustimmung des schwierigen Verständnis von Politik treffen insbesondere solche Personen, welche sich ebenfalls nur geringfügig für Politik interessieren.

Ebenfalls bestätigt sich, dass die Wahl des Access-Panel auch die Stärke, Existenz und Richtung von Zusammenhängen beeinflussen kann. Findet sich keine Verbindung der „External Efficacy“ zum Verständnis des und Interesse an Politik, sowie der empfundenen Wahlnorm innerhalb der Erhebungen auf der Basis des Respondi-Panel, zeigt sich solch eine bei einer Nutzung des LINK-Panel. Es handelt sich somit bei diesen Stichproben um einen durchgehend anderen Personenkreis: Eine empfundene Komplexität von Politik geht mit

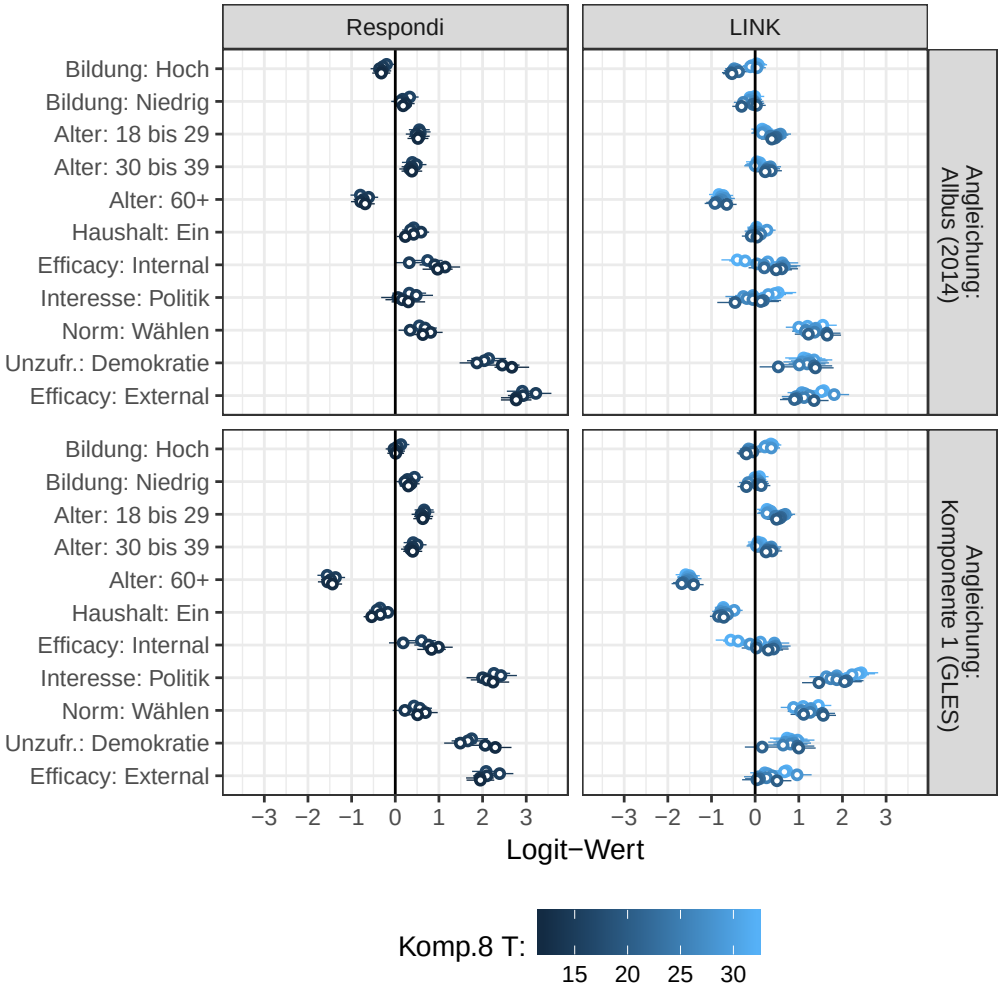
einem geringen Glaube an eine Responsivititität der Akteure einher. Innerhalb der Stichproben des Respondi-Panels überwiegt hingegen der mangelnde Glaube an die Responsivität systematisch über alle Gruppen hinweg. Ein Zusammenhang zur Komplexität von Politik zeigt sich nicht, auch Personen mit einem guten Verständnis der Komplexität von Politik weisen eine kritische Grundhaltung gegenüber den politischen Akteuren auf.

Dies bestätigen die Ergebnisse des kombinierten „Kompetenz-/Norm-“ und „Einstellungs“-Modells in Abbildung 10.18. Die Erhebungen des jeweiligen Access-Panels zeigen eine hohe interne Homogenität der Unterschiede zu den beiden gewählten Referenzumfragen. Es findet sich keine Umfrage innerhalb der Erhebungen des *jeweiligen* Access-Panel, welche systematisch andere Abweichungen aufzeigt. Auch bestätigen sich bisherige Ergebnisse: Der Personenkreis onlinebasierter Befragungen weist kein abweichendes Kompetenzverständnis von Politik zu denjenigen Personen einer persönlichen Referenzbefragung auf. Erfolgt die Rekrutierung einer Onlineumfrage offline über Telefonstichproben, ist die angegebene Zufriedenheit mit der Situation der Demokratie in der Bundesrepublik und die Einschätzung der Responsivität der Akteure des politischen Systems vergleichbar zu ebenfalls offline rekrutierten persönlichen Befragungen. Hingegen ist die empfundene Norm des Wählens als Bürgerpflicht im Vergleich zu diesen erhöht.

Die Ziehung von Stichproben aus einem onlinerekrutierten Access-Panel zeichnet ein anderes Bild. Die Unzufriedenheit mit der Demokratie in der Bundesrepublik und die pessimistische Einschätzung der Responsivität der politischen Akteure ist klar erhöht, wenn diese mit offline rekrutierten persönlichen Befragungen verglichen werden. Diese Unterschiede hinsichtlich der verwendeten Einstellungsfragen finden sich im Vergleich zu beiden Referenzumfrageprogrammen, auch wenn diese im Vergleich zum Allbus etwas stärker ist.

Hinsichtlich des politischen Interesses ändert sich dieses Bild. Im Vergleich zum Allbus zeigen die befragten Personen beider Access-Panel ein vergleichbares Interesse an Politik, schätzen diese (und ihre politischen Akteure) jedoch nur pessimistischer ein. Wird die Komponente 1 der GLES genutzt, ist das politische Interesse hingegen innerhalb der Onlineumfragen stärker ausgeprägt. Da im Vergleich zu den vorherigen Erhebungen des Allbus die Durchführung 2014 einen stark erhöhten Anteil an Personen mit einem starken oder sogar sehr starken Interesse an Politik aufweist, kann dies jedoch ebenfalls ein Problem der Qualität des Allbus zu sein (Abbildung 7.1 in Kapitel 7.2.3). Es ist daher nicht unwahrscheinlich, dass dieser selbst mit einer Verzerrung konfrontiert ist.

Abbildung 10.18.: Unterschiede Komponente 8 2009 bis 2016 in Referenz zur Komponente 1 2013 und Allbus 2014 („Kompetenz/Norm“- und „Einstellungs“-Modell)



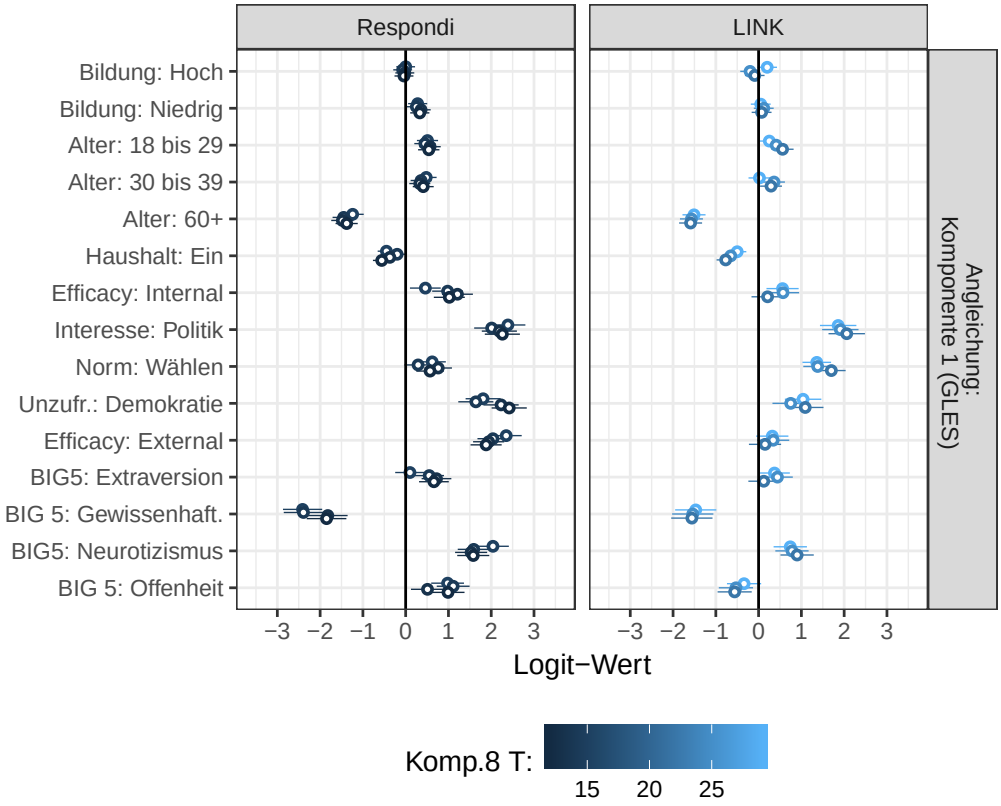
Quelle: Eigene Darstellung. Allbus mit V870 („Personenbezogenes Ost-West-Gewicht“) gewichtet, Komponente 1 mit w_trow. T12 bis T16, T20 bis T22 und T25 bis T33 der Komponente 8 berücksichtigt. Variablen auf einen Wertebereich von 0 bis 1 reskaliert über Minimal- und Maximalwerte, sofern notwendig.

Da dies jedoch nicht auf die Einstellungsfragen zutrifft, welche durchgängig ein vergleichbares Niveau zur Komponente 1 aufweisen, stellt der Allbus auf der Grundlage dieser weiterhin eine adäquate Referenzumfrage dar.

Im Rahmen der Erhebungen T12 bis T15, der T22, T25, T29 und T33 der Komponente 8 wurden ebenfalls die Persönlichkeitsmerkmale der „BIG Five“ abgefragt. Die Formulierungen sind durchgehend identisch zu denjenigen in Tabelle 10.3 des vorangegangenen Kapitels. Ebenfalls wurde nun auch über alle diese Erhebungen die Dimension der „Verträglichkeit“ durch die Formulierung „Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.“

abgefragt. Um das Modell jedoch konstant zur Korrektur im Kontext der Bundestagswahl 2013 zu halten (Abbildung 10.8), wurde auf eine Aufnahme dieser Dimension verzichtet. Die Abbildung 10.19 zeigt die Ergebnisse der Integration der Persönlichkeitsdimensionen in das „Einstellungs“-Modell. Der Allbus als Referenzdatensatz ist nicht aufgeführt, da dieser keine Operationalisierung und Abfrage der Persönlichkeitsmerkmale der BIG 5 enthält.

Abbildung 10.19.: Unterschiede Komponente 8 2009 bis 2016 in Referenz zur Komponente 1 2013 und Allbus 2014 („Persönlichkeits“-Modell)



Quelle: Eigene Darstellung. Allbus mit V870 („Personenbezogenes Ost-West-Gewicht“) gewichtet, Komponente 1 mit w_trow. T12 bis T15, T22, T25, T29 und T33 der Komponente 8 berücksichtigt. Variablen auf einen Wertebereich von 0 bis 1 reskaliert über Minimal- und Maximalwerte, sofern notwendig.

Erneut zeigt sich eine hohe Homogenität der aufgefundenen Unterschiede zwischen den Onlineumfragen und der persönlichen Referenzbefragung, welche die Effekte des Modells 10.8 des vorherigen Kapitel nun auch für einen längeren Zeitraum validieren. Stichproben auf der Basis des Respondi-Panel neigen zur Auswahl eines Personenkreises, welcher im Vergleich zur Komponente 1 einen erhöhten Grad an Offenheit, einen ausgeprägten Neurotizismus und eine stark verringerte Gewissenhaftigkeit kombiniert. Die Extraversion ist hingegen nur

geringfügig erhöht. Auch die Personen der Stichproben auf der Basis des LINK-Panels kombinieren eine verringerte Gewissenhaftigkeit mit einem erhöhten Neurotizismus, paaren dies jedoch mit einer leicht verringerten Offenheit.

Die formulierten Korrekturmodelle für onlinebasierte Befragungen auf der Basis von Access-Panels des vorangegangenen Kapitels 10.1 lassen sich somit auch langfristig anwenden. Die Stärke und Richtung der gefundenen Unterschiede zu einer *einmalig* durchgeführten Referenzumfrage sind auch noch in Stichproben präsent, welche Jahre vor oder nach dieser Referenzbefragung durchgeführt wurden. Doch gilt dies auch für den Einfluss dieser Korrektur? Findet sich ebenfalls eine konstante Verzerrung der Beteiligungsabsicht und des Wahlverhaltens? Und lässt sich dies ebenfalls durch die aufgeführten Modelle korrigieren und somit die erfasste Realität der Stichproben der onlinebasierten Befragungen stärker an die reale Meinung und Verhalten des gesamten Elektorats innerhalb der Bundesrepublik angleichen? Diese Fragen beantworten nun die nachfolgenden Kapitel.

10.2.2. Bestimmung der Bedeutungsgewichte

Um dies zu überprüfen, wurden auf der Basis dieser Korrekturmodelle ebenfalls Bedeutungsgewichte bestimmt. In allen Fällen werden durch diese Unterschiede der onlinebasierten Befragungen zu der persönlichen Referenzbefragung verringert. Neben einfachen PS-Gewichten wurde auch erneut ein PSM mit einer Subklassifikation und dem Nearest-, CEM- und Full-Algorithmus durchgeführt. Für das PS-Gewicht wurde die Inverse der resultierenden PS der angewandten logistischen Regression genutzt und auf ein arithmetisches Mittel von 1 normalisiert. Dies garantiert eine konstante Anzahl von befragten Personen bei einer Anwendung. Für das CEM wurde alle abgefragten fünfstufigen Likert-Skalen in die Kategorien „Ablehnung“, „Neutral“ und „Zustimmung“ zusammengefasst („coarsened“). Dies erleichtert die Existenz von exakten Paaren für diesen Algorithmus. Die Einteilung der Subklassifikation an Hand der PS basiert auf sechs Klassen, welche eine möglichst geringe Varianz der PS zueinander aufweisen. Die Definition des „caliper“ als Angabe der Nähe zweier Personen der Vergleichs- und Referenzumfrage für den Nearest-Algorithmus ist das 0.25-fache der Standardabweichung der PS. Weisen zwei Personen eine stärkere Distanz auf, sind diese nicht mehr „ähnlich“ genug. Der Full-Algorithmus weist ein Minimum von 0.1 und ein Maximum von zwei Kontrolleinheiten (onlinebasierte Befragungen) pro Treatmenteinheit (persönliche Referenzbefragung) auf. Um die Fallzahl konstant zu halten, sind alle Gewichtungsvariablen des PSM

ebenfalls auf ein arithmetisches Mittel von 1 normiert.

Schafft es das PSM mit den jeweiligen zur Verfügung stehenden Algorithmen eine ausreichende Balancierung der beiden am Matching beteiligten Gruppen zu erreichen? Probleme ergeben sich für die Stichproben der Komponente 8 auf der Grundlage des „Minimal"-Modells, wenn der Allbus als Referenzdatensatz genutzt wird und die Betrachtung über die standardisierten Differenzen der Balancierung der PS hinausgeht. Eine Angleichung an die Komponente 1 der GLES ist in einem besseren Maß möglich (Abbildung 10.20 nachfolgend).

Auch zu hohe Maximalgewichte können ein Problem darstellen. In diesem Fall wird das Antwortverhalten einer Person als stellvertretend für eine größere Gruppe angenommen. Erfasst das Korrekturmodell alle notwendigen Informationen, kann dies plausibel sein. Ist dies nicht der Fall, kann die gewichtete Betrachtung für einzelne Variablen ein stärker verzerrtes Bild erzeugen. Hohe Maximalgewichte von ca. 30 ergeben sich jedoch lediglich bei der Anwendung des CEM-Algorithmus. Die Subklassifikation erreicht bessere Werte von knapp unter 20, die direkte Nutzung der PS-Gewichte von knapp unter 10. Positiv sind erneut die resultierenden Gewichte auf der Basis des Full-Algorithmus. Hier zeigen sich konstante Maximalgewichte von lediglich knapp über fünf in Referenz zum Allbus und knapp unter fünf in Referenz zur Komponente 1. Einzelnen vorherigen Personen wird somit ein maximales inhaltliches Äquivalent der Bedeutung von fünf Personen nach der Anwendung dieser zugewiesen (Abbildung 10.21 nachfolgend).

Der Ausschluss von Personen durch fehlende Werte ist generell sehr gering und stellt kein nennenswertes Problem dar. Die prozentualen Anteile übersteigen in nur sehr wenigen Fällen 7.5 % der ursprünglichen befragten Personen der anzugleichenden Stichproben (Abbildung 10.22 nachfolgend). Nachfolgend werden nun diese unterschiedlichen Bedeutungsgewichte angewandt und ihr Einfluss auf die Wahlentscheidung und Beteiligungsabsicht untersucht.

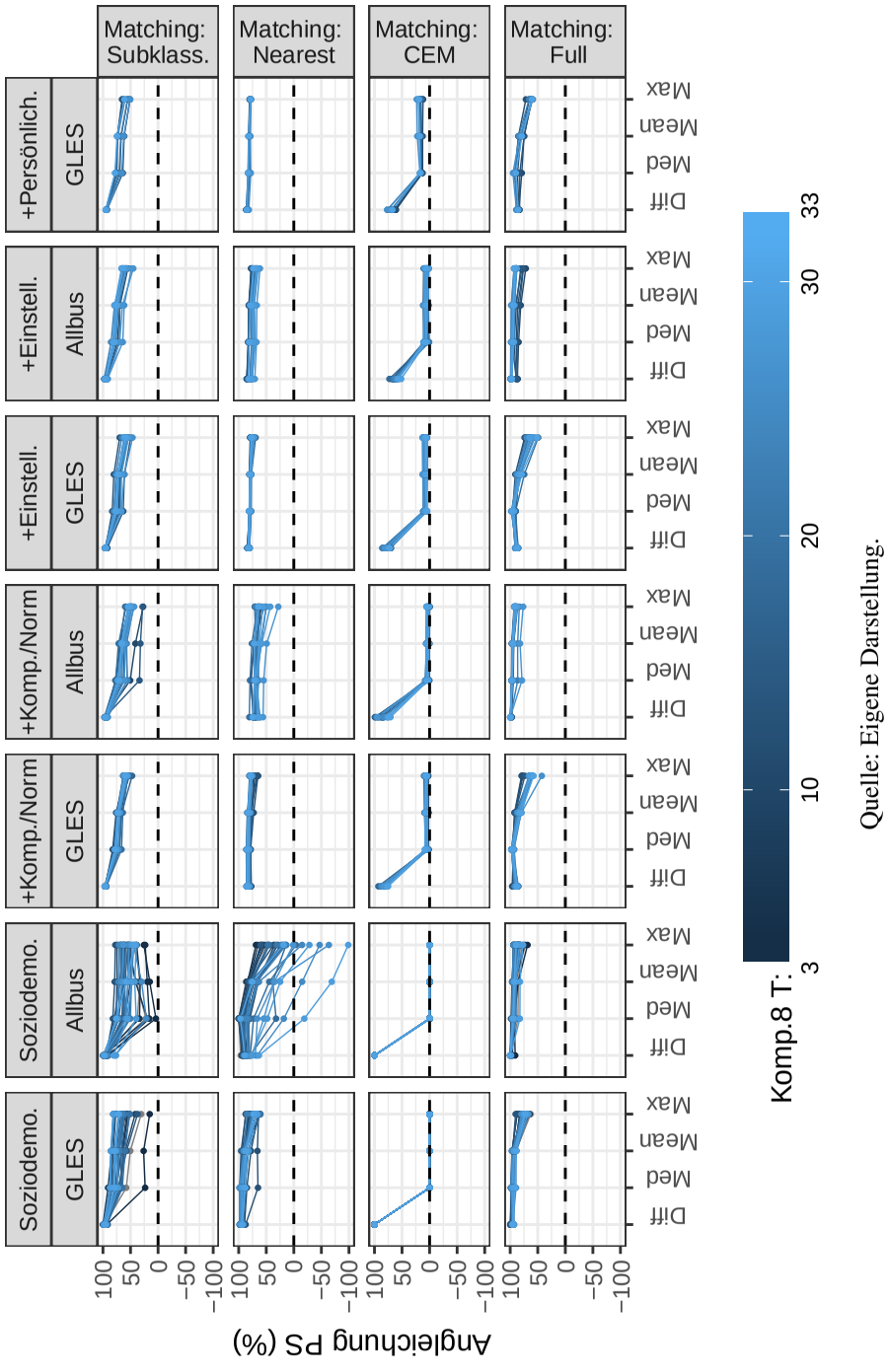
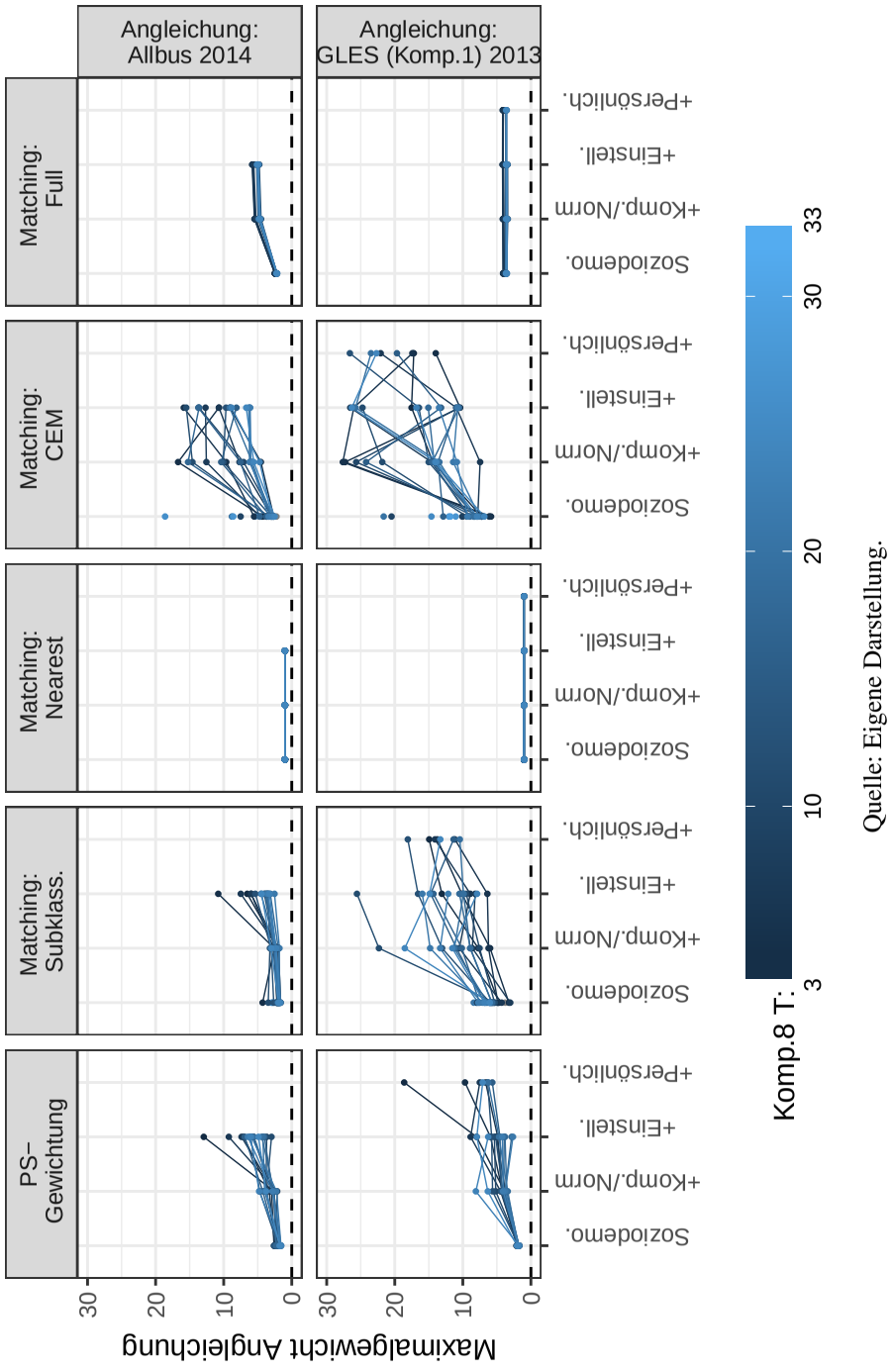


Abbildung 10.20.: Balancierung: Komponente 8 (Kapitel 10.2)

Quelle: Eigene Darstellung.



Quelle: Eigene Darstellung.

Abbildung 10.21.: Maximalgewichte: Komponente 8 (Kapitel 10.2)

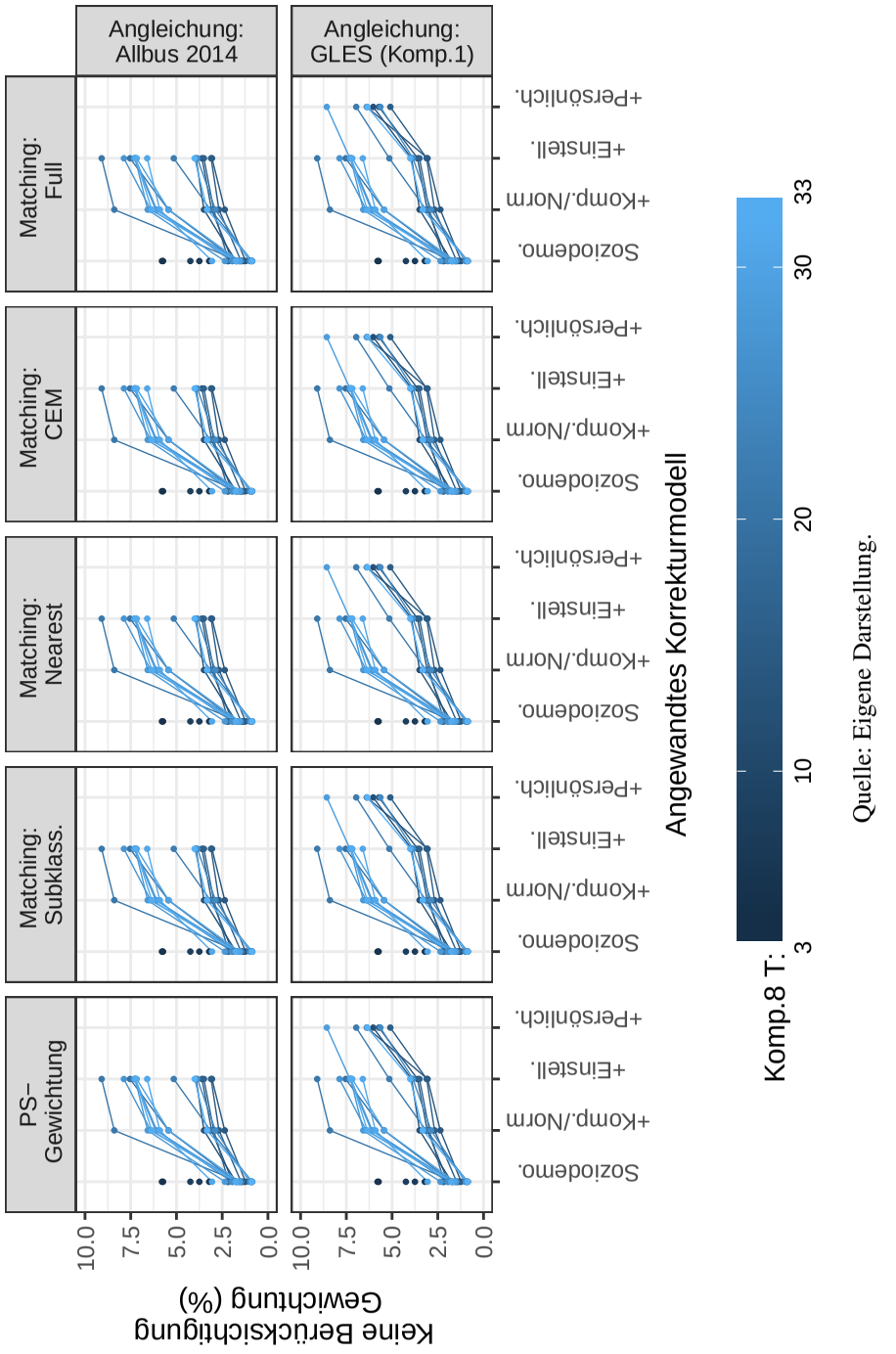


Abbildung 10.22.: Fehlende Werte Korrekturmodell: Komponente 8 (Kapitel 10.2)

Quelle: Eigene Darstellung.

10.2.3. Wahlverhalten und Beteiligungsabsicht 2009 bis 2016

Sorgt die Anwendung der bestimmten Gewichtungsvariablen für eine verbesserte Abbildung der wahlberechtigten Bevölkerung der Bundesrepublik durch die onlinebasierten Befragungen? Ob Veränderungen des Zweitstimmenanteils dieser Umfragen durch eine angewandte Gewichtung der Personen eine Verbesserung oder doch eine schlechtere Erfassung bedeuten, kann nur mit Hilfe einer geeigneten Referenzverteilung der Zweitstimmen bewertet werden. Das vorherige Kapitel konnte hierzu das amtliche Ergebnis der zeitlich naheliegenden Bundestagswahl 2013 nutzen. Da die (betrachteten) Erhebungen der Komponente 8 jedoch in einen Zeitraum von Juli 2009 (T3) bis August 2016 (T33) durchgeführt wurden, sind die Ergebnisse der Bundestagswahlen 2009 und 2013 nicht geeignet, da die Zweitstimmenpräferenz im Aggregat sich zu schnell verändern kann (siehe zu dieser Volatilität auch das Kapitel 7.3). Das Kapitel 9.2 hat jedoch eine vergleichbare Qualität der zufallsbasierten Erhebungen der persönlichen und telefonischen Befragungen der GLES im Kontext der Bundestagswahlen 2009 und 2013 herausgestellt, auch wenn die letztere Befragungsform zunehmende Probleme der korrekten Erfassung soziodemografischer Rahmenvariablen hat.

Eine mögliche Referenz stellen daher die (kommerziellen) telefonischen Erhebungen der Wahlumfragen von Emnid, Forsa, der Forschungsgruppe Wahlen und Infratest Dimap, sowie der persönlichen Befragung von Allensbach, dar. Deren Momentaufnahmen und Einschätzungen der Entwicklung der Zweitstimmenanteile sind über die Homepage wahlrecht.de offen beziehbar und decken den gesamten Zeitraum der Erhebungen der Komponente 8 ab.[17] Hierdurch sind die Ergebnisse von Befragungen mit einer zufallsbasierten Auswahl verfügbar und eine Aggregation der Ergebnisse dieser Umfragen erlaubt die im Durchschnitt vorhandene Einschätzung der Institute als eine zeitlich variierende Referenz.

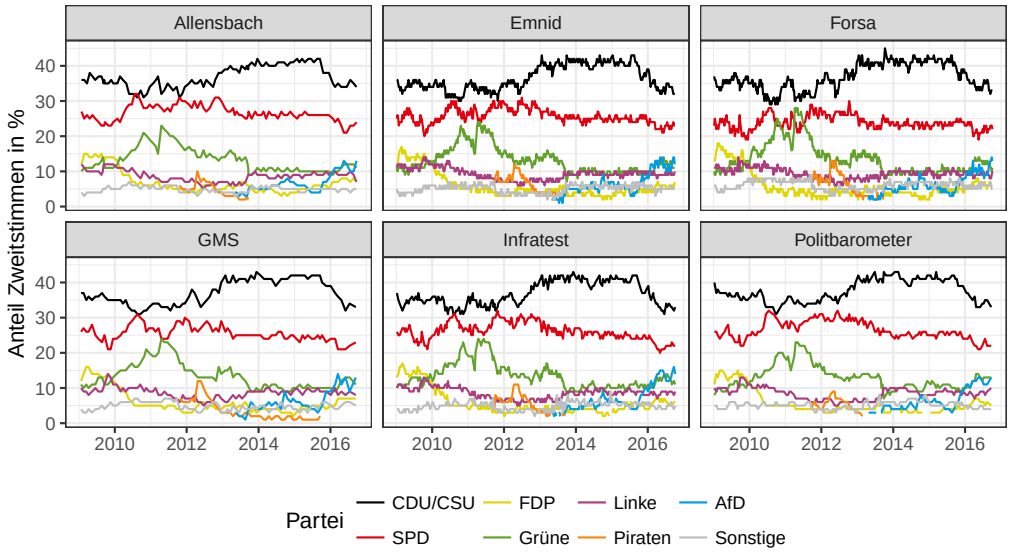
Unterschiede ergeben sich durch die zufallsbasierte Schwankung zwischen den Erhebungen, vor allem aber durch die Anwendung von Gewichtsverfahren nach einem institutsinternen Schlüssel. Die Berechnung der arithmetischen Mittel der Anteile über alle Erhebungen der Institute hinweg ergibt somit die durchschnittliche kollektive Einschätzung der führenden Meinungsforschungsinstitute in der Bundesrepublik zu einem bestimmten Zeitpunkt. Als Referenz der korrekten Erfassung über einen längeren Zeitraum hinweg ist dies die beste zur Verfügung stehende Möglichkeit für den Rahmen dieser Arbeit, wenn die Erfassung der zeitlichen Dynamik der Veränderung der Zweitstimmenanteile es-

[17] Beispielweise für Forsa über die Unterseite http://www.wahlrecht.de/umfragen/forsa.htm (Stand: 11. Mai 2018).

sentiell ist. Durch die Erhebungen von Allensbach wird zudem ein persönliches Frageprogramm berücksichtigt, alle anderen genutzten Institute nutzen die telefonische Befragungsform.

Für jeden genutzten Zeitpunkt wird das arithmetische Mittel der angegebenen Anteilswerte *einer* Partei über *alle* Institute hinweg bestimmt. Wie der visuelle Vergleich der Verlauf der Zweitstimmenanteile in Abbildung 10.23 zeigt, weisen alle Institute eine ähnliche Einschätzung des Verlaufs und Entwicklung der Zweitstimmenpräferenz des Elektorats auf, wenn auch in unterschiedlicher Intensität und variierenden Anteilswerten.

Abbildung 10.23.: Institute: Zweitstimmenanteile

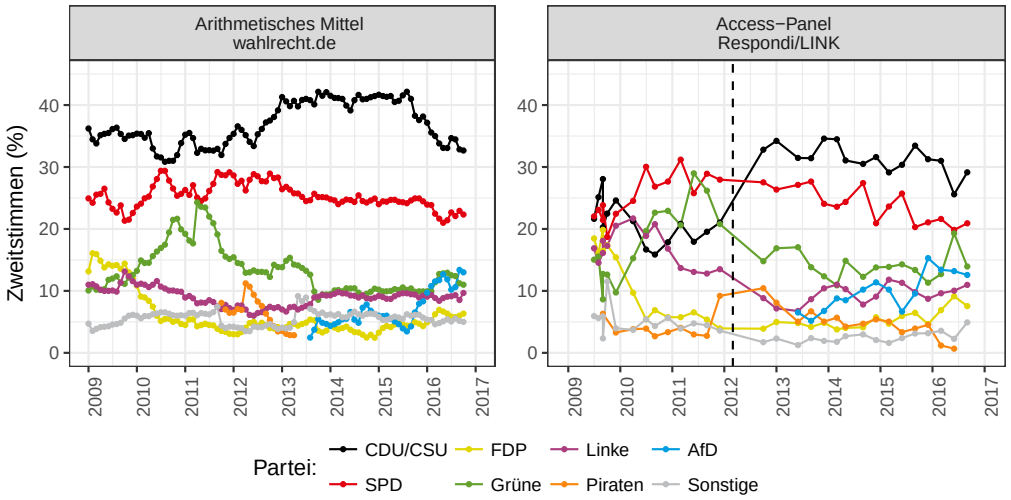


Quelle: Eigene Darstellung. Anteile sind über wahlrecht.de bezogen. GMS wurde zur Aggregation nicht berücksichtigt.

Als Zeitintervalle zusammenhängender Erhebungen wird der jeweilige Monat von Januar 2009 bis Dezember 2016 definiert. Zum Beispiel wurden für den Anteil der SPD im Februar 2014 die Anteilswerte *aller* angegebenen Umfragen auf wahlrecht.de aggregiert, deren Feldzeit ganz oder teilweise in diesen Monat fällt. Das resultierende arithmetische Mittel wird dann als ein näherungsweise wahrer Wert für diesen Monat angenommen und hierdurch mögliche Schwankungen und „Institutseffekte“ durch die jeweiligen individuellen Gewichtungsverfahren der Institute ausgeglichen.

Die Abbildung 10.24 vergleicht den Verlauf dieser gemittelten Zweitstimmenpräferenz des Elektorats mit der Einschätzung dieser auf der Basis der Erhebungen der Komponente 8 der GLES. Wird die Stichprobe aus der Grundgesamtheit des Respondi-Panel ausgewählt, bestätigt sich ein durchgehend zu geringer

Abbildung 10.24.: Zweitstimmenpräferenz Meinungsforschungsinstitute ($\bar{x}$) und Komponente 8 2009 bis 2016



Quelle: Eigene Darstellung. Durchschnitt bezieht sich auf alle Angaben von Umfragen eines Monats über wahlrecht.de

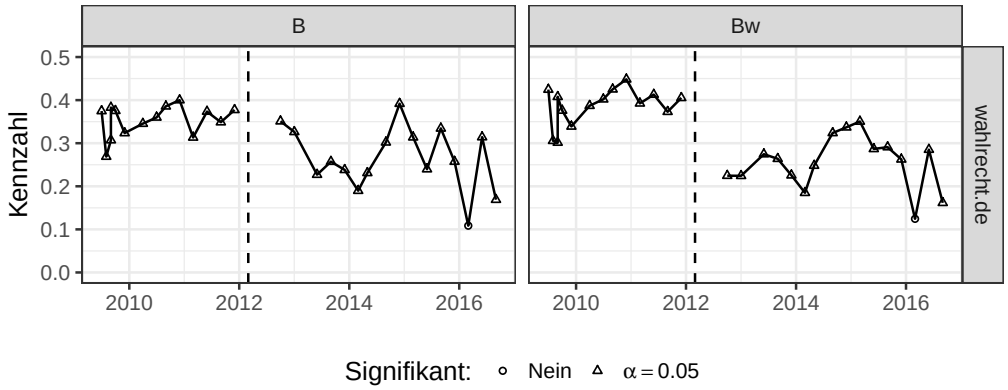
Anteil an Personen mit einer Präferenz für die Union. Auf der Grundlage dieser Erhebungen stellt sich vielmehr für die SPD die Frage, wer ihr Juniorpartner bei einer Regierungsbildung wird. Von dieser relativen Schwäche der Union profitiert vor allem „Die Linke“, welche 2010 einen Zweistimmenanteil von 20 % auf sich vereinbaren kann. Erst das, durch die Reaktorkatastrophe von Fukushima bedingte, Allzeithoch von „B90 / Die Grünen“ führt zu einem Absinken dieser auf unter 15 %. Eine Beobachtung, welche sich auf der Basis der aggregierten Referenzverteilung der Meinungsforschungsinstitute nicht konstatieren lässt. Dort geht der Anstieg Anfang 2011 mit einem Absinken der Anteile von Union und SPD einher.

Dass diese unterschiedlichen Einschätzungen durch das Erhebungsdesign bedingt sind, zeigt der Wechsel des Panelbetreibers Anfang 2012. Die Union kommt nun sprunghaft auf Werte von über 30 % und stellt nun auch durchgängig die größte Partei hinsichtlich der Zweitstimmenpräferenz. Weiterhin ist jedoch ebenfalls ein klare Diskrepanz von ungefähr zehn Prozentpunkten zu den Einschätzungen der führenden Meinungsforschungsinstitute vorhanden. Dieser Unterschied geht zu Gunsten der kleineren Parteien, vor allem die Grünen, aber auch z.B. der Piraten, welche von dieser geringeren Einschätzung des Potentials der Union profitieren können.

Wie lassen sich diese visuell bereits offensichtlichen Abweichungen zusammenfassen? Die Abbildung 10.25 aggregiert diese Unterschiede durch die Darstellung der Entwicklungen von B und B_w für den betrachteten Zeitraum.

Zur Berechnung wurden die jeweiligen gemittelten Anteile der Umfrageinstitute eines Monats als eine Referenzverteilung genutzt.[18] Der Zeitraum der Berücksichtigung der Piraten und der AfD als eigenständige Kategorie der Zweitstimmenpräferenz unterscheidet sich zwischen den verschiedenen Instituten, ebenso auch zur Komponente 8. Diese wurden daher nur als eine eigenständige Kategorie berücksichtigt, wenn sowohl *alle* Meinungsforschungsinstitute, als auch die Komponente 8 in einem Monat diese berücksichtigt haben. Andernfalls wurde die Berücksichtigung in manchen Umfragen dort der Kategorie „Sonstige“ zugerechnet.

Abbildung 10.25.: Zweitstimmenpräferenz Abweichung Komponente 8 zu Meinungsforschungsinstituten ($\bar{x}$) für 2009 bis 2016



Quelle: Eigene Darstellung. Referenzverteilung: Arithmetisches Mittel der Institute des jeweiligen Monats. Piraten und AfD nur berücksichtigt, sofern in allen Datenquellen vorhanden. Ansonsten Zuordnung der expliziten Berücksichtigung zu „Sonstige“. Eingezeichnete Referenzlinie ist der Zeitpunkte des Wechsel des Panelbetreibers.

Der Verlauf von B und B_w offenbart drei Aspekte: Unterschiede zwischen den Prognosen der Umfrageinstitute und der Entwicklung der Komponente 8 sind vorhanden, beide Kennzahlen weichen eindeutig und durchgängig von Null ab. Lediglich die erste Erhebung 2016 auf der Basis des LINK Panels sorgt für vergleichbare Verteilungen zu denjenigen der Meinungsforschungsinsitute. Diese Unterschiede sind zudem stärker ausgeprägt, wenn die Auswahl der Stichproben durch das Respondi-Panel vorgenommen wird. Diese weisen zudem höhere Werte für B_w aus und zeigen somit systematische Probleme der Berücksichtigung der Wählerschaft der relativ größeren Parteien, was in diesem

[18] Die Erhebungen der Komponente 8 basieren meist auf einem Monat. Wurde eine Erhebung über zwei Monate vollzogen, wurde der Monat mit den meisten Tagen als Grundlage genommen. Die Einteilung basiert auf den angegebenen Feldzeiten über die Variable field der jeweiligen Datensätze. Für eine Darstellung der Erhebungszeiträume siehe ebenfalls http://www.gesis.org/wahlen/gles/daten/ (Stand: 11. Mai 2018).

Fall auf die bereits aufgedeckten zu geringen Anteile der Union zurückzuführen ist (Abbildung 10.24). Finden sich bis Anfang 2012 Werte für B_w von ca. 0.4 für das Respondi-Panel, sinken diese bei einer Nutzung desjenigen von LINK auf unterhalb 0.3. Diese sind dort zudem im Durchschnitt geringer als die jeweiligen berechneten Werte von B.[19]

Die Abweichungen der angegebenen Zweitstimmenpräferenz der Stichproben des Link-Panels sind jedoch ebenfalls markant und die Anteilswerte beinahe durchgängig nicht vergleichbar zu denjenigen der (nach Anwendung der Institutsgewichte) aggregierten Zweitstimmenpräferenzen der kommerziellen Meinungsforschungsinstitute. Die jeweiligen anvisierten Grundgesamtheiten, welche für die Meinungsforschungsinstitute das Elektorat und für die nicht-zufallsbasierten und quotierten onlinebasierten Befragungen das jeweilige Access-Panel darstellen, unterscheiden sich also markant hinsichtlich ihrer jeweiligen Zweitstimmenpräferenz einer hypothetisch (oder in wenigen Fällen real) stattfindenden Bundestagswahl. Erfolgt die Rekrutierung in das Access-Panel offline, sind diese Unterschiede resultierender Stichproben zwar geringer, jedoch immer noch eindeutig vorhanden.

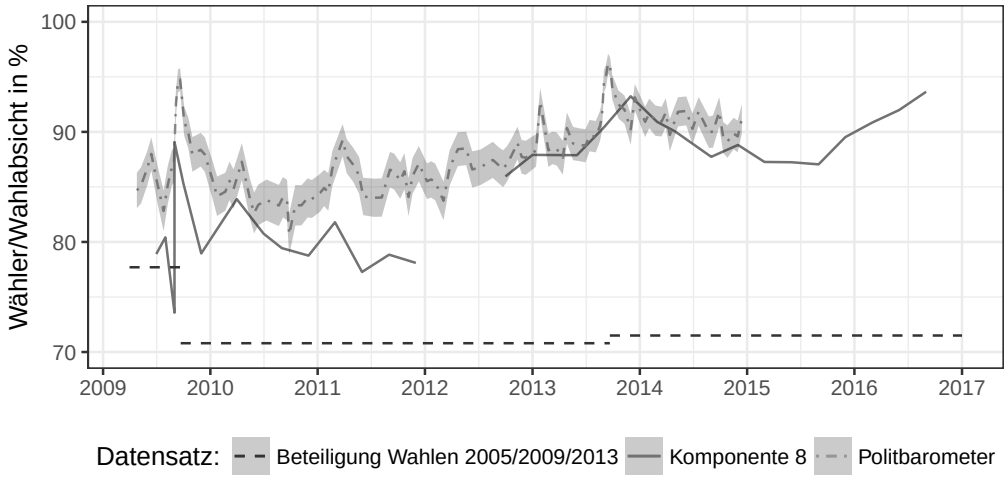
Telefonische Befragungen weisen jedoch ebenfalls Probleme auf, deren eigentlich anvisierte Grundgesamtheit des Elektorats adäquat zu erfassen: Sie tendieren zu einer Überschätzung der Beteiligungsabsicht an Wahlen. Das Kapitel 9 konnte bereits zeigen, dass dies (unter anderem) auf systematisch erhöhte Werte des angegebenen politischen Interesses der befragten Personen zurückzuführen ist. Eine Korrektur dieser zu hohen Werte zu vergleichbaren persönlichen Befragungen resultiert daher auch in einer Reduktion der erfassten Beteiligungsraten (siehe Abbildung 9.4 des Kapitel 9.1). Die Rekrutierung des Access-Panel von LINK über Telefonziehungen erzeugt ebenfalls Stichproben, welche ein erhöhtes politische Interesse in Relation zu persönlichen Befragungen aufweisen. Auch lassen sich erhöhte Zustimmungsraten zur Wahlnorm finden, welche sich innerhalb der Stichproben des Respondi-Panel nicht in dieser Stärke vorhanden sind (siehe hierzu Abbildung 10.18).

Auch hinsichtlich der Beteiligungsraten weisen diese Erhebungen erhöhte Werte gegenüber den vollständig online durchgeführten Erhebungen der Komponente 8 auf, wie der Verlauf der klassifizierten Wähler der Komponente 8

[19]Das arithmetische Mittel der Differenz von B_w und B beträgt 0.033 (Respondi) und -0.01 (LINK). Das LINK-Panel weist eine größere Bandbreite auf, die Standardabweichungen dieser Abweichungen betragen 0.02 (Respondi) und 0.05 (LINK). Dies lässt sich, bereits visuell, auf die Erhebungen in 2013 zurückführen.

in Abbildung 10.26 zeigt.[20] Die verwendeten Daten des Politbarometer sind identisch zu denjenigen aus Abbildung 9.4 des Kapitel 9.1.

Abbildung 10.26.: Entwicklung Wahlbeteiligungsabsicht Komponente 8 und Politbarometer für 2009 bis 2016



Quelle: Eigene Darstellung. Politbarometer: Ost-West-Ungleichgewichte korrigiert. Verlauf des 95 %-Konfidenzintervalls der Stichproben des Politbarometer an Hand der Auffassung als einfache Zufallsziehung nach Formel 6.5.

Die Erhebungen auf der Basis des Respondi-Panel ergeben Raten, welche überwiegend unterhalb derjenigen des Politbarometers liegen. Die Nutzung des LINK-Panel resultiert hingegen in Stichproben, welche vergleichbar hohe Beteiligungsraten zu denjenigen des Politbarometers aufweisen. Die bessere Abbildung der Zweitstimmenpräferenz durch die zufallsbasierte Rekrutierung in dieses Access-Panel hinein, geht daher zu Lasten einer zu hohen Einschätzung der Beteiligungsabsicht durch eine selektive Auswahl der Personen.

Sind diese bisherigen Beobachtungen des Kapitels, wie bereits in den vorherigen Kapiteln für telefonische und onlinebasierte Befragungen im Kontext der Bundestagwswahl 2013, korreliert? In diesem Fall müsste eine Korrektur der zu hohen Raten des Interesse an Politik und der empfundenen Wahlnorm durch die in Kapitel 10.2.2 bestimmten Bedeutungsgewichte auch zu einer Reduktion der sich zu hoch ergebenden Beteiligungsraten in den offlinerekrutierten Varianten führen (siehe Abbildung 10.26). Auch sollte sich in den onlinerekrutierten Erhebungen der Komponente 8 eine Reduzierung der Unterschiede zu den durchschnittlich reportierten Zweitstimmenanteilen der führenden Meinungsforschungsinstitute aus Abbildung 10.25 ergeben und somit B und B_w sinken. Dies

[20] Eine Person wurde als ein solcher eingeordnet, wenn diese „bestimmt“ oder „wahrscheinlich“ zur Wahl gehen wird.

bedingt sich durch eine Angleichung der onlinebasierten Befragungen an die anvisierte Grundgesamtheit der telefonischen Erhebungen der Institute, welche auf der Grundlage zufallsbasierter Auswahldesigns das Elektorat der Bundesrepublik anvisieren.

10.2.4. Auswirkung der Bedeutungsgewichte auf das Wahlverhalten und die Beteiligungsabsicht

Der Effekt der Anwendung der in Kapitel 10.2.2 bestimmten Bedeutungsgewichte wird nun nachfolgend in zwei Schritte gezeigt. Erstens wird untersucht, ob die Auswirkung einer Gewichtung auf die Zielvariablen der Wahlentscheidung und Wahlbeteiligung abhängig vom formulierten Korrekturmodell ist. Hierzu weist die Tabelle 10.4 die Ergebnisse sechs verschiedener Regressionsmodelle aus. Separat für jedes Access-Panel werden die Veränderungen von B, B_w und der definierten Wahlbeteiligungsrate (WBT) in Abhängigkeit des jeweils angewandten Korrekturmodells dargestellt. Tabelle 10.5 überträgt dieses Prinzip anschließend, um den den Einfluss des gewählten Algorithmus des PSM auf den Erfolg der Reduktion von B, B_w und der Wahlbeteiligungsrate zu untersuchen. Durch diese separaten Analysen wird beantwortet, ob erstens das angewandte Modell sowie der unterliegende Algorithmus generell überhaupt eine Verbesserung bewirkt und zweitens, welches Modell und welcher Algorithmus die besten Ergebnisse über die vorliegenden Stichproben erreichen kann.

Weiterhin kann die getroffene Wahl der persönlichen zufallsbasierten Referenzumfrage einen zusätzlichen Effekt auf die Korrektur haben. Hierzu wird zwischen dem (durchschnittlichen) Effekt der Angleichung an den Allbus und die Erhebung der Komponente 1 der GLES differenziert. Da ebenfalls die Anzahl der befragten Personen (n) und der zeitliche Kontext der Befragung einen Einfluss auf die Qualität einer Befragung haben können (siehe z.B. das Kapitel 9 der zeitlich zunehmenden Probleme telefonischer Befragungen), werden auch diese als Kontrollvariablen mit in das Modell aufgenommen. Für letzteres wird das Startjahr der Erhebungen des jeweiligen Access-Panels als Ausgangsbasis verwendet und fortgeschrieben. Die Erhebungen von LINK starteten in 2012. Wurde eine Erhebung nun 2014 durchgeführt, erhält sie somit den Wert +2.

Die Tabelle 10.4 zeigt den durchschnittlichen Effekt der Anwendung der unterschiedlichen Korrekturmodelle und jeweiligen gewählten Referenzumfrage aus dem Kapitel 10.2.1, unabhängig vom jeweils gewählten Matchingalgorithmus.

Tabelle 10.4.: Modell: Veränderung Komponente 8 für B, B_w (Meinungsforschungsinstitute) und Wahlbeteiligungsabsicht durch Gewichtung (Korrekturmodelle)

	Respondi			LINK		
	B	B_w	WBT	B	B_w	WBT
Konstante	0.31^{***}	0.38^{***}	82.22^{***}	0.28^{***}	0.33^{***}	93.48^{***}
	(0.03)	(0.03)	(1.95)	(0.05)	(0.05)	(1.42)
n	-0.0005	-0.004	-0.15	0.01^{*}	0.001	-0.40^{***}
	(0.002)	(0.003)	(0.16)	(0.003)	(0.004)	(0.12)
Jahr	0.005^{***}	0.005^{***}		-0.003^{***}	-0.003^{***}	
	(0.001)	(0.001)		(0.001)	(0.001)	
Vergleich: Allbus			Referenz			
Vergleich: Komp. 1	-0.01^{*}	-0.004	-2.34^{***}	0.0004	-0.003	-1.59^{***}
	(0.01)	(0.01)	(0.41)	(0.01)	(0.01)	(0.26)
Keine Korrektur:			Referenz			
Minimalmodell	-0.02	-0.02	3.15^{***}	-0.01	-0.01	1.66^{**}
	(0.01)	(0.02)	(0.88)	(0.02)	(0.02)	(0.68)
Kompetenz/Norm	-0.01	-0.01	-1.05	0.01	0.02	-1.14
	(0.01)	(0.02)	(0.96)	(0.02)	(0.02)	(0.70)
Einstellungen	-0.06^{***}	-0.12^{***}	0.65	-0.004	-0.02	-0.11
	(0.02)	(0.02)	(1.01)	(0.02)	(0.02)	(0.71)
Persönlichkeit	-0.07^{***}	-0.11^{***}	-0.36	-0.001	-0.06^{**}	-0.26
	(0.02)	(0.02)	(1.25)	(0.03)	(0.03)	(0.93)
Beobachtungen	275	275	275	436	436	436
R^2	0.23	0.35	0.31	0.05	0.10	0.21
Korrigiertes R^2	0.21	0.33	0.29	0.04	0.08	0.20

Note: *p<0.1; **p<0.05; ***p<0.01

Quelle: Eigene Darstellung.

Hinsichtlich B und B_w der Zweitstimmenpräferenz sind die positiven Effekte der Korrektur innerhalb der Respondi-Erhebungen stärker und gehen auch von einem höheren Ausgangsniveau der Verzerrung aus („Konstante“). Dies haben auch die bereits in Abbildung 10.24 und 10.25 gezeigten visuell stärkeren Unterschiede der Erhebungen bis 2012 auf der Basis des Respondi Access-Panels angedeutet.

Zudem findet sich nur ein sehr geringer Effekt der Kontrollvariable der Erhöhung der Stichprobengröße, was sich jedoch durch die geringe Variation dieser zwischen den Erhebungen erklären lässt. Es wurden $\bar{x} = 1081$ Personen mit $s_x = 56$ Personen über alle Erhebungen der Komponente 8 befragt. Auch zeigt sich eine steigende zeitliche Verzerrung innerhalb der Respondi-Stichproben. Innerhalb derjenigen von LINK ist hingegen eine Verringerung der Abweichungen zu den telefonischen Befragungen der Meinungsforschungsinstitute feststellbar. Ist die Wahl der persönlichen Referenzumfrage relevant? Der

Effekt des Wechsels der Angleichung vom Allbus hin zur persönlichen Befragung der Komponente 1 ist nur sehr gering und deren Qualität der Angleichung ist grundsätzlich gleichwertig.

Wichtiger für eine erfolgreiche Angleichung an die Referenzverteilungen der Zweitstimmen von wahlrecht.de ist die Wahl des formulierten Korrekturmodells. Der Einbezug der soziodemografischen Merkmale („Minimal"-Modell) oder das Interesse und Verständnis von Politik und der empfundenen Wahlnorm („Kompetenz/Norm") lassen keine Reduktion der Unterschiede der Wahl*entscheidung* zu. Wie bereits im Kontext der Bundestagswahl 2013 aufgedeckt (siehe Kapitel 10.1), ist für eine erfolgreiche Korrektur der Respondi-Stichproben erneut der Einbezug der Einschätzung der Responsivität des politischen Systems und der Demokratieunzufriedenheit („Einstellungen") notwendig. Erst durch diese Informationen über die befragten Personen der Erhebungen von Respondi und der jeweiligen Referenzumfragen zur Angleichung lässt sich eine Verbesserung, gegenüber einer ungewichteten Betrachtung, erreichen. Zudem ist der Effekt auf B_w stärker, eine Verringerung der Abweichungen ist also insbesondere durch eine Angleichung der Anteilswerte der relativ größeren Parteien erfolgt.

Der zusätzliche Einbezug der Persönlichkeitsmerkmale sorgt hingegen erneut nur für eine geringfügige weitere Verbesserung innerhalb der Respondi-Stichproben. Anders stellt sich dies bei einer Betrachtung der LINK-Erhebungen dar. Hier sorgt erst der Einbezug der Unterschiede der Persönlichkeitsmerkmale zwischen den befragten Personen der persönlichen Befragung der Komponente 1 (der Allbus hat diese Information nicht erhoben) und den Onlineumfragen für eine starke Verbesserung von B_w. Die vorherigen Korrekturmodelle haben hingegen keine Verbesserung ergeben.

Ist zudem das Ziel die zu hohen Wahlbeteiligungsraten der onlinebasierten Befragungen zu reduzieren, eignet sich die persönliche Befragung der Komponente 1 über alle onlinebasierten Befragungen besser als eine Referenz. Dies ist auf die geringen Unterschiede zwischen den Onlineumfragen der Komponente 8 und des Allbus hinsichtlich des Interesse an Politik und der empfundenen Wahlnorm zurückzuführen (Abbildung 10.18). Durch das insbesondere zu hohe Interesse an Politik im Allbus 2014 gegenüber den vorherigen Erhebungen, lassen sich die zu hohen Raten der Onlineumfragen technisch nicht ausreichend reduzieren.

Doch lässt sich, neben dieser elementaren Funktion eines gut spezifizierten Korrekturmodells für den Erfolg, ein weiterer Vorteil der jeweiligen technischen Art der Angleichung herausstellen? Hat also die Wahl des verwendeten Algorithmus des PSM einen Einfluss auf den Erfolg der Korrektur? Die Tabelle 10.5

stellt dies nun heraus. An Stelle der verwendeten Korrekturmodelle wird nun der Effekt der jeweiligen Algorithmen betrachtet.

Tabelle 10.5.: Modell: Veränderung Komponente 8 für B, B_w (Meinungsforschungsinstitute) und Wahlbeteiligungsabsicht durch Gewichtung (Matchingalgorithmen)

	Respondi			LINK		
	B	B_w	WBT	B	B_w	WBT
Konstante	0.18***	0.12***	77.42***	0.32***	0.30***	86.73***
	(0.03)	(0.04)	(2.36)	(0.06)	(0.06)	(1.93)
n	0.01***	0.02***	0.27	0.002	0.004	0.26
	(0.003)	(0.003)	(0.19)	(0.005)	(0.01)	(0.18)
Jahr	0.004***	0.003***		−0.003***	−0.003***	
	(0.001)	(0.001)		(0.001)	(0.001)	
Vergleich: Komp. 1	−0.01*	−0.001	−2.44***	−0.002	−0.01	−1.31***
	(0.01)	(0.01)	(0.45)	(0.01)	(0.01)	(0.27)
PS-Gewichtung	−0.003	0.03	2.68**	−0.01	−0.01	1.46*
	(0.02)	(0.02)	(1.17)	(0.02)	(0.02)	(0.86)
Matching: Subklass.	−0.03**	−0.04*	1.90*	−0.003	−0.01	−0.29
	(0.01)	(0.02)	(1.05)	(0.02)	(0.02)	(0.74)
Matching: Nearest	0.01	0.02	1.12	−0.003	0.01	1.58**
	(0.01)	(0.02)	(1.08)	(0.02)	(0.02)	(0.75)
Matching: CEM	−0.01	−0.01	3.14***	0.01	0.0001	1.79**
	(0.01)	(0.02)	(1.06)	(0.02)	(0.02)	(0.75)
Matching: Full	−0.02*	−0.03*	1.99*	−0.004	−0.01	−0.21
	(0.01)	(0.02)	(1.05)	(0.02)	(0.02)	(0.74)
Beobachtungen	275	275	275	436	436	436
R^2	0.18	0.17	0.15	0.05	0.05	0.15
Korrigiertes R^2	0.16	0.15	0.13	0.03	0.03	0.13

Note: *p<0.1; **p<0.05; ***p<0.01

Quelle: Eigene Darstellung.

Zwei Algorithmen zeigen hinsichtlich einer Korrektur der Stichproben des Respondi-Panel einen im Durchschnitt besseren Erfolg im Vergleich zu einer einfachen PS-Gewichtung: Die Subklassifikation und der Full-Algorithmus. B und B_w sinken bei einer Anwendung dieser, unabhängig vom angewandten Korrekturmodell, am stärksten. Auf die Wahlbeteiligungsrate wirkt sich die Korrektur aller Algorithmen bei einer direkten Betrachtung der Koeffizienten nachteilig aus, diese steigt. Absolut gilt dies jedoch nur für eine Nutzung des Allbus, da der reduzierende Effekt der Verwendung der Komponente 1 an Stelle des Allbus mit -2.44 größer ist also diejenigen der meisten Matchingalgorithmen. Insbesondere die Subklassifikation und der Full-Algorithmus, neben einem Nearest-Matching, sind erneut erfolgreich.

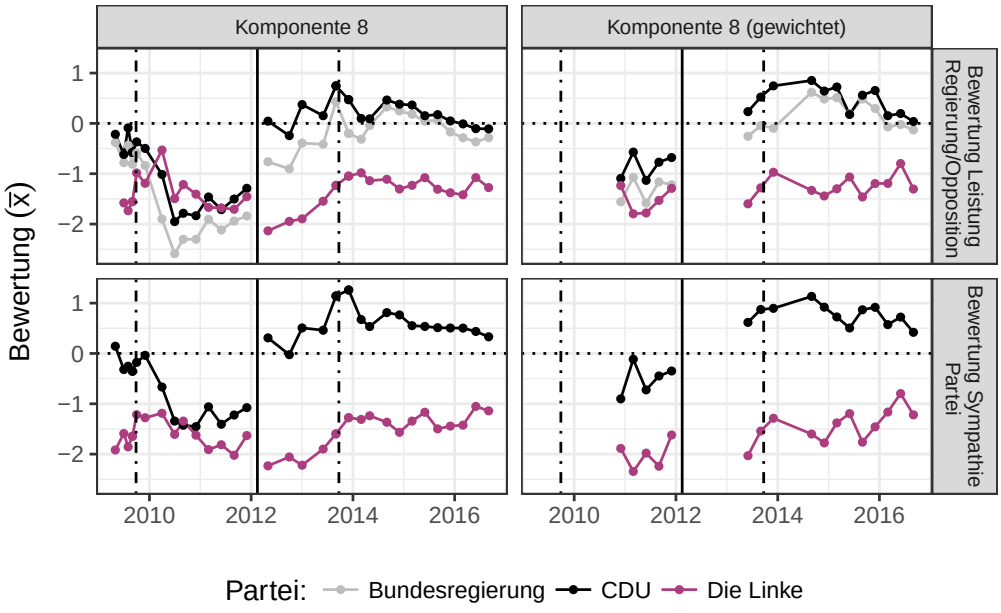
Generell lassen sich diese Ergebnisse auch für das Link-Panel validieren, die Auswirkungen der technischen Wahl des Algorithmus sind dort jedoch schwächer ausgeprägt, was auch die geringen Werte der Modellgütekriterien

unterstreichen. Dies ist wenig verwunderlich, haben doch bereits die Korrekturmodelle kaum eine Verbesserung dieser ermöglicht. Bei einer Betrachtung der Wahlbeteiligungsraten stellt sich jedoch auch ein nachhaltiger positiver Effekt ein. Hierbei glänzen erneut die Subklassifikation und der Full-Algorithmus.

Wirkt sich dies nun ebenfalls auf weitere Zielvariablen einer Wahlumfrage aus, wie die Sympathie der beiden Parteien CDU und Linke, sowie die Bewertung ihrer Arbeit innerhalb des politischen Systems? Lassen sich die visuell bereits erkennbaren Unterschiede zum Politbarometer und der Effekt des verwendeten Access-Panels reduzieren, wie sie die Abbildung 10.15 bereits am Anfang des Kapitels aufgezeigt hat?

Die Abbildung 10.27 stellt den Verlauf der arithmetischen Mittel der Erhebungen der Komponente 8 erneut dar. Im Kontrast zur ungewichteten Betrachtung

Abbildung 10.27.: Veränderung Bewertung Leistung und Parteiensympathie Bundesregierung, CDU und Die Linke durch Gewichtung für Komponente 8 2010 bis 2016



Quelle: Eigene Darstellung. Gestrichelte Linien stellen die Termine der Bundestagswahlen 2009 und 2013 dar. Durchgezogene Linie markiert den Wechsel des Access-Panel. Anzahl Umfragen sind ungewichtet: 14 (Respondi), 19 (LINK). Gewichtet: 5 (Respondi), 12 (LINK).

werden nun als Beispiel die Bedeutungsgewichte des kombinierten „Kompetenz/Norm“- und „Einstellungs“-Modells genutzt, welche technisch durch einen Full-Algorithmus im Rahmen des PSM bestimmt wurden. Dies kann jedoch nur für die Umfragen erfolgen, welche auch die Variablen dieses Korrekturmodells abgefragt haben. Insbesondere die Anzahl der Stichproben des Respondi-Panel reduziert sich hierdurch von anfänglich 15 auf fünf.

Bereits grafisch ist für diese ein positiver Effekt der Anwendung der Bedeutungsgewichte eindeutig erkennbar: Die Bewertung der Leistung der Bundesregierung und derjenigen der CDU als Regierungspartei verzeichnet einen Anstieg der arithmetischen Mittel durch die Gewichtung, die Bewertung der Leistung der Linken als Oppositionspartei sowie deren Sympathie sinkt hingegen. Die Erhebungen des LINK-Panel zeigen hingegen nur geringfügige Änderungen, was durch die vorab bereits präsentierten Ergebnisse wenig überraschend ist.

Um eine präzisere Angabe der Auswirkung der Anwendung dieses Bedeutungsgewichts auf die arithmetischen Mittel zu ermöglichen, weist die Tabelle 10.6 die Veränderungen dieser durch die Differenz der ungewichteten und gewichteten Werte aus. Zeigt sich für die LINK-Erhebungen ein nur geringfügig von Null abweichender Effekt, ist dieser für die Umfragen auf der Basis von Respondi eindeutig vorhanden. Im arithmetischen Mittel steigt durch die Anwendung des Bedeutungsgewichts sowohl die Bewertung der Leistung der Bundesregierung und der CDU als Regierungspartei, als auch empfundene Sympathie gegenüber der CDU um durchschnittlich 0.7 Skalenpunkte auf dieser elfstufigen Skala.

Tabelle 10.6.: Veränderung Bewertung Leistung und Parteiensympathie Bundesregierung, CDU und Die Linke durch Gewichtung für Komponente 8 2010 bis 2016 (Durchschnitt)

Institut:	Respondi	Respondi	Respondi	LINK	LINK	LINK
Kennzahl:	$\bar{x}$	Min	Max	$\bar{x}$	Min	Max
Bundesreg.	0.70	0.53	0.83	-0.02	-0.42	0.39
CDU (Reg.)	0.71	0.58	0.89	-0.19	-0.60	0.22
Linke (Opp.)	0.06	-0.13	0.18	0.01	-0.22	0.28
CDU (Symp.)	0.74	0.55	0.95	-0.11	-0.41	0.36
Linke (Symp.)	-0.21	-0.43	0.01	-0.04	-0.26	0.26

Quelle: Eigene Darstellung.

Diese Veränderung ist zudem sehr konstant über die fünf verfügbaren Stichproben, die Minimal- und Maximalwerte weichen nicht sehr stark voneinander ab und ein positiver Effekt um mindestens einen halben Skalenpunkt stellt sich in allen Fällen ein. Dies zeigt die letztendlich gemeinsame Ursache des Wirkungskanals: Die Nutzung des Access-Panel von Respondi als eine synthetische Grundgesamtheit zur Stichprobenziehung ergibt einen befragten Personenkreis, welcher im Vergleich zu persönlichen Referenzumfragen und ihrer Grundgesamtheit des gesamten Elektorats sehr unzufrieden mit der Gesamtsitutation der Demokratie und der Responsivität der Akteure des politischen Systems ist.

Diese Gruppe an Personen hat zudem eine geringere Wahrscheinlichkeit der Union ihre Zweitstimme zu geben. Werden diese Unterschiede zu persönlichen Befragungen jedoch über Bedeutungsgewichte im Rahmen eines PSM negiert, wirkt sich dies auch auf das Zweitstimmenverhalten aus. Die Anteilswerte der Union erhöhen sich, somit steigt dann auch beispielsweise die Zufriedenheit mit der Leistung der CDU innerhalb der Regierung und die empfundene Sympathie ihr gegenüber. Die Stichproben des telefonrekrutierten Access-Panels von LINK zeigen hingegen keine Veränderung bei einer Anwendung dieser Korrektur, da diese Unterschiede nicht vorhanden oder nicht so stark ausgeprägt sind.

Eine letzte Einschränkung ist der hier betrachtete Zeitpunkt des Erfolgs: Die vorliegenden Respondi-Erhebungen wurden vor allem in 2011 durchgeführt. In dieser Phase führte die Union die schwarz-gelbe Koalition an und der Erfolg der Korrektur und die aufgezeigten Verbindungen könnten an diesem spezifischen Fall liegen. Auch die Analysen im Kontext der Bundestagswahl 2013 haben die Korrektur als robust gezeigt, wurden ihrerseits jedoch ebenfalls unter den Impressionen der schwarz-gelben Koalition durchgeführt. Dies gilt ebenso für die Nachwahlbefragung des KQ der Welle 7 des Wahlkampfpanels der Komponente 3. Die Korrektur funktionierte ebenso gut wie auf der Basis der Rekrutierungswelle 1 des Wahlkampfpanels (siehe Abbildung 10.14). Während des Erhebungszeitraums vom 24.09.2013 bis zum 03.10.2013 war jedoch weiterhin die schwarz-gelbe Koalition im Amt. Da die Große Koalition erst am 17.12.2013 ihre Arbeit aufgenommen hat, ist der mögliche Effekt dieses Wechsels auf die Qualität der Stichproben des Respondi-Panels nicht erfassbar gewesen. Hierdurch wäre feststellbar, ob dies auch zu einer Verschiebung des Bezugsobjekts der Bundesregierung geführt hat. Als eine Folge sollte sich die Auswirkung der Korrektur auf die Sympathie der CDU weiterhin in dem Ausmaß zeigen, wie diese aufgefunden wurde. Der Unterschied der Bewertung der Leistung der Bundesregierung insgesamt sollte sich jedoch abschwächen, da sich das Bezugsobjekt geändert hat: Nicht mehr die FDP stellt von da an den Juniorpartner, sondern die SPD.

Als ein Fazit lässt sich jedoch, trotz dieser Einschränkung, festhalten: Unterschiede der Komponente 8 bestehen in Relation zu den Verteilungen der zufallsbasierten telefonischen Befragungen der Meinungsinstitute der Bundesrepublik an Hand der Zweitstimmenpräferenz. Dies betrifft die Stichproben auf der Basis des onlinerekrutierem Access-Panel von Respondi stärker, die Wahlbeteiligungsabsicht wird hingegen durch die Stichproben des offlinerekrutierten Access-Panels von LINK stärker überschätzt. Diese erreicht mit 90 %

ein Niveau, welches vergleichbar zu denjenigen in telefonischen Befragungen ist. Die Angleichung an persönliche Referenzumfragen durch die Bestimmung von Bedeutungsgewichten auf der Basis verschiedener Korrekturmodellen kann diese Unterschiede verringern und die berücksichtigten Informationen sind für einen Erfolg entscheidend.

Inhaltlich müssen Unterschiede zwischen den zu verbessernden Umfragen und der Referenzumfrage vorhanden sein, um überhaupt eine Angleichung technisch möglich zu machen. Ebenfalls müssen diese auch mit den inhaltlichen Zielvariablen korrelieren, um eine verbesserte Abbildung dieser durch eine Gewichtung zu erlauben. Für Wahlumfragen sind hierfür kontextspezifische Variablen wie das Interesse an Politik, die empfundene Wahlnorm und die Zufriedenheit mit dem politischen System und seinen Akteuren entscheidend. Die Wahl der technischen Art der Angleichung durch die Nutzung von PS kann eine weitere Verbesserung erbringen. Die Unterschiede sind jedoch geringer als diejenigen der zu Grunde liegenden Korrekturmodelle, ihre Wahl stellt also einen weiteren *zusätzlichen* Bonus zur wichtigen Spezifikation eines guten Korrekturmodells dar. Wird dies berücksichtigt, ist die Verbesserung der Zweitstimmenanteile und weiterer Zielvariablen onlinerekrutierter Erhebungen im Rahmen eines PSM möglich. Für telefonisch offlinerekrutierte onlinebasierte Wahlumfragen, welche vielmehr mit einem Beteiligungsbias konfroniert sind, lässt sich hingegen über eine Angleichung die zu hohen Beteiligungsraten reduzieren.

11. Diskussion und Ausblick

Ziel der vorliegenden Arbeit war die Analyse der Qualität derjenigen Wahlumfragen, für deren Stichproben auf die konstruierte Grundgesamtheit eines Access-Panels zurückgegriffen wird. Hierzu wurde in Kapitel 2 der „Total Survey Error“ (TSE) als ein theoretischer Rahmen präsentiert, durch welchen konzeptionell die verschiedenen Fehlerquellen in Umfragen erfassbar werden. Auf der Grundlage der Unterteilung in einen Beobachtungs- und einen Nicht-Beobachtungs-Fehler wurde der Fokus auf zweiteren gelegt: Nonresponse- und Noncoverage-Probleme haben in Umfragen das Potential, einen systematischen Fehler hervorzurufen. Das Kapitel 3 hat den gesamten Fehler einer Umfrage in Form eines statistischen Bias definitorisch erfasst. Ebenso wurden Möglichkeiten zur entsprechenden Messung dieses gesamten Fehlers einer Umfrage aufgezeigt. In Kapitel 4 wurde die Erstellung von Bedeutungsgewichten auf der Basis eines „Propensity Score Matching“ (PSM) als Lösung präsentiert, um möglichen Verzerrungen in Umfragen zu entgegnen. Für den spezifischen Kontext von Wahlumfragen wurden dazu als Korrekturfaktoren die Individualmerkmale der Soziodemografie, das Zustimmungsverhalten zu Kompetenz- und Einstellungsfragen sowie die Persönlichkeit einer Person als relevant herausgestellt. Sie erweisen sich als bedeutend für das Teilnahme- und Antwortverhalten in Wahlumfragen.

Wie lassen sich nun die einleitend gestellten Fragen beantworten? Welches Qualitätsniveau der Erfassung des Wahl- und Beteiligungsverhaltens der wahlberechtigten Bevölkerung lässt sich für nicht-zufallsbasierte Onlineumfragen auf der Basis von Stichproben aus Access-Panels bescheinigen? Wie verhält sich dieser, immer noch relativ neue, Typus der Datenerhebung in Relation zu den zufallsbasierten Klassikern der telefonischen und persönlichen Befragung? Lässt sich zudem die Qualität von Onlineumfragen, welche auf Erhebungen aus einem Access-Panel basieren, über die Angleichung an einen geeigneten Referenzdatensatz auf der Basis von Bedeutungsgewichten durch ein PSM steigern? Ist hierfür relevant, wie die Art der Rekrutierung in dieses hinein zu Stande kommt? Das nachfolgende Fazit der Arbeit widmet sich diesen Fragen und ihrer Beantwortung.

11.1. Resumee der Arbeit

Eine Korrektur ist nur dann erfolgreich, wenn eine qualitativ hochwertige Referenzumfrage bereitsteht. Diese muss sich im Vergleich zur angleichenden Umfrage als weniger anfällig für die Fehlerquellen des TSE erweisen. Um die Entscheidung für eine solche Referenzumfrage treffen zu können, wurden in einem ersten Schritt die zufallsbasierten Erhebungen der persönlichen Befragungen des Allbus und des ESS zu denjenigen der telefonischen Erhebungen von Forsa und des Politbarometers verglichen. Haben sich für diese zunehmende Probleme der Abdeckung der (wahlberechtigten) Bevölkerung und ein starkes Potential für einen Nonresponse-Bias ergeben, zeigte sich dieser Trend bei den persönlichen Befragungen in einem geringerem Ausmaß. Hierzu wurde belegt, dass persönliche Befragungen seit 1990 ein Niveau in Relation zur amtlichen Statistik aufweisen, welches telefonbasierte Wahlumfragen noch nie erreicht haben. Zudem beginnt sich der komparative Nachteil telefonischer Befragungen zu verstärken. Das erfasste Bildungsniveau und das Durchschnittsalter hat sich kontinuierlich im Vergleich zu den Verteilungen der amtlichen Statistik erhöht. Eine Mischung aus einem Nonresponse- und einem Noncoverage-Fehler, letzteres bedingt durch die zunehmende Mobilfunknutzung der jüngeren Generation, macht sich daher bemerkbar. Die zufallsbasierte persönliche Befragung und die mit ihr assoziierten Auswahldesigns stellen somit einen klaren „Goldstandard der Soziodemografie“ dar.

Aus diesen Gründen haben sich telefonische Befragungen auch als systematisch abweichend von diesem „Goldstandard“ der persönlichen Befragung gezeigt. Hierzu wurden alle verfügbaren Forsa- und Politbarometer-Erhebungen seit 1994 in Relation zum Allbus und dem ESS gesetzt. Sowohl für die 894 betrachteten Stichproben von Forsa, als auch für die 278 von Politbarometer, zeigte sich ein vergleichbares Bild: Der Personenkreis telefonischer Befragungen, als eigentlich angedachte Zufallsrealisation der (wahlberechtigten) Bevölkerung, wird im Vergleich zu den Stichproben des Allbus immer älter und höher gebildet, auch wenn sich die Tendenz der Entwicklungen in unterschiedlichen zeitlichen Perioden seit 1994 unterscheidet.

Seit 2010 hat zudem eine Intensivierung der Entwicklung eingesetzt. Die Wahrscheinlichkeit jüngerer Personen Teil einer telefonischen Befragung auf der Basis von Festnetzstichproben zu sein, sinkt von Jahr zu Jahr. Als maßgeblicher Unterschied zwischen den persönlichen und telefonischen Befragungen hat sich zudem *durchgängig* das Interesse an Politik herausgestellt. Seit 1994 findet sich in den 200 dafür genutzten Erhebungen des Politbarometers so gut wie

keine Stichprobe, welche *nicht* gegenüber dem Allbus ein erhöhtes politisches Interesse aufzeigt. Diese Selektivität der Erfassung in telefonischen Befragungen induziert einen Nonresponse- und Noncoverage-*Bias*, wenn sie auch ursächlich für das inhaltliche Antwortverhalten ist.

Die gezeigten Auswirkungen haben sich als ambivalent herausgestellt: Einerseits hat sich das erfasste Wahl*verhalten* durch die Erhebungen des Politbarometer als robust gegenüber der Korrektur dieser Verzerrungen gezeigt und ist somit von *diesen nicht* betroffen. Andererseits lässt sich die, traditionell in telefonischen Wahlumfragen offenbarende, zu hohe Wahl*beteiligungsabsicht* auf diese Selektivität der telefonischen Befragung zurückführen. Die durchgehend zu hohen Anteile klassifizierter Wähler von bis zu 90 % in den Erhebungen des Politbarometers seit 1994 reduzieren sich um bis zu acht Prozentpunkte durch eine Korrektur im Rahmen des angewandten PSM, wofür auf einer Subklassifikation und den Nearest-, CEM- und den Full-Algorithmus genutzt wurde. Neben der Angabe der Absicht der Wahlbeteiligung als Folge einer sozialen Erwünschtheit durch den Interviewer, erklärt somit die selektive Teilnahme durch einen Nonresponse-Fehler, ebenfalls einen Teil der zu hohen Beteiligungsabsicht. Erst das Aufkommen der „Alternative für Deutschland“ (AfD) im Zuge der Bundestagswahl 2013 hat diese erfolgreiche Korrektur eingeschränkt. Im Rahmen der Arbeit wurde hierfür folgende Erklärung getroffen: Das Auftreten der AfD führt zu einer erhöhten Beteiligungsabsicht politisch weniger interessierter Personen. Daher ist die Differenzierungsmöglichkeit auf der Grundlage des Interesses an Politik nicht mehr gegeben, die Korrektur funktioniert nicht mehr, da die wahre Wahlbeteiligungsabsicht gestiegen ist.

Dies bestätigte und konkretisierte auch die Betrachtung der telefonisch erhobenen Komponente 2 der „German Longitudinal Election Study“ (GLES) im Kontext der Bundestagswahlen 2009 und 2013. Durch diese war eine bessere Evaluation möglich, da durch die zeitlich nahe stattfindenden Wahlen das wahre Wahlverhalten des Elektorats als Referenzwerte zur Verfügung steht. Zeigten sich im Vergleich zu den Erhebungen der persönlichen Befragung der Komponente 1 und des Allbus ähnliche grundlegende soziodemografische Verzerrungen wie durch die vorherige Betrachtung von Forsa und Politbarometers, spiegelte sich dies auch in identischen Auswirkungen wider. Die Erfassung des Wahlverhaltens des Elektorats im Vorfeld einer Bundestagswahl lässt sich durch beide zufallsbasierten Befragungsformenen auf einem qualitativ annähernd ähnlichem Niveau erfassen. Zudem weisen beide Erhebungsmodi vergleichbare Probleme auf: Im Kontext der Bundestagswahl 2013 konnten diese nicht das Potential

der AfD und die relative Bedeutung der Kategorie „Sonstige“ erkennen. Neben einem Nonresponse-Problem ist hierfür ein Messfehler durch die interviewerad-ministrierte Durchführung wahrscheinlich.

Ebenso zeigte sich auf der Grundlage der genutzten Wochenstichproben des „Rolling-Cross-Section“ (RCS)-Designs der Komponente 2 ein durchgängig zu hohes Beteiligungsniveau, welches sich auf einen systematischen Fehler durch die Selektivität der Teilnahme an der Befragung und die ungleiche Abdeckung der anvisierten Grundgesamtheit zurückführen lässt. Dies galt ebenso für das angegebene Interesse am Wahlkampf der beiden Bundestagswahlen. Die Beobachtung der Intensivierung des Wahlkampfinteresses mit zunehmender Nähe zur kommenden Wahl konnte, über eine ebenfalls durchgeführte Korrektur im Rahmen eines PSM, teilweise auf eine Selektivität der Teilnahme zurückgeführt werden. Nicht das Interesse am Wahlkampf steigt, sondern die politisch interessierten Personen, welche sich in der Tendenz eher für Wahlkämpfe interessieren, nehmen zunehmend an Wahlumfragen teil, je näher die kommende Bundestagswahl ist.

Zusammenfassend sind Telefonumfragen daher nicht per se verzerrt durch einen Nonresponse- und Noncoverage-Fehler, die zufallsbasierte Auswahlgrundlage scheint vielmehr ein weiterhin (noch) akzeptables Qualitätsniveau für bestimmte Zielvariablen einer Wahlumfrage zu garantieren. Telefonische Befragungen sind jedoch mit einem zu hohen Bildungsniveau, zu hohen politischen Interesse und einem zu hohen Durchschnittsalter konfrontiert. Korrelieren diese Variablen mit einer inhaltlich interessierenden Variable, wie beispielsweise der Beteiligungsabsicht, kommt es zu einem systematischen Nonresponse- und Noncoverage-Fehler. Daher hat sich alleine die persönliche Befragung als Möglichkeit der Referenz zur Angleichung onlinebasierter Befragungen erwiesen.

Onlineumfragen hingegen bleiben, trotz ihrer ökonomischen Vorteile, ein Erhebungsmodus mit vielschichtigen Problemen. Ein hohe Gefahr eines Noncoverage-Bias qua Design geht einher mit einem Nonresponse-Fehler, welcher sich durch die ungleiche Nutzungsintensität eines vorhandenen Internetzugangs und des damit verbundenen Einflusses auf die Teilnahmewahrscheinlichkeit an Onlineumfragen ergibt. Die Datenqualität dieser Erhebungsform ist daher a priori beeinträchtigt. Die Konstruktion von Access-Panels als Lösung kann funktionieren, sofern die Rekrutierung zufallsbasiert und damit offline, stattfindet. Zudem muss ein Internetzugang gestellt werden, sofern dieser nicht verfügbar ist. Ein Noncoverage-Fehler lässt sich hierdurch vermeiden, wenn die

Wahrscheinlichkeit der Annahme dieses Angebots nicht in einer Abhängigkeit zur Nutzungskompetenz steht.

In der Praxis treffen diese Charakteristika auf die gängigen Access-Panels nicht zu. Beide im Rahmen der Arbeit betrachteten Varianten stellen keine technische Infrastruktur für Personen ohne einen Internetzugang bereit. Als Resultat ergeben sich konstruierte Pools als Grundgesamtheit für daraus zu ziehende Zufallsstichproben, welche nicht die allgemeine oder wahlberechtigte Bevölkerung in einem verkleinerten Abbild widerspiegeln. Die Art der Rekrutierung in das Access-Panel hat zudem einen entscheidenden Einfluss auf die Qualität der korrekten Abbildung der Bevölkerung. Das betrachtete LINK Internet Panel weist eine aktiv offlinebasierte Form der Rekrutierung auf, während das Respondi-Panel auf rein onlinebasierten Kanälen Personen aktiv zur Registrierung erfasst. Hierdurch war die Arbeit in der Lage zu zeigen, das sich Stichproben aus diesen beiden Varianten eines Access-Panel zu persönlichen Referenzumfragen unterschiedlich verhalten, um über diese gezeigten Unterschiede Korrekturen in Form von Bedeutungsgewichten abzuleiten.

Ist dies möglich, lassen sich die Vorteile dieses Befragungsformats nutzen. Im vorliegenden Fall hat sich, bedingt durch die Selbstadministration der Erhebung, die bessere Möglichkeit der Erfassung des relativen Anteils der Zweitstimmen der AfD auf Basis *aller* genutzten Onlineumfragen im Kontext der Bundestagswahl 2013 gezeigt. Dieser Vorteil wurde jedoch von einem hohen Preis begleitet: Die restlichen angetretenen Parteien der Wahl waren teilweise stark verzerrt abgebildet. Insbesondere die zu geringen Anteile der präferierten Zweitstimmenanteile für die Union ließen eine adäquate Abbildung des Wahlverhaltens im direkten Vor- und Nachgang der Bundestagswahl nicht zu. Auf der Basis des onlinerekrutierten Access-Panels von Respondi haben sich diese Verzerrungen als ausgeprägter gezeigt, als dies auf der Basis der offlinerekrutierten Variante von LINK der Fall war. Die Arbeit war in der Lage, einen Teil der dafür verantwortlichen Gründe theoretisch hergeleitet zu benennen und aufzuzeigen, diese zu modellieren und somit abschließend darauf aufbauend zu korrigieren.

Die Ursache der schwierigen Erfassung durch das Respondi-Panel konnte auf einen bestimmten Typus von Person zurückgeführt werden, welcher durch die Rekrutierungsform systematisch bevorteilt wird: Jung, politisch interessiert, offen für neue Erfahrungen und mit einem expressiven Charakter. Dies kombiniert dieser Idealtypus durch eine stark ausgeprägte Unzufriedenheit mit dem Status der Demokratie in der Bundesrepublik und einer geringen Einschätzung der Responsivität der politischen Akteure. Ebenso lässt sich eine neurotische

Tendenz ausmachen. Durch die erfolgreich angewandten Korrekturen im Rahmen eines PSM war es möglich, die zu hohe Repräsentanz dieser Gruppe in den Erhebungen auf der Basis des Respondi-Panel zu reduzieren.

Als Folge dieser angewandten Korrektur hat sich eine klare Verbesserung der Abbildung des Wahlverhaltens der wahlberechtigten Bevölkerung im Kontext der Bundestagswahl 2013 gezeigt, welche mit dem Niveau der Befragungen des offlinerekrutierten LINK-Panel und der Qualität der Erfassung durch die persönlichen und telefonischen Befragungen der GLES vergleichbar ist. Ebenso hat sich die Wahlbeteiligungsabsicht auf ein Niveau abgesenkt, welches mit der Erfassung durch die persönliche Befragung der Komponente 1 konkurrieren kann.

Auch der Wirkungskanal der Korrektur auf weitere interessierende Zielvariablen einer Wahlumfrage wurde aufgezeigt: Das Niveau der empfundenen Sympathie für die CDU und deren Bewertung als maßgebliche Regierungspartei sowie die Bewertung der Leistung der Bundesregierung insgesamt, haben sich durch die Anwendung der Bedeutungsgewichte denjenigen der zufallsbasierten Befragung der Komponente 1 angeglichen. Deren arithmetische Mittel äußerten eine zu pessimistische Grundhaltung, da die Meinungen des skizzierten überpräsentierten Typus zu stark in die Erhebungen auf der Basis des Respondi-Panel vertreten sind. Veränderungen von über einer halben Kategorie der elfstufigen Bewertungsskalen waren daher möglich. Das offlinerekrutierte LINK-Panel zeigte hingegen bereits ohne eine Korrektur ein Niveau auf, welches vergleichbar war.

Dass es sich bei diesem aufgezeigtem Wirkungskanal der Korrektur nicht nur um einen einmaligen Effekt im Kontext der Bundestagswahl 2013 handelt, wurde anschließend durch eine Ausweitung des Zeitraums aufgezeigt. Hierzu wurden 30 kontinuierliche Erhebungen der Komponente 8 der GLES von 2009 bis 2016 genutzt. Die systematisch zu niedrigen Bewertungen der Stichproben auf Basis des Respondi-Panels hinsichtlich der Sympathie für die CDU, der Bewertung ihrer Regierungsleistung und derjenigen der Bundesregierung insgesamt, konnten auf einen Effekt des genutzten Access-Panel zurückgeführt werden. Der Wechsel des Betreiber von Respondi hin zu LINK erzeugte einen abrupten Anstieg der Bewertungen dieser Bezugsobjekte durch die erhobenen Stichproben. Gleichzeitig musste „Die Linke“ ein Absinken der Bewertung ihrer Oppositionsarbeit und ihrer Sympathie hinnehmen, da ihre Anhängerschaft innerhalb der Erhebungen der onlinerekrutierten Variante überpräsentiert war. Dass die Art der Rekrutierung diesen Effekt verursacht, konnte über einen externen

Vergleich des Verlaufs der Bewertungen identischer Fragen auf der Basis des Politbarometers für den identischen Zeitraum aufgezeigt werden.

Die Gründe, welche ursächlich für diese Unterschiede der Bewertungen sind, konnten auch in einem erweiterten Rahmen auf die Unterschiede der Stichproben der onlinebasierten Befragungen, in Relation zu persönlichen Referenzumfragen, zurückgeführt werden. Gemeinsam war *allen* onlinebasierten Stichproben die Tendenz, vor allem die ältere Bevölkerung zu benachteiligen sowie denjenigen mit einem erhöhten Interesse an Politik eine höhere Chance der Partizipation zuzugestehen. Innerhalb der Erhebungen des LINK-Panel zeigte sich zudem eine erhöhte Zustimmung des Wählens als eine staatsbürgerliche Pflicht und eine (leichte) Erhöhung der Demokratieunzufriedenheit, welche sich mit erhöhten Zustimmungsraten zu dem Persönlichkeitsmerkmal des Neurotizismus kombiniert hat. Die Gewissenhaftigkeit war hingegen klar verringert. Der Unterschied dieser beiden Persönlichkeitsmerkmale hat sich, mit einer erhöhten Rate, auch innerhalb der Befragungen auf der Basis des Respondi-Panels gezeigt. Auch ein etwas erhöhter Grad an Offenheit war dort erkennbar. In Kontrast zu den Befragungen des LINK-Panel fanden sich jedoch durchgehend stark erhöhte Zustimmungsraten einer negativen Einschätzungen der Offenheit des politischen Systems durch eine zu geringe Responsivität der maßgeblichen Akteure, wohingegen die Wahlnorm keine Differenzierungsmöglichkeit zum Allbus 2014 und der Komponente 1 der GLES erlaubt hat.

Die darauf aufbauende Korrektur dieser Unterschiede durch ein PSM hat die Variablen des formulierten *„Einstellungs"-Modells* als die relevanten Ursachen herausstellen können, welche das Wahl*verhalten* systematisch beeinflussen. Dies konnte an Hand eines Vergleichs zu dem aggregierten Verlauf der Zweitstimmenanteile der kommerziellen Meinungsforschungsinstitute in der Bundesrepublik gezeigt werden. Erreichte die Berücksichtigung der Selektivität der Stichproben der Onlinebefragungen auf der Basis des Respondi-Panel durch die Anwendung der Bedeutungsgewichte dahingehend eine Reduktion, galt dies ebenfalls für die Berücksichtigung der Persönlichkeitsmerkmale innerhalb der Befragungen auf der Basis von LINK.

Die Variablen des *„Kompetenz/Norm"-Modells* sind hingegen relevant für die abweichend reportierte Wahl*beteiligungsabsicht*. Auf der Basis dieses Korrekturmodells sank der angegebene Anteil der Personen mit einer angegebenen Wahlbeteiligungsabsicht. Hierbei hat sich der Vergleich zur Komponente 1 als vorteilhafter gegenüber der Nutzung des Allbus als Referenz erwiesen. Dies begründet sich durch das dort erhöhte politische Interesse der Durchführung in

2014, welche (bisher) einmalig bisher aufgetreten ist. Die Wahl des korrekten Modells zur Bestimmung der „Propensity Scores“ (PS) als Basis des Matching-prozess hat sich hierbei zudem als relevanter herausgestellt, als die Wahl der technischen Ausführung in Form des gewählten Algorithmus. In der Tendenz lässt jedoch die Wahl einer Subklassifikation mit sechs Klassen oder des Full-Algorithmus in seinen durchgehend verwendeten Spezifikationen einen leichten Vorteil gegenüber den anderen getesten Varianten erkennen.

Letztlich war die Arbeit daher in der Lage zu zeigen, dass die Kombination des „Einstellungs“-Modells mit einer Balancierung in Bezug zur Komponente 1 der GLES zu einer durchweg homogenen aufsteigenden Bewertung der Leistung der Bundesregierung und der Regierungsarbeit der CDU führt, sofern die Stichproben auf der Basis eines onlinerekrutierten Access-Panel erhoben werden. Dies ist möglich gewesen, obwohl der Referenzdatensatz in 2013 erhoben wurde, die onlinebasierten Befragungen jedoch größtenteils in 2011 durchgeführt wurden. Die Qualität der Korrektur war vergleichbar zu dem Erfolg, welche auf der Basis des identischen Referenzdatensatz im Kontext der Bundestagswahl 2013 möglich war. Der größere zeitliche Abstand beeinträchtigte also nicht die Qualität der Korrektur, da die berücksichtigten Variablen sich durch eine hohe Stabilität auszeichnen. Die Effekte auf die Bewertung der Oppositionsarbeit für „Die Linke“ sind im Vergleich geringer. Die gemessene Sympathie ist jedoch ebenfalls im arithmetischen Mittel durchweg gesunken.

Dass es zu einer starken gemeinsamen Veränderung der Sympathie für die CDU, die Bewertung ihrer Regierungsarbeit und die Bewertung der Leistung der Bundesregierung insgesamt kommt, bestätigt den langfristig aufgezeigten Trend der verknüpften der Bewertung der Regierungsarbeit mit der Sympathie für die maßgebliche Regierungspartei in Kapitel 7.3, sie bilden ein Bezugsobjekt.

11.2. Quo vadis? Eine Prognose über die zukünftigen Entwicklungen

Wie lässt sich nun auf der Basis dieser Erkenntnisse eine *Prognose* über die kommenden Entwicklungen der Erhebungslandschaft der Wahlforschung zeichnen? Telefonische Befragungen sind das zentrale Erhebungsmittel von Wahlumfragen. Es wird sich mit jedem weiteren Jahr jedoch die Frage stellen, ob sie dies bleiben können. Die beschriebenen Entwicklungen der Zusammensetzungen der Stichproben lassen sich – ohne hierbei zu übertreiben – als bedrohlich hinsichtlich der weiteren Rechtfertigung der Nutzung bezeichnen. Mit jedem

Jahr nehmen die Verzerrungen weiter zu. Offensichtlich betrifft dies grundlegende soziodemografische Variablen, welche häufig im Fokus liegen. Dies kombiniert sich, nachgewiesen an Hand des Interesses an Politik, jedoch mit weiteren interessierenden Merkmalen. Noch ergibt sich die Möglichkeit, auf Basis der zufallsbasierten *Auswahl* auch ein, mehr oder weniger, zufallsbasiertes Teilnahmeverhalten zu erlangen. Es ergibt sich momentan *nicht* der Schluss einer markanten Beeinflussung des Wahlverhaltens und den damit einhergehenden weiteren Verzerrungen durch die selektive Teilnahme.

Was das onlinebasierte Erhebungsformat als ein Medium der Zukunft definiert, ist daher die *Prognose der Entwicklung telefonischener Befragungen*. Im Rahmen der Arbeit konnte skizziert werden, dass sich das Abdeckungsproblem dieser mit jedem Jahr seit 2010 in einer solchen Form intensiviert hat, dass die Zukunft hier keine Verbesserung verspricht. Das Gegenteil wird vielmehr der Fall sein: Personen unter 30 Jahren werden über ihre Nutzung von Smartphones so sozialisiert, dass dieses nicht als Mittel der telefonischen Kommunikation gesehen wird. Es ist ein Gerät, um eine mobile Datennutzung zu garantieren. Dieser Personenkreis wird, wahrscheinlich, nicht mit 35 Jahren beginnen dies fundamental zu ändern. Momentan mögen die Auswirkungen noch geringfügig sein, dass Wahlverhalten wird weiter adäquat durch die zufallsbasierte telefonische Befragung erfasst. In Zukunft wird sich dieses Format jedoch ernsteren Problemen stellen müssen, wenn ein immer größerer Teil der Bevölkerung über Festnetzstichproben nicht mehr erreichbar sein wird. Die aufgezeigte Stärke der Tendenz dieser Entwicklung belegt dies.

Die Sicht der Meinungsforschungsinstitute bewirkt hierbei den Eindruck, die Nutzung von „Dual Frame“-Ansätzen ist eine Option, welche beliebig gezogen werden kann. Werden jedoch Festnetzstichproben mit diesen kombiniert, tritt ein zusätzlicher Effekt des Erhebungsmodus auf. Die Befragung über das Festnetz und den Mobilfunk sind auf Grund der unterschiedlichen Rahmenbedingungen nicht ohne weitere Überlegungen vergleichbar. Ein Interview im öffentlichen Raum über den Mobilfunk erzeugt eine noch höhere Wahrscheinlichkeit sozial erwünschter Antworten oder anderer Umwelteinflüsse. Zudem sorgt der Versuch der Kontaktaufnahme über den Mobilfunk für einen noch stärkeren Eingriff in die Privatsphäre, als es bei einer Kontaktaufnahme über das Festnetz der Fall ist. Dieses dient gerade zu einer „fernmündlichen“ Kontaktierung.

Aus dem Grund stellt sich für die Zukunft der Wahlforschung nicht die Frage *ob* onlinebasierte Umfragen valide sind, sondern *wie* onlinebasierte Umfragen endlich valide werden könn(t)en. Die im Rahmen der Arbeit vorgestellten

Lösungen der Angleichung an externe Referenzdatensätze sind hierzu ein Stück des Puzzles. Es lässt sich eine Verbesserung der Abbildung der wahlberechtigten Bevölkerung erreichen, weshalb die standardisierten qualitativ hochwertigen persönlichen Umfrageprogramme der Sozialwissenschaften, wie der Allbus, der ESS oder die GLES, ein notwendiges Kriterium zur Sicherung der zukünftigen Möglichkeit der breiten Durchführung von Umfragen innerhalb der empirischen Sozialforschung sind. Nur diese können eine valide Referenz bilden, um die eigene vorliegende Datenlage qualitativ einzuordnen und eventuell korrigierend einzugreifen. Die aufgezeigten Korrekturen können aber auch nicht mehr als dieses Puzzle-Stück sein. Angleichungen über ein PSM sind immer kontextabhängig, das Auffinden eines geeigneten Korrekturmodells *und* die Verfügbarkeit der relevanten Informationen zeit- und kostenaufwendig. Es wird keine „Weltformel" existierten, welche die Stichproben onlinebasierter Befragungen technisch auf ein qualitativ hochwertiges Niveau in einem universellen Kontext bringt.

Zwei weitere Lösungen sind daher notwendig. Die erste funktioniert automatisch, die Zeit heilt einige Wunden: Im Zuge der sich ausweitenden *Nutzung* des Internet wird es auch zu einer steigenden Professionalisierung im Umgang mit dem Medium kommen, welches sich gerade durch die Entwicklung von Tablets intensiviert hat. Die heutige Generation der Bevölkerung zwischen 40 und 59 Jahren stellt in naher Zukunft die Alterskohorte ab 60 Jahren. Deren heutige Sozialisation mit dem Internet wird deren Akzeptanz und Nutzung von onlinebasierten Befragungen steigern. Ein Teil der heute aufgefunden Probleme hinsichtlich des „Coverage" werden sich demnach verringern.

Dass die onlinebasierte Befragung eine zunehmende Option darstellt, nicht zuletzt im wissenschaftlichen Berreich, trifft jedoch nicht zu weil sie besser geworden ist. Der Grund liegt vielmehr darin, dass sich die telefonische Befragung zunehmend verschlechtert – sie droht marginalisiert zu werden. Probleme bestehen bei Onlineumfragen jedoch weiter und müssen gelöst werden. Aus der Sicht des Autors kann dies nur die weitere Optimierung und Verbesserung von Access-Panels sein. Nur durch diese ist ein klar abgegrenzter Personenkreis vorhanden, welcher mit den Referenzverteilungen der amtlichen Statistik kontinuierlich verglichen werden kann. Zudem steht auch für diese Referenzumfragen, wie dem genutzten Allbus und dem ESS, zur Verfügung, welche den Bedarf der Optimierung durch das Aufzeigen von Unterschieden zu diesen aufzeigen können. Das Qualitätsmanagement muss also noch ausgefeilter werden.

Dann kann die Option bestehen, dass Access-Panels (irgendwann) wirklich ein adäquates Abbild der allgemeinen Bevölkerung bereitstellen. Sie sind dann

als Pool zur Rekrutierung von Stichproben verfügbar, über welche wiederum der Inferenzschluss auf das zu Grunde liegende Access-Panel getroffen werden kann. Hierdurch wären letztendlich Aussagen über eine Population möglich, welche derjenigen der *allgemeinen* wahlberechtigten Bevölkerung sehr nahe kommt.

Literaturverzeichnis

AAPOR (2008): Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, 5. Aufl., Lenexa, Kansas: The American Association for Public Opinion Research.

AAPOR (2011): Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, 7. Aufl., Lenexa, Kansas: The American Association for Public Opinion Research.

Abadie, Alberto (2002): Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models, in: Journal of the American Statistical Association, 97, S. 284–292.

Abendschön, Simone/Roßteutscher, Sigrid (2016): Wahlbeteiligung junger Erwachsener – Steigt die soziale und politische Ungleichheit?, in: Roßteutscher, Sigrid/Faas, Thorsten/Rosar, Ulrich (Hrsg.): Bürgerinnen und Bürger im Wandel der Zeit: 25 Jahre Wahl- und Einstellungsforschung in Deutschland, Wiesbaden: Springer VS, S. 67–91.

ADM (2012): ADM-Forschungsprojekt "Dual-Frame-Ansätze" 2011/2012, Forschungsbericht, Frankfurt: Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.

ADM (2014): Jahresbericht 2013, Frankfurt: Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute.

Andersen, Ronald/Frankel, Martin R./Kasper, Judith (1979): Total survey error: Applications to Improve Health Surveys, San Francisco (CA): Jossey-Bass.

Ansolabehere, Stephen/Schaffner, Brian F. (2014): Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison, in: Political Analysis, 22, S. 285–303.

Arzheimer, Kai (2009): Mehr Nutzen als Schaden? Wirkung von Gewichtungsverfahren, in: Schoen, Harald/Rattinger, Hans/Gabriel, Oscar W. (Hrsg.): Vom Interview zur Analyse. Methodische Aspekte der Einstellungs- und Wahlforschung, Baden-Baden: Nomos, S. 361–388.

Arzheimer, Kai (2016): Strukturgleichungsmodelle: Eine anwendungsorientierte Einführung, Wiesbaden: Springer VS.

Arzheimer, Kai/Evans, Jocelyn (2014): A New Multinomial Accuracy Measure for Polling Bias, in: Political Analysis, 22, S. 31–44.

Arzheimer, Kai/Evans, Jocelyn (2016): Estimating Polling Accuracy in Multi-Party Elections Using Surveybias, in: The Stata Journal, 16, S. 139–158.

Asendorpf, Jens B./Neyer, Franz J. (2012): Psychologie der Persönlichkeit, 5. Aufl., Berlin, Heidelberg: Springer.

Atrostic, B. K./Bates, Nancy A./Burt, Geraldine/Silberstein, Adriana (2001): Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights, in: Journal of Official Statistics, 17, S. 209–226.

Austin, Peter C. (2008a): A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003, in: Statistics in Medicine, 27, S. 2037–2049.

Austin, Peter C. (2008b): Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review, in: Circulation. Cardiovascular quality and outcomes, 1, S. 62–67.

Austin, Peter C. (2014): A comparison of 12 algorithms for matching on the propensity score, in: Statistics in Medicine, 33, S. 1057–1069.

Bacher, Johann (2009): Analyse komplexer Stichproben, in: Weichbold, Martin/Bacher, Johann/Wolf, Christof (Hrsg.): Umfrageforschung: Herausforderungen und Grenzen, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 253–274.

Bai, Haiyan (2015): Methodological Considerations in Implementing Propensity Score Matching, in: Pan, Wei/Bai, Haiyan (Hrsg.): Propensity Score Analysis: Fundamentals and Developments, New York und London: The Guilford Press, S. 74–88.

Baker, Reg/Blumberg, Stephen J./Brick, J. Michael/Couper, Michael P./Courtright, Melanie/Dennis, J. Michael/Dillman, Don A./Frankel, Martin R./Garland, Philip/Groves, Robert M./Kennedy, Courtney/Krosnick, Jon A./Lavrakas, Paul J./Lee, S./Link, M./Piekarski, L./Rao, K./Thomas, R. K./Zahs, D. (2010): Research Synthesis: AAPOR Report on Online Panels, in: Public Opinion Quarterly, 74, S. 711–781.

Baker, Reg/Brick, J. Michael/Bates, Nancy A./Battaglia, Mike/Couper, Michael P./Dever, Jill A./Gile, Krista J./Tourangeau, Roger (2013): Summary Report of the AAPOR Task Force on Non-probability Sampling, in: Journal of Survey Statistics and Methodology, 1, S. 90–143.

Bandilla, Wolfgang/Kaczmirek, Lars/Blohm, Michael/Neubarth, Wolfgang (2009): Coverage- und Nonresponse-Effekte bei Online-Bevölkerungsumfragen, in: Jackob, Nikolaus/Schoen, Harald/Zerback, Thomas (Hrsg.): Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung, Wiesbaden: Springer VS, S. 129–143.

Battaglia, Michael P./Hoaglin, David/Frankel, Martin R. (2009): Practical Considerations in Raking Survey Data, in: Survey Practice, 2.

Beierlein, Constanze/Kemper, Christoph J./Kovaleva, Anastassiya/Rammstedt, Beatrice (2012): Ein Messinstrument zur Erfassung politischer Kompetenz- und Einflussüberzeugungen: Political Efficacy Kurzskala (PEKS), GESIS-Working Papers 2012/18, Mannheim: GESIS.

Bergmann, Michael (2015): Panel Conditioning: Wirkungsmechanismen und Konsequenzen wiederholter Befragungen, Baden-Baden: Nomos.

Bethlehem, Jelke (2010): Selection Bias in Web Surveys, in: International Statistical Review, 78, S. 161–188.

Bethlehem, Jelke (2015): Web Surveys in Official Statistics, in: Engel, Uwe/ Jann, Ben/Lynn, Peter/Scherpenzeel, Annette/Sturgis, Patrick (Hrsg.): Improving Survey Methods: Lessons from recent research, New York/London: Routledge, S. 156–169.

Bethlehem, Jelke G. (1988): Reduction of Nonresponse Bias Through Regression Estimation, in: Journal of Official Statistics, 4, S. 251–260.

Bethlehem, Jelke G. (2002): Weighting nonresponse adjustments based on auxiliary information, in: Groves, Robert M./Dillman, Don A./Eltlinge, John L./ Little, Roderick J.A (Hrsg.): Survey nonresponse, New York: Wiley, S. 265–287.

Bethlehem, Jelke G. (2009): Applied survey methods: A statistical perspective, Hoboken und New Jersey: Wiley.

Bethlehem, Jelke G./Biffignandi, Silvia (2012): Wiley Handbook of Web Surveys, Hoboken, NJ: Wiley.

Bethlehem, Jelke G./Cobben, Fannie/Schouten, Barry (2011): Handbook of Nonresponse in Household Surveys, Hoboken/New Jersey: Wiley.

Bethlehem, Jelke G./Schouten, Barry (2016): Nonresponse Error: Detection and Correction, in: Wolf, Christof/Joye, Dominique/Smith, Tom E. C./Fu, Yang-chih (Hrsg.): The SAGE Handbook of Survey Methodology, London: Sage Publications, S. 558–578.

Bethlehem, Jelke/Cobben, Fannie/Schouten, Barry (2009): Indicators for the representativeness of survey response, in: Survey Methodology, 35, S. 101–113.

Beullens, Koen/Loosveldt, Geert (2012): Should high response rates really be a primary objective?, in: Survey Practice, 5, S. 1–5.

Beullens, Koen/Loosveldt, Geert (2016): Interviewer Effects in the European Social Survey, in: Survey Research Methods, 10, S. 103–118.

Biemer, Paul P. (2010): Total Survey Error: Design, Implementation, and Evaluation, in: Public Opinion Quarterly, 74, S. 817–848.

Biemer, Paul P./Lyberg, Lars (2003): Introduction to survey quality, Hoboken/ New Jersey: Wiley.

Billiet, Jaak B./Davidov, Eldad (2008): Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design, in: Sociological Methods & Research, 36, S. 542–562.

Billiet, Jaak/Philippens, Michel/Fitzgerald, Rory/Stoop, Ineke (2007): Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents, in: Journal of Official Statistics, 23, S. 135–162.

Blasius, Jörg/Brandt, Maurice (2009): Repräsentativität in Online-Befragungen, in: Weichbold, Martin/Bacher, Johann/Wolf, Christof (Hrsg.): Umfrageforschung: Herausforderungen und Grenzen, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 157–177.

Blohm, Michael (2006): Datenqualität durch Stichprobenverfahren bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften – ALLBUS, in: Faulbaum, Frank/Wolf, Christof (Hrsg.): Stichprobenqualität in Bevölkerungsumfragen, Bonn: Informationszentrum Sozialwissenschaften, S. 37–54.

Blom, Annelies G./Bosnjak, Michael/Cornilleau, Anne/Cousteaux, Anne-Sophie/Das, Marcel/Douhou, Salima/Krieger, Ulrich (2016): A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe, in: Social Science Computer Review, 34, S. 8–25.

Blom, Annelies G./Gathmann, Christina/Krieger, Ulrich (2015): Setting Up an Online Panel Representative of the General Population: The German Internet Panel, in: Field Methods, 27, S. 391–408.

Blom, Annelies G./Herzing, Jessica M. E./Cornesse, Carina/Sakshaug, Joseph W./Krieger, Ulrich/Bossert, Dayana (2016): Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel, in: Social Science Computer Review, S. 1–23.

Blumberg, Stephen J./Luke, Julian V. (2014): Wireless substitution: Early release of estimates from the National Health Interview Survey, National Health Interview Survey Early Release Program.

Blumenberg, Manuela S./Gummer, Tobias (2013): Gewichtung in der German Longitudinal Election Study 2009, GESIS - Technical Reports 19, Mannheim: GESIS.

Börsch-Supan, Axel/Winter, Joachim/Elsner, Detlev/Faßbender, Heino/Kiefer, Rainer/McFadden, DAniel (2004): How to make internet surveys representative: A case study of a two-step weighting procedure, MEA Discussion Papers 067-04, München: Munich Center for the Economics of Aging.

Bosnjak, Michael/Haas, Iris/Galesic, Mirta/Kaczmirek, Lars/Bandilla, Wolfgang/Couper, Michael P. (2013): Sample Composition Discrepancies in Different Stages of a Probability-based Online Panel, in: Field Methods, 25, S. 339–360.

Bowley, Arthur L. (1906): Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science, York, 1906, in: Journal of the Royal Statistical Society, 69, S. 540–558.

Brady, Henry E./Verba, Sidney/Schlozman, Kay Lehman (1995): Beyond Ses: A Resource Model of Political Participation, in: The American Political Science Review, 89, S. 271–294.

Brandtzæg, Petter Bae/Heim, Jan/Karahasanović, Amela (2011): Understanding the new digital divide—A typology of Internet users in Europe, in: International Journal of Human-Computer Studies, 69, S. 123–138.

Brehm, John (1994): Stubbing our Toes for a Foot in the Door? Prior Contact, Incentives and Survey Response, in: International Journal of Public Opinion Research, 6, S. 45–63.

Brick, J. Michael (2011): The Future of Survey Sampling, in: Public Opinion Quarterly, 75, S. 872–888.

Brick, J. Michael (2013): Unit Nonresponse and Weighting Adjustments: A Critical Review, in: Journal of Official Statistics, 29, S. 329–353.

Brick, J. Michael/Kalton, Graham (1996): Handling missing data in survey research, in: Statistical Methods in Medical Research, 5, S. 215–238.

Brick, J. Michael/Williams, Douglas (2012): Explaining Rising Nonresponse Rates in Cross-Sectional Surveys, in: The ANNALS of the American Academy of Political and Social Science, 645, S. 36–59.

Brüggen, Elisabeth/Dholakia, Utpal M. (2010): Determinants of Participation and Response Effort in Web Panel Surveys, in: Journal of Interactive Marketing, 24, S. 239–250.

Brunton-Smith, Ian/Sturgis, Patrick/Williams, Joel (2012): Is Success in Obtaining Contact and Cooperation Correlated With the Magnitude of Interviewer Variance?, in: Public Opinion Quarterly, 76, S. 265–286.

Buskirk, Trent/Andres, Charles (2013): Smart Surveys for Smart Phones: Exploring Various Approaches for Conducting Online Mobile Surveys via Smartphones, in: Survey Practice, 5, S. 1–11.

Busse, Britta/Fuchs, Marek (2012): The components of landline telephone survey coverage bias. The relative importance of no-phone and mobile-only populations, in: Quality & Quantity, 46, S. 1209–1225.

Busse, Britta/Laub, Simon/Fuchs, Marek (2015): Exit Questions: Bestimmung und Analyse des Nonresponse Bias in einer Panelbefragung im Mobilfunknetz, in: Schupp, Jürgen/Wolf, Christof (Hrsg.): Nonresponse Bias, Wiesbaden: Springer VS, S. 327–358.

Caliendo, Marco/Kopeinig, Sabine (2008): Some Practical Guidance for the Implementation of Propensity Score Matching, in: Journal of Economic Surveys, 22, S. 31–72.

Calinescu, Melania/Bhulai, Sandjai/Schouten, Barry (2013): Optimal resource allocation in survey designs, in: European Journal of Operational Research, 226, S. 115–121.

Calinescu, Melania/Schouten, Barry (2016): Adaptive Survey Designs for Nonresponse and Measurement Error in Multi-Purpose Surveys, in: Survey Research Methods, 10, S. 35–47.

Callegaro, Mario/Baker, Reg/Bethlehem, Jelke G./Göritz, Anja S./Krosnick, John A./Lavrakas, Paul J. (2014): Online panel research: History, concepts, applications and a look at the future, in: Callegaro, Mario/Baker, Reg/Bethlehem, Jelke/Göritz, Anja S./Krosnick, John A./Lavrakas, Paul J. (Hrsg.): Online panel research: A data quality perspective, Chichester, West Sussex: Wiley, S. 1–22.

Callegaro, Mario/Villar, Ana/Yeager, David/Krosnick, Jon A. (2014): A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples1, in: Callegaro, Mario/Baker, Reg/Bethlehem, Jelke/Göritz, Anja S./Krosnick, John A./Lavrakas, Paul J. (Hrsg.): Online panel research: A data quality perspective, Chichester, West Sussex: Wiley, S. 23–53.

Campbell, Angus (1954): The voter decides, Row, Peterson.

Capara, Gian Vittorio/Barbaranelli, Claudio/Zimbardo, Philip G. (1999): Personality Profiles and Political Parties, in: Political Psychology, 20, S. 175–197.

Carney, Dana R./Jost, John T./Gosling, Samuel D./Potter, Jeff (2008): The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind, in: Political Psychology, 29, S. 807–840.

Carpini, Michael X. Delli/Keeter, Scott (1993): Measuring Political Knowledge: Putting First Things First, in: American Journal of Political Science, 37, S. 1179–1206.

Caspi, Avshalom/Roberts, Brent W./Shiner, Rebecca L. (2005): Personality development: stability and change, in: Annual Review of Psychology, 56: S. 453–484.

Chang, Linchiat/Krosnick, Jon A. (2009): National Surveys Via Rdd Telephone Interviewing Versus the Internet: Comparing Sample Representativeness and Response Quality, in: Public Opinion Quarterly, 73, S. 641–678.

Chang, Linchiat/Krosnick, Jon A. (2010): Comparing Oral Interviewing with Self-Administered Computerized QuestionnairesAn Experiment, in: Public Opinion Quarterly, 74, S. 154–167.

Chatfield, Chris (1988): Apples, oranges and mean square error, in: International Journal of Forecasting, 4, S. 515–518.

Church, Allan H. (1993): Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis, in: Public Opinion Quarterly, 57, S. 62.

Clark, M.H. (2015): Propensity Score Ajustment Methods, in: Pan, Wei/Bai, Haiyan (Hrsg.): Propensity Score Analysis: Fundamentals and Developments, New York und London: The Guilford Press, S. 115–140.

Clarke, Harold D./Sanders, David/Stewart, Marianne C./Whiteley, Paul (2008): Internet Surveys and National Election Studies: A Symposium, in: Journal of Elections, Public Opinion and Parties, 18, S. 327–330.

Cobben, Fannie (2009): Nonresponse in sample surveys: Methods for analysis and adjustment, The Hague: Statistics Netherlands.

Cochran, W. G./Chambers, S. Paul (1965): The Planning of Observational Studies of Human Populations, in: Journal of the Royal Statistical Society. Series A (General), 128, S. 234.

Cochran, William G. (1977): Sampling Techniques, 3. Aufl., New York: Wiley.

Cochran, William G./Rubin, Donald B. (1973): Controlling Bias in Observational Studies: A Review, in: Sankhya: The Indian Journal of Statistics, Series A (1961-2002), 35, S. 417–446.

Couper, Mick P. (2000): Web Surveys, in: Public Opinion Quarterly, 64, S. 464–494.

Couper, Mick P. (2011): The Future of Modes of Data Collection, in: Public Opinion Quarterly, 75, S. 889–908.

Couper, Mick P. (2013): Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys, in: Survey Research Methods, 7, S. 145–156.

Couper, Mick P./Kapteyn, Arie/Schonlau, Matthias/Winter, Joachim (2007): Noncoverage and nonresponse in an Internet survey, in: Social Science Research, 36, S. 131–148.

Couper, Mick P./Miller, P. V. (2009): Web Survey Methods: Introduction, in: Public Opinion Quarterly, 72, S. 831–835.

Crawford, Scott D./Couper, Michael P./Lamias, Mark J. (2001): Web Surveys: Perceptions of Burden, in: Social Science Computer Review, 19, S. 146–162.

Crump, Richard K./Hotz, V. Joseph/Imbens, Guido W./Mitnik, Oscar A. (2009): Dealing with limited overlap in estimation of average treatment effects, in: Biometrika, 96, S. 187–199.

Curtin, R./Presser, S./Singer, Eleanor (2005): Changes in Telephone Survey Nonresponse over the Past Quarter Century, in: Public Opinion Quarterly, 69, S. 87–98.

Curtin, Richard/Presser, Stanley/Singer, Eleanor (2000): The Effects of Response Rate Changes on the Index of Consumer Sentiment, in: Public Opinion Quarterly, 64, S. 413–428.

Czado, Claudia/Schmidt, Thorsten (2011): Mathematische Statistik, Berlin: Springer.

Dehejia, Rajeev H./Wahba, Sadek (2002): Propensity Score-Matching Methods for Nonexperimental Causal Studies, in: Review of Economics and Statistics, 84, S. 151–161.

Deming, W. Edwards (1944): On Errors in Surveys, in: American Sociological Review, 9, S. 359–369.

Deming, W. Edwards (1950): Some Theory of Sampling, New York: Dover Publications.

Deming, W. Edwards/Stephan, Frederick F. (1940): On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, in: The Annals of Mathematical Statistics, 11, S. 427–444.

Diamond, Alexis/Sekhon, Jasjeet S. (2013): Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Ba-

lance in Observational Studies, in: Review of Economics and Statistics, 95, S. 932–945.

Diekmann, Andreas (2007): Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen, 18. Aufl., Reinbek bei Hamburg: Rowohlt-Taschenbuch-Verlag.

Digman, John M. (1990): Personality Structure: Emergence of the Five-Factor Model, in: Annual Review of Psychology, 41, S. 417–440.

Dillman, Don A. (1978): Mail and telephone surveys: The total design method, New York: Wiley.

Dillman, Don A. (2002): Survey nonresponse in design, data collection, and analysis, in: Groves, Robert M./Dillman, Don A./Eltlinge, John L./Little, Roderick J.A (Hrsg.): Survey nonresponse, New York: Wiley, S. 8–12.

Dillman, Don A./Smyth, Jolene D./Christian, Leah Melani (2009): Internet, mail, and mixed-mode surveys: The tailored design method, 3. Auflage, Hoboken/ New Jersey: Wiley.

Dollinger, Stephen J./Leong, Frederick T. L. (1993): Volunteer Bias and the Five-Factor Model, in: The Journal of Psychology, 127, S. 29–36.

Duffy, Bobby/Smith, Kate/Terhanian, George/Bremer, John (2005): Comparing data from online and face-to-face surveys, in: International Journal of Market Research, 47, S. 615–639.

DuGoff, Eva H./Schuler, Megan/Stuart, Elizabeth A. (2014): Generalizing Observational Study Results: Applying Propensity Score Methods to Complex Surveys, in: Health Services Research, 49, S. 284–303.

Durand, Claire (2008): The Polls of the 2007 French Presidential Campaign: Were Lessons Learned from the 2002 Catastrophe?, in: International Journal of Public Opinion Research, 20, S. 275–298.

Eckman, Stephanie/Kreuter, Frauke (2013): Undercoverage Rates and Undercoverage Bias in Traditional Housing Unit Listing, in: Sociological Methods & Research, 42, S. 264–293.

Ehlen, John/Ehlen, Patrick (2007): Cellular-Only Substitution in the United States as Lifestyle Adoption: Implications for Telephone Survey Coverage, in: Public Opinion Quarterly, 71, S. 717–733.

Elliot, Marc N./Haviland, Amelia (2007): Use of a webbased convenience sample to supplement a probability sample, in: Survey Methodology, 33, S. 211–215.

Elliot, Michael R./Little, Roderick J.A (2000): Model-Based Alternatives to Trimming Survey Weights, in: Journal of Official Statistics, 16, S. 191–209.

Eltinge, John L. (2002): Diagnostics for the Practical Effects of Nonresponse Adjustment Methods, in: Groves, Robert M./Dillman, Don A./Eltlinge, John L./Little, Roderick J.A (Hrsg.): Survey nonresponse, New York: Wiley, S. 431–443.

Erens, Bob/Burkill, Sarah/Couper, P. Mick/Conrad, Frederick/Clifton, Soazig/Tanton, Clare/Phelps, Andrew/Datta, Jessica/Mercer, H. Catherine/Sonnenberg, Pam/Prah, Philip/Mitchell, R. Kirstin/Wellings, Kaye/Johnson, M. Anne/Copas, J. Andrew (2014): Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison With a Probability Sample Interview Survey, in: Journal of Medical Internet Research, 16, S. 1–14.

European Commission (2014): E-Communications and Telecom Single Market Household Survey: Bericht, Special Eurobarometer 414, Brüssel.

European Social Survey (2014): ESS Round 7 Project Instructions (PAPI), London: ESS ERIC Headquarters.

Evans, Joel R./Mathur, Anil (2005): The value of online surveys, in: Internet Research, 15, S. 195–219.

Faas, Thorsten (2004): Online or Not Online? A Comparison of Offline and Online Surveys Conducted in the Context of the 2002 German Federal Election, in: Bulletin de Méthodologie Sociologique, 82, S. 42–57.

Faas, Thorsten (2014): Zur Wahrnehmung und Wirkung von Meinungsumfragen, in: Aus Politik und Zeitgeschichte, B43-45: S. 3–10.

Faas, Thorsten (2017): Demoskopische Befunde - ihre Hintergründe, ihre Verarbeitung, ihre Folgen: einige (ein)leitende Überlegungen, in: Faas, Thorsten/Molthagen, Dietmar/Mörschel, Tobias (Hrsg.): Demokratie und Demoskopie, Wiesbaden/s.l.: Springer VS, S. 7–24.

Faas, Thorsten/Huber, Sascha (2010): Experimente in der Politikwissenschaft: Vom Mauerblümchen zum Mainstream, in: Politische Vierteljahresschrift, 51, S. 721–749.

Faas, Thorsten/Huber, Sascha (2015): Haben die Demoskopen die FDP aus dem Bundestag vertrieben? Ergebnisse einer experimentellen Studie, in: Zeitschrift für Parlamentsfragen, 4: S. 746–759.

Faas, Thorsten/Schoen, Harald (2006): Putting a Questionnaire on the Web ist not Enough - A Comparison of Online and Offline Surveys Conducted in the Context of the German Federal Election 2002, in: Journal of Official Statistics, 22, S. 177–190.

Fan, Weimiao/Yan, Zheng (2010): Factors affecting response rates of the web survey: A systematic review, in: Computers in Human Behavior, 26, S. 132–139.

Faulbaum, Frank (2014): Total Survey Error, in: Baur, Nina/Blasius, Jörg (Hrsg.): Handbuch Methoden der empirischen Sozialforschung, Springer VS, S. 439–453.

Faulbaum, Frank/Prüfer, Peter/Rexroth, Margrit (2009): Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität, Wiesbaden: Springer VS.

Feather, Norman T. (1977): Value importance, conservatism, and age, in: European Journal of Social Psychology, 7, S. 241–245.

Fishbein, Martin (1963): An Investigation of the Relationships between Beliefs about an Object and the Attitude toward that Object, in: Human Relations, 16, S. 233–239.

Fowler, Floyd J./Mangione, Thomas W. (1990): Standardized Survey Interviewing: Minimizing Interviewer-Related Error, London/New Delhi: Sage Publications.

Frankel, Martin R. (2010): Sampling Theory, in: Marsden, Peter V./Wright, James D. (Hrsg.): Handbook of Survey Research, Bingley, UK: Emerald, S. 83–138.

Fricker, R. D./Schonlau, M. (2002): Advantages and Disadvantages of Internet Research Surveys: Evidence from the Literature, in: Field Methods, 14, S. 347–367.

Fricker, Scott/Tourangeau, Roger (2011): Examining the Relationship Between Nonresponse Propensity and Data Quality in Two National Household Surveys, in: Public Opinion Quarterly, 74, S. 934–955.

Fuchs, Marek (2010): Herausforderungen der Umfrageforschung, in: Faulbaum, Frank/Wolf, Christof (Hrsg.): Gesellschaftliche Entwicklungen im Spiegel der empirischen Sozialforschung, Wiesbaden: VS-Verl., S. 227–252.

Fuchs, Marek (2012): Der Einsatz von Mobiltelefonen in der Umfrageforschung Methoden zur Verbesserung der Datenqualität, in: Faulbaum, Frank/Stahl, Matthias/Wiegand, Erich (Hrsg.): Qualitätssicherung in der Umfrageforschung, Wiesbaden: Springer VS, S. 51–73.

Fuchs, Marek/Bossert, Dayana/Stukowski, Sabrina (2013): Response Rate and Nonresponse Bias - Impact of the Number of Contact Attempts on Data Quality in the European Social Survey, in: Bulletin of Sociological Methodology, 117, S. 26–45.

Gangl, Markus (2010): Causal Inference in Sociological Research, in: Annual Review of Sociology, 36, S. 21–47.

Gehring, Uwe W./Weins, Cornelia (2010): Grundkurs Statistik für Politologen und Soziologen, Wiesbaden: VS Verlag für Sozialwissenschaften.

Gelman, Andrew (2007): Struggles with Survey Weighting and Regression Modeling, in: Statistical Science, 22, S. 153–164.

Gerber, Alan S./Huber, Gregory A./Doherty, David/Dowling, Conor M. (2011): The Big Five Personality Traits in the Political Arena, in: Annual Review of Political Science, 14, S. 265–287.

Gerber, Alan S./Huber, Gregory A./Doherty, David/Dowling, Conor M. (2012): Personality and the Strength and Direction of Partisan Identification, in: Political Behavior, 34, S. 653–688.

Gerber, Alan S./Huber, Gregory A./Doherty, David/Dowling, Conor M./Ha, Shang E. (2010): Personality and Political Attitudes: Relationships across Issue Domains and Political Contexts, in: American Political Science Review, 104, S. 111–133.

Goldberg, Lewis R. (1990): An alternative "description of personality": The Big-Five factor structure, in: Journal of Personality and Social Psychology, 59, S. 1216–1229.

Goritz, A. S. (2008): The Long-Term Effect of Material Incentives on Participation in Online Panels, in: Field Methods, 20, S. 211–225.

Göritz, Anja S. (2004): The impact of material incentives on response quantity, response quality, sample composition, survey outcome, and cost in online access panels, in: International Journal of Market Research, 46, S. 327–345.

Göritz, Anja S. (2006): Incentives in Web Studies: Methodological Issues and a Review, in: International Journal of Internet Science, 1, S. 58–70.

Gosling, Samuel D./Rentfrow, Peter J./Swann, William B. (2003): A very brief measure of the Big-Five personality domains, in: Journal of Research in Personality, 37, S. 504–528.

Goyder, John (1987): The silent minority: Nonrespondents on sample surveys, Cambridge: Polity Press.

Graeske, Jennifer/Kunz, Tanja (2009): Stichprobenqualität der CELLA-Studie unter besonderer Berücksichtigung der Mobile-onlys, in: Häder, Michael/ Häder, Sabine (Hrsg.): Telefonbefragungen über das Mobilfunknetz: Konzept, Design und Umsetzung einer Strategie zur Datenerhebung, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 57–70.

Green, Jane/Prosser, Christopher (2016): Party system fragmentation and single-party government: The British general election of 2015, in: West European Politics, 39, S. 1299–1310.

Groves, R. M. (2006): Nonresponse Rates and Nonresponse Bias in Household Surveys, in: Public Opinion Quarterly, 70, S. 646–675.

Groves, R. M./Peytcheva, Emilia (2008): The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis, in: Public Opinion Quarterly, 72, S. 167–189.

Groves, R. M./Presser, S./Dipko, S. (2004): The Role of Topic Interest in Survey Participation Decisions, in: Public Opinion Quarterly, 68, S. 2–31.

Groves, Robert M. (1989): Survey Errors and Survey Costs, New York: Wiley.

Groves, Robert M. (2004): Survey Errors and Survey Costs, Hoboken, NJ: Wiley-Interscience.

Groves, Robert M. (2011): Three Eras of Survey Researchs, in: Public Opinion Quarterly, 75, S. 861–871.

Groves, Robert M./Cialdini, Robert B./Couper, Mick P. (1992): Understanding The Decision to Participate in a Survey, in: Public Opinion Quarterly, 56, S. 475–495.

Groves, Robert M./Couper, Mick (1998): Nonresponse in household interview surveys, New York: Wiley.

Survey nonresponse (2002), New York: Wiley.

Groves, Robert M./Fowler, Floyd J./Couper, Mick/Lepkowski, James M./Singer, Eleanor/Tourangeau, Roger (2009): Survey Methodology, 2. Aufl., Hoboken, N.J.: Wiley.

Groves, Robert M./Heeringa, Steven G. (2006): Responsive design for household surveys: tools for actively controlling survey errors and costs, in: Journal of the Royal Statistical Society: Series A (Statistics in Society), 169, S. 439–457.

Groves, Robert M./Lyberg, Lars (2010): Total Survey Error: Past, Present, and Future, in: Public Opinion Quarterly, 74, S. 849–879.

Groves, Robert M./Singer, Eleanor/Corning, Amy (2000): Leverage-Saliency Theory of Survey Participation, in: Public Opinion Quarterly, 64, S. 299–308.

Gruijter, Dato N. de/Kamp, Leo J. Th. van der (2008): Statistical Test Theory for the Behavioral Sciences, Boca Raton: Chapman & Hall/CRC.

Gu, Xing Sam/Rosenbaum, Paul R. (1993): Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms, in: Journal of Computational and Graphical Statistics, 2, S. 405–420.

Gummer, Tobias (2015): Multiple Panels in der empirischen Sozialforschung: Evaluation eines Forschungsdesigns mit Beispielen aus der Wahlsoziologie, Wiesbaden: Springer VS.

Guo, Shenyang/Fraser, Mark W. (2010): Propensity Score Analysis: Statistical Methods and Applications, Thousand Oaks/Calif: Sage Publications.

Guo, Shenyang/Fraser, Mark W. (2015): Propensity Score Analysis: Statistical Methods and Applications, 2. Aufl., Thousand Oaks/Calif: Sage Publications.

Häder, Michael (2015): Empirische Sozialforschung: Eine Einführung, 3. Aufl., Wiesbaden: Springer VS.

Häder, Sabine/Ganniner, Matthias/Gabler, Siegfried (2009): Die Stichprobenziehung für den European Social Survey - Prinzipien und Ergebnisse, in: Weichbold, Martin/Bacher, Johann/Wolf, Christof (Hrsg.): Umfrageforschung: Herausforderungen und Grenzen, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 181–193.

Hansen, Ben B. (2004): Full Matching in an Observational Study of Coaching for the SAT, in: Journal of the American Statistical Association, 99, S. 609–618.

Hansen, Ben B. (2007): Optmatch: Flexible, Optimal Matching for Observational Studies, in: R News, 7, S. 18–24.

Hansen, Ben B./Bowers, Jake (2008): Covariate Balance in Simple, Stratified and Clustered Comparative Studies, in: Statistical Science, 23, S. 219–236.

Hansen, Morris H./Hurwitz, William N. (1946): The Problem of Non-Response in Sample Surveys, in: Journal of the American Statistical Association, 41, S. 517–529.

Hartmann, Peter H./Schimpl-Neimanns, Bernhard (1992): Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 44, S. 315–340.

Heckel, Christiane/Hofman, Oliver (2014): Das ADM-Stichproben-System (F2F) ab 1997, in: Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V (Hrsg.): Stichproben-Verfahren in der Umfrageforschung, Wiesbaden: Springer VS, S. 85–116.

Heckel, Christiane/Wiese, Kathrin (2012): Sampling Frames for Telephone Surveys in Europe, in: Häder, Sabine/Häder, Michael/Kühne, Mike (Hrsg.): Telephone surveys in Europe: Research and practice, Berlin/New York: Springer.

Heer, Wim de (1999): International Response Trends: Results of an International Survey, in: Journal of Official Statistics, 15, S. 129–142.

Heerwegh, Dirk (2009): Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects, in: International Journal of Public Opinion Research, 21, S. 111–121.

Heerwegh, Dirk/Abts, Koen/Loosveldt, Geert (2007): Minimizing survey refusal and noncontact rates: do our efforts pay off?, in: Survey Research Methods, 1, S. 3–10.

Heerwegh, Dirk/Loosveldt, Geert (2009): Face-to-Face versus Web Surveying in a High-Internet-Coverage Population: Differences in Response Quality, in: Public Opinion Quarterly, 72, S. 836–846.

Heij, Vincent de/Schouten, Barry/Shlomo, Natalie (2014): RISQ manual 2.0: Tools in SAS and R for the computation of R-indicators and partial R-indicators, 7th Framework Programma (FP7) of the European Union Work package 8, Deliverable 2.1.

Hillygus, D. Sunshine (2005): The MISSING LINK: Exploring the Relationship Between Higher Education and Political Engagement, in: Political Behavior, 27, S. 25–47.

Hillygus, D. Sunshine/Jackson, Natalie/Young, McKenzie (2014): Professional respondents in nonprobability online panels, in: Callegaro, Mario/Baker, Reg/Bethlehem, Jelke/Göritz, Anja S./Krosnick, John A./Lavrakas, Paul J. (Hrsg.): Online panel research: A data quality perspective, Chichester, West Sussex: Wiley, S. 219–237.

Ho, Daniel E./Imai, Kosuke/King, Gary/Stuart, Elizabeth A. (2007): Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, in: Political Analysis, 15, S. 199–236.

Ho, Daniel/Imai, Kosuke/King, Gary/Stuart, Elizabeth A. (2011): MatchIt: Nonparametric Preprocessing for Parametric Causal Inference, in: Journal of Statistical Software, 42, S. 1–28.

Holbrook, A. L./Krosnick, Jon A. (2010): Social desirability bias in voter turnout reports: Tests using the item count technique, in: Public Opinion Quarterly, 74, S. 37–67.

Holbrook, Allyson L./Green, Melanie C./Krosnick, Jon A. (2003): Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires, in: Public Opinion Quarterly, 67, S. 79–125.

Holbrook, Allyson L./Krosnick, Jon A./Pfent, Alison (2007): The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms, in: Lepkowski, James M./Tucker, Clyde/Brick, J. Michael/Leeuw, Edith de/Japec, Lilli/Lavrakas, Paul J./Link,

Michael W./Sangster, Roberta L. (Hrsg.): Advances in Telephone Survey Methodology, Hoboken, N.J.: Wiley, S. 499–528.

Holland, Paul W. (1986): Statistics and Causal Inference, in: Journal of the American Statistical Association, 81, S. 945–960.

Holland, Paul W. (1988): Causal Inference, Path Analysis, and Recursive Structural Equations Models, in: Sociological Methodology, 18: S. 449–484.

Hoonakker, Peter/Carayon, Pascale (2009): Questionnaire Survey Nonresponse: A Comparison of Postal Mail and Internet Surveys, in: International Journal of Human-Computer Interaction, 25, S. 348–373.

Horvitz, Daniel G./Thompson, Donovan J. (1952): A Generalization of Sampling Without Replacement From a Finite Universe, in: Journal of the American Statistical Association, 47, S. 663–685.

Hox, Joop J./Leeuw, Edith D. de (1994): A comparison of nonresponse in mail, telephone, and face-to-face surveys, in: Quality and Quantity, 28, S. 329–344.

Hox, Joop/Mohorko, Anja/Leeuw, Edith de (2013): Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage across Countries and over Time, Survey Methods: Insights from the Field.

Hunsicker, Stefan/Schroth, Yvonne (2014): Dual-Frame-Ansatz in politischen Umfragen, Arbeitspapiere der Forschungsgruppe Wahlen e.V 2, Mannheim: Institut für Wahlanalysen und Gesellschaftsbeobachtung.

Hyndman, Rob J./Koehler, Anne B. (2006): Another look at measures of forecast accuracy, in: International Journal of Forecasting, 22, S. 679–688.

Iacus, Stefano M./King, G./Porro, Giuseppe (2009): CEM: Software for Coarsened Exact Matching, in: Journal of Statistical Software, 30, S. 1–27.

Iacus, Stefano M./King, Gary/Porro, Giuseppe (2011): Multivariate Matching Methods That are Monotonic Imbalance Bounding, in: Journal of the American Statistical Association, 106: S. 345–361.

Iacus, Stefano M./King, Gary/Porro, Guiseppe (2012): Causal Inference without Balance Checking: Coarsened Exact Matching, in: Political Analysis, 20, S. 1–24.

Imai, Kosuke/King, Gary/Stuart, Elizabeth (2008): Misunderstandings Among Experimentalists and Observationalists about Causal Inference, in: Journal of the Royal Statistical Society (A), 171, S. 481–502.

Imai, Kosuke/Ratkovic, Marc (2014): Covariate balancing propensity score, in: Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, S. 243–263.

Initiative D21 (2015): D21-Digital-Index 2015: Die Gesellschaft in der digitalen Transformation.

Jäckle, Annette/Lynn, Peter/Sinibaldi, Jennifer/Tipping, Sarah (2012): The Effect of Interviewer Experience, Attitudes, Personality and Skills on Respondent Co-operation with Face-to-Face Surveys, in: Survey Research Methods, 7, S. 1–15.

Jann, Ben (2005): Einführung in die Statistik, 2., bearb. Aufl., München/Wien: Oldenbourg.

Jann, Ben (2015): Asking Sensitive Questions: Overview and Introduction, in: Engel, Uwe/Jann, Ben/Lynn, Peter/Scherpenzeel, Annette/Sturgis, Patrick (Hrsg.): Improving Survey Methods: Lessons from recent research, New York/ London: Routledge.

Johnston, Richard/Brady, Henry E. (2002): The rolling cross-section design, in: Electoral Studies, 21, S. 283–295.

Joye, Dominique/Pollien, Alexandre/Sapin, Marlène/Stähli, Michèle Ernst (2012): Who Can Be Contacted by Phone? Lessons from Switzerland, in: Häder, Sabine/Häder, Michael/Kühne, Mike (Hrsg.): Telephone surveys in Europe: Research and practice, Berlin/New York: Springer, S. 85–102.

Kalton, Graham (1983): Compensating for missing survey data, Ann Arbor, MI: Survey Research Center, Institute for Social Research, the University of Michigan.

Kalton, Graham/Flores-Cervantes, Ismael (2003): Weighting Methods, in: Journal of Official Statistics, 19, S. 81–97.

Kalton, Graham/Maligalig, Dalisay S. (1991): A Comparison of Methods of Weighting Adjustment for Nonresponse, in: Proceedings of the US Bureau of the Census Annual Research Conference: S. 409–428.

Kauermann, Göran/Küchenhoff, Helmut (2011): Stichproben: Methoden und praktische Umsetzung mit R, Berlin, Heidelberg: Springer.

Keeter, Scott/Kennedy, Courtney/Clark, April/Tompson, Trevor/Mokrzycki, Mike (2007): What's Missing from National Landline RDD Surveys?: The Impact of the Growing Cell-Only Population, in: Public Opinion Quarterly, 71, S. 772–792.

Keeter, Scott/Kennedy, Courtney/Dimock, Michael/Best, Jonathan/Craighill, Peyton (2006): Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey, in: Public Opinion Quarterly, 70, S. 759–779.

Keeter, Scott/Miller, Carolyn/Kohut, Andrew/Groves, Robert M./Presser, Stanley (2000): Consequences of Reducing Nonresponse in a National Telephone Survey, in: Public Opinion Quarterly, 64, S. 125–148.

Kersten, Hubert M.P./Bethlehem, Jelke G. (1984): Exploring and reducing the nonresponse bias by asking the basic question, in: Statistical Journal of the United Nations Economic Commission for Europe, 2, S. 369–380.

Keusch, Florian (2015): Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection, in: Management Review Quarterly, 65, S. 183–216.

Kish, Leslie (1962): Studies of Interviewer Variance for Attitudinal Variables, in: Journal of the American Statistical Association, 57, S. 92–115.

Kish, Leslie (1965): Survey Sampling, New York: Wiley.

Kish, Leslie (1995): Methods for Design Effects, in: Journal of Official Statistics, 11, S. 55–77.

Koch, Achim (1997): ADM-Design und Einwohnermelderegister-Stichprobe: Stichprobenverfahren bei mündlichen Bevölkerungsumfragen, in: Gabler, Siegfrid/Hoffmeyer-Zlotnik, Jürgen H. P. (Hrsg.): Stichproben in der Umfragepraxis, Opladen: Westdt. Verlag, S. 99–116.

Koch, Achim (1998): Wenn ”mehr” nicht gleichbedeutend mit ”besser” ist: Ausschöpfungsquoten und Stichprobenverzerrungen in allgemeinen Bevölkerungsumfragen, in: ZUMA Nachrichten, 22, S. 66–90.

Koch, Achim (2002): 20 Jahre Feldarbeit im ALLBUS: Ein Blick in die Blackbox, in: ZUMA Nachrichten, 26, S. 9–37.

Kohler, Ulrich/Kreuter, Frauke (2012): Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung, 4., aktual. und überarb. Aufl., München: Oldenbourg.

Kosinski, Michal/Bachrach, Yoram/Kohli, Pushmeet/Stillwell, David/Graepel, Thore (2014): Manifestations of user personality in website choice and behaviour on online social networks, in: Machine Learning, 95, S. 357–380.

Kreuter, F. (2012): Facing the Nonresponse Challenge, in: The ANNALS of the American Academy of Political and Social Science, 645, S. 23–35.

Kreuter, Frauke (2013): Improving surveys with paradata: Analytic uses of process information, Hoboken, N.J.: Wiley.

Kreuter, Frauke/Olson, Kristen/Wagner, James R./Yan, Ting/Ezzati-Rice, Trena M./Casas-Cordero, Carolina/Lemay, Michael/Peytchev, Andy/Groves, Robert M./Raghunathan, Trivellore E. (2010): Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple

surveys, in: Journal of the Royal Statistical Society: Series A (Statistics in Society), 173, S. 389–407.

Kreuter, Frauke/Presser, Stanley/Tourangeau, Roger (2008): Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity, in: Public Opinion Quarterly, 72, S. 847–865.

Krosnick, Jon A. (1991): Response strategies for coping with the cognitive demands of attitude measures in surveys, in: Applied Cognitive Psychology, 5, S. 213–236.

Krosnick, Jon A. (1999): Survey research, in: Annual Review of Psychology, 50, S. 537–567.

Krueger, Brian S./West, Brady T. (2014): Assessing the Potential of Paradata and other Auxiliary Data for Nonresponse Adjustments, in: Public Opinion Quarterly, 78, S. 795–831.

Kruskal, William/Mosteller, Frederick (1979a): Representative Sampling, I: Non-Scientific Literature, in: International Statistical Review, 47, S. 13–24.

Kruskal, William/Mosteller, Frederick (1979b): Representative Sampling, II: Scientific Literature, Excluding Statistics, in: International Statistical Review, 47, S. 111–127.

Kruskal, William/Mosteller, Frederick (1979c): Representative Sampling, III: The Current Statistical Literature, in: International Statistical Review, 47, S. 245–265.

Lavrakas, Paul J. (2010): Telephone Surveys, in: Marsden, Peter V./Wright, James D. (Hrsg.): Handbook of Survey Research, Bingley, UK: Emerald, S. 471–498.

Lavrakas, Paul J./Shuttles, Charles D./Steeh, Charlotte/Fienberg, Howard (2007): The State of Surveying Cell Phone Numbers in the United States: 2007 and Beyond, in: Public Opinion Quarterly, 71, S. 840–854.

Lazarsfeld, Paul/Fiske, Marjorie (1938): The ”Panel” as a New Tool for Measuring Opinion, in: The Public Opinion Quarterly, 2, S. 596–612.

Lee, Brian K./Lessler, Justin/Stuart, Elizabeth A./Biondi-Zoccai, Giuseppe (2011): Weight Trimming and Propensity Score Weighting, in: PLoS ONE, 6, e18174.

Lee, Sunghee (2006): Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys, in: Journal of Official Statistics, 22, S. 329–349.

Lee, Sunghee/Valliant, Richard (2009): Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment, in: Sociological Methods & Research, 37, S. 319–343.

Leenheer, Jorna/Scherpenzeel, Annette (2013): Does It Pay Off to Include Non-Internet Households in an Internet Panel?, in: International Journal of Internet Science, 8, S. 17–29.

Leeuw, Edith de (2005): To Mix or Not to Mix Data Collection Modes in Surveys, in: Journal of Official Statistics, 21, S. 233–255.

Leeuw, Edith de/Heer, Wim de (2002): Trends in household survey nonresponse: A longitudinal and international comparison, in: Groves, Robert M./Dillman, Don A./Eltlinge, John L./Little, Roderick J.A (Hrsg.): Survey nonresponse, New York: Wiley, S. 41–54.

Leeuw, Edith de/Hox, Joop J. (2011): Internet Surveys as Part of a Mixed-Mode Design, in: Das, Marcel/Ester, Peter/Kaczmirek, Lars (Hrsg.): Social and behavioral research and the internet: Advances in applied methods and research strategies, New York: Routledge, S. 45–76.

International Handbook of Survey Methodology (2008), New York: Psychology Press.

Lessler, Judith T./Kalsbeek, William D. (1992): Nonsampling error in surveys, New York: Wiley.

Little, Roderick J.A (1986): Survey Nonresponse Adjustments for Estimates of Means, in: International Statistical Review, 54, S. 139–157.

Little, Roderick J.A. (1988): Missing-Data Adjustments in Large Surveys, in: Journal of Business & Economic Statistics, 6, S. 287–296.

Little, Roderick J.A/Rubin, Donald B. (2002): Statistical analysis with missing data, 2. Aufl., Hoboken/N. J: Wiley-Interscience.

Little, Roderick J.A/Vartivarian, Sonya (2005): Does Weighting for Nonresponse Increase the Variance of Survey Means?, in: Survey Methodology, 31, S. 161–168.

Lohr, Sharon L. (2010): Sampling: Design and Analysis, 2. Aufl., Boston, Mass.: Brooks/Cole.

Lönnqvist, Jan-Erik/Paunonen, Sampo/Verkasalo, Markku/Leikas, Sointu/Tuulio-Henriksson, Annamari/Lönnqvist, Jouko (2007): Personality characteristics of research volunteers, in: European Journal of Personality, 21, S. 1017–1030.

Loosveldt, Geert/Sonck, Nathalie (2008): An evaluation of the weighting procedures for an online access panel survey, in: Survey Research Methods, 2, S. 93–105.

Luellen, Jason K./Shadish, William R./Clark, M.H. (2005): Propensity scores: an introduction and experimental test, in: Evaluation Review, 29, S. 530–558.

Lundquist, Peter/Särndal, Carl-Erik (2013): Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey, in: Journal of Official Statistics, 29.

Lynn, Peter (2009): Methodology of longitudinal surveys, Chichester, UK: Wiley.

Lynn, Peter/Clarke, Paul/Martin, Jean/Sturgis, Patrick (2002): The effects of extended interviewer efforts on non-response bias, in: Groves, Robert M./Dillman, Don A./Eltlinge, John L./Little, Roderick J.A (Hrsg.): Survey nonresponse, New York: Wiley, S. 135–148.

Lynn, Peter/Kaminska, Olena (2012): Factors Affecting Measurement Error in Mobile Phone Interviews, in: Häder, Sabine/Häder, Michael/Kühne, Mike (Hrsg.): Telephone surveys in Europe: Research and practice, Berlin/New York: Springer, S. 211–228.

Makridakis, Spyros (1993): Accuracy measures: theoretical and practical concerns, in: International Journal of Forecasting, 9, S. 527–529.

Malgorzata, Turner/Sturgis, Patrick/Martin, David/Skinner, Chris (2015): Can Interviewer Personality, Attitudes and Experience Explain the Design Effect in Face-to-Face Surveys?, in: Engel, Uwe/Jann, Ben/Lynn, Peter/Scherpenzeel, Annette/Sturgis, Patrick (Hrsg.): Improving Survey Methods: Lessons from recent research, New York/London: Routledge, S. 72–85.

Malhotra, Neil/Krosnick, Jon A. (2007): The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples, in: Political Analysis, 15, S. 286–323.

Manfreda, Katja Lozar/Bosnjak, Michael/Berzelak, Jernej/Haas, Iris/Vehovar, Vasja (2008): Web surveys versus other survey modes, in: International Journal of Market Research, 50, S. 79–104.

Mangione, Thomas W./Fowler, Floyd J./Louis, Thomas A. (1992): Question Characteristics and Interviewer Effects, in: Journal of Official Statistics, 8, S. 293–307.

Marcus, Bernd/Schutz, Astrid (2005): Who Are the People Reluctant to Participate in Research? Personality Correlates of Four Different Types of Nonresponse as Inferred from Self- and Observer Ratings, in: Journal of Personality, 73, S. 959–984.

Martin, Elizabeth A./Traugott, Michael W./Kennedy, Courtney (2005): A Review and Proposal for a New Measure of Poll Accuracy, in: Public Opinion Quarterly, 69, S. 342–369.

Massey, Douglas S./Tourangeau, Roger (2012): Where Do We Go from Here? Nonresponse and Social Measurement, in: The ANNALS of the American Academy of Political and Social Science, 645, S. 222–236.

Matsuo, Hideko/Billiet, Jaak/Loosveldt, Geert/Berglund, Frode/Kleven, Øyvin (2010): Measurement and adjustment of non-response bias based on non-response surveys: the case of Belgium and Norway in the European Social Survey Round 3, in: Survey Research Methods, 4, S. 165–178.

Maurer, Marcus/Jandura, Olaf (2009): Masse statt Klasse? Einige kritische Anmerkungen zu Repräsentativität und Validität von Online-Befragungen, in: Jackob, Nikolaus/Schoen, Harald/Zerback, Thomas (Hrsg.): Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung, Wiesbaden: Springer VS, S. 61–73.

Mavletova, A. (2013): Data Quality in PC and Mobile Web Surveys, in: Social Science Computer Review, 31, S. 725–743.

Mebane, Walter R./Sekhon, Jasjeet S. (2011): Genetic Optimization Using Derivatives: The rgenoud Package for R, in: Journal of Statistical Software, 42, S. 1–26.

Merkle, Daniel M./Edelman, Murray (2002): Nonresponse in Exit Polls: A Comprehensive Analysis, in: Groves, Robert M./Dillman, Don A./Eltlinge, John L./Little, Roderick J.A (Hrsg.): Survey nonresponse, New York: Wiley, S. 243–257.

Messer, B. L./Dillman, Don A. (2011): Surveying the General Public over the Internet Using Address-Based Sampling and Mail Contact Procedures, in: Public Opinion Quarterly, 75, S. 429–457.

Mitofsky, Warren J. (1998): Review: Was 1996 a Worse Year for Polls Than 1948?, in: Public Opinion Quarterly, 62, S. 230.

Mohorko, Anja/Leeuw, Edith de/Hox, Joop (2013): Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time, in: Journal of Official Statistics, 29, S. 609–622.

Mondak, Jeffery J./Halperin, Karen D. (2008): A Framework for the Study of Personality and Political Behaviour, in: British Journal of Political Science, 38, S. 335–362.

Mosteller, Frederick/Hyman, Herbert/McCarthy, Philip J./Marks, Eli S./Truman, David B. (1949): The Pre-Election Polls of 1948: Report to the Committee on Analysis of Pre-Election Polls and Forecasts, New York: Social Science Research Council.

Neller, Katja (2005): Kooperation und Verweigerung: Eine Non-Response-Studie, in: ZUMA-Nachrichten, 57: S. 9–36.

Neyman, J. (1937): Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability, in: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 236, S. 333–380.

Neyman, Jerzy (1923): On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9: Translated with an introduction by D. M. Dabrowska and T. P. Speed. 1990, in: Statistical Science, 5, S. 465–472.

Neyman, Jerzy (1934): On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, in: Journal of the Royal Statistical Society, 97, S. 558–625.

Norris, Pippa (2004): Electoral Engineering: Voting Rules and Political Behavior, Cambridge: Cambridge Univ. Press.

Olson, Kirsten (2006): Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias, in: Public Opinion Quarterly, 70, S. 737–758.

Olson, Kirsten/Bilgen, Ipek (2011): The Role Of Interviewer Experience on Acquiescence, in: Public Opinion Quarterly, 75, S. 99–114.

Olson, Kristen (2012): Paradata for Nonresponse Adjustment, in: The ANNALS of the American Academy of Political and Social Science, 645, S. 142–170.

O'Muircheartaigh, Colm/Campanelli, Pamela (1999): A multilevel exploration of the role of interviewers in survey non-response, in: Journal of the Royal Statistical Society: Series A (Statistics in Society), 162, S. 437–446.

Ostendorf, Fritz/Angleitner, Alois (1994): The Five-Factor taxonomy. Robust dimensions of personality description, in: Psychologica Belgica, 34, S. 175–194.

Pan, Wei/Bai, Haiyan (2015): Propensity Score Analysis: Concepts and Issues, in: Pan, Wei/Bai, Haiyan (Hrsg.): Propensity Score Analysis: Fundamentals and Developments, New York und London: The Guilford Press, S. 3–19.

Pasek, Josh (2015): When will Nonprobability Surveys Mirror Probability Surveys? Considering Types of Inference and Weighting Strategies as Criteria for Correspondence, in: International Journal of Public Opinion Research, 28, S. 1–23.

Petrolia, Daniel R./Bhattacharjee, Sanjoy (2009): Revisiting Incentive Effects: Evidence from a Random-Sample Mail Survey on Consumer Preferences for Fuel Ethanol, in: Public Opinion Quarterly, 73, S. 537–550.

Peytchev, Andy (2012): Multiple Imputation for Unit Nonresponse and Measurement Error, in: Public Opinion Quarterly, 76, S. 214–237.

Peytchev, Andy (2013): Consequences of Survey Nonresponse, in: The ANNALS of the American Academy of Political and Social Science, 645, S. 88–111.

Peytchev, Andy/Baxter, Rodney K./Carley-Baxter, Lisa R. (2009): Not All Survey Effort is Equal: Reduction of Nonresponse Bias and Nonresponse Error, in: Public Opinion Quarterly, 73, S. 785–806.

Peytchev, Andy/Carley-Baxter, Lisa R./Black, Michele C. (2010): Coverage Bias in Variances, Associations, and Total Error From Exclusion of the Cell Phone-Only Population in the United States, in: Social Science Computer Review, 28, S. 287–302.

Peytchev, Andy/Carley-Baxter, Lisa R./Black, Michele C. (2011): Multiple Sources of Nonobservation Error in Telephone Surveys: Coverage and Nonresponse, in: Sociological Methods & Research, 40, S. 138–168.

Peytchev, Andy/Riley, Sarah/Rosen, Jeff/Murphy, Joe/Lindblad, Mark (2010): Reduction of Nonresponse Bias through Case Prioritization, in: Survey Research Methods, 4, S. 21–29.

Peytcheva, Emilia/Groves, Robert M. (2009): Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates, in: Journal of Official Statistics, 25, S. 193–201.

Preisendorfer, Peter/Wolter, Felix (2014): Who Is Telling the Truth? A Validation Study on Determinants of Response Behavior in Surveys, in: Public Opinion Quarterly, 78, S. 126–146.

Proner, Hanna (2011): Ist keine Antwort auch eine Antwort? Die Teilnahme an politischen Umfragen, 1. Aufl., Wiesbaden: VS Verl. für Sozialwissenschaften.

Quatember, Andreas (2015): Datenqualität in Stichprobenerhebungen: Eine verständnisorientierte Einführung in Stichprobenverfahren und verwandte Themen, Berlin, Heidelberg: Springer Spektrum.

Raento, Mika/Oulasvirta, Antti/Eagle, Nathan (2009): Smartphones: An Emerging Tool for Social Scientists, in: Sociological Methods & Research, 37, S. 426–454.

Ragnedda, Massimo/Muschert, Glenn W. (2013): The digital divide: The internet and social inequality in international perspective, Abingdon, Oxon/New York, NY: Routledge.

Rainer Schnell (2013): Linking Surveys and Administrative Data, WP-GRLC-2013-03, Nürnberg: German Record Linkage Center.

Rammstedt, Beatrice (2007): The 10-Item Big Five Inventory: Norm values and investigation of sociodemographic effects based on a German population representative sample, in: European Journal of Psychological Assessment, 23, S. 193–201.

Rammstedt, Beatrice/Goldberg, Lewis R./Borg, Ingwer (2010): The measurement equivalence of Big Five factor markers for persons with different levels of education, in: Journal of Research in Personality, 44, S. 53–61.

Rammstedt, Beatrice/Kemper, Christoph J./Klein, Mira Céline/Beierlein, Constanze/Kovaleva, Anastassiya (2013): Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: 10 Item Big Five Inventory (BFI-10), in: Methoden, Daten, Analysen (mda), 7, S. 233–249.

Rammstedt, Beatrice/Kemper, Christoph J./Schupp, Jürgen (2013): Einleitung – Standardisierte Kurzskalen zur Erfassung psychologischer Merkmale in Umfragen, in: Methoden, Daten, Analysen (mda), 7, S. 145–152.

Ramsey, J. B. (1969): Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis, in: Journal of the Royal Statistical Society. Series B (Methodological), 31, S. 350–371.

Rattinger, Hans/Roßdeutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2011): Vorwahl-Querschnitt (GLES 2009), hrsg. von GESIS Datenarchiv, Köln.

Rattinger, Hans/Roßdeutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2014a): Vorwahl-Querschnitt (GLES 2013), hrsg. von GESIS Datenarchiv, Köln.

Rattinger, Hans/Roßdeutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2015a): Langfrist-Online-Tracking, T17 (GLES), Köln: GESIS Datenarchiv.

Rattinger, Hans/Roßdeutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2016a): German Longitudinal Election Study – Wahlkampf-Panel 2013, 20.06.-04.10.2013, hrsg. von GESIS – Leibniz-Institut für Sozialwissenschaften, Köln.

Rattinger, Hans/Roßteutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard (2013): Rolling Cross-Section Campaign Survey with Post-election Panel Wave (GLES 2013): ZA5703 Data file, hrsg. von GESIS Data Archive, Köln.

Rattinger, Hans/Roßteutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2014b): Rolling Cross-Section-Wahlkampfstudie mit Nachwahl-Panelwelle (GLES 2013), hrsg. von GESIS Datenarchiv, Köln.

Rattinger, Hans/Roßteutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2015b): Langfrist-Online-Tracking, T16 (GLES), hrsg. von GESIS Datenarchiv, Köln.

Rattinger, Hans/Roßteutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2015c): Langfrist-Online-Tracking, T29 (GLES), hrsg. von GESIS Datenarchiv, Köln.

Rattinger, Hans/Roßteutscher, Sigrid/Schmitt-Beck, Rüdiger/Weßels, Bernhard/Wolf, Christof (2016b): Langfrist-Online-Tracking, T17 (GLES), hrsg. von GESIS Datenarchiv, Köln.

Rentfrow, Peter J./Gosling, Samuel D./Jokela, Markus/Stillwell, David J./Kosinski, Michal/Potter, Jeff (2013): Divided we stand: three psychological regions of the United States and their political, economic, social, and health correlates, in: Journal of Personality and Social Psychology, 105, S. 996–1012.

Roberts, Caroline/Vandenplas, Caroline/Stähli, Michèle Ernst (2014): Evaluating the impact of response enhancement methods on the risk of nonresponse bias and survey costs, in: Survey Research Methods, 8, S. 67–80.

Robins, James (1986): A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, in: Mathematical Modelling, 7, S. 1393–1512.

Rogelberg, Steven G./Conway, James M./Sederburg, Matthew E./Spitzmuller, Christiane/Aziz, Shanaz/Knight, William E. (2003): Profiling active and passive nonrespondents to an organizational survey, in: The Journal of applied psychology, 88, S. 1104–1114.

Rosenbaum, Paul R. (1989): Optimal Matching for Observational Studies, in: Journal of the American Statistical Association, 84, S. 1024–1032.

Rosenbaum, Paul R./Rubin, Donald B. (1983): The central role of the propensity score in observational studies for causal effects, in: Biometrika, 70, S. 41–55.

Rosenbaum, Paul R./Rubin, Donald B. (1984): Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, in: Journal of the American Statistical Association, 79, S. 516.

Rosenbaum, Paul R./Rubin, Donald B. (1985): Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score, in: The American Statistician, 39, S. 33–38.

Roßdeutscher, Sigrid/Schäfer, Armin (2015): Räumliche Unterschiede der Wahlbeteiligung bei der Bundestagswahl 2013: Die soziale Topografie der Nichtwahl, in: Korte, Karl-Rudolf (Hrsg.): Die Bundestagswahl 2013, Wiesbaden: Springer VS, S. 99–118.

Roth, Dieter (2008): Empirische Wahlforschung: Ursprung, Theorien, Instrumente und Methoden, 2., akt., Wiesbaden: VS Verlag für Sozialwissenschaften.

Rubin, Donald B. (1974): Estimating causal effects of treatments in randomized and nonrandomized studies, in: Journal of Educational Psychology, 66, S. 688–701.

Rubin, Donald B. (1976a): Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples, in: Biometrics, 32, S. 109–120.

Rubin, Donald B. (1976b): Multivariate Matching Methods That are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes, in: Biometrics, 32, S. 121–132.

Rubin, Donald B. (1977): Assignment to Treatment Group on the Basis of a Covariate, in: Journal of Educational and Behavioral Statistics, 2, S. 1–26.

Rubin, Donald B. (1978): Bayesian Inference for Causal Effects: The Role of Randomization, in: The Annals of Statistics, 6, S. 34–58.

Rubin, Donald B. (2006): Matched Sampling for Causal Effects, Cambridge: Cambridge University Press.

Rubin, Donald B./Thomas, Neal (1992a): Affinely Invariant Matching Methods with Ellipsoidal Distributions, in: The Annals of Statistics, 20, S. 1079–1093.

Rubin, Donald B./Thomas, Neal (1992b): Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions, in: Biometrika, 79, S. 797–809.

Ryu, Erica/Couper, Mick P./Marans, Robert W. (2005): Survey Incentives: Cash vs. In-Kind; Face-to-Face vs. Mail; Response Rate vs. Nonresponse Error, in: International Journal of Public Opinion Research, 18, S. 89–106.

Sánchez-Fernández, Juan/Muñoz-Leiva, Francisco/Montoro-Ríos, Francisco J./Ibáñez-Zapata, José Ángel (2010): An analysis of the effect of pre-incentives and post-incentives based on draws on response to web surveys, in: Quality & Quantity, 44, S. 357–373.

Sanders, David/Clarke, Harold D./Stewart, Marianne C./Whiteley, Paul (2006): Does Mode Matter For Modeling Political Choice? Evidence From the 2005 British Election Study, in: Political Analysis, 15, S. 257–285.

Särndal, Carl-Erik/Lundström, Sixten (2005): Estimation in Surveys with Nonresponse, Hoboken, NJ: Wiley.

Särndal, Carl-Erik/Swensson, Bengt/Wretman, Jan Hakan (1992): Model Assisted Survey Sampling, New York: Springer-Verlag.

Saßenroth, Denise (2012): Der Einfluss von Persönlichkeitseigenschaften auf die Kooperationsbereitschaft in Umfragen: Befunde der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 2004, 2006 und 2008, in: Methoden, Daten, Analysen (mda), 6, S. 21–44.

Schafer, Joseph L./Kang, Joseph (2008): Average causal effects from nonrandomized studies: A practical guide and simulated example, in: Psychological Methods, 13, S. 279–313.

Schneiderat, Götz/Schlinzig, Tino (2012): Mobile- and Landline-Onlys in Dual-Frame-Approaches: Effects on Sample Quality, in: Häder, Sabine/Häder, Michael/Kühne, Mike (Hrsg.): Telephone surveys in Europe: Research and practice, Berlin/New York: Springer, S. 121–143.

Schnell, Rainer (1997): Nonresponse in Bevölkerungsumfragen: Ausmass, Entwicklung und Ursachen, Opladen: Leske + Budrich.

Schnell, Rainer (1998): Besuchs- und Berichtsverhalten der Interviewer, in: Statistisches Bundesamt (Hrsg.): Interviewereinsatz und -qualifikation, Stuttgart: Metzler-Poeschel, S. 156–170.

Schnell, Rainer (2012): Survey-Interviews. Standardisierte Befragungen in den Sozialwissenschaften, Wiesbaden: Springer VS.

Schnell, Rainer (2013): Linking Surveys and Administrative Data, WP-GRLC-2013-03, Nürnberg: German Record Linkage Center.

Schnell, Rainer/Esser, Elke/Hill, Paul B. (2013): Methoden der empirischen Sozialforschung, 10., überarb. Aufl., München [u.a.]: Oldenbourg.

Schnell, Rainer/Kreuter, Frauke (2005): Separating Interviewer and Sampling-Point Effects, in: Journal of Official Statistics, 21, S. 389–410.

Schnell, Rainer/Noack, Marcel (2014): The Accuracy of Pre-Election Polling of German General Elections, in: methods, data, analyses (mda), 8, S. 5–24.

Schoen, Harald/Schumann, Siegfried (2007): Personality Traits, Partisan Attitudes, and Voting Behavior. Evidence from Germany, in: Political Psychology, 28, S. 471–498.

Schonlau, Matthias/Zapert, Kinga/Simon, Lisa Payne/Sanstad, Katherine Haynes/Marcus, Sue M./Adams, John/Spranca, Mark/Kan, Hongjun/Turner, Rachel/Berry, Sandra H. (2004): A Comparison Between Responses From a

Propensity-Weighted Web Survey and an Identical RDD Survey, in: Social Science Computer Review, 22, S. 128–138.

Schouten, Barry/Bethlehem, Jelke/Beullens, Koen/Kleven, Øyvin/Loosveldt, Geert/Luiten, Annemieke/Rutar, Katja/Shlomo, Natalie/Skinner, Chris (2012): Evaluating, Comparing, Monitoring, and Improving Representativeness of Survey Response Through R-Indicators and Partial R-Indicators, in: International Statistical Review, 80, S. 382–399.

Schouten, Barry/Shlomo, Natalie/Skinner, Chris (2011): Indicators for Monitoring and Improving Representativeness of Response, in: Journal of Official Statistics, 27, S. 231–253.

Schuler, Megan (2015): Overview of Implementing Propensity Score Analysis in Statistical Software, in: Pan, Wei/Bai, Haiyan (Hrsg.): Propensity Score Analysis: Fundamentals and Developments, New York und London: The Guilford Press, S. 20–48.

Schumann, Siegfried (2001): Persönlichkeitsbedingte Einstellungen zu Parteien: Der Einfluß von Persönlichkeitseigenschaften auf Einstellungen zu politischen Parteien, Reprint 2016, Berlin, Boston: Oldenbourg Wissenschaftsverlag.

Schumann, Siegfried (2002): Prägen Persönlichkeitseigenschaften Einstellungen zu Parteien? Ergebnisse einer empirischen Untersuchungsreihe, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 54, S. 64–84.

Schumann, Siegfried (2014): Persönlichkeit und Wahlverhalten, in: Falter, Jürgen W./Schoen, Harald (Hrsg.): Handbuch Wahlforschung, Wiesbaden: Springer VS, S. 591–624.

Schumann, Siegfried/Schoen, Harald (2005): Persönlichkeit: Eine vergessene Größe der empirischen Sozialforschung, Wiesbaden: VS Verlag für Sozialwissenschaften.

Sciarini, Pascal/Goldberg, Andreas C. (2016): Turnout Bias in Postelection Surveys: Political Involvement, Survey Participation, and Vote Overreporting, in: Journal of Survey Statistics and Methodology, 4, S. 110–137.

Sekhon, Jasjeet S. (2011): Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R, in: Journal of Statistical Software, 42, S. 1–52.

Shadish, William R./Clark, M.H./Steiner, Peter M. (2008): Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments, in: Journal of the American Statistical Association, 103, S. 1334–1344.

Shih, Tse-Hua/Xitao, Fan (2008): Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis, in: Field Methods, 20, S. 249–271.

Shin, E./Johnson, T. P./Rao, K. (2012): Survey Mode Effects on Data Quality: Comparison of Web and Mail Modes in a U.S. National Panel Survey, in: Social Science Computer Review, 30, S. 212–228.

Shlomo, Natalie/Skinner, Chris/Schouten, Barry (2012): Estimation of an indicator of the representativeness of survey response, in: Journal of Statistical Planning and Inference, 142, S. 201–211.

Singer, Eleanor (2006): Introduction: Nonresponse Bias in Household Surveys, in: Public Opinion Quarterly, 70, S. 637–645.

Singer, Eleanor/Groves, Robert M./Corning, Amy D. (1999): Differential Incentives: Beliefs About Practices, Perceptions of Equity, and Effects on Survey Participation, in: Public Opinion Quarterly, 63, S. 251.

Singer, Eleanor/Ye, Cong (2012): The Use and Effects of Incentives in Surveys, in: The ANNALS of the American Academy of Political and Social Science, 645, S. 112–141.

Skitka, Linda J./Sargis, Edward G. (2006): The internet as psychological laboratory, in: Annual Review of Psychology, 57: S. 529–555.

Smith, Jeffrey A./Todd, Petra E. (2005): Does matching overcome LaLonde's critique of nonexperimental estimators?, in: Journal of Econometrics, 125, S. 305–353.

Smith, Tom W. (2011): Refining the Total Survey Error Perspective, in: International Journal of Public Opinion Research, 23, S. 464–484.

Sniderman, Paul M./Grob, Douglas B. (1996): Innovations in Experimental Design in Attitude Surveys, in: Annual Review of Sociology, 22, S. 377–399.

Steeh, Charlotte G. (1981): Trends in Nonresponse Rates, 1952-1979, in: Public Opinion Quarterly, 45, S. 40–57.

Steeh, Charlotte/Kirgis, Nicole/Cannon, Brian/DeWitt, Jeff (2001): Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century, in: Journal of Official Statistics, 17, S. 227–247.

Steinbrecher, Markus/Roßmann, Joss/Bergmann, Michael (2015): Das Wahlkampf-Panel der German Longitudinal Election Study 2009: Konzeption, Durchführung, Aufbereitung und Archivierung. Version 5.0.0, GESIS Papers 03.

Stern, Michael J./Bilgen, Ipek/Dillman, Don A. (2014): The State of Survey Methodology: Challenges, Dilemmas, and New Frontiers in the Era of the Tailored Design, in: Field Methods, 26, S. 284–301.

Stoop, Ineke A. (2004): Surveying Nonrespondents, in: Field Methods, 16, S. 23–54.

Stoop, Ineke A. (2005): The Hunt for the Last Respondent: Nonresponse in Sample Surveys, The Hague: SCP, Social and Cultural Planning Office of the Netherlands.

Stoop, Ineke A. (2016): Unit Nonresponse, in: Wolf, Christof/Joye, Dominique/Smith, Tom E. C./Fu, Yang-chih (Hrsg.): The SAGE Handbook of Survey Methodology, London: Sage Publications, S. 409–424.

Stoop, Ineke A./Billiet, Jaak/Koch, Achim/Fitzgerald, Rory (2010): Reducing survey nonresponse: Lessons learned from the European Social Survey, Oxford: Wiley-Blackwell.

Stuart, Elizabeth A. (2010): Matching Methods for Causal Inference: A Review and a Look Forward, in: Statistical Science, 25, S. 1–21.

Stuart, Elizabeth A./Lalongo, Nicholas S. (2010): Matching Methods for Selection of Participants for Follow-Up, in: Multivariate Behavioral Research, 45, S. 746–765.

Sturgis, Patrick/Baker, Nick/Callegaro, Mario/Fisher, Stephen/Green, Jane/Jennings, Will/Kuha, Jouni/Lauderdale, Ben/Smith, Patten (2016): Report of the Inquiry into the 2015 British general election opinion polls, London: Market Research Society and British Polling.

Taddicken, Monika (2009): Methodeneffekte von Web-Befragungen: Soziale Erwünschtheit vs. Soziale Entkontextualisierung, in: Weichbold, Martin/Bacher, Johann/Wolf, Christof (Hrsg.): Umfrageforschung: Herausforderungen und Grenzen, Wiesbaden: VS Verlag für Sozialwissenschaften, S. 85–104.

Terhanian, George/Bremer, John (2000): Confronting the Selection-Bias and Learning Effects Problems Associated with Internet Research, Rochester, NY: Harris Interactive.

Terhanian, George/Bremer, John/Smith, Renee/Thomas, Randy (2000): Correcting Data from Online Surveys for the Effects of Nonrandom Selection and Nonrandom Assignment, White paper, Rochester, NY: Harris Interactive.

Terwey, Michael/Baltzer, Stefan (2014): ALLBUS 1980-2012 Variable Report: Studien-Nr. 4578 Version: 1.0.0, VARIABLE Reports 07, GESIS Datenarchiv für Sozialwissenschaften.

Thoemmes, Felix J./Kim, Eun Sook (2011): A Systematic Review of Propensity Score Methods in the Social Sciences, in: Multivariate Behavioral Research, 46, S. 90–118.

Thompson, Steven K. (2012): Sampling, 3. Aufl., Hoboken, N.J.: Wiley.

Tilley, James/Evans, Geoffrey (2014): Ageing and generational effects on vote choice: Combining cross-sectional and panel data to estimate APC effects, in: Electoral Studies, 33: S. 19–27.

Toepoel, Vera (2012): Effects of Incentives in Surveys, in: Gideon, Lior (Hrsg.): Handbook of Survey Methodology for the Social Sciences, New York: Springer, S. 209–223.

Toepoel, Vera (2016): Doing surveys online, Los Angeles, Calif.: Sage Publications.

Toepoel, Vera/Das, Marcel/van Soest, Arthur (2009): Relating Question Type to Panel Conditioning: Comparing Trained and Fresh Respondents, in: Survey Research Methods, 3, S. 73–80.

Toepoel, Vera/Lugtig, Peter (2014): What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users, in: Social Science Computer Review, 32, S. 544–560.

Tourangeau, Roger/Conrad, Frederick G./Couper, Mick P. (2013): The Science of Web Surveys, New York: Oxford University Press.

Tourangeau, Roger/Kreuter, Frauke/Eckman, Stephanie (2012): Motivated Underreporting in Screening Interviews, in: Public Opinion Quarterly, 76, S. 453–469.

Tourangeau, Roger/Rips, Lance J./Rasinski, Kenneth A. (2000): The Psychology of Survey Response, New York: Cambridge University Press.

Tourangeau, Roger/Shapiro, Gary/Kearney, Anne/Ernst, Lawrence (1997): Who Lives Here? Survey Undercoverage and Household Roster Questions, in: Journal of Official Statistics, 13, S. 1–18.

Tourangeau, Roger/Yan, Ting (2007): Sensitive Questions in Surveys, in: Psychological Bulletin, 133, S. 859–883.

Trafimow, David/Marks, Michael (2014): Editorial, in: Basic and Applied Social Psychology, 37, S. 1–2.

Traugott, Michael W. (2001): Trends: Assessing Poll Performance in the 2000 Campaign, in: The Public Opinion Quarterly, 65, S. 389–419.

Truett, Kim R. (1993): Age Differences in Conservatism, in: Personality and Individual Differences, 14, S. 405–411.

Valliant, Richard/Dever, Jill A./Kreuter, Frauke (2013): Practical Tools for Designing and Weighting Survey Samples, New York: Springer.

van Hiel, Alain/Cornelis, I./Roets, A. (2007): The intervening role of social worldviews in the relationship between the five-factor model of personality and social attitudes, in: European Journal of Personality, 21, S. 131–148.

van Selm, Martine/Jankowski, Nicholas W. (2006): Conducting Online Surveys, in: Quality and Quantity, 40, S. 435–456.

Vandenplas, Caroline/Joye, Dominique/Staehli, Michèle/Pollien, Alexandre (2015): Identifying Pertinent Variables for Nonresponse Follow-Up Surveys. Lessons Learned from 4 Cases in Switzerland, in: Survey Research Methods, 9, S. 141–158.

Vecchione, Michele/Schoen, Harald/Castro, José Luis González/Cieciuch, Jan/Pavlopoulos, Vassilis/Caprara, Gian Vittorio (2011): Personality correlates of party preference: The Big Five in five big European countries, in: Personality and Individual Differences, 51, S. 737–742.

Vehovar, Vasja/Toepoel, Vera/Steinmetz, Stephanie (2016): Non-probability Sampling, in: Wolf, Christof/Joye, Dominique/Smith, Tom E. C./Fu, Yang-chih (Hrsg.): The SAGE Handbook of Survey Methodology, London: Sage Publications, S. 329–345.

Vehre, Helen (2012): "Sie wollen mir doch was verkaufen!": Analyse der Umfrageteilnahme in einem offline rekrutierten Access Panel, Wiesbaden: Springer VS.

Vetter, Angelika (1997a): Political Efficacy: Alte und neue Meßmodelle im Vergleich, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 49, S. 53–73.

Vetter, Angelika (1997b): Political Efficacy - Reliabilität und Validität: Alte und neue Meßmodelle im Vergleich, Wiesbaden: Deutscher Universitätsverlag.

Voogt, Robert J. J./Saris, Willem E. (2003): To Participate or Not to Participate: The Link Between Survey Participation, Electoral Participation, and Political Interest, in: Political Analysis, 11, S. 164–179.

Wagner, James R. (2010): The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data, in: Public Opinion Quarterly, 74, S. 223–243.

Wagner, James R. (2012a): A Comparison of Alternative Indicators for the Risk of Nonresponse Bias, in: Public Opinion Quarterly, 76, S. 555–575.

Wagner, James R. (2012b): Adaptive Contact Strategies in Telephone and Face-to-Face Surveys, in: Survey Research Methods, 7, S. 45–55.

Wagner, James R./West, Brady T./Kirgis, Nicole/Lepkowski, James M./Axinn, William G./Ndiaye, Shonda K. (2012): Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection, in: Journal of Official Statistics, 28, S. 477–499.

Wasmer, Martina/Blohm, Michael/Walter, Jessica/Scholz, Evi/Jutz, Regina (2014): Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften": (ALLBUS) 2012, GESIS-Technical Reports 22, Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften.

Wasserstein, Ronald L./Lazar, Nicole A. (2016): The ASA's Statement on p - Values: Context, Process, and Purpose, in: The American Statistician, 70, S. 129–133.

Watson, Dorothy (1992): Correcting for Acquiescent Response Bias in the Absence of a Balanced Scale: An Application to Class Consciousness, in: Sociological Methods & Research, 21, S. 52–88.

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

West, Brady T./Kreuter, Frauke/Jaenichen, Ursula (2013): "Interviewer" Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?, in: Journal of Official Statistics, 29, S. 277–297.

West, Brady T./Olson, Kristen (2011): How Much of Interviewer Variance is Really Nonresponse Error Variance?, in: Public Opinion Quarterly, 74, S. 1004–1026.

Wright, Malcolm J./Farrar, David P./Russell, Deborah F. (2014): Polling Accuracy in a Multiparty Election, in: International Journal of Public Opinion Research, 26, S. 113–124.

Yan, Ting/Tourangeau, Roger (2008): Fast times and easy questions: The effects of age, experience and question complexity on web survey response times, in: Applied Cognitive Psychology, 22, S. 51–68.

Yeager, David S./Krosnick, Jon A./Chang, Linchiat/Javitz, Harold S./Levendusky, Matthew S./Simpser, Alberto/Wang, Rui (2011): Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples, in: Public Opinion Quarterly, 75, S. 709–747.

Yu, Julie/Cooper, Harris (1983): A Quantitative Review of Research Design Effects on Response Rates to Questionnaires, in: Journal of Marketing Research, 20, S. 36–44.

A. Berechnete Design-Gewichte: Politbarometer

Tabelle A1.: Anhang: Berechnete Design-Gewichte Politbarometer Jahresstichproben

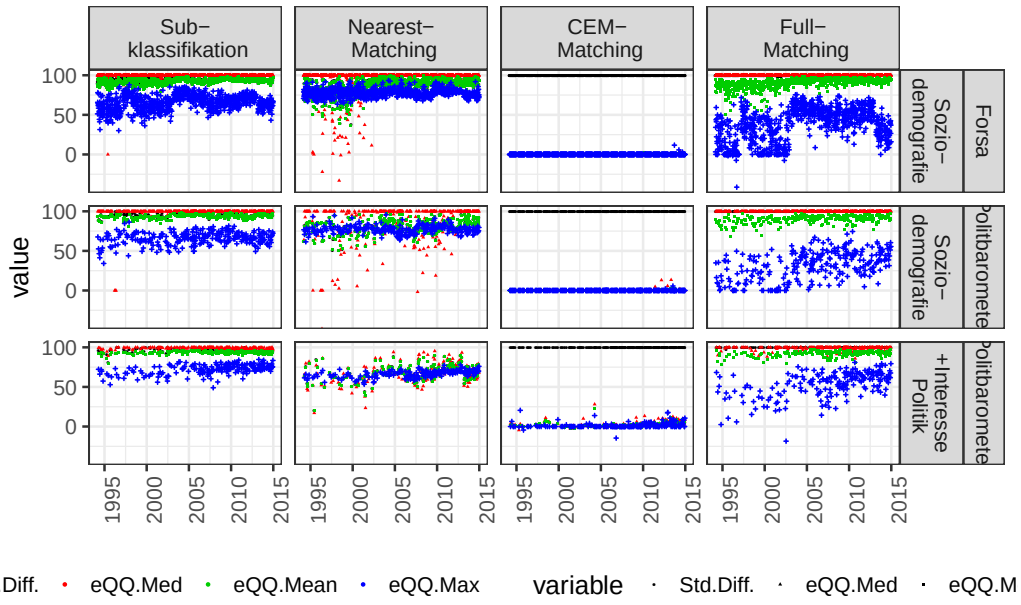
Jahr	West: Zensus	Ost: Zensus	West: Polit.	Ost: Polit.	Gewicht: West	Gewicht: Ost
1991	80.73	19.27	48.59	51.41	1.66	0.37
1992	81.03	18.97	48.68	51.32	1.66	0.37
1993	81.23	18.77	48.99	51.01	1.66	0.37
1994	81.36	18.64	48.80	51.20	1.67	0.36
1995	81.49	18.51	49.65	50.35	1.64	0.37
1996	81.59	18.41	79.35	20.65	1.03	0.89
1997	81.66	18.34	80.08	19.92	1.02	0.92
1998	81.75	18.25	80.22	19.78	1.02	0.92
1999	81.86	18.14	58.53	41.47	1.40	0.44
2000	82.00	18.00	52.84	47.16	1.55	0.38
2001	82.18	17.82	52.64	47.36	1.56	0.38
2002	82.33	17.67	53.38	46.62	1.54	0.38
2003	82.45	17.55	59.20	40.80	1.39	0.43
2004	82.55	17.45	59.35	40.65	1.39	0.43
2005	82.64	17.36	59.34	40.66	1.39	0.43
2006	82.73	17.27	59.67	40.33	1.39	0.43
2007	82.83	17.17	59.15	40.85	1.40	0.42
2008	82.92	17.08	58.84	41.16	1.41	0.42
2009	82.98	17.02	59.75	40.25	1.39	0.42
2010	83.05	16.95	55.62	44.38	1.49	0.38
2011	83.16	16.84	59.52	40.48	1.40	0.42
2012	83.24	16.76	59.89	40.11	1.39	0.42
2013	83.32	16.68	60.37	39.63	1.38	0.42
2014	83.38	16.62	60.24	39.76	1.38	0.42

Quelle: Eigene Darstellung. Daten: Politbarometer-Stichproben über die GESIS bezogen (http://www.gesis.org/wahlen/politbarometer, Stand: 11. Mai 2018). Amtliche Referenzwerte: Destatis über die „Gemeinsames neues statistisches Informationssystem“ (GENESIS)-Datenbank, Indikator 12411-0009.

B. Matching: Telefonumfragen

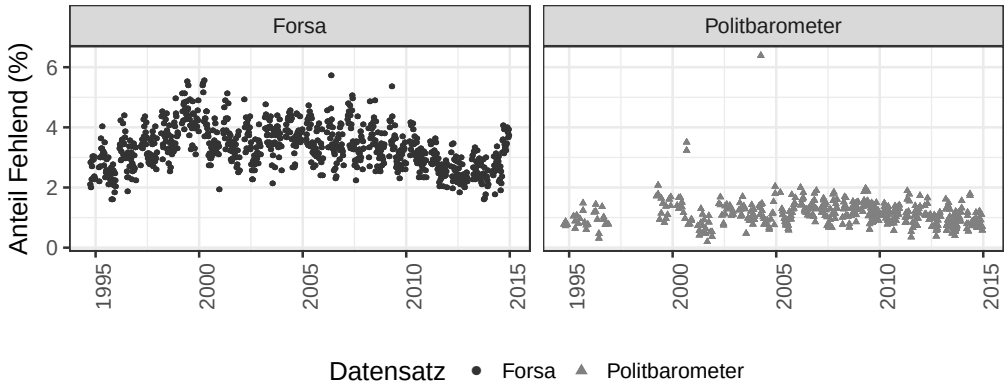
Nachfolgend sind jeweils die fehlenden Werte des jeweiligen Korrekturmodells, die erreichten Balancierungen des durchgeführten „Propensity Score Matching“ (PSM) an Hand verschiedener Kennzahlen und die Maximalgewichte der daraus resultierenden Bedeutungsgewichte für das Kapitel 9.1 und das Kapitel 9.2 dargestellt.

Abbildung B1.: Balancierung: Forsa und Politbarometer (Kapitel 9.1)



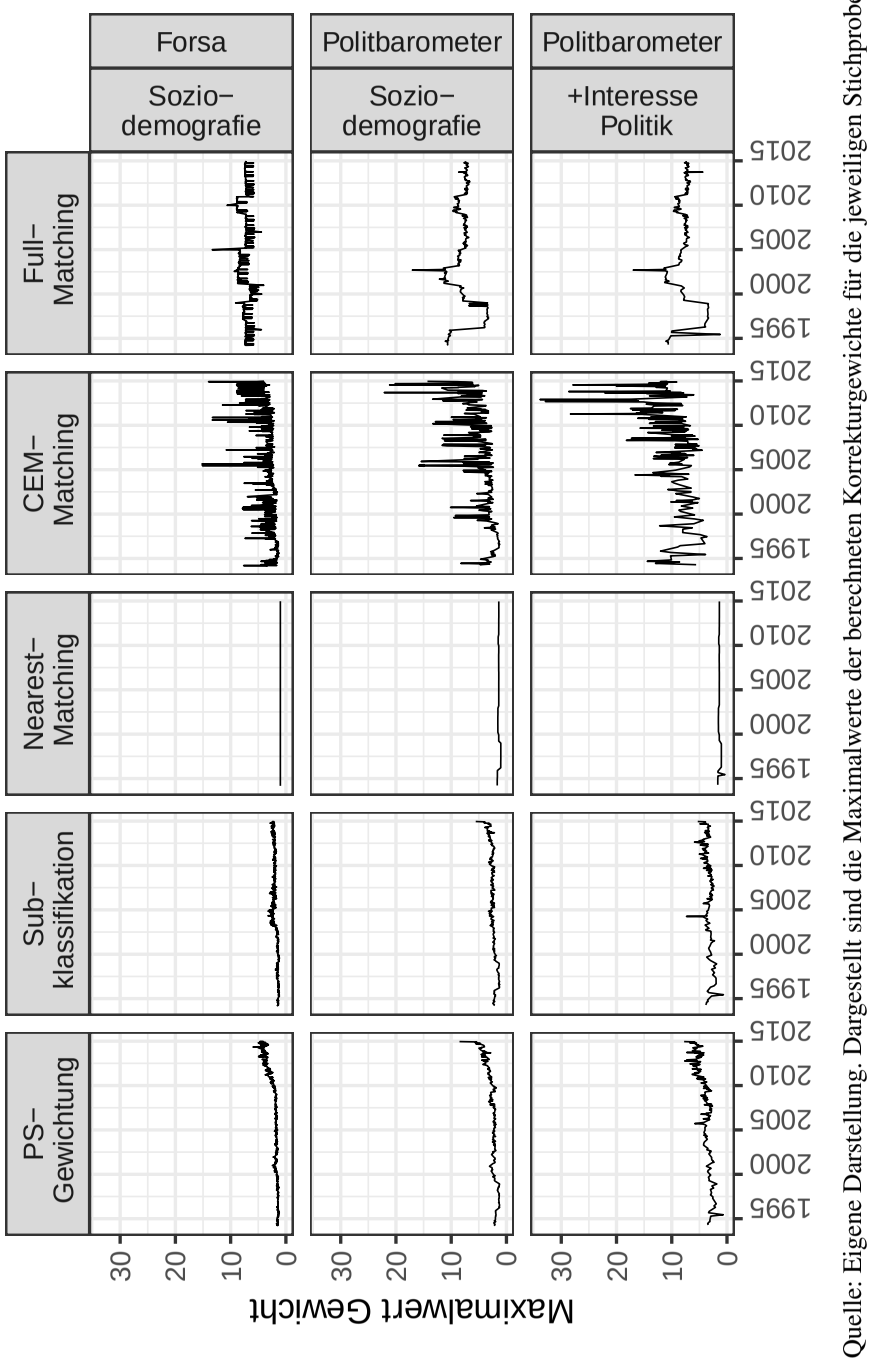
Quelle: Eigene Darstellung. Dargestellt ist die erreichte Balancierung der jeweiligen Stichproben

Abbildung B2.: Fehlende Werte Korrekturmodell: Forsa und Politbarometer (Kapitel 9.1)



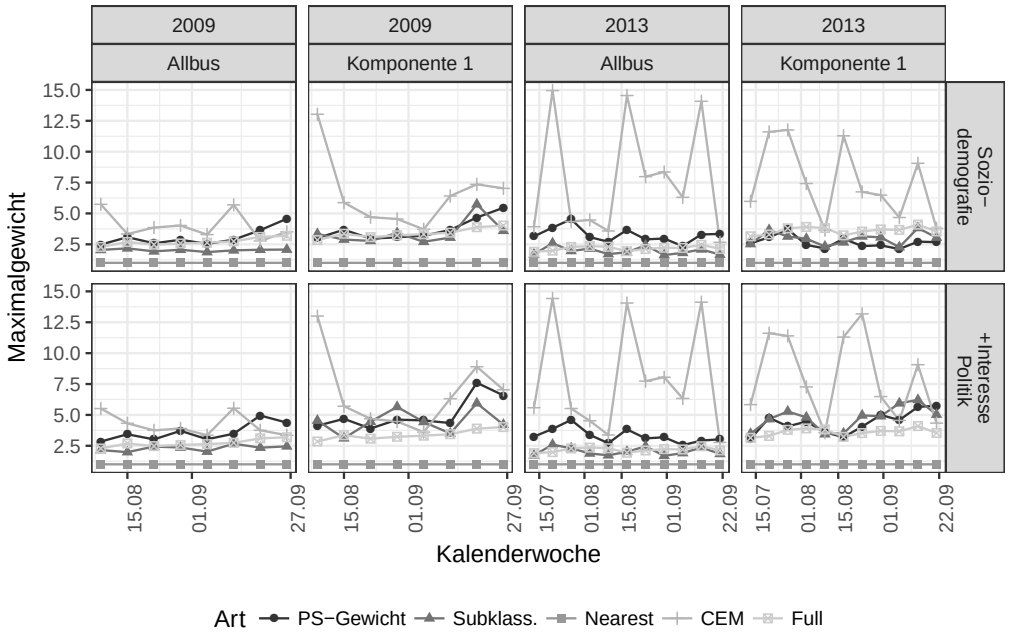
Quelle: Eigene Darstellung. Dargestellt sind die prozentualen Anteile der Beobachtungen der jeweiligen Stichprobe, welche keinen gültigen Gewichtungswert erhalten haben.

Abbildung B3.: Maximalgewichte: Forsa und Politbarometer (Kapitel 9.1)



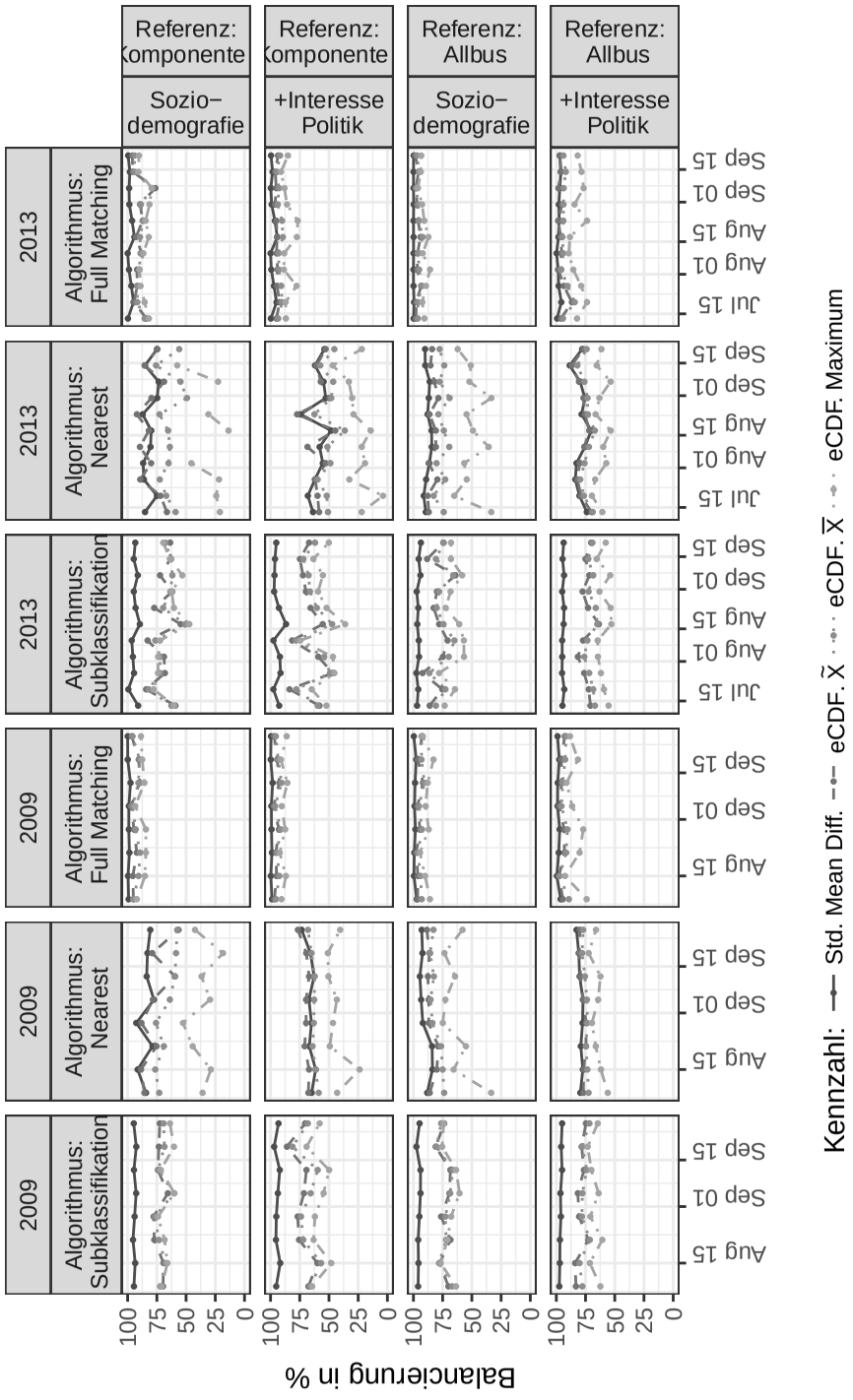
Quelle: Eigene Darstellung. Dargestellt sind die Maximalwerte der berechneten Korrekturgewichte für die jeweiligen Stichproben.

Abbildung B4.: Maximalgewichte: Komponente 2 (Kapitel 9.2)



Quelle: Eigene Darstellung.

Abbildung B5.: Balancierung: Komponente 2 (Kapitel 9.2)



Quelle: Eigene Darstellung.