Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute

Jürgen Schupp
Christof Wolf *Hrsg.*

Nonresponse Bias

Qualitätssicherung sozialwissenschaftlicher Umfragen

Springer VS

Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute

ASI

Jürgen Schupp • Christof Wolf (Hrsg.)

Nonresponse Bias

Qualitätssicherung sozialwissenschaftlicher Umfragen

Springer VS

Herausgeber
Jürgen Schupp
DIW Berlin, Deutschland

Christof Wolf
GESIS – Leibniz-Institut für Sozialwissenschaften
Mannheim, Deutschland

Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute
ISBN 978-3-658-10458-0 ISBN 978-3-658-10459-7 (eBook)
DOI 10.1007/978-3-658-10459-7

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer VS

Lektorat: Katrin Emmerich, Katharina Gonsior

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Fachmedien Wiesbaden ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Inhalt

Mixed Mode Design

Attrition und Nonresponse Bias in Panel-Studien

Vorwort

Die Ausschöpfungsraten von sozialwissenschaftlichen Erhebungen in Deutschland befinden sich im Vergleich zur Situation vor 25 Jahren auf einem – auch im internationalen Vergleich - niedrigen Niveau. Anspruchsvolle Umfragen mit genauer Überwachung des Feldes sowie kontrollierten Bruttostichproben erreichen derzeit selten eine höhere Ausschöpfung als etwa 35 %. Damit wird die Frage nach der Qualität dieser Umfragen immer drängender. Bilden solche Stichproben mit vergleichsweise geringer Beteiligungsquote noch die Lebenslagen, Einstellungen und Verhaltensweisen der jeweiligen Grundgesamtheit ab? Wie groß ist ihre Selektivität bzw. ihre Verzerrung, also der Nonresponse Bias? Wie geeignet sind Gewichtungs- und Hochrechnungsverfahren, um trotz geringer Ausschöpfungsquoten gleichwohl verallgemeinerbare Schlüsse auf die Grundgesamtheit ziehen zu können? Diese Fragen stehen im Mittelpunkt des Bandes, der sich in fünf Abschnitte gliedert.

Der Reader beginnt im ersten Abschnitt zu innovativen statistischen Verfahren mit einem Beitrag von Jan Eric Blumenstiel & Tobias Gummer, zu neueren statistischen Ansätzen um den Nonresponse Bias präventiv zu verhindern.

Der zweite Abschnitt widmet sich der Entwicklung des Teilnahmeverhaltens in zwei replikativen sozialwissenschaftlichen Querschnittstudien. Beide Surveys, der ESS und der ALLBUS, werden mit Hilfe von geschulten Interviewern und durch persönlich-mündliche computergestützte Verfahren (CAPI) durchgeführt. Dass die Ausschöpfung in sozialwissenschaftlichen Erhebungen zurückgegangen ist, ist unstrittig und betrifft auch diese beiden Erhebungsprogramme. Weitgehend wenig erforscht ist jedoch, ob dieser Rückgang mit einer Vergrößerung der Stichprobenverzerrung einhergeht. Michael Weinhardt & Stefan Liebig stellen die Ergebnisse der deutschen Stichprobe im European Social Survey (ESS) vor und beleuchten Teilnahmeverhalten und Stichprobenverzerrung. Michael Blohm & Achim Koch präsentieren die Ergebnisse einer experimentellen Studie im Rahmen des ALLBUS 2008 und beantworten die Frage, ob eine höhere Ausschöpfung zu anderen Umfrageergebnissen führt.

Im dritten Abschnitt werden die spezifischen Nonresponse-Herausforderungen bei telefonischen und postalischen Befragungen diskutiert. Matthias Sand stellt anhand einer Dual-Frame-Telefonstichprobe und devicespezifischem Nonresponse mögliche Gewichtungsstrategien vor. Nathalie Guzy präsentiert in ihrem Beitrag die Ergebnisse einer Nonresponseanalyse in einer telefonischen Opferbefragung. Karl-Heinz Reuband diskutiert die Probleme von Nonresponse für postalische Befragungen. Er stellt dabei Strategien wie Incentives, Fragebogenlänge und Anonymität zur Maximierung von Ausschöpfungsquoten anhand einer empirischen Studie vor.

Im vierten Abschnitt werden Erfahrungen aus Mixed-Mode-Designs präsentiert, also dem Einsatz von unterschiedlichen Erhebungs- und Befragungstechniken im Verlauf einer Studie. Uwe Engel & Björn Oliver Schmidt stellen die Ergebnisse des Einsatzes unterschiedlicher Verfahren während der Rekrutierungsphase, Kontaktierung sowie schließlich der Datenerhebung einer empirischen Studie vor und präsentieren zudem Ergebnisse unterschiedlicher Incentivierungsstrategien. Der Beitrag von Patrick Schmich stellt die methodischen Ergebnisse einer Projektstudie im Mixed Mode Design vor. Auch die Frage der Verknüpfung von Individualangaben aus Register-Daten mit Befragungsdaten aus Umfragen stellt eine besondere Herausforderung dar, da hierfür in Deutschland spezielle datenschutzrechtlich relevante Einverständniserklärungen eingeholt werden müssen. Der Beitrag von Arne Jonas Warnke geht der Frage zu Verzerrungen durch selektive Stichproben am Beispiel eines verknüpften Arbeitgeber-Arbeitnehmer Datensatzes mit Zugang zu administrativen Quellen nach.

Im fünften Abschnitt werden schließlich Erfahrungen aus Längsschnittstudien zu Ausfällen über die Zeit sowie die Gütekriterien bei Panelstudien vorgestellt. Der Beitrag von Britta Busse, Simon Laub & Marek Fuchs basiert auf Ergebnissen einer Panelbefragung im Mobilfunknetz und geht der Frage nach, ob Exit Questions ein adäquates Mittel darstellen können, einen Nonresponse Bias zu identifizieren. Der Aufsatz von Corinna Kleinert, Bernhard Christoph & Michael Ruland untersucht die Nonresponse Problematik in der Startkohorte 6 des Nationalen Bildungspanels (NEPS) und präsentiert Ergebnisse eines Experiments zu Auswirkungen der Administration von Kompetenztests bei Erwachsenen. Bettina Müller & Laura Castiglioni stellen die Entwicklung von Ausfällen aus mehreren Erhebungswellen des Beziehungs- und Familienpanels pairfam vor und diskutieren Hinweise auf systematische Ausfälle. Martin Kroh, Rainer Siegers & Simon Kühne prä-

sentieren im Rahmen der Haushaltlängsschnittstudie Sozio-oekonomisches Panel (SOEP) die Herausforderungen bei der Integration von Auffrischungsstichproben sowie der hierfür möglichen Gewichtungsverfahren. Der Reader schließt mit einem Beitrag von Thomas Glaser & Elisabeth Kafka, die einen Vorschlag zur Behebung von selektivem Bias in der österreichischen Stichprobe des europäischen Survey of Income and Living Conditions (EU-SILC Österreich) zur Diskussion stellen.

Insgesamt dokumentieren die Beiträge eindrucksvoll, dass die Qualitätssicherung sozialwissenschaftlicher Umfragen zu einem großen Thema gereift ist und der Stand der Forschung in diesem Bereich, der sich an internationalen Standards orientiert, einen zentralen Beitrag zur gesamten surveydatengestützten quantitativen empirischen Sozialforschung leistet. Die Ergebnisse dieses Readers vereinen die schriftlichen Ausarbeitungen einer gemeinsamen Tagung der Arbeitsgemeinschaft sozialwissenschaftlicher Institute (ASI e.V.) und der Sektion Methoden der Empirischen Sozialforschung in der Deutschen Gesellschaft für Soziologie (DGS), die im November 2013 in Berlin stattfand. Eingeladen waren Vertreterinnen und Vertreter einschlägiger sozialwissenschaftlicher Umfragen, um deren spezifische Erfahrungen zu präsentieren und zur Diskussion zu stellen. Der Band wurde zudem ergänzt durch einige wenige Beiträge von Autorinnen und Autoren, die nicht an der Tagung teilnehmen konnten.

Unser Dank an dieser Stelle geht an alle Autorinnen und Autoren für ihre Kooperationsbereitschaft bei diesem Buchprojekt. Ein besonderes Dankeschön richten wir an Frau Bettina Zacharias (GESIS) für die Integration der überarbeiteten Beiträge in den Satzspiegel des Verlags.

Berlin und Mannheim im März 2015

Jürgen Schupp & Christof Wolf

Innovative statistische Verfahren

Prävention, Korrektur oder beides?

Drei Wege zur Reduzierung von Nonresponse Bias mit Propensity Scores

Jan Eric Blumenstiel[1] *& Tobias Gummer*[2]

[1] Universität Mannheim
[2] GESIS – Leibniz-Institut für Sozialwissenschaften

1 Einleitung

Die Ausschöpfungsquote galt lange Zeit als wichtigster Indikator der Datenqualität in Umfragen, denn sie ist relativ einfach zu berechnen und zu vergleichen. Sie kann jedoch auch sehr leicht fehlinterpretiert werden, da eine hohe Ausschöpfung nicht immer mit einem niedrigen Nonresponse Bias einhergeht. Deshalb wurden in den letzten Jahren verschiedene Maßnahmen entwickelt, wie der Nonresponse Bias unabhängig von der Ausschöpfungsquote reduziert werden kann. In diesem Beitrag werden dazu drei Wege auf Grundlage von geschätzten Teilnahmewahrscheinlichkeiten („Propensity Scores") vorgestellt und anhand einer Simulationsstudie diskutiert.

In den letzten Jahren war ein Trend zurückgehender Ausschöpfungsquoten zu beobachten (de Leeuw und de Heer 2002; Wasmer et al. 2012). Beispielsweise sank die Ausschöpfungsquote im ALLBUS allein zwischen 2004 und 2010 um zehn Prozentpunkte von 45 auf 35 Prozent (konnte 2012 aber wieder leicht gesteigert werden). In der Umfrageforschung bestand die wesentliche Reaktion auf diesen Trend zunächst in der Entwicklung geeigneter Maßnahmen, um die Ausschöpfung wieder zu erhöhen, beispielsweise durch Mixed-Mode Designs und Incentives (z.B. Dillman et al. 2009; Singer et al. 2000), Ankündigungsschreiben (de Leeuw et al. 2007) oder einen

J. Schupp, C. Wolf (Hrsg.), *Nonresponse Bias,* Schriftenreihe der ASI – Arbeitsgemeinschaft Sozialwissenschaftlicher Institute, DOI 10.1007/978-3-658-10459-7_1,

umfassenden Maßnahmenkatalog (Laurie et al. 1999). In der Praxis zeigen viele der vorgeschlagenen Maßnahmen jedoch nur einen relativ geringen Effekt, verursachen z.T. aber erhebliche Mehrkosten, wie etwa vorausbezahlte Incentives.

Gleichsam rückt im Zuge der Diskussion um zurückgehende Ausschöpfungsquoten auch die Grundannahme wieder stärker in den Blick, wonach eine hohe Ausschöpfung mit einer geringen Verzerrung einherginge. Mehrere Studien haben gezeigt (vgl. z.B. die Meta-Analyse von Groves 2006; Groves und Peytcheva 2008; Schouten et al. 2009), dass diese Annahme so nicht zutrifft und die Ausschöpfung grundsätzlich kein geeignetes Maß für den Nonresponse Bias darstellt. Zwar besteht bei niedrigen Ausschöpfungsquoten ein höheres Potential für Verzerrungen, über das tatsächliche Ausmaß des Nonresponse Bias sagen diese aber wenig aus. Dies gilt umso mehr als gezeigt wurde, dass der Bias keinesfalls ein konstantes Merkmal einer bestimmten Umfrage ist, sondern eher als variablenspezifisches Konzept verstanden werden sollte (Groves 2006; Luiten und Schouten 2013; Rosen et al. 2014; Wagner 2012). Beullens und Lossveldt (2012) bezeichnen die Konzentration auf die Ausschöpfungsquote daher als Zielverschiebung, weil das eigentliche Ziel (niedriger Bias) durch ein einfacher zu verfolgendes Ziel (Ausschöpfung erhöhen) ersetzt wird, wodurch jedoch das zugrundeliegende Problem Gefahr läuft, aus den Augen verloren zu werden.

Die Herausforderung für die Umfrageforschung besteht in dieser Situation darin, dass einerseits mit Ausschöpfungsquoten deutlich unter 50 Prozent in face-to-face-Umfragen (McCarty et al. 2006; Wasmer et al. 2012) und z.T. unter 20 Prozent in telefonischen Umfragen (Partheymüller et al. 2013) ohne Zweifel ein nicht zu vernachlässigendes Potential für Nonresponse Bias besteht. Gleichzeitig kann dieser Bias aber weder direkt gemessen werden, noch wäre die „blinde" Maximierung der Ausschöpfung ein geeignetes Gegenmittel (Beullens und Loosveldt 2012; Stoop 2005). Anzustreben sind stattdessen differenziertere Forschungsdesigns, die dem Paradigmenwechsel von blinder Maximierung der Erfolgsquote hin zu informierteren Strategien zur Reduzierung des Bias Rechnung tragen. Zusätzlich unterliegen Designentscheidungen in zunehmendem Maße auch finanziellen Restriktionen. Ein Rückgang der Ausschöpfung von 45 auf 35 Prozent wie im ALLBUS zwischen 2004 und 2010 geschehen bedeutet, dass für 1000 Interviews über 600 Personen mehr kontaktiert werden müssen als zuvor. Dieser höhere Aufwand schlägt sich ebenso in höheren Fallpreisen nie-

der wie die zunehmende Marktkonzentration bundesweit leistungsfähiger CAPI-Felder infolge der abnehmenden Bedeutung persönlicher Befragungen für die meisten kommerziellen Auftraggeber (ADM 2013).

Vor diesem Hintergrund sollen im Folgenden drei mögliche Verfahren zur Reduzierung des Nonresponse Bias auf Basis von Propensity Scores vorgestellt und anhand einer Simulationsstudie diskutiert werden. Als Ausgangsüberlegung kann Nonresponse Bias dazu nach Bethlehem (2002) approximativ verstanden werden als Verhältnis der Kovarianz einer Variable y und der Teilnahmewahrscheinlichkeit p zu der mittleren Teilnahmewahrscheinlichkeit in der Stichprobe:

$$bias_{\bar{y}} \sim \frac{\sigma_{y,p}}{\bar{p}}$$

Wie beschrieben bestand das Ziel traditionellerweise darin, die mittlere Teilnahmewahrscheinlichkeit und somit den Nenner dieses Bruchs zu erhöhen. Diese Strategie kann jedoch ineffektiv sein oder den Bias sogar vergrößern, denn es besteht die Gefahr, dass die Erhöhung der Ausschöpfung nur zu „more of the same" (Peytchev et al. 2009) führt. In diesem Fall werden zusätzliche Befragte erreicht, die sehr ähnliche Eigenschaften aufweisen wie die bereits zuvor interviewten Personen. Strategien auf Basis von Propensity Scores zielen deshalb explizit darauf ab, die Varianz dieser Teilnahmewahrscheinlichkeiten zu reduzieren. Gelingt dies, so verringert sich die Kovarianz im Zähler der Formel und damit auch der Nonresponse Bias. Im folgenden Abschnitt werden dazu drei Varianten vorgestellt, wie dieser Schritt gelingen kann. Anschließend werden das Vorgehen und die Datengrundlage der Simulationsstudie erläutert, deren Ergebnisse dann in Abschnitt fünf diskutiert werden.

2 Verfahren zur Reduzierung des Nonresponse Bias auf Basis von Propensity Scores

Im Zusammenhang mit der Abkehr von der Konzentration auf Ausschöpfungsquoten wurden unter den Begriffen „responsive“ und „adaptive“ Design flexiblere und komplexere Vorgehensweisen für die Durchführung von Umfragen vorgeschlagen (z.B. Groves und Heeringa 2006; Couper und Wagner 2011; Wagner et al. 2012). Diese Flexibilität kann sich beispielsweise in der unterschiedlichen Gestaltung bestimmter Designelemente einer Umfrage wie dem Incentive für vorher definierte Subgruppen oder aber der Anpassung bestimmter Designelemente während der Feldphase in Abhängigkeit der Entwicklung vorher definierter Indikatoren äußern. Um Befragte oder Gruppen von Befragten zu identifizieren, die gesondert behandelt werden sollen, können Propensity Scores bzw. Teilnahmewahrscheinlichkeiten geschätzt werden. Je nachdem, ob die geschätzten Wahrscheinlichkeiten korrektiv, präventiv oder beides in Kombination eingesetzt werden, um den Nonresponse Bias zu reduzieren, unterscheiden wir zwischen drei verschiedenen Verfahren.

2.1 Propensity Score Gewichtung

Die nachträgliche, korrektive Verwendung von Propensity Scores geht auf Rosenbaum und Rubin (1983) zurück. Ursprüngliches Einsatzgebiet waren bei diesen Autoren Beobachtungsstudien, bei denen eine Randomisierung des Treatments unmöglich oder ethisch unangebracht ist, sodass sich die Vergleichsgruppe auch in weiteren Eigenschaften systematisch von der Treatment-Gruppe unterscheiden kann. Um diesem Problem zu begegnen, wird die Wahrscheinlichkeit des Treatments zunächst durch geeignete Indikatoren geschätzt. Im zweiten Schritt werden Personen mit und ohne Treatment mit ähnlicher Wahrscheinlichkeit gematcht. Die Analyse erfolgt dann auf Grundlage der auf diese Weise zusammengespielten Stichprobe, sodass sich Personen bei gegebener Propensity nur noch durch das Auftreten des Treatments unterscheiden.

Die gleiche Logik kann auch dazu verwendet werden, um den Coverage-, Selection- oder Nonresponse-Bias in Umfragen durch Gewichtung zu reduzieren (vgl. Bergmann 2011; Bremer et al. 2000; Lee 2006; Little 1984; Schonlau et al. 2009). Um beispielsweise die Verzerrungen durch Non-Coverage und Self-Selection zu reduzieren, die mit nicht-zufälligen Online-

Erhebungen verbunden sind, kann eine auf diese Weise erhobene Stichprobe mithilfe von Propensity Scores an eine repräsentative Referenzstudie angeglichen werden (vgl. zum Vorgehen Bergmann 2011). Dazu werden zunächst beide Stichproben zusammengespielt und anschließend die Wahrscheinlichkeit der Teilnahme an der Referenzstudie mittels Variablen vorhergesagt, die in beiden Befragungen erhoben wurden. Anschließend wird die Verteilung der (üblicherweise klassierten) Propensities durch Gewichtung zwischen beiden Stichproben angeglichen. Das bedeutet insbesondere, dass Befragte der Online-Erhebung mit einer hohen vorhergesagten Teilnahmewahrscheinlichkeit an der Referenzstudie hochgewichtet werden, während Online-Befragte mit einer geringen Teilnahmewahrscheinlichkeit an der Referenzstudie heruntergewichtet werden.

Das Verfahren kann ebenfalls zur Korrektur des Nonresponse Bias angewendet werden. Die abhängige Variable des nach Abschluss der Befragung berechneten logistischen Vorhersagemodells bildet dann die Teilnahme an einer Befragung in einer Querschnittsstudie oder an einer Befragungswelle in einer Panelstudie. Anschließend werden die Befragten mit einer niedrigen vorhergesagten Teilnahmewahrscheinlichkeit hoch-, diejenigen mit einer hohen geschätzten Wahrscheinlichkeit heruntergewichtet. Auf diese Weise soll die Varianz der Teilnahmewahrscheinlichkeiten im Nachhinein reduziert und somit der Zähler der oben präsentierten Formel zur Berechnung des Nonresponse Bias verkleinert werden.

Voraussetzung für die Anwendung der Propensity Score Gewichtung ist folglich, dass die Variablen zur Vorhersage der Teilnahmewahrscheinlichkeit für Respondenten und Nicht-Respondenten vorliegen. Zudem sollen die verwendeten Variablen nicht nur prädiktiv für die Teilnahme sein, sondern gleichzeitig mit den inhaltlich interessierenden Variablen korreliert sein (Bergmann 2011; Kreuter und Olson 2011). Diese Voraussetzungen können vor allem in Querschnittsstudien nicht erfüllt sein, bei denen oft nichts oder nur sehr wenige demographische Eigenschaften der Nicht-Teilnehmer bekannt sind. Im Kontext einer Panelerhebung liegen für die Nicht-Respondenten späterer Panelwellen jedoch die Angaben aus der ersten Befragung vor, sodass sich das Verfahren theoretisch besonders dazu eignet, den durch Panel Attrition entstandenen Nonresponse Bias zu reduzieren. Wie gut dies tatsächlich gelingt, hängt entscheidend davon ab, wie gut der Ausfallprozess der Panelteilnahme durch das Vorhersagemodell erklärt werden kann und wie die interessierenden Variablen mit den eingesetzten Variablen kor-

reliert sind. Im Folgenden wird dies anhand einer Simulationsstudie überprüft, die auf Daten aus einer realen Panelerhebung basiert.

2.2 Case Prioritization

Die Schätzung der Teilnahmewahrscheinlichkeiten kann jedoch auch vor dem Feldbeginn erfolgen, mit dem Ziel eines präventiven Eingriffs in die Datenerhebung, um die Teilnahmewahrscheinlichkeit von Personen mit niedriger vorhergesagter Propensity zu erhöhen und somit die Varianz der Propensities zu verringern.

Ein entsprechendes Verfahren wurde von Peytchev et al. (2010) unter dem Begriff „Case Prioritization" vorgeschlagen. Diese Methode umfasst zwei Schritte. Zunächst wird vor Beginn der Datenerhebung für jede Zielperson die Teilnahmewahrscheinlichkeit geschätzt. In einer Querschnittsstudie kann sich wieder das Problem stellen, dass nur wenige Informationen über alle Zielpersonen bekannt sind, beispielsweise aus dem Sampling Frame. Im Kontext einer Panelstudie können dagegen Informationen aus der vorherigen Welle herangezogen werden, wobei neben inhaltlichen und demographischen Variablen auch Paradaten oder Interviewereinschätzungen infrage kommen (vgl. Blumenstiel 2013). Als abhängige Variable kann beispielsweise die Antwort auf die Frage nach der Wiederbefragungsbereitschaft in der folgenden Welle verwendet werden, die in manchen Panelstudien am Ende der ersten Befragungswelle gestellt wird – unter der Annahme, dass einer Verweigerung unmittelbar im Anschluss an die erste Welle ähnliche Faktoren zugrunde liegen wie einer späteren Verweigerung in der Folgewelle.

Im zweiten Schritt wird ein Erhebungsdesign entworfen, das eine Intervention für Befragte mit niedriger geschätzter Teilnahmewahrscheinlichkeit vorsieht. Ist diese erfolgreich, so wird die Ausschöpfung bei Befragten mit niedriger geschätzter Propensity erhöht, wodurch sich die die Varianz der Propensities in der Stichprobe und gleichsam die Kovarianz zwischen inhaltlichen Variablen und Propensities reduzieren. Auf diese Weise wird eine Reduzierung des Nonresponse Bias erreicht. Die Intervention kann sowohl auf Befragten- als auch auf Interviewerebene erfolgen, geeignete Maßnahmen könnten u.a. höhere Incentives, zusätzliche Kontakte oder eine Verkürzung des Fragebogens umfassen.

Darüber hinaus nennen Peytchev et al. (2010) zwei weitere mögliche positive Effekte dieses Verfahrens. Erstens kann Case Prioritization die Varianz von nachträglich berechneten Anpassungsgewichten reduzieren (die dem Kehrwert der geschätzten Teilnahmewahrscheinlichkeit entsprechen) und damit die effektive Stichprobengröße erhöhen. Zweitens kann das Verfahren die Repräsentativität einer Panelstichprobe gemessen am sog. R-Indikator erhöhen (Schouten et al. 2009), der ebenfalls auf der Varianz der Teilnahmewahrscheinlichkeit basiert und damit einer sehr ähnlichen Logik folgt.

Dem Verfahren der Case Prioritization liegt implizit die Annahme zugrunde, dass die Teilnahmewahrscheinlichkeiten durch einen Eingriff in die Erhebung verändert werden können. Auf Befragten-Ebene ist dies ohne weiteres vorstellbar, da etwa Veränderungen in der Höhe der Incentives oder der Länge der Befragung das Verhältnis aus Kosten und Nutzen für eine Zielperson entscheidend verändern können. Auf Interviewer-Ebene wäre ein Eingriff allerdings zwecklos, wenn ohnehin jeder Zielperson der maximal mögliche Aufwand geschenkt würde. In der Praxis ist dies für Interviewer in face-to-face-Umfragen jedoch unmöglich. Werden Interviewer – wie in Deutschland üblich – auf Fallbasis bezahlt und auf Grundlage ihrer erreichten Ausschöpfung evaluiert (s. West und Groves 2013 für ein besseres Evaluationsverfahren, bei dem ebenfalls Propensity Scores zum Einsatz kommen), so ist die Maximierung der Ausschöpfungsquote aus ihrer Sicht die rationale Strategie. Um dies zu erreichen, konzentrieren sich Interviewer üblicherweise auf die aus ihrer Sicht am leichtesten zu erreichenden Fälle (Stoop 2005), also diejenigen mit hoher Teilnahmewahrscheinlichkeit. Berücksichtigt man die jüngsten Erkenntnisse über den Zusammenhang zwischen Ausschöpfung und Stichprobenverzerrung, so läuft diese Strategie jedoch Gefahr „more of the same“ zu produzieren, also die Ausschöpfung zu maximieren ohne den Bias zu reduzieren (Peytchev et al. 2009).

Bisherige Anwendung des Case Prioritization-Verfahrens haben gezeigt, dass die Teilnahmewahrscheinlichkeiten im Vorfeld der Befragung relativ präzise geschätzt werden können, die Durchführung einer erfolgreichen Intervention jedoch nicht immer gelingt (Peytchev et al. 2010). Allerdings gibt es auch Beispiele von Studien, in denen die Ausschöpfung bei Zielpersonen mit niedriger geschätzter Teilnahmewahrscheinlichkeit erfolgreich erhöht werden konnte (Blumenstiel 2013; Rosen et al. 2014).

2.3 Kombination beider Verfahren

Wenngleich beide beschriebene Verfahren in der Literatur bisher nur separat diskutiert wurden, so scheint es für die Praxis naheliegend, Prävention und Korrektur zu kombinieren. Was die Propensity Score Gewichtung betrifft, so konnte gezeigt werden, dass diese erfolgreich zur Reduzierung des Bias beitragen kann (z.B. Lee 2006; Bergmann 2011). Allerdings kann die nachträgliche Gewichtung die Verzerrung einer Stichprobe nicht vollständig reduzieren und funktioniert überdies nicht in allen Fällen, wobei normalerweise unbekannt ist ob und wie gut die Verzerrung einer bestimmten Variablen reduziert werden konnte. Zudem bringt dieses Verfahren unter Umständen weitere Nachteile mit sich, auf die in der Diskussion der Ergebnisse näher eingegangen wird.

Das Verfahren der Case Prioritization konnte zwar bereits erfolgreich angewendet werden, auch in diesem Fall ist aber eine vollständige Reduzierung des Bias unrealistisch. Selbst wenn die Ausschöpfung bei Zielpersonen mit niedriger Teilnahmewahrscheinlichkeit gesteigert werden kann, so scheint es nach bisherigen Erkenntnissen kaum möglich, diese soweit zu steigern, dass sie auf dem Niveau derjenigen mit hoher Teilnahmewahrscheinlichkeit liegt.

Die zusätzliche Anwendung der korrektiven Propensity Score Gewichtung nach bereits erfolgter präventiver Case Prioritization verursacht keinen erheblichen Mehraufwand, weder in finanzieller noch in zeitlicher Hinsicht. Theoretisch scheint es deshalb sinnvoll, beide Verfahren zu kombinieren, also zunächst durch Case Prioritization präventiv auf die Varianz der Propensities einzuwirken und nach Abschluss der Befragung Gewichte auf Grundlage der – dann erneut geschätzten – Propensity Scores zu berechnen. Bisher ist jedoch unbekannt, ob eine solche Kombination tatsächlich zu einer stärkeren Reduktion der Verzerrung beitragen kann. Dies soll im Folgenden anhand einer Simulationsstudie überprüft werden.

3 Simulationsaufbau und Datengrundlage

Mit Propensity Score Gewichtung, Case Prioritization und der Kombination beider Verfahren wurden drei mögliche Wege skizziert, um dem Nonresponse Bias zu begegnen. Zum Test der Effekte der jeweiligen Verfahren liegt als Untersuchungsdesign – besonders für Case Prioritization – zunächst ein Experiment nahe. Um den Effekt der verschiedenen Methoden auf Nonresponse zu messen, müssten allerdings mehrere Herausforderungen überwunden werden:

Erstens müssen genügend Ressourcen vorhanden sein, um das Experiment mit einer ausreichend großen Fallzahl durchzuführen. Der Hinzugewinn zusätzlicher eigentlich schlecht erreichbarer Fälle stellt den wesentlichen Vorteil der Case Prioritization-Methode dar. Dies hat, neben der Homogenisierung der Teilnahmewahrscheinlichkeit verschiedener Gruppen, den Effekt, dass die Schätzungen des Datensatzes durch die erhöhte Fallzahl präziser werden. Für nennenswerte Verbesserungen setzt das eine entsprechend große Experimentalgruppe voraus. Je nach gewählten Interventionsmaßnahmen kann dies auch eine erhebliche Belastung des finanziellen Budgets bedeuten.

Zweitens hängt der Erfolg des Experiments entscheidend davon ab, ob die Priorisierung von Fällen mit einer niedrigen Teilnahmewahrscheinlichkeit auch zu einer Erhöhung der Ausschöpfung in dieser Gruppe führt, andernfalls könnte der Nonresponse Bias nicht reduziert werden. Die bisherigen Studien zu Case Prioritization berichten hinsichtlich des Erfolgs der Intervention jedoch unterschiedliche Befunde (Blumenstiel 2013; Luiten und Schouten 2013; Peytchev et al. 2010; Rosen et al. 2014).

Drittens erschwert der Vergleich verschiedener Methoden die Konstruktion eines Experiments noch weiter. In einem solchen Fall müssten neben vier Untersuchungsgruppen für die Folgen von Case Prioritization noch weitere Untersuchungsbedingungen zur Bestimmung der Effekte der Gewichtungsverfahren sowie der Kombination von Gewichtung und Priorisierung berücksichtigt werden.

Auf Grund dieser Herausforderungen werden im Folgenden mittels einer realdatengestützen Simulation die Effekte einer erfolgreichen Case Prioritization, Propensity Score Gewichtung und der Kombination beider Verfahren getestet.

Eine solche Simulationsstudie bietet mehrere Vorteile. In einem Experiment ist der Effekt der Case Prioritization-Methode abhängig von der erreichten Erhöhung der Ausschöpfung in der Gruppe der Befragten mit geringer Teilnahmewahrscheinlichkeit. Hat die Priorisierung keinen Erfolg, kann nichts über den Effekt der Methode an sich gesagt werden. Wenn eine gewisse Erhöhung der Ausschöpfung gelingt, aber keine Reduzierung des Bias, bleibt unklar, ob dies gegen das Verfahren spricht oder ob bei einem höheren Interventionserfolg die Verzerrung hätte reduziert werden können. In einer Simulation kann dies getestet werden, indem Annahmen über die Erhöhung der Ausschöpfung in der Gruppe der Befragten mit niedriger Propensity getroffen und simuliert werden. Anschließend kann dann untersucht werden, ob eine möglicherweise zur Reduzierung des Bias notwendige Erhöhung in der Praxis realistisch wäre.

Weiterhin müssen im hier vorgeschlagenen Verfahren keine Experimentalgruppen gebildet werden, sowohl die einzelnen Verfahren als auch die Kombination beider Verfahren kann mit der gesamten Stichprobe getestet werden. Da die Simulation auf Realdaten basiert, ergibt sich der Vorteil, dass den Daten ein echtes Ausfallmuster zu Grunde liegt. Bei einem Experiment müsste der Ausfall dagegen aktiv manipuliert werden. Schlussendlich kann die Datenbasis in einer Simulationsstudie frei gewählt werden. Es können also Daten eingesetzt werden, welche die Voraussetzungen hinsichtlich vorhandenem Unit Nonresponse und dem Bedarf an korrektiven oder präventiven Methoden bestmöglich erfüllen.

Als Datengrundlage für die folgende Simulationsstudie dient das 1998 im Rahmen des DFG-Projekts *Politische Einstellungen, politische Partizipation und Wählerverhalten im vereinigten Deutschland* gestartete dreiwellige Langfrist-Panel (ZA4662)[1]. Ausgehend von einer zur Bundestagswahl 1998 im Adress-Random Verfahren erhobenen Querschnittsstudie wurden teilnahmebereite Respondenten zu zwei weiteren Wahlen befragt (2002, 2005). Die gesamte Befragung war als face-to-face Studie angelegt, auf Grund der verkürzten Vorlaufzeit durch die vorgezogene Bundestagswahl 2005 musste die dritte Welle (2005) jedoch telefonisch durchgeführt werden. Die Analysen konzentrieren sich auf den Übergang der Jahre 1998 auf 2002, da hier

1 Das Projekt wurde von den Primärforschern Jürgen W. Falter, Oscar W. Gabriel und Hans Rattinger durchgeführt. Die 1998 begonnene Panelstudie wurde von Hans Rattinger nach Ende des Projekts in 2002 bis 2005 fortgeführt.

zum einen genug Fälle für die notwendigen Auswertungen vorliegen und zum anderen, um trunkierte Fälle zu vermeiden. Das sind Befragte, die 1998 und 2005 teilnahmen und nur 2002 nicht erreicht wurden. Gleichzeitig bedeutet diese Auswahl eine Beschränkung auf die als face-to-face Interviews durchgeführten Wellen, eine mögliche modusbedingte Verzerrung der Ergebnisse wird damit vermieden. Da zu Zwecken der Evaluation der drei Verfahren nicht anzunehmen ist, dass ihre Wirkung vom betrachteten Zeitraum abhängig ist, gestalten sich die Analysen durch diese Auswahl besser nachvollziehbar.

In der ersten Welle des Panels (1998) konnten 3.337 Personen befragt werden, davon gaben 2.629 (78,8%) an für weitere Befragungen teilnahmebereit zu sein. Auf Basis dieses Adressbestandes konnten im Jahr 2002 noch 1.744 Interviews realisiert werden, was einer Ausschöpfung von 66,3% der wiederbefragungsbereiten Personen entspricht, bzw. 52,3% aller Querschnittsfälle. Um zu evaluieren wie gut die vorgestellten Korrektur- und Präventionsverfahren funktionieren, ist ein gewisses Maß an Nonresponse notwendig, andernfalls wäre der mögliche Effekt der Korrekturverfahren durch einen Ceiling-Effekt stark begrenzt und negative Befunde wären vermutlich eher auf das Fehlen einer Verzerrung denn auf ein Versagen der Methoden zurückzuführen (Peytchev et al. 2010). Der Übergang des Panels in den Jahren 1998 auf 2002 bietet dies zumindest rein quantitativ. Gleichsam sollte der Ausfall systematisch, also durch ein geeignetes Modell zu erklären sein, ansonsten stünden die geschätzten Propensities und der Bias nicht im Zusammenhang und die Korrekturverfahren blieben wirkungslos.

Weiterhin erlaubt die Verwendung einer Panelstudie bei der Bestimmung der Propensity Scores auf eine breite Auswahl an geeigneten Variablen zurückzugreifen, welche in der ersten Welle befragt wurden. Dies ermöglicht ein umfangreiches Schätzmodell zum Teilnahmeverhalten aufzustellen, was dem Idealfall der Anwendung entspricht.

Um den Erfolg der drei Verfahren beim Ausgleich der Verzerrung zu bestimmen, ist ein Referenzdatensatz notwendig. Da für Analysen inhaltliche Variablen in aller Regel mindestens ebenso relevant sind wie soziodemographische Variablen, dient die im selben Projekt zur Bundestagswahl 2002 durchgeführte Querschnittsstudie (ZA3861) als Referenzdatensatz. Wie das Panel wurde die Studie als face-to-face Befragung erhoben und weist in weiten Teilen einen ähnlichen Fragebogen auf. Es ist daher anzunehmen, dass bei einem Vergleich der beiden Datensätze keine Verzerrun-

gen durch abweichende Umfragemethodik auftreten. Weiterhin erlaubt die Wahl des Datensatzes eine Vielzahl inhaltlich relevanter Merkmale zu vergleichen und die Analyse nicht nur auf einige wenige Variablen limitieren zu müssen. Denn gerade bei großen mehrthematischen Umfrageprojekten stehen nicht einzelne Merkmale im Zentrum der Studie, sondern eine große Zahl verschiedener Konstrukte, um Mehrwerte für die wissenschaftliche Gemeinschaft zu produzieren. Je nach Fragestellung sind unterschiedliche Merkmale für die Forschenden von Interesse. Aus diesem Grund werden der Nonresponse Bias und die Folgen von Korrektur- und Präventionsverfahren im Folgenden anhand einer ganzen Reihe an Variablen und nicht lediglich einer kleinen Auswahl evaluiert.

Der durch Panel Attrition erzeugte Nonresponse Bias ergibt sich durch einen Merkmalsvergleich zwischen der zweiten Welle des Panels (2002) und dem Referenz-Querschnitt aus dem selben Jahr. Die meisten Studien zu Unit Nonresponse (querschnittlich sowie im Panel) beschränken sich dabei auf eine kleine Auswahl an zentralen Variablen (z.B. Koch 1998; Loosveldt und Sonck 2008; Vandecasteele und Debels 2007), die sich meist größtenteils aus sozio-demographischen Variablen zusammensetzt.[2]

Soll eine größere Anzahl an Merkmalen verglichen werden, ist es problematisch jede einzelne Verteilung zwischen Datensatz und Referenz gegenüberzustellen. Das schränkt die Übersichtlichkeit und Nachvollziehbarkeit der Ergebnisse erheblich ein. Im Folgenden wird das Problem gelöst, indem für jedes zu vergleichende Merkmal ein Dissimilaritätsindex berechnet wird (Duncan und Duncan 1955, S. 211):

$$D = \frac{1}{2} \sum_{1}^{i} \left| x_i^p - x_i^r \right|$$

Dabei sei x_i^p der relative Anteil der Befragten der i-ten Kategorie des Merkmals X im Panel. x_i^r ist dagegen der relative Anteil der Befragten der i-ten

2 Von der hier genannten Literatur weichen Loosveldt und Sonck (2008) von der etwas einseitigen Auswahl ab und präsentieren auch Verteilungen einiger inhaltlicher Variablen. Die Tendenz Nonresponse Bias eher für sozio-demographische Merkmale festzustellen liegt in der Verfügbarkeit geeigneter Referenzen begründet. Durch amtliche Daten sind vergleichbare, verlässliche und frei zugängliche Referenzverteilungen verfügbar, allerdings beschränkt auf sozio-demographische Merkmale.

Kategorie des Merkmals *X* im Referenz-Querschnitt. Aus den summierten Beträgen der Differenzen ergibt sich, wie stark die Verteilungen der beiden Datensätze voneinander abweichen, unabhängig von der Richtung der Abweichung. Der Dissimilaritätsindex beschreibt, um wie viele Prozentpunkte die Merkmalsverteilung im Panel umverteilt werden müsste, um der Referenzverteilung zu entsprechen. Da mit Ausnahme des Nonresponse kein Unterschied zwischen den hier verwendeten Datensätzen (Panel und Querschnitt) bestehen sollte, können die berechneten Indizes als merkmalsspezifische Indikatoren für den Nonresponse Bias verstanden werden.

Zur Evaluation der drei Verfahren zur Reduktion der Verzerrung in Panels kommen drei verschiedene Strategien zum Einsatz:

Um die Folgen der Propensity Score Gewichtung (d.h. der Korrektur) zu testen, werden zunächst die Gewichte für die Befragten der zweiten Welle des Panels berechnet. Danach werden die merkmalsspezifischen Dissimilaritätsindizes mit den gewichteten Verteilungen des Panels bestimmt und mit den Indizes verglichen, welche auf ungewichteten Verteilungen basieren.

Die Berechnung der Gewichte entspricht dem in Abschnitt 2 dargelegten Vorgehen.[3] Zunächst wird dazu ein Schätzmodell für die Teilnahme an der zweiten Welle der Befragung aufgestellt. Dieses beinhaltet neben sozio-demographischen auch inhaltliche Variablen (Tabelle 1). Das verwendete Erklärungsmodell richtet sich in erster Linie nach der Arbeit von Watson und Wooden (2009), welche Determinanten longitudinalen Teilnahmeverhaltens in einen Antwort- oder Befragungsprozess mit mehreren Phasen einteilen. Dabei orientieren sich die Autoren an der Konzeption von Lepkowski und Couper (2002). Aus diesen Arbeiten und der Konzeption von Groves et al. (2004) ergibt sich das hier vorgestellte Modell, welches in ähnlicher Form auch in der German Longitudinal Election Study (GLES) zur Modellierung von Panelattrition verwendet wird (vgl. Blumenberg und Gummer 2013; Blumenstiel und Gummer 2013).

Als abhängige Variable für das Erklärungsmodell dient die tatsächliche Teilnahme in der zweiten Panelwelle (2002). Diese ist entsprechend als dichotome abhängige Variable codiert.

---

3 Dabei erfolgt keine weitere Stratifizierung der Teilnahmewahrscheinlichkeiten, da die Folgen einer reinen Propensity Score Gewichtung evaluiert werden sollen und nicht noch weitere Anpassungsmechanismen.

Um die Propensity Scores der Befragten zu bestimmten, wird im nächsten Schritt eine logistische Regression geschätzt. Aus den Ergebnissen der logistischen Regression (Tabelle 1) kann für jeden Fall die individuelle Teilnahmewahrscheinlichkeit berechnet werden. Diese dient invertiert als Propensity Score Gewicht (vgl. Horvitz und Thompson 1952).

Tabelle 1: Logistische Regressionen zur Teilnahme und Teilnahmebereitschaft an der zweiten Panelwelle (Logit-Koeffizienten).

	Teilnahme	reduziertes Modell	Teilnahme-bereitschaft	reduziertes Modell
	Koeff. / (S.E.)	Koeff. / (S.E.)	Koeff. / (S.E.)	Koeff. / (S.E.)
Geschlecht: Frau	0,098	0,062	-0,078	-0,126
	(0,079)	(0,075)	(0,096)	(0,091)
Alter	0,083***	0,082***	0,038*	0,036*
	(0,014)	(0,013)	(0,016)	(0,015)
Alter^2	-0,001***	-0,001***	-0,000**	-0,000*
	(0,000)	(0,000)	(0,000)	(0,000)
Bildung: mittel	0,119	0,166[a]	-0,051	-0,029
	(0,093)	(0,091)	(0,110)	(0,107)
Bildung: hoch	0,204[a]	0,254*	0,383**	0,445***
	(0,106)	(0,102)	(0,140)	(0,135)
Region: Ostdeutschland	-0,329***	-0,289***	-0,308**	-0,285**
	(0,085)	(0,079)	(0,104)	(0,097)
Beschäftigung: Hausfrau-/mann	-0,048		0,073	
	(0,145)		(0,171)	
Beschäftigung: Rentner	0,080		0,527***	
	(0,132)		(0,158)	
Familienstand: verheiratet	0,157[a]	0,176[a]	-0,131	-0,117
	(0,095)	(0,093)	(0,117)	(0,115)
Haushaltsgröße	0,112**	0,114**	0,189***	0,186***
	(0,038)	(0,038)	(0,050)	(0,049)
Wahlbeteiligungs-absicht	0,244*		0,444***	
	(0,105)		(0,116)	
Parteiverdrossenheit	-0,064		0,041	
	(0,093)		(0,117)	
Kanzlerwahl: unentschieden	-0,137		-0,117	
	(0,097)		(0,113)	

	Teilnahme	reduziertes Modell	Teilnahme-bereitschaft	reduziertes Modell
	Koeff. / (S.E.)	Koeff. / (S.E.)	Koeff. / (S.E.)	Koeff. / (S.E.)
Wissen: Anzahl Bundesländer	0,111		0,091	
	(0,076)		(0,094)	
Polit. Interesse: mittel	0,331***	0,394***	0,398***	0,469***
	(0,092)	(0,089)	(0,104)	(0,101)
Polit. Interesse: hoch	0,705***	0,779***	0,736***	0,847***
	(0,111)	(0,105)	(0,138)	(0,129)
Item Nonresponse	0,011	-0,025	-0,052	-0,130***
	(0,033)	(0,028)	(0,035)	(0,029)
Unzufriedenheit: Demokratie	-0,055		-0,022	
	(0,041)		(0,050)	
Polit. Wissen: Zweitstimme	-0,097		0,067	
	(0,078)		(0,095)	
Ortsgröße	-0,396***	-0,380***	0,304**	0,302**
	(0,078)	(0,077)	(0,093)	(0,092)
Dauer der Befragung	0,000[a]		0,000	
	(0,000)		(0,000)	
Efficacy: internal	-0,217*		-0,077	
	(0,084)		(0,099)	
Efficacy: external	0,079		-0,022	
	(0,075)		(0,091)	
Konstante	-2,113***	-2,198***	-0,411	-0,230
	(0,372)	(0,333)	(0,439)	(0,391)
N	3313	3334	3313	3334
Pseudo R^2	0,060	0,054	0,059	0,049

[a] p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Bedingt durch Item Nonresponse kann nicht für jeden Fall ein Propensity Score geschätzt werden. In diesem Fall werden reduzierte Modelle verwendet, welche hauptsächlich sozio-demographische Variablen ohne nennenswerte Anteile fehlender Werte einschließen (Spalte 3 in Tabelle 1). Würden Fälle mit Item Nonresponse aus den Analysen ausgeschlossen, bspw. durch listwise deletion, könnte das zu einer Verzerrung der Ergebnisse führen, da Item Nonresponse in die Analyse als Unit Nonresponse eingehen würde: Die

betroffenen Fälle würden nicht bei der Berechnung der Merkmalsverteilungen berücksichtigt. Das gewählte Vorgehen erlaubt Item und Unit Nonresponse zu trennen und mit den Dissimilaritätsindizes den Nonresponse Bias durch Attrition zu beschreiben.

Die nach der beschriebenen Methode bestimmten Gewichte werden zur Berechnung der merkmalsspezifischen Verteilungen verwendet. Auf dieser Basis können wiederum die Dissimilaritätsindizes berechnet werden. Als Vergleich dienen Indizes auf Basis der ungewichteten Verteilungen. Die Differenz zwischen gewichteten und ungewichteten Indizes entspricht dem Korrektureffekt der Gewichtung.

Um die Folgen der Case Prioritization evaluieren zu können, werden erhöhte Ausschöpfungsquoten für Fälle mit niedrigen Teilnahmewahrscheinlichkeiten simuliert. Diese werden in einem ersten Schritt mittels logistischer Regressionen vorhergesagt. Als Erklärung der Teilnahme dient auch hier das in Tabelle 1 spezifizierte Modell. Im Gegensatz zur Berechnung der Panelgewichte ist allerdings nicht die tatsächliche Teilnahme an Welle 2 die abhängige Variable (denn diese wäre realiter vor dem Feldstart der zweiten Welle unbekannt), sondern die in der ersten Welle vom Befragten geäußerte Wiederbefragungsbereitschaft. Dies entspricht der Information, welche im realen Anwendungsfall vorliegt und für die Durchführung der Priorisierung verwendet wird. Um Probleme durch Item Nonresponse auszuschließen, wird auch hier das reduzierte Modell geschätzt (Spalte 5 in Tabelle 1), wenn andernfalls keine individuelle Teilnahmewahrscheinlichkeit zu ermitteln ist.

Die geschätzten individuellen Teilnahmewahrscheinlichkeiten auf Basis der im Jahr 1998 berichteten Teilnahmebereitschaft werden im nächsten Schritt in niedrige und hohe Wahrscheinlichkeiten unterteilt. Die Gruppe der niedrigen Wahrscheinlichkeiten (d.h. die niedrige Propensity Gruppe) umfasst die 30% aller Fälle mit den niedrigsten Teilnahmewahrscheinlichkeiten. Für die Bestimmung der Zielgruppe der Priorisierung über Perzentile gegenüber einer Bestimmung über vordefinierte Wahrscheinlichkeitsschwellen spricht der Realitätsbezug: Je nach den Randbedingungen und methodischen Charakteristika einer Studie sollte die Wahrscheinlichkeitsverteilung variieren. Das kann dazu führen, dass ein vorgegebener Absolutwert in einem Fall für 30 Prozent der Personen zutrifft und in einem anderen Fall nur für 10 Prozent. Da es allerdings im Interesse des Forschenden sein sollte, möglichst viele Fälle zu erreichen, würde man versuchen so

viele kritische Fälle durch Priorisierung zur Teilnahme zu bewegen wie angesichts der finanziellen Ausstattung der Studie möglich. Eine Auswahl einer vorab festgelegten Anzahl an Fällen ist daher die realistische Basis für unsere Simulation. Andererseits erscheint eine zu starke Aufweichung des Kriteriums für niedrige Wahrscheinlichkeiten wie etwa der bei Peytchev et al. (2010) eingesetzte Split am Median unzweckmäßig, da dann die Gefahr besteht, dass sich die beiden Gruppen nicht ausreichend unterscheiden.

Für Befragte mit niedriger Teilnahmewahrscheinlichkeit werden im zweiten Schritt erhöhte Ausschöpfungsquoten simuliert, d.h. eine erfolgreiche Priorisierung der Fälle, indem Häufigkeitsgewichte berechnet werden. Da die tatsächliche Teilnahme in der Wiederbefragung 2002 bekannt ist, kann die Ausschöpfung sowohl für Befragte mit niedriger als auch hoher Teilnahmewahrscheinlichkeit bestimmt werden (Tabelle 2). Die Gewichte wurden so gewählt, dass die Ausschöpfung sich im Falle niedriger Wahrscheinlichkeiten erhöht. Das schließt zwei Szenarien ein: Zum Ersten eine Halbierung der Differenz in den Ausschöpfungsquoten (d.h. eine Erhöhung um 9,5 Prozentpunkte) und zum Zweiten eine vollständige Angleichung (d.h. eine Erhöhung um 18,9 Prozentpunkte). Der deutliche Unterschied von knapp 19 Prozentpunkten in der Ausschöpfung zwischen beiden Gruppen deutet zudem darauf hin, dass die Teilnahmewahrscheinlichkeit durch das verwendete Modell gut vorhergesagt werden kann und bekräftigt zudem die Wahl des Kriteriums für niedrige Wahrscheinlichkeiten.

Tabelle 2: Geschätzte Teilnahme und wahre Ausschöpfungsquote in der zweiten Panelwelle.

	niedrige Propensity	
Teilnahme Welle 2	nein	ja
nein	42,04%	60,97%
ja	57,96%	39,03%

Mit Hilfe der so bestimmten Gewichtungsfaktoren lassen sich im dritten Schritt die Folgen der Case Prioritization für zwei Szenarien erfolgreicher Anwendung bestimmen (Halbierung und vollständige Angleichung der

Ausschöpfungsquoten). Dazu werden die Merkmalsverteilungen gewichtet und der Nonresponse Bias über einzelne Dissimilaritätsindizes bestimmt.

Um die Wirkung einer Kombination der Korrektur- und Präventionsverfahren zu testen, kommen die Häufigkeitsgewichte zum Einsatz, welche im Rahmen der Priorisierung berechnet wurden. Diese werden im Zuge der Berechnung der logistischen Regressionen bei der Panelgewichtung aktiviert. Das entspricht dem Fall, dass die Ausschöpfungsquote der niedrigen Propensity-Fälle erfolgreich erhöht wurde (Szenario Priorisierung 1 und Priorisierung 2) und der verbliebene Rest an Nonresponse Bias durch Propensity Score Gewichtung korrigiert werden soll.

Da zwei Szenarien für die Priorisierung untersucht werden, ergeben sich ebenfalls zwei Kombinationsmöglichkeiten: Gewichtung nach einer um 9,5 Prozentpunkte erhöhten Ausschöpfungsquote der Befragten mit niedrigem Propensity Score und nach einer Erhöhung um 18,9 Prozentpunkte. Für jede der Kombinationen werden die gewichteten Merkmalsverteilungen verwendet, um den Nonresponse Bias zu bestimmen.

Die vorgestellte Methode erlaubt es, den Nonresponse Bias vor und nach dem Einsatz der verschiedenen Verfahren zu beurteilen und die Wirkung zwischen ihnen zu vergleichen. Da es sich bei den Dissimilaritätsindizes um standardisierte Maße handelt, können die Verfahren in ihrer Wirkung direkt miteinander verglichen werden.

4 Ergebnisse

Die Effekte der Propensity Score Gewichtung, der Case Prioritization und der Kombination beider Verfahren sind in den Abbildungen 1 bis 4 für verschiedene Variablen abgetragen. Die Auswahl beschränkt sich dabei nicht nur auf Variablen, welche in den Schätzmodellen zur Bestimmung der Propensity Scores berücksichtigt wurden, sondern geht darüber hinaus. Es lassen sich vier analytische Variablen-Gruppen unterscheiden:

1. Sozio-demographische Variablen, berücksichtigt im Modell
2. Sozio-demographische Variablen, nicht berücksichtigt im Modell
3. Inhaltliche Variablen, berücksichtigt im Modell
4. Inhaltliche Variablen, nicht berücksichtigt im Modell

Die analytische Trennung dieser Gruppen ist sinnvoll, um zu prüfen, welche Folgen der Einsatz der drei Strategien auf einen Datensatz im Gesamten hat. Das schließt selbstverständlich auch Merkmale ein, welche zunächst als nicht mit dem Ausfallmechanismus verbunden betrachtet wurden und daher nicht in die Bestimmung der Propensity Scores eingehen. Gleichzeitig stehen für spätere Analysen nicht nur sozio-demographische Variablen im Vordergrund, sondern (im Gegenteil) meist inhaltliche. Gerade bei letzteren fehlt es in der bisherigen Literatur an einer umfänglicheren empirischen Überprüfung der Folgen von Verfahren zur Minderung des Nonresponse Bias.

Abbildung 1 zeigt den ermittelten Bias in Form von Dissimilaritätsindizes für die erste Gruppe an Merkmalen (sozio-demographische Variablen, berücksichtigt im Modell). Neben dem durch Panelattrition verursachten Nonresponse Bias sind die Effekte der drei Verfahren abgetragen. Aus dem Abstand zwischen dem Bias ohne Einsatz eines Verfahrens und dem Bias nach Einsatz eines Verfahrens ergibt sich der jeweilige Erfolg oder Misserfolg.

Für die fünf hier untersuchten Variablen zeigt sich erstens ein stark variierender Nonresponse Bias. Das bestätigt den Befund, dass es sich bei Bias um eine merkmalsspezifische Eigenschaft handelt (Bethlehem 2002). Zweitens führen die eingesetzten Verfahren in zwei von fünf Fällen (Geschlecht, Ortsgröße) zu einer Verstärkung des vorliegenden Bias, was die teilweise ambivalenten Befunde zur Wirkung von Propensity Score Gewichten bestätigt (Bergmann 2011). Drittens zeigt sich je nach Variable ein unterschiedlich stark ausgeprägter Effekt der Verfahren. Während die Verschlechterung bei der Variablen „Geschlecht“ gering ausfällt, reduzieren die Verfahren bei der Variable mit dem größten Bias, Bildung, den Nonresponse Bias in einem Umfang von nahezu 5 Prozentpunkten. Die Verzerrung der Variable Region kann schließlich durch die Gewichtung und die Kombination beinahe vollständig ausgeglichen werden.