Erschienen in:
Jackob, Nikolaus/Schoen, Harald/Zerback, Thomas (Hrsg.), Sozialforschung im Internet: Methodologie und Praxis der Online-Befragung, Wiesbaden: VS Verlag für Sozialwissenschaften, 2009, 145-157.

Fallen Gewichte ins Gewicht? Eine Analyse am Beispiel dreier Umfragen zur Bundestagswahl 2002

Von Thorsten Faas & Harald Schoen

1. Einleitung

Die Nutzung des Internet ist weltweit auf dem Vormarsch – und damit einhergehend auch der Einsatz von Online-Umfragen. Eine Umfrage via Internet durchzuführen ist für den Forscher vergleichsweise wenig kosten- und zeitintensiv. Allerdings haben diese Vorteile ihren Preis: Die größte Herausforderung besteht noch immer darin, bevölkerungsrepräsentative Ergebnisse zu gewinnen. Es ist eine Trivialität – aber dennoch wahr: Da die Durchführung von Online-Umfragen zwingend einen Zugang zum Internet voraussetzt, ein solcher Zugang in der Bevölkerung aber nicht universell gegeben ist, leiden einfache Online-Umfragen unter einem erheblichen Coverage-Problem. Die Herausforderungen beschränken sich allerdings keineswegs auf diesen Aspekt. Die Ziehung von Stichproben und die Rekrutierung von Probanden stellen ebenfalls potenzielle Problemquellen dar – die auch dann noch bestünden, wenn in der Bevölkerung universeller Zugang zum Internet gegeben wäre.[1]

Diese Probleme beschneiden die Aussagekraft von Online-Umfragen empfindlich. Denn sofern die Befragten keine Zufallsstichprobe aus der allgemeinen Bevölkerung oder auch nur aus der Gruppe der Internetnutzer darstellen, lassen sich aus Erkenntnissen über die online Interviewten keine belastbaren Schlüsse auf diese eigentlich interessierenden Populationen ableiten. Um diesem Problem beizukommen, wurden verschiedene Strategien vorgeschlagen: Ein Rat lautet, man müsse Daten aus Online-Umfragen nur geeignet gewichten, und schon erhalte man Ergebnisse, die auf interessierende Populationen übertragbar seien. Beispielsweise verwendet die bekannte Umfrage „Perspektive Deutschland“ eine solche Technik. Dank Gewichtung, so ist dort zu lesen, „[...] können die nicht repräsentativen Aussagen der Online-Stichprobe so umgewichtet werden, dass sich die Ergebnisse wie die einer repräsentativen Umfrage in der Altersgruppe der 16- bis 69-Jährigen interpretieren lassen.“[2] Allerdings sind auch Stimmen zu vernehmen, die diese und ähnliche Aussagen als allzu optimistisch erscheinen lassen. Sie weisen darauf hin, dass derartige Gewichtungsverfahren voraussetzungsreich sind und nicht zwangsläufig zum Ziel führen.[3] Allerdings liegen bislang nur wenige empirische Befunde zu dieser wichtigen Frage vor[4], im deutschen Sprachraum hat sie bislang kaum Beachtung gefunden.[5]

Einen Beitrag, diesen Mangel an empirischen Ergebnissen in bescheidenem Maße zu reduzieren, soll der vorliegende Aufsatz leisten. Er diskutiert Gewichtungsprozeduren zunächst theoretisch. Anschließend untersucht er an einem ausgewählten Beispiel aus der

1 Vgl. u.a. Couper 2000; Kemmerzell & Heckel 2001; Wildner & Conklin 2001; Hauptmanns & Lander 2003.
2 Perspektive Deutschland 2006: 128.
3 Vgl. z.B. Schonlau et al. 2007.
4 Vgl. u.a. Taylor 2000; Schonlau et al. 2004, 2007; Lee 2006.
5 Siehe aber Faas & Rattinger 2004.

politikwissenschaftlichen Forschung empirisch, inwieweit Gewichtungsverfahren bei Online-Umfragen auftretende Verzerrungen mindern können. Dazu werden die Ergebnisse einer offline durchgeführten Befragung von Internet-Nutzern mit den Befunden zweier Online-Erhebungen verglichen. Eine davon wurde auf der Basis eines (offline rekrutierten) Pools von Befragungswilligen durchgeführt, die zweite wurde als offene, für jedermann via WWW zugängliche Online-Umfrage realisiert. In diesem Vergleich werden schrittweise die Wirkungen unterschiedlich anspruchsvoller Gewichtungsverfahren untersucht. Abschließend werden wir die Ergebnisse kurz resümieren.

2. Repräsentativitätsprobleme und Gewichtungsverfahren

Für traditionelle Offline-Umfragen – ganz gleich, ob persönlich oder telefonisch durchgeführt – gibt es etablierte Verfahren zur Ziehung bevölkerungsrepräsentativer Stichproben und folglich zur Rekrutierung von Probanden. Umfrageforscher können sich offizieller Melderegister bedienen, um zunächst zufällig regionale Einheiten und innerhalb dieser eine Zufallsstichprobe der Bevölkerung zu ziehen. Alternativ können sich Forscher der so genannten Methode der *Random Routes* bedienen. Ebenfalls ausgehend von der zufälligen Auswahl regionaler Einheiten werden innerhalb dieser Einheiten zufällig Haushalte ausgewählt, indem zur Identifikation der Zielhaushalte vorher festgelegte Begehungsinstruktionen verwendet werden. Auch innerhalb der Haushalte erfolgt eine zufällige Auswahl der Zielperson, etwa anhand der Verteilung der Geburtstage im Haushalt. Telefonumfragen schließlich kombinieren Elemente aus beiden Ansätzen. Auf der Basis vorhandener Datenbanken mit Telefonnummern werden Telefonnummern gezogen. Durch zufällige Variation der letzten Ziffern der so ausgewählten Nummern können auch nicht aufgelistete Teilnehmer in die Stichprobe aufgenommen werden. Dies entspricht dem Register-basierten Ansatz, der allerdings an dieser Stelle nur Haushalte umfasst. Innerhalb der Haushalte werden Zielpersonen (analog zum Vorgehen bei *Random Routes*) zufällig ausgewählt.[6]

Die Crux der Online-Forschung besteht nun darin, dass keiner der skizzierten Ansätze sich auf das Internet und seine Nutzer übertragen lässt. Es existiert keine Liste der Internet-Nutzer, ebenso wenig ein dem Ansatz der *Random Routes* vergleichbares Verfahren. Und die zufällige Generierung etwa von E-Mail-Adressen mag zwar für Spammer in Frage kommen, sie scheidet aber als seriöses Verfahren zur Rekrutierung von Umfrageteilnehmern ebenfalls aus. Vor diesem Hintergrund kann es nicht verwundern, dass die Arbeitsgemeinschaft Deutscher Marktforschungsinstitute zu folgendem, noch immer gültigen Fazit kommt: „Online-Befragungen, die für die Zielgruppe der *Internetnutzer insgesamt* Anspruch auf Repräsentativität erheben, sollten gegenwärtig auf der Grundlage einer vorherigen Offline-Auswahl bzw. Offline-Rekrutierung mittels geeigneter Screening-Techniken durchgeführt werden, da gegenwärtig keine eindeutig definierte Online-Auswahlgrundlage von Internetnutzern existiert. Das heißt konkret: Es liegt weder eine vollständige und aktuelle Liste aller Internetnutzer vor, noch existieren Websites, deren Besucherstrukturen für die der Internetnutzer insgesamt repräsentativ sind. Deshalb ist eine Online-Auswahl bzw. Online-Rekrutierung der Teilnehmer nach dem Zufallsverfahren nicht möglich.“[7]

---

[6] Vgl. dazu etwa Häder & Glemser 2004.
[7] ADM et al. 2001.

In der Praxis bedeutet dies, dass Befragte, die über die etablierten und gerade skizzierten Verfahren der Zufallsauswahl der klassischen Umfrageforschung ausgewählt werden, gefragt werden, ob sie einen Zugang zum Internet haben. Ist das der Fall, werden sie weiterhin eingeladen, sich als Befragte für zukünftige Online-Umfragen zur Verfügung zu stellen. Nehmen sie diese Einladung an, gehören sie ab sofort zu einem Pool von Internet-Nutzern, die als potenzielle Teilnehmer an künftigen Online-Umfragen zur Verfügung stehen. Daher spricht man auch von Access-Panels. Steht dann eine konkrete Online-Umfrage an, werden aus diesem Access-Panel zufällig Befragte ausgewählt und zur Teilnahme eingeladen. In einer idealen Welt liefert dieses mehrstufige Auswahlverfahren eine repräsentative Stichprobe der Gemeinde der Internet-Nutzer: Zunächst werden die Zielpersonen für die herkömmlichen Offline-Umfragen ausgewählt, die Onliner dieser Gruppe stellen sich für zukünftige Online-Befragungen zur Verfügung, aus denen dann wiederum für konkrete Umfragen zufällig Respondenten ausgewählt werden.

Allerdings ist vergleichsweise wenig darüber bekannt, ob und wie sich dieses Verfahren in der Praxis bewährt. Gewisse Zweifel scheinen angebracht, schließlich kann die Rekrutierungskette an jeder einzelnen Stelle brechen – mit der Folge systematischer Fehler. Diese setzen potenziell schon bei der Rekrutierung der Befragten für die ursprüngliche Offline-Befragung an: Die Ausschöpfungsquoten solcher Umfragen liegen heute selbst im besten Fall bei nur rund 50 Prozent. Wer von diesen Personen sich schließlich bereit erklärt, zukünftig an Online-Umfragen teilzunehmen, ist ebenfalls eine offene Frage. Wer gibt schon gern seine E-Mail-Adresse weiter? Die Ergebnisse von Kemmerzell und Heckel zeigen, dass nur etwa ein Viertel der Befragten diesen Schritt ging.[8] Dabei sind es, wie an anderer Stelle gezeigt werden konnte[9], vor allem erfahrene Internet-Nutzer, die sich für Online-Umfragen zur Verfügung stellen.

Angesichts all dieser Probleme kann es nicht verwundern, dass Umfrageforscher auch nach alternativen Wegen gesucht haben, um Stichproben von Internetnutzern zu ziehen und zu repräsentativen Ergebnissen zu gelangen. Eine einfache Lösung besteht natürlich darin, die Rekrutierung den (potenziellen) Befragten selbst zu überlassen. Man stelle eine Umfrage ins Netz und hoffe auf Teilnehmer. An solchen offenen Umfragen kann folglich jeder Internet-Nutzer, der davon erfährt, teilnehmen – potenziell sogar mehrfach. Auf den ersten Blick mag diese Art der „Rekrutierung“ an das historische Fiasko von Literary Digest erinnern. Vor der US-Präsidentenwahl 1936 hatte diese Zeitschrift ebenfalls darauf verzichtet, Probanden systematisch zu rekrutieren. Stattdessen hatte man die Frage der Rekrutierung in die Hände der Zeitschrift der Leser gelegt. Auf der Basis der so „gezogenen“ Stichprobe resultierte eine falsche Prognose des damaligen Wahlausgangs – im Gegensatz zur Prognose von George Gallup, der auf der Basis einer Zufallsstichprobe die Richtung des Ergebnisses traf und damit den Durchbruch für die moderne, auf Zufallsstichproben basierende Umfrageforschung schaffte.

Dennoch erleben solche offenen Umfragen im digitalen Zeitalter ein Revival. Prominentes Beispiel in Deutschland ist die schon erwähnte Umfrage „Perspektive Deutschland“. Dabei handelt es sich um eine Reihe von Umfragen, die von einem Konsortium rund um die Unternehmensberatung McKinsey durchgeführt wurde. Jedermann kann während der jeweiligen Feldzeit unter www.perspektive-deutschland.de an den Umfragen teilnehmen.

---

8 Vgl. Kemmerzell & Heckel 2001.

9 Vgl. Faas 2003b.

Die Teilnehmerzahlen sind – auch dank massiver Werbung – beeindruckend. 2005 etwa nahmen mehr als 600.000 Internet-Nutzer an der Befragung teil.

Um keinen falschen Eindruck entstehen zu lassen: Natürlich sind sich die Verantwortlichen dieser (und vergleichbarer) Umfragen der damit verbundenen Probleme bewusst. Sie argumentieren allerdings, dass die Probleme beherrschbar seien. Terhanian und Bremer etwa, die ähnliche Umfragen in den USA durchgeführt haben, sehen das Problem der Literary-Digest-Umfrage weniger in der verzerrten Stichprobe selbst, sondern im Versäumnis der Verantwortlichen, „(...) to weight the characteristics of the final sample of respondents to reflect the characteristics of likely voters."[10] Bei entsprechender Gewichtung hätte, so das Argument, auch Literary Digest mit seiner Prognose das richtige Ergebnis getroffen.

Gewichtungsverfahren sind keine Erfindung des Internet-Zeitalters, sondern wurden und werden auch für offline durchgeführte Umfragen genutzt. Dabei lassen sich zwei Arten von Gewichten unterscheiden: Designgewichte und Anpassungsgewichte.[11] Erstere dienen dazu, aus dem Erhebungsdesign resultierende Verzerrungen zu korrigieren, so dass die Verteilungen nach Gewichtung jenen entsprechen, die sich bei einer Zufallsstichprobe mit gleich verteilten Auswahlwahrscheinlichkeiten ergeben würden. Ein klassisches Beispiel dafür sind Transformationsgewichte in zweistufigen Zufallsauswahlen. Auf der ersten Stufe werden – wie schon oben skizziert – zufällig Haushalte ausgewählt, anschließend innerhalb dieser Haushalte Personen per Zufall interviewt. Es liegt auf der Hand, dass die Auswahlwahrscheinlichkeit einer Person innerhalb eines Haushalts mit zunehmender Haushaltsgröße abnimmt. Im Ergebnis sind Personen aus Einpersonenhaushalten im Vergleich zu solchen aus Vier- oder Fünfpersonenhaushalten systematisch überrepräsentiert. Um dies zu korrigieren, werden aus den designbedingt unterschiedlichen Auswahlwahrscheinlichkeiten Designgewichte berechnet.

Anpassungsgewichte dienen demgegenüber nicht dazu, wohlbekannte Eigenschaften des Erhebungsdesigns zu kompensieren. Vielmehr sollen damit selektive Stichprobenausfälle ausgeglichen werden. Diese können beispielsweise darauf beruhen, dass bestimmte Personen schlechter erreichbar sind als andere oder es überproportional häufig ablehnen, ein Interview zu führen. Mit Anpassungsgewichten wird nun versucht, die gemeinsame Verteilung ausgewählter Merkmale in der Stichprobe mit der entsprechenden Verteilung in der Grundgesamtheit zur Deckung zu bringen. Häufig werden in Umfragen etwa die gemeinsamen Verteilungen soziodemographischer Merkmale in der Stichprobe an jene in der Grundgesamtheit angepasst. Das Ziel besteht darin, mit Hilfe dieser Gewichtungsprozedur trotz selektiver Ausfälle eine repräsentative Stichprobe zu gewinnen, die es erlaubt, Schlussfolgerungen auf die angezielte Grundgesamtheit zu ziehen.[12]

Bei den für Online-Erhebungen vorgeschlagenen Gewichtungen handelt es sich somit um Anpassungsgewichte. Die Wahrscheinlichkeit, überhaupt von einer Online-Umfrage zu erfahren, ist ebenso unterschiedlich in der Bevölkerung verteilt, wie die technischen Voraussetzungen für eine Beteiligung an solchen Umfragen – gleiches gilt für die Bereitschaft, an der Umfrage überhaupt teilzunehmen. Diese potenziellen Verzerrungen werden dadurch zu korrigieren versucht, dass die gemeinsame Verteilung sozialstruktureller und „psycho-

[10] Terhanian & Bremer 2002: 3.
[11] Vgl. u.a. Kish 1990; Gabler 2004.
[12] Vgl. u.a. Rubin & Thomas 1996; Smith et al. 2000; Terhanian & Bremer 2002.

grafischer“ oder „webografischer“[13] Merkmale in der Stichprobe an jene in einer offline befragten Stichprobe angepasst wird.[14] Wenn diese Prozedur zum Erfolg führen soll, müssen die Gewichtungsmerkmale mit der Teilnahmewahrscheinlichkeit stark zusammenhängen. Darüber hinaus müssen sich Teilnehmer und Nicht-Interviewte innerhalb der durch die Gewichtungsvariablen abgegrenzten Gruppen sehr ähnlich sein.[15] Ist dies nicht der Fall, unterscheiden sich also die befragten von den nicht befragten Personen in einer Gruppe, kann die Gewichtung zu ernsthaften Verzerrungen führen.[16] Die Tatsache, dass für Webumfragen, wie im obigen Beispiel, neben sozialstrukturellen auch „psychografische“ Merkmale zur Gewichtung herangezogen werden, kann man als Versuch interpretieren, die Annahme plausibler zu machen, dass die zweite Bedingung tatsächlich erfüllt sei und die Gewichtung zu einer Anpassung der Randverteilungen an jene in der Grundgesamtheit führe.[17]

Ob dies tatsächlich gelingt, ist freilich eine empirische Frage. Denn auch wenn die Personen in einer Zelle der Gewichtungsmatrix in vielerlei Hinsicht homogen sind, können sie sich im Hinblick auf das jeweils betrachtete Merkmal doch unterscheiden. Eine sonst zur Anpassung der Randverteilungen gut geeignete Gewichtungsprozedur kann also bei der interessierenden Variablen erfolglos bleiben oder gar zu größeren Verzerrungen führen. Vor diesem Hintergrund kann es kaum überraschen, dass vorliegende Untersuchungen zu recht unterschiedlichen Aussagen über die Leistungsfähigkeit von Anpassungsgewichten für Online-Umfragen gelangen. So finden Varedian und Forsman in einer Marketingumfrage kaum positive Gewichtungseffekte.[18] Andere Autoren stellen dagegen in Untersuchungen zu diversen Themenfeldern fest, dass Anpassungsgewichtungen Verzerrungen in Randverteilungen reduzieren, aber nicht unbedingt völlig beseitigen.[19]

Der vorliegende Beitrag reiht sich in die Versuche ein, die Leistungsfähigkeit von Anpassungsgewichtungen bei Online-Umfragen empirisch zu untersuchen. Er betrachtet die Wirksamkeit unterschiedlich anspruchsvoller Gewichtungsprozeduren bei unterschiedlichen Arten von Umfragen unter Internet-Nutzern und vergleicht die resultierenden Ergebnisse miteinander. Einerseits handelt es sich um eine traditionelle Offline-Umfrage, in deren Rahmen auf der Basis von *Random Routes* Befragte ausgewählt wurden, die dann persönlich befragt wurden. Aus dieser Umfrage werden allerdings – um die Vergleichbarkeit zu erhöhen (und damit sämtliche Coverage-Probleme in der Folge des so genannten Digital Divides auszublenden) – nur die Internet-Nutzer berücksichtigt. Diese Umfrage wird als Vergleichsmaßstab für zwei Online-Umfragen verwendet. Eine davon wurde auf der Basis eines (offline rekrutierten) Pools von Befragungswilligen durchgeführt, die zweite wurde als offene, für jedermann via WWW zugängliche Online-Umfrage realisiert. Als Maßstab

---

13 Schonlau et al. 2007.

14 Diese Prozedur führt offensichtlich nur dann zu einer Anpassung an die Verteilungen in der Grundgesamtheit, wenn entweder Informationen aus einer Vollerhebung der Grundgesamtheit oder zumindest einer echten Zufallsstichprobe aus dieser vorliegen.

15 Vgl. Arzheimer 2008.

16 Problematisch sind vor allem Gewichte mit sehr hohen oder niedrigen Werten (Gabler 2004). Auf weitere technische Fragen, etwa die dürftige Besetzung von Zellen in der Gewichtungsmatrix, können wir hier nicht näher eingehen.

17 Wirkungen von Gewichtungsprozeduren auf andere Schätzparameter behandeln wir hier nicht ausführlich. Es sei lediglich darauf verwiesen, dass Gewichtungen häufig mit erhöhten Varianzen und einer geringeren Präzision von Schätzungen einhergehen (Kish 1990: 127).

18 Vgl. Varedian & Forsman 2003.

19 Vgl. Taylor 2000; Isaksson & Forsman 2003; Schonlau et al. 2004, 2007; Lee 2006.

für den Vergleich wollen wir die Frage nach dem beabsichtigten Wahlverhalten heranziehen, da alle Umfragen im Umfeld der Bundestagswahl 2002 durchgeführt wurden.

3. Datengrundlage: Drei Umfragen vor der Bundestagswahl 2002

Ehe wir zu den empirischen Analysen kommen, sind zunächst noch einige Informationen zu den drei Umfragen zu geben. Tabelle 1 liefert Details zu den drei Umfragen. Grundlage der ersten Umfrage ist eine repräsentative Stichprobe der deutschen Staatsbürger ab einem Alter von 16 Jahren. Zwischen dem 12. August und dem 8. November 2002 wurden über Sample Points, *Random Routes* und – innerhalb der Haushalte – letzte Geburtstage insgesamt 3.263 Personen zufällig ausgewählt. Die Rücklaufquote dieser Befragung lag bei 63,8 Prozent. 1.076 der so ausgewählten Personen nutzten nach eigenen Angaben das Internet. Wir wollen mit Blick auf diese Umfrage im Weiteren von der „Offline-Umfrage" sprechen.

Die zweite Umfrage umfasst ebenfalls eine Zufallsstichprobe der deutschen Internet-Nutzer. Ingesamt 1.165 Personen wurden aus einem offline rekrutieren Access-Panel zufällig ausgewählt.[20] Die Feldzeit dieser Umfrage umfasste den Zeitraum vom 13. September bis zum 1. Oktober 2002; die Rücklaufquote (bezogen auf die zur Befragung eingeladenen Mitglieder des Access-Panels) lag in diesem Fall bei 73,5 Prozent. Hier wollen wir im Weiteren kurz vom „Access-Panel" sprechen.

Bei der dritten Umfrage handelt es sich um eine offene Online-Umfrage, die im Zeitraum vom 20. August bis zum 22. September 2002 unter www.wahlumfrage2002.de allen Internet-Nutzern offenstand.[21] Die Teilnehmer dieser Umfrage rekrutierten sich selbst – insgesamt 34.098 Personen taten dies. Der Aufbau der Umfrage war dabei modular: Nach einem Basismodul, das zu Beginn auszufüllen war, wurden die Teilnehmer eingeladen, drei weitere Zusatzmodule auszufüllen. Da einige der Fragen, die im weiteren Verlauf verwendet werden, erst in zwei Zusatzmodulen erhoben wurden, müssen wir uns auf Befragte beschränken, die das Basismodul sowie die beiden Zusatzmodule bearbeitet haben. Dies waren 9.189 Befragte.[22] Wenn wir in den folgenden Abschnitten auf diese Umfrage verweisen, sprechen wir nur kurz von der „Wahlumfrage".

Im Folgenden liegt der Fokus der Analyse auf dem (beabsichtigten und berichteten) Wahlverhalten der Befragten – zweifelsohne die zentrale Frage der Wahlforschung. In der empirischen Analyse vergleichen wir die Verteilungen dieser Variable, die resultieren, wenn man unterschiedlich anspruchsvolle Gewichtungsprozeduren einsetzt. An erster Stelle steht ein Vergleich, in der lediglich für die Offline-Umfrage ein Designgewicht verwendet wird: Da ostdeutsche Befragte dort bewusst überrepräsentiert wurden, wird nur diese Verzerrung korrigiert. Im zweiten Schritt wird die sozialstrukturelle Zusammensetzung der drei Stichproben (Alter, Geschlecht, Bildung) harmonisiert. Harmonisierung bedeutet in diesem wie den folgenden Schritten, dass die Verteilung der beiden Online-Umfragen an die Verteilung der Offline-Stichprobe angepasst wird. Eine vereinfachende Annahme muss dabei getroffen werden: Alle Variablen werden dichotomisiert, da sonst einzelne Zellen der zugrunde liegenden Matrix zu schwach besetzt wären (siehe hierzu auch den Anhang). Im

[20] Siehe auch Faas 2003b.

[21] Siehe auch Faas 2003a.

[22] Wie dieser Übersicht zu entnehmen ist, unterscheiden sich die drei Umfragen nicht nur hinsichtlich ihres Befragungs- und Rekrutierungsmodus, sondern auch hinsichtlich ihrer Feldzeit. Allerdings bleiben die verschiedenen Feldzeiten ohne Effekt, wie zeitlich differenzierte Analysen zeigen.

nächsten Schritt wird die Häufigkeit der Internet-Nutzung kontrolliert, da frühere Analysen gezeigt haben, dass gerade in Access-Panels routinierte Nutzer überrepräsentiert sind.[23] Schließlich werden im letzten Schritt auch substanzielle Variablen in die Gewichtung mit einbezogen, nämlich solche, die die Teilnahme an offenen Online-Umfragen wahrscheinlicher machen. Dazu wird die Internal Efficacy herangezogen.[24] Die Annahme dabei ist, dass Befragte, die sich für politisch einflussreich halten, auch häufiger an politischen Online-Umfragen teilnehmen.

Tabelle 1: Details zu den drei im Umfeld der Bundestagswahl 2002 durchgeführten Umfragen

	Offline-Umfrage	**Access Panel**	**Wahlumfrage**
Feldzeit	12. August bis 8. November	13. September bis 1. Oktober	20. August bis 22. September
Teilnehmer	3.263	1.165	9.189
Teilnehmer mit Internet-Nutzung	1.076	1.165	9.189
Rekrutierung	Zufällige Auswahl über Sample Points, Random Route und Last Birthday	Zufällige Auswahl aus einem Access-Panel	Selbstrekrutierung
Befragungsmodus	PAPI*	CASI*	CASI
Ausschöpfung	63,8%	76,1%	—

* PAPI = Paper and Pencil Interview, CASI = Computer Assisted Self-Administered Interview

4. Empirische Ergebnisse

Im ersten Schritt betrachten wir die Wahlabsicht der Respondenten in den drei Umfragen, nachdem die ungleichen Auswahlwahrscheinlichkeiten Ost- und Westdeutscher in der Offline-Umfrage mittels Designgewichtung korrigiert wurden. Die Ergebnisse für diesen (wie auch die folgenden Vergleiche) liefert Tabelle 2. Demnach existieren beträchtliche Unterschiede zwischen den drei Erhebungen – obwohl sie alle nur Internet-Nutzer umfassen. Die Union schneidet am besten auf der Basis der Offline-Umfrage ab, sie erreicht dort 31,6 Prozent. In den Online-Umfragen dagegen kommt sie nicht über ein Fünftel (in der offenen WWW-Umfrage) bzw. ein Viertel (im Access-Panel) der Stimmen hinaus. Die SPD auf der anderen Seite wird die Ergebnisse auf der Basis des Access-Panels bevorzugen: Dort bekommt sie 41,1 Prozent, während sie in der Wahlumfrage2002 unter die Dreißigprozentmarke rutscht. Die kleinen Parteien schließlich – FDP, Grüne und PDS – schneiden in der offenen Online-Umfrage am besten ab. Gerade die Grünen erzielen dort mit rund einem Viertel der Stimmen ein außerordentlich gutes Ergebnis und avancieren damit zur zweitstärksten Partei.

[23] Vgl. Faas 2003b.
[24] Vgl. Campbell et al. 1954; Balch 1974; Vetter 1997.

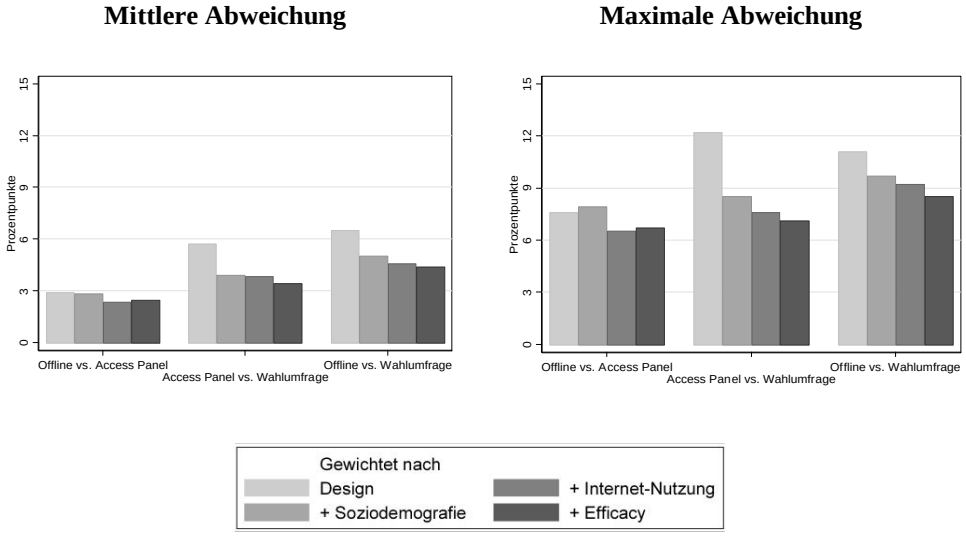
Tabelle 2: Verteilung der Wahlabsichten in den drei Umfragen in Abhängigkeit vom Gewichtungsverfahren (Prozent)

	CDU/CSU	SPD	Grüne	FDP	PDS	Sonstige
(1) Designgewicht						
Offline-Umfrage	31,6	37,7	14,6	9,0	5,3	1,8
Access-Panel	24,0	41,1	14,1	11,4	4,7	4,7
Wahlumfrage	20,9	28,9	25,7	16,1	5,5	2,9
(2) Designgewicht + Soziodemographie						
Access-Panel	23,7	40,8	14,4	11,5	5,0	4,7
	(-0,3)	(-0,3)	(+0,3)	(+0,1)	(+0,3)	(+0,0)
Wahlumfrage	21,9	32,4	22,9	14,4	5,3	3,2
	(+1,0)	(+3,5)	(-2,8)	(-1,7)	(-0,2)	(+0,3)
(3) Designgewicht + Soziodemographie + Internet-Nutzung						
Access-Panel	25,1	40,3	14,3	10,4	5,1	4,9
	(+1,4)	(-0,5)	(-0,1)	(-1,1)	(+0,1)	(+0,2)
Wahlumfrage	22,4	33,3	21,9	14,1	5,1	3,0
	(+0,5)	(+0,9)	(-1,0)	(-0,3)	(-0,2)	(-0,2)
(4) Designgewicht + Soziodemographie + Internet-Nutzung + Efficacy						
Access-Panel	24,9	40,5	14,7	10,6	4,7	4,7
	(-0,2)	(+0,2)	(+0,4)	(+0,2)	(-0,4)	(-0,2)
Wahlumfrage	23,1	33,4	21,5	13,8	5,0	3,3
	(+0,7)	(+0,1)	(-0,4)	(-0,3)	(-0,1)	(+0,3)

Die in Klammern angegebenen Werte geben die Unterschiede zur jeweils vorangegangenen Gewichtungsstufe an. Der Wert „+1,4“ in der Spalte CDU/CSU im Abschnitt (3) bedeutet zum Beispiel, dass der Stimmenanteil für CDU und CSU bei zusätzlicher Berücksichtigung der Internetnutzung im Vergleich zu einem Design- und Soziodemographie-Gewicht um 1,4 Punkte ansteigt.

Abbildung 1 liefert anhand zweier Maßzahlen eine Visualisierung der Unterschiede der resultierenden Verteilungen: Die durchschnittliche sowie die maximale Abweichung einzelner Parteianteile für jedes Paar von Umfragen. Erwartungsgemäß erweist sich insbesondere die offene Internet-Umfrage als Ausreißer – sie weicht von den beiden anderen Umfragen deutlich ab. Aber auch zwischen den beiden anderen Umfragen ergeben sich deutliche Unterschiede – und dies, obwohl doch beide für sich in Anspruch nehmen wollen, eine Zufallsstichprobe von Internet-Nutzern zu liefern. Die oben skizzierten Annahmen, die dazu erfüllt sein müssten, scheinen sich in der Praxis aber nicht zu bewähren. Gerade das Access-Panel, dies konnte an anderer Stelle bereits gezeigt werden, scheint davon betroffen. Weitere Gewichtungsmaßnahmen jedenfalls scheinen vor diesem Hintergrund unerlässlich.

Abbildung 1: Überblick der Auswirkungen der verschiedenen Gewichtungsmaßnahmen



Soziodemografische Gewichtungen gehören zur etablierten Praxis in der Umfrageforschung. Dass sich bestimmte, sozialstrukturell definierte Personengruppen weniger gut als andere erreichen lassen, ist hinlänglich bekannt. Als Beispiel kann man auf ältere Frauen oder auch auf Personen mit geringerer formaler Bildung verweisen. Daher soll auch hier an erster Stelle für die – tatsächlich sehr unterschiedliche – Sozialstruktur der drei Stichproben kontrolliert werden. Allerdings genügt die Harmonisierung der Stichproben nach Alter, Geschlecht und Bildung nicht, um die zuvor beobachtbaren Unterschiede in der Verteilung der Wahlabsicht zum Verschwinden zu bringen. Obwohl sich die Werte aufgrund der Gewichtung erheblich verändern (und bis zu einem gewissen Grad auch annähern), bleiben die Muster der Unterschiede dennoch erhalten.

Die deutlichsten Unterschiede ergeben sich dabei für die offene Internet-Umfrage – wenig überraschend, denn in dieser Umfrage weichen die Verteilungen der nun harmonisierten sozialstrukturellen Variablen ursprünglich am deutlichsten ab.[25] Insbesondere die Anteile der Grünen und der FDP gehen um drei bzw. zwei Prozentpunkte zurück, während auf der anderen Seite Union und SPD nun besser abschneiden. Aus diesen Entwicklungen ergeben sich die geringer gewordenen Unterschiede zwischen dieser und den anderen beiden Umfragen. Die Unterschiede zwischen diesen beiden Umfragen wiederum bleiben von der zusätzlichen Gewichtung nahezu unberührt. Die zwischen ihnen beobachtbaren Unterschiede haben also keine sozialstrukturellen Ursachen, sondern müssen andere Quellen haben: Gerade die Intensität der Internet-Nutzung unterscheidet sich zwischen diesen beiden Gruppen – entsprechend wird im nächsten Schritt der Einfluss dieser Variablen näher analysiert.

[25] Im Anhang finden sich die Verteilungen der entsprechenden Variablen.

Tatsächlich verändern sich die Ergebnisse auf der Basis des Access-Panels durch die zusätzliche Berücksichtigung der Intensität der Internet-Nutzung deutlich. Inhaltlich bewegen sich dadurch erneut die Anteile von Grünen und FDP nach unten, während Union und SPD (letztere vor allem im Falle der offenen Internet-Umfrage) hinzugewinnen. Angesichts der Tatsache, dass sich sowohl im Access-Panel als auch der offenen Internet-Umfrage mehr Internet-affine Befragte befinden, kann man aus den Veränderungen folgern, dass diese Nutzer eher die beiden kleinen Parteien unterstützen, weniger dagegen SPD und Union.

Insgesamt – das zeigt Abbildung 1 – haben sich die Verteilungen der Wahlabsichten weiter angenähert. Die durchschnittliche Abweichung zwischen der Offline-Umfrage und der Umfrage auf der Basis des Access-Panels liegt nun bei rund zwei Prozentpunkten, während die mittlere Abweichung der offenen Internet-Umfrage zu den beiden anderen immer noch bei rund vier Prozentpunkten liegt. Dies ist weiterhin eine ganz erhebliche (mittlere) Abweichung. Noch viel größer fallen natürlich die maximalen Unterschiede aus – sie erreichen weiterhin Größenordnungen von bis zu neun Prozentpunkten. Bleibt die Frage zu klären, ob diese Unterschiede in der unterschiedlichen Bereitschaft, an Umfragen teilzunehmen, ihren Ursprung haben. Um diese (ansatzweise) testen zu können, wollen wir im letzten Schritt zusätzlich noch das politische Effektivitätsbewusstsein in die Gewichtungen einbeziehen.

Diese zusätzliche Maßnahme berührt erwartungsgemäß die offene Internet-Umfrage am stärksten – schließlich soll diese Variable die Teilnahme-Bereitschaft messen, welche sich wiederum am stärksten auf die Teilnahme an der offenen Internet-Umfrage auswirken sollte: Dort nämlich müssen die Befragten die Umfrage selbst finden und sich für die Teilnahme motivieren. Die Größenordnung der beobachtbaren Unterschiede ist allerdings vergleichsweise moderat – und die Folgen bemerkenswert: Die Unterschiede zwischen den Umfragen sind auch nach diesen umfangreichen Gewichtungsmaßnahmen immer noch beträchtlich. Das Unterfangen, die Unterschiede zwischen den Umfragen zu beseitigen, ist gescheitert, auch wenn sich die Verteilungen angenähert haben.

5. Fazit

Ziel dieses Beitrags war es, verschiedene Verfahren zur Rekrutierung von Internet-Nutzern für Umfragen zu vergleichen und die Folgen der verschiedenen Verfahren für resultierende Verteilungen substanzieller Variablen zu beobachten. Ausgehend davon sollten dann verschiedene Gewichtungsmaßnahmen hinsichtlich ihrer Leistungsfähigkeit geprüft werden, um feststellen zu können, ob diese Maßnahmen Verzerrungen auffangen und korrigieren können. Wie die Ergebnisse zeigen, können – ausgehend von erheblichen Unterschieden in den Ausgangsverteilungen – die Gewichtungsmaßnahmen zwar einige dieser Unterschiede auffangen, ohne sie aber gänzlich beseitigen zu können. Die auf der Basis der offenen Internet-Umfrage basierenden Ergebnisse reagieren am stärksten auf die sozialstrukturelle Gewichtung, während die Intensität der Internet-Nutzung vor allem Unterschiede zwischen der Offline-Umfrage (in der „light“ Internet-Nutzer präsenter sind) und den beiden Online-Umfragen verkleinert. Im letzten Schritt bewirkte auch die Aufnahme des politischen Effektivitätsbewusstseins noch einmal eine Annäherung der offenen Internet-Umfrage an die beiden anderen Umfragen.

Die Gewichtungen hatten also (positive) Folgen. Aber offenkundig waren die Zusammenhänge zwischen den Gewichtungsvariablen und den substanziellen Variablen (hier also der Wahlabsicht) nicht stark genug, um die Unterschiede vollständig auszugleichen. Das überrascht nicht. Wählen ist ein komplexer Prozess, der nicht durch einige wenige, ausgewählte Variablen determiniert wird. Es bleibt zu untersuchen, ob die Einbeziehung anderer substantieller Variablen die Gewichtungsprozeduren noch effektiver machen könnte. Auch ist zu prüfen, wie sich die vorgestellten sowie weiter gehende Gewichtungsverfahren auf bi- und multivariate Zusammenhänge sowie die Präzision von Schätzungen auswirken. Diese Fragen stellen sich natürlich auch im Hinblick auf andere als die hier exemplarisch betrachtete Wahlabsichtsfrage, die in der Politikwissenschaft von erheblichem Interesse ist. Auf diese Weise könnte die Forschung wichtige Erkenntnisse über die merkmals- und kontextspezifische, aber auch die generelle Leistungsfähigkeit von Gewichtungsverfahren gewinnen. Sie könnten dazu beitragen, die Möglichkeiten und Grenzen dieses Ansatzes, die Aussagekraft von Online-Umfragen zu steigern, besser abschätzen zu können.

Literatur

ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., ASI Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V., BVM Berufsverband Deutscher Markt- und Sozialforscher e.V. & D.G.O.F. Deutsche Gesellschaft für Online-Forschung e.V., 2001: Standards for Quality Assurance for Online Surveys. Frankfurt am Main.

Arzheimer, K. (2008): Gewichtungsvariation. In: Schoen, H., Rattinger, H. & Gabriel, O. W. (Hrsg.): Methodische Probleme der Wahl- und Einstellungsforschung, Baden-Baden (im Erscheinen).

Balch, G. I. (1974): Multiple Indicators in Survey Research: The Concept of 'Sense of Political Efficacy'. In: Political Methodology, 1, S. 1-43.

Campbell, A., Gurin, G. & Miller, W. E. (1954): The Voter Decides, New York.

Couper, M.P. (2000): Web Surveys: A Review of Issues and Approaches. In: Public Opinion Quarterly, 64, S. 464-494.

Faas, T. (2003a): www.wahlumfrage2002.de – Ergebnisse und Analysen, Bamberger Beiträge zur Politikwissenschaft: Forschungsschwerpunkt Politische Einstellungen und Verhalten, Nr. II-11.

Faas, T. (2003b): Offline rekrutierte Access Panels: Königsweg der Online-Forschung? In: ZUMA-Nachrichten, 53, S. 58-76.

Faas, T. & Rattinger, H. (2004): Drei Umfragen, ein Ergebnis? Ergebnisse von Offline- und Online-Umfragen anlässlich der Bundestagswahl 2002 im Vergleich. In: Brettschneider, F., van Deth, J. & Roller, E. (Hrsg.): Die Bundestagswahl 2002. Analysen der Wahlergebnisse und des Wahlkampfes, Wiesbaden, S. 277-299.

Gabler, S., (2004): Gewichtungsprobleme in der Datenanalyse. In: Diekmann, A. (Hrsg.), Methoden der Sozialforschung, Wiesbaden, S. 128-147.

Häder, S. & Glemser, A. (2004): Stichprobenziehung für Telefonumfragen in Deutschland. In: Diekmann, A. (Hrsg.): Methoden der Sozialforschung, Wiesbaden, S. 148-171.

Hauptmanns, P. & Lander, B. (2003): Zur Problematik von Internet-Stichproben. In: Theobald, A., Dreyer, M. & Starsetzki, T. (Hrsg): Online-Marktforschung: Theoretische Grundlagen und praktische Erfahrungen, Wiesbaden.

Isaksson A. & Forsman, G. (2004): A Comparison between Using the Web and Using the Telephone to Survey Political Opinions. Paper presented at the annual meeting of the American Association for Public Opinion Research, Sheraton Music City, Nashville, TN, Aug 16, 2003.

Kemmerzell, P. & Heckel, C. (2001): Grundgesamtheit und Stichprobe bei Online-Befragungen, repräsentativ zu was? In: Planung & Analyse, 4, S. 52-60.

Kish, L., (1990): Weighting: Why, When, and How. In: Proceedings of the Survey Research Methods Section. http://www.amstat.org/sections/srms/Proceedings/papers/1990_018.pdf (10.3.2008), S. 121-130.

Lee, S. (2006): Propensity Score Adjustment as Weighting Scheme for Volunteer Panel Web Surveys. In: Journal of Official Statistics, 22, S. 329-349.

Perspektive Deutschland (2006): Projektbericht Perspektive-Deutschland 2005/06. Die größte gesellschaftspolitische Online-Umfrage (URL: http://www.perspektive-deutschland.de/files/presse_2006/pd5-Projektbericht.pdf) (10.3.2008).
Rubin, D. B. & Thomas, N. (1996): Matching Using Estimated Propensity Scores: Relating Theory to Practice. In: Biometrics, 52, S. 254-268.
Schonlau, M., Fricker, R. D. Jr. & Elliott, M. N. (2002): Conducting Research Surveys via E-mail and the Web. Santa Monica, CA.
Schonlau, M., van Soest, A., Kapteyn, A., Couper, M. & Winter, M. (2004): Adjusting for Selection Bias in Web Surveys Using Propensity Scores: The Case of the Health and Retirement Study. In: Proceedings of the Section on Survey Statistics, American Statistical Association, S. 4326-4333.
Schonlau, M., Van Soest, A. & Kapteyn, A. (2007): Are ‚Webographic' or Attitudinal Questions Useful for Adjusting Estimates from Web Surveys Using Propensity Scoring? In: Survey Research Methods, 1, S. 155-163.
Smith, P. J., Rao, J. N. K., Battaglia. M. P., Daniels, D. & Ezzati-Rice, T. (2000): Compensating for Nonresponse Bias in the National Immunization Survey Using Response Propensities. In: Proceedings of the Section on Survey Research Methods, American Statistical Association, S. 641-646.
Taylor, H. (2000): Does Internet Research 'Work'? Comparing Online Survey Results With Telephone Surveys. In: International Journal of Market Research, 42, S. 51-63.
Terhanian, G. & Bremer, J. (2003): A Multi-Method Approach for Reducing Error in Internet-Based Surveys of Non-Probability Samples. Paper presented at the 98th Annual Meeting of the American Political Science Association (APSA) in Boston, August 26th to September 1st, 2002.
Varedian, M. & Forsman, G. (2003): Comparing Propensity Score Weighting with other Weighting Methods: A Case Study on Web Data (Presented at the 2003 AAPOR meetings).
Vetter, A. (1997): Political Efficacy – Reliabilität und Validität, Wiesbaden.
Wildner, R. & Conklin, M. (2001): Stichprobenbildung für Marktforschung im Internet. In: Planung & Analyse, 2, S. 18-27.

Anhang – Verteilungen der Variablen, die in die Gewichte eingeflossen sind

Hinweis: Die folgenden Verteilungen sind ungewichtet – bei einer Ausnahme: Für die Offline-Umfrage wurde angesichts des Oversamplings für Ostdeutschland ein entsprechendes Designgewicht eingesetzt.

	Offline-Umfrage	Access Panel	Wahlumfrage
	Geschlecht		
Mann	56,3	58,8	75,7
Frau	43,7	41,2	24,3
	Alter		
Geburtsjahr < 1963	45,2	40,3	30,4
Geburtsjahr > 1962	54,8	59,7	69,6
	Bildung		
Bis Realschule	52,1	49,9	20,4
Abitur und mehr	47,9	50,1	79,6

	Offline-Umfrage	Access Panel	Wahlumfrage
Internet-Nutzung			
Häufigkeit der Nutzung			
Mehrmals am Tag (5)	18,4	39,1	58,7
(Fast) Jeden Tag (4)	28,2	36,1	26,9
Mehrmals pro Woche (3)	38,0	21,0	11,3
Ein paar Mal im Monat (2)	12,8	3,0	2,1
Seltener (1)	2,6	0,8	0,9
Internet-Nutzung seit ...			
1998 oder früher (5)	29,2	47,9	59,2
1999 (4)	18,6	21,3	15,4
2000 (3)	26,2	21,9	15,9
2001 (2)	18,0	7,8	7,0
2002 (1)	8,0	1,2	2,5
Index der Internet-Nutzung			
Summe der beiden Indikatoren, dichotomisiert			
Weniger erfahrener Nutzer (2-6)	42,2	15,4	12,0
Erfahrene Nutzer (7-10)	57,8	84,6	88,0
Politisches Effektivitätsbewusstsein			
"*Könnte führende Rolle in einer politischen Gruppe übernehmen*"			
Stimme voll und ganz zu (5)	15,7	12,5	33,8
Stimme eher zu (4)	31,0	29,5	36,3
Teils teils (3)	21,7	27,1	16,0
Stimme eher nicht zu (2)	22,2	24,5	11,0
Stimme überhaupt nicht zu (1)	9,5	6,5	2,9
"*Politische Fragen kann ich gut verstehen*"			
Stimme voll und ganz zu (5)	25,9	17,4	40,2
Stimme eher zu (4)	44,0	44,2	46,1
Teils teils (3)	23,6	31,6	12,3
Stimme eher nicht zu (2)	5,3	5,7	1,2
Stimme überhaupt nicht zu (1)	1,2	1,1	0,2
Index des Effektivitätsbewusstseins			
Summe der beiden Indikatoren, dichotomisiert			
Geringes EB (2-7)	53,5	60,5	30,3
Hohes EB (8-10)	46,5	39,5	69,7