methoden daten analysen
ZEITSCHRIFT FÜR EMPIRISCHE SOZIALFORSCHUNG

mda

2009, Jahrgang 3, Heft 2

Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung

Die Zeitschrift wird herausgegeben von GESIS – Leibniz-Institut für Sozialwissenschaften.

Redaktion: Paul **Lüttinger**
GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Tel.: 0621 – 1246-268
E-Mail: mda@gesis.org
Internet: www.gesis.org/MDA/

Die MDA deckt alle Fragestellungen aus dem Bereich der Empirischen Sozialforschung ab, insbesondere aus dem Bereich der Umfragemethodik. Im Vordergrund stehen Artikel, die die methodischen und/oder statistischen Kenntnisse der Profession erweitern, sowie Beiträge, die sich mit der Anwendung der Methoden der Empirischen Sozialforschung in der Forschungspraxis beschäftigen, oder solche, in denen ein statistisches Verfahren exemplarisch angewandt wird. Obwohl der Schwerpunkt auf Umfragemethoden liegt, sind Beiträge zu anderen methodischen Bereichen willkommen.

Alle Beiträge, die zur Veröffentlichung in der MDA eingereicht werden, werden von mindestens zwei unabhängigen Gutachtern blind begutachtet.

Der Nachdruck von Beiträgen ist nach Absprache möglich. Die MDA erscheint zweimal im Jahr und steht als Printversion und online zur Verfügung. Die Registrierung für den Bezug der MDA erfolgt über die Web-Seiten von GESIS:

http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/

Druck: Concordia-Druckerei König oHG, Mannheim-Sandhofen
Gedruckt auf chlorfrei gebleichtem Papier.

ISSN 1864-6956

3. Jahrgang 2009

Inhalt

REZENSIONEN

ANKÜNDIGUNG

Editorial

Liebe Leserin, lieber Leser,

dieses Heft beschließt den dritten Jahrgang der MDA. In diesen drei Jahren wurden 26 wissenschaftliche Artikel veröffentlicht und 22 Bücher besprochen. Dabei waren über 120 Personen als Autoren bzw. Autorinnen oder als Gutachter bzw. Gutachterinnen beteiligt. Wie Tabelle 1 zeigt, hat sich die Zahl der eingereichten Beiträge im Vergleich zum letzten Jahr positiv entwickelt. Statt 14 sind uns bis Ende Oktober 2009 bereits 23 Manuskripte zur Veröffentlichung angeboten worden. Dies sind zwar noch etwas weniger als für den ersten Jahrgang, allerdings zeigt die hohe Ablehnungsquote für 2007 auch, dass viele dieser Beiträge den Ansprüchen der Herausgeber und Gutachter nicht gerecht wurden.

Tabelle 1 Manuskripteingang und Verbleib, 2007-2009

	Eingegangen	Angenommen	Abgelehnt	Noch offen	Annahmequote
2007	28	9	19	-	32 %
2008	14	8	6	-	57 %
2009	23	9	10	4	
Gesamt	65[a)]	26	35	4	

a) Davon 10 Manuskripte in englischer Sprache (4 angenommen, 5 abgelehnt, 1 offen).

Die in diesem Heft veröffentlichten Arbeiten bieten wieder einen vielfältigen Einblick in den aktuellen Stand der akademischen Umfrageforschung. Jan Marcus untersucht in seinem Beitrag „Der Einfluss von Erhebungsformen auf den Postmaterialismus-Index", ob der Anteil der Postmaterialisten systematisch zwischen verschiedenen Erhebungsprogrammen in Deutschland variiert. Dazu führt er eine Metaanalyse von 99 deutschen Bevölkerungssurveys durch, die zeigt, dass sich der Anteil der Postmaterialisten unter Kontrolle von Kovariaten überzufällig zwischen den Surveys unterscheidet. Marcus führt die beobachteten Unterschiede u. a. auf unterschiedliche Erhebungsmodi, unterschiedliche Erhebungsinstitute und unterschiedliche Stichprobenpläne zurück.

Im zweiten Beitrag untersucht Sven Stadtmüller die „Erwünschten und unerwünschten Effekte monetärer Anreize bei postalischen Befragungen". Der Autor fragt, ob monetäre Anreize neben der Rücklaufquote auch die Rücklaufgeschwindigkeit und die Zusammensetzung der realisierten Stichprobe beeinflussen. Die Ergebnisse Stadtmüllers sind ermutigend: Belohnungen sparen Zeit und Geld, da sie neben der Rücklaufquote auch die Rücklaufgeschwindigkeit erhöhen. Für unerwünschte Effekte monetärer Anreize findet er dagegen keine Belege.

Auf besonderes Interesse dürfte der dritte Beitrag dieses Heftes „Zur Analyse von Wahlergebnissen in Parteienhochburgen unter Berücksichtigung von Regressionsphänomenen" stoßen. Die Autoren, Thomas Ostermann und Rainer Lüdtke, untersuchen, inwieweit es sich bei den Verlusten, die alle großen demokratischen Parteien bei der Bundestagswahl 2005 und der Landtagswahl in Hessen 2008 in ihren Hochburgen erlitten haben, um ein Regressionsphänomen zur Mitte handelt. Entsprechende statistische Verfahren und Modelle werden in diesem Zusammenhang vorgestellt, angewandt und diskutiert.

Der letzte Beitrag unter der Rubrik „Forschungsberichte" widmet sich der „Entwicklung einer neuen fehlertoleranten Methode bei der Verknüpfung von personenbezogenen Datenbanken unter Gewährleistung des Datenschutzes". Rainer Schnell, Tobias Bachteler und Jörg Reiher setzen dabei an der immer häufiger von der Forschung verlangten Verknüpfung unterschiedlicher Datenbasen an. Dabei müssen die Identifikatoren der jeweiligen Einheiten (Personen, Betriebe etc.) aus Datenschutzgründen vor der Zusammenführung oftmals verschlüsselt werden. Vor besondere Anforderungen wird eine entsprechende Verschlüsselung durch fehlerhafte Identifikatoren gestellt (Wolff statt Wolf, Christoph statt Christof etc.). Die Autoren stellen ein neues Verfahren vor, das trotz starker Verschlüsselung Fehler in den Identifikatoren toleriert und ähnlich gute Ergebnisse erbringt wie unverschlüsselte Identifikatoren.

Wie auch in den vergangenen Heften haben wir in dieser Ausgabe der MDA einen Bericht aus der Praxis. Dieses Mal berichten Achim Koch, Annelies G. Blom,

Ineke Stoop und Joost Kappelhof unter dem Titel „Data Collection Quality Assurance in Cross-National Surveys: The Example of the ESS" über die im European Social Survey vorgegebenen Qualitätsstandards. Dabei konzentrieren sich die Autoren auf die Qualitätsstandards für die Datenerhebung und deren Überwachung. Diese Standards zielen darauf, die Vergleichbarkeit der Daten über alle im ESS erhobenen Ländersamples hinweg zu gewährleisten.

Zum ersten Mal in der kurzen Geschichte der MDA hat ein früher veröffentlichter Beitrag zu einem Kommentar geführt, den wir selbstverständlich gemeinsam mit einer Replik der Autoren abdrucken. Den Abschluss dieses Heftes bildet wie immer ein umfangreicher Rezensionsteil, der dieses Mal fünf Besprechungen umfasst.

Im Namen der Herausgeber danke ich allen Autoren, Rezensenten und Gutachtern, die an diesem Jahrgang mitgewirkt haben. Außerdem gilt mein Dank der Redaktion der MDA, insbesondere Paul Lüttinger und Christa von Briel, ohne deren tatkräftige Unterstützung kein Heft der MDA erschienen wäre.

Mannheim, 1. Dezember 2009 Christof Wolf

Der Einfluss von Erhebungsformen auf den Postmaterialismus-Index

The Effect of Survey Methods on the Postmaterialism Index

Jan Marcus

Zusammenfassung
Der vorliegende Beitrag ist eine Metaanalyse von 99 deutschen Bevölkerungssurveys aus dem GESIS-Datenarchiv, in denen die postmaterialistische Einstellung der Befragten ermittelt wird. Es wird anhand von Mehrebenenmodellen gezeigt, dass sich der Anteil der Postmaterialisten unter Kontrolle von Geburtsjahr der Befragten und Erhebungsjahr des Surveys überzufällig zwischen den Surveys unterscheidet. Diese Differenzen sind zum Teil auf unterschiedliche Formen der Datenerhebung zurückzuführen. Insbesondere die ALLBUS- und Eurobarometer-Serien differieren stark. Zudem sind ein Institutseffekt und ein Effekt durch Quotenstichproben auszumachen.

Abstract
This article reports on a meta-analysis of 99 German population surveys from the GESIS Data Archive, all containing the postmaterialism question. Using multi-level modelling, it can be shown that the share of postmaterialists significantly varies between the surveys when controlling for respondents' year of birth and survey year. These differences can partly be attributed to varying forms of data collection. In particular the ALLBUS and Eurobarometer surveys differ strongly. In addition, house effects and an effect of quota sampling can be found.

1 Einleitung[1]

Ronald Ingleharts Postmaterialismus-Theorie ist eine der bekanntesten sozialwissenschaftlichen Theorien. Sie hat nicht nur im akademischen Kontext eine intensive Debatte ausgelöst, sondern genießt auch außerhalb der Wissenschaft einen vergleichsweise hohen Bekanntheitsgrad. Allerdings gibt es eine Reihe von Postmaterialismus-Studien, die nicht mit dieser Theorie übereinstimmen bzw. sich gegenseitig widersprechen. Wie Böltken und Jagodzinski (1984: 70) aufschlüsseln, kann dies daran liegen, dass die Theorie empirisch falsch, das Messinstrument nicht angemessen, oder die Datenbasis verzerrt ist. Da eine zuverlässige Datenlage unabdingbar ist, um die Gültigkeit der Theorie zu prüfen, wirkt es um so erstaunlicher, dass die Datenbasis, die den vielen Analysen zum Postmaterialismus zugrunde liegt, nur äußerst selten Gegenstand eingehender Untersuchung gewesen ist.

Ausnahmen hiervon bilden zwei Aufsätze aus den 1980er Jahren (Böltken/Gehring 1984; Krebs/Hofrichter 1989), deren Autoren beobachten, dass sich die Anteile der Postmaterialisten zwischen verschiedenen zeitnahen Erhebungen in Deutschland erheblich unterscheiden. Die vorliegende Metaanalyse greift Hypothesen über den Einfluss von Erhebungsformen, welche in den beiden Aufsätzen aufgestellt wurden, auf und entwickelt die dort geäußerten Ideen fort. Während in diesen Aufsätzen wenige Studien untersucht werden und primär deskriptiv vorgegangen wird, soll in diesem Beitrag anhand von 99 Surveys aus dem GESIS-Datenarchiv der mögliche Einfluss von Erhebungsmethoden auf den Postmaterialismus-Index mit anspruchsvolleren Mitteln der Datenanalyse untersucht werden.

Der vorliegende Artikel stellt zunächst das Konzept Postmaterialismus und den Index dar, mit dem Postmaterialismus empirisch ermittelt wird. Weiterhin werden verschiedene Hypothesen über den Einfluss von Erhebungsformen auf den Postmaterialismus-Index abgeleitet. Der darauf folgende Abschnitt erläutert die Kriterien, nach denen Surveys und Befragungspersonen in den Surveys ausgewählt wurden. Der Test der Hypothesen erfolgt anschließend mittels eines zweistufigen Mehrebenenmodells.

1 Die Daten, die in diesem Beitrag benutzt werden, wurden durch GESIS zugänglich gemacht. Sie wurden von GESIS für die Analyse aufbereitet und dokumentiert. Eine Übersicht über die verwendeten Studien findet sich in Tabelle A1 im Anhang. Für das Bereitstellen der Datensätze sei Markus Cziesla vom GESIS-Datenarchiv gedankt.
Weiterhin gilt mein Dank Rainer Schnell für die Anregung zur Beschäftigung mit diesem Thema sowie Elena Engelhardt, Dominic Fritz, Nicolas Grießhaber, Thomas Hinz, Peter Selb, Zacharias Ziegelhöfer und zwei anonymen Gutachtern für hilfreiche Anmerkungen zu früheren Versionen dieses Artikels.

2 Postmaterialismus: Konzept und Operationalisierung

In einer kaum mehr zu überschauenden Masse an Büchern, Artikeln und Sammelbandbeiträgen hat Ronald Inglehart (u. a. 1971, 1977, 1981) seine Theorie des Postmaterialismus dargelegt und erläutert. Da es sich bei dem vorliegenden Beitrag um eine primär methodisch orientierte Arbeit handelt, soll hier lediglich kurz auf diese Theorie eingegangen werden.

Ingleharts Theorie des Wertewandels, die auf einer Knappheits- und einer Sozialisationshypothese basiert, geht davon aus, dass Generationen, die in Zeiten materieller Sicherheit aufgewachsen sind, eher postmaterialistische Wertorientierungen (wie Selbstverwirklichung und gesellschaftliche Partizipation) entwickeln. Materielle Güter sind nämlich reichlich vorhanden, während postmaterialistische Güter knapp sind. Generationen, die nicht in Zeiten wirtschaftlichen Wohlstandes aufgewachsen sind, tendieren hingegen eher zu materialistischen Werten (wie Streben nach Sicherheit und ökonomischem Wohlstand).

Jüngere Menschen sollten – nach Ingleharts Theorie – tendenziell eher Postmaterialisten sein, weil es seit dem Zweiten Weltkrieg in den westlichen Industrienationen einen (fast) durchgängigen wirtschaftlichen Aufschwung gab.

Da eine gute Theorie nicht nur plausibel erscheinen, sondern auch empirischen Prüfungen standhalten sollte, muss das Konzept Postmaterialismus beobachtbar gemacht werden. Als Operationalisierung schlägt Inglehart (1971: 994) einen Index vor, den er aus der folgenden Frage ableitet:[2]

> „If you had to choose among the following things, which are the two that seem most desirable to you?
> - Maintaining order in the nation.
> - Giving people more say in important political decisions.
> - Fighting rising prices.
> - Protecting freedom of speech."

Befragte, die die Items 1 und 3 nennen, werden als Materialisten bezeichnet; werden die Items 2 und 4 ausgewählt, so handelt es sich bei den Befragten um Postmaterialisten. Personen, die sowohl ein postmaterialistisches als auch ein materialistisches Item auswählen, werden in Mischkategorien eingeordnet. Diese Kategorien werden häufig in eine ordinale Folge gebracht und die entstandene Skala als Postmaterialismus-Index bezeichnet. Der Begriff Postmaterialismus-Index

2 Dargestellt ist die englische Originalversion, da es bei den deutschen Formulierungen Unterschiede gibt, auf die weiter unten näher eingegangen wird. Anzumerken ist jedoch, dass auch im Englischen Varianten existieren, die in den Formulierungen von der Originalversion abweichen (Davis/Davenport 1999: 650f.).

wird allerdings nicht immer auf die gleiche Art verwendet: Es werden sowohl die Zusammenfassung der o. g. Items als auch die daraus resultierende Ordinalskala als Postmaterialismus-Index bezeichnet.[3] Wenn im Folgenden von Postmaterialismus-Index die Rede ist, so ist damit die Konstruktion einer Postmaterialisten-Variablen aus den Items 2 und 4 gemeint. Für diese Ausarbeitung wird angenommen, dass der Postmaterialismus-Index ein valides Messinstrument für das theoretische Konstrukt Postmaterialismus darstellt.[4]

3 Hypothesen über den Einfluss diverser Erhebungsformen

In ihren Aufsätzen zeigen Böltken und Gehring (1984) und Krebs und Hofrichter (1989), dass der Anteil der Postmaterialisten in Deutschland sich zwischen verschiedenen zeitnahen Erhebungen deutlich unterscheidet. Die Autoren thematisieren jeweils verschiedene Erhebungsformen, die einen Einfluss auf den Anteil der Postmaterialisten haben könnten. Im Folgenden werden diese Ideen aufgegriffen und weiter entwickelt.

Dazu werden verschiedene Erhebungsformen, also Details der Sammlung von Survey-Daten, dargestellt und ihr möglicher Einfluss auf den Postmaterialismus-Index herausgearbeitet. Es werden nur Erhebungsformen berücksichtigt, die bei den für diese Studie ausgewählten Surveys zur Anwendung kommen. Beispielsweise wird der mögliche Einfluss von telefonischen Interviews hier nicht diskutiert, da alle 99 Surveys als face-to-face Interviews durchgeführt wurden. Weiterhin werden nur die Auswirkungen von Erhebungsformen auf postmaterialistische Wertprioritäten untersucht und nicht auf materialistische und gemischte. Dies geschieht einerseits, um die Modelle möglichst einfach zu halten, und andererseits, da schlechter theoretisch ableitbar ist, ob ein – und wenn ja, welcher – Einfluss der verschiedenen Erhebungsformen auf die anderen Wertprioritäten zu erwarten ist.

Viele der in diesem Beitrag präsentierten Hypothesen fußen auf der Prämisse der schwierigeren Erreichbarkeit von Postmaterialisten. Daher wird diese zunächst theoretisch begründet und empirisch geprüft: Postmaterialisten weisen aufgrund ihres tendenziell höheren Bildungsabschlusses (Böltken/Gehring 1984: 46; Schnell 1993: 27; Kroh 2008: 484) und des damit verbundenen tendenziell größeren Einkommens vermutlich eine höhere Mobilität auf. Das kann sich z. B. in längeren Arbeitszeiten, häufigeren arbeitsbedingten Ortswechseln oder der verstärkten

3 Ein Index entsteht durch das Zusammenfassen von zwei oder mehr Einzelindikatoren zu einer neuen Variablen.

4 Die Validität wurde verschiedentlich angezweifelt (z. B. Davis/Davenport 1999; Kohler 1998).

Möglichkeit der Freizeitgestaltung außerhalb des eigenen Hauses äußern. All diese Faktoren erschweren es, Postmaterialisten bei einer Befragung anzutreffen. Weiterhin soll mit dem Postmaterialismus ein Wertewandel einhergehen, der weg von der patriarchalischen Gesellschaftsform geht. Somit steht zu vermuten, dass unter den Postmaterialisten mehr Haushalte mit berufstätigen Frauen zu finden sind. Das wiederum macht die Haushalte schwieriger zu erreichen (Schnell 1997: 219). Auch empirische Ergebnisse sprechen dafür, dass Postmaterialisten seltener anzutreffen sind: Für den ALLBUS 1980 (Schnell 1993: 24f.) und den ALLBUS 2000 (Schnell 2002) wurde dies bereits anhand von Grafiken gezeigt.

Im Folgenden wird die Prämisse der schwereren Erreichbarkeit der Postmaterialisten anhand dieser ALLBUS-Studien und weiterer Surveys überprüft, die Angaben zur Anzahl der Kontaktversuche enthalten.[5] Dazu wird der von Cuzick (1985) entwickelte nicht-parametrische Test auf Trend über angeordnete Gruppen verwendet. Diejenigen Befragten, bei denen das Interview nach der gleichen Anzahl von Kontaktversuchen stattfand, werden jeweils zu einer Gruppe zusammengefasst und als von den anderen Gruppen unabhängige Stichprobe betrachtet. Die Gruppen werden anschließend entsprechend ihrer Kontaktzahl in aufsteigender Reihenfolge sortiert. Fasst man die Befragungspersonen der genannten fünf Surveys zusammen, so ergibt sich für die Teststatistik Z ein Wert von $z = 8{,}23$ und damit ein auf dem 1 %-Niveau signifikanter, steigender Trend der Postmaterialismus-Variablen über die Anzahl der Kontaktversuche.

3.1 Stichprobendesign

In den folgenden Unterabschnitten werden für den Postmaterialismus-Index möglicherweise relevante Faktoren betrachtet, die mit dem Stichprobendesign zusammenhängen, also mit der Art und Weise, wie die Befragungspersonen ausgewählt wurden.

Quotenstichproben

Quotenstichproben sind bewusste Auswahlen, bei denen die Interviewer bei der Selektion der Befragten Quoten bestimmter Merkmale erfüllen müssen, deren Verteilung in der Population aufgrund von Referenzstatistiken bekannt ist. So verwendet z. B. das Institut für Demoskopie in Allensbach für den deutschen Teil der European Values Study (1981) Alter, Geschlecht und Beruf als Quotierungsmerkmale; als

5 Das sind die Studien Außenpolitische Einstellungen (1992), Politische Einstellungen (1995) und Persönlichkeit und Wahlverhalten (2003).

Referenzstatistik dient hier der Zensus. Quotenstichproben werden häufig in der Markt- und Meinungsforschung angewandt, da sie sich als billiger im Vergleich zu Zufallsauswahlen erwiesen haben. Allerdings sind sie mit einer ganzen Reihe von Problemen behaftet.

Erstens wird – meist implizit – angenommen, dass die mittels der Quotierungsmerkmale gebildeten Klassen sich hinsichtlich der interessierenden Variable(n) nicht unterscheiden (Schnell 1993: 29). Auf den konkreten Fall übertragen, würde das bedeuten, dass z. B. alle 18- bis 25-jährigen Studentinnen Postmaterialisten wären.

Zweitens können Verzerrungen dadurch auftreten, dass Interviewer sich Personen aus ihrem Bekanntenkreis auswählen (Schnell/Hill/Esser 2005: 303f.) oder „sich möglichst in der (eigenen) Mittelschicht nach aufgeschlossenen, informierten und auskunftswilligen Interviewpartnern umsehen" (Böltken/Gehring 1984: 44). Es wurde bereits mehrfach gezeigt, dass höher Gebildete stärker zum Postmaterialismus tendieren (z. B. Böltken/Gehring 1984: 46; Küchler 1984: 227). Da Postmaterialismus somit mit der gesellschaftlichen Schicht zusammenhängt, wäre zu erwarten, dass Quotenstichproben einen höheren Anteil von Postmaterialisten produzieren.

Drittens haben Personen, die leichter zu erreichen sind, eine erhöhte Auswahlwahrscheinlichkeit. Es wäre demnach zu erwarten, dass bei Quotenstichproben weniger Postmaterialisten befragt werden, weil sie – wie oben begründet – seltener zu Hause anzutreffen sind.

Bei Betrachtung der genannten Probleme steht zu vermuten, dass Quotenstichproben andere Ergebnisse hinsichtlich Postmaterialismus liefern als Zufallsauswahlen. Da es sowohl Argumente dafür gibt, dass durch das Quotaverfahren Postmaterialisten bevorzugt ausgewählt werden, als auch dafür, dass weniger Postmaterialisten befragt werden, und nicht klar ist, ob diese Effekte sich aufheben oder ob ein Effekt den anderen überlagert, ist die folgende Hypothese ungerichtet formuliert.

H1: *Die Wahrscheinlichkeit einen Postmaterialisten zu befragen unterscheidet sich zwischen Quoten- und Zufallsauswahlen.*

Die Informationen bezüglich des Stichprobendesigns der in diesem Beitrag verwendeten Studien stammen aus der Kurzbeschreibung des jeweiligen Surveys im Datenbestandskatalog sowie den Codebüchern und Methodenberichten zu den Studien, sofern sie im GESIS-Datenarchiv vorliegen.[6]

6 Über die Studie Wählerverhalten (1990) heißt es im Datenbestandskatalog, dass keine näheren Angaben über das Auswahlverfahren vorliegen. Um diese Studie nicht aus den Analysen fallen zu lassen, wurde eine Modus-Imputation vorgenommen. Das heißt, es wird angenommen, dass dieser Studie ein ADM-Design zugrunde liegt.

Stichproben aus Einwohnermelderegistern

Bei der Auswahl für face-to-face Bevölkerungsbefragungen kommen in Deutschland in der Regel nur zwei Zufallsstichproben-Designs vor: Zum einen auf Einwohnermelderegistern basierende Zufallsauswahlen und zum anderen verschiedene Versionen des Random Walks (Hoffmeyer-Zlotnik 2006: 19). Beide Verfahren stellen mehrstufige, geschichtete Zufallsauswahlen dar. Bei ersterem werden auf der ersten Stufe zunächst Gemeinden als Primary Sampling Units (PSUs) ausgewählt und anschließend Personen aus dem Einwohnermelderegister der jeweiligen Gemeinde. Random Walk ist ein dreistufiges Verfahren, bei dem auf der ersten Stufe synthetische Stimmbezirke[7] als PSUs, auf der zweiten Stufe mittels Random Walk Haushalte und auf der dritten Stufe mittels Schweden-Schlüssel[8] Personen ausgesucht werden.

Diese auf den ersten Blick eher technisch anmutenden Differenzen können durchaus einen Einfluss auf den Postmaterialismus-Index haben. Bei Random-Walk-Stichproben können die Interviewer ihren Spielraum entgegen der Vorschriften nutzen, um das Interview in einem per Random-Walk-Anweisung vorgeschriebenen – aber nicht erreichten – Haushalt oder mit derjenigen Person durchzuführen, die sie zwar im vorgeschriebenen Haushalt antreffen, aber die nicht die mittels Schweden-Schlüssel vorgegebene ist (Schnell 1997: 59). Dass dies nicht nur eine theoretische Möglichkeit ist, sondern von den Interviewern auch sehr wahrscheinlich genutzt wird, hat Hoffmeyer-Zlotnik (2006) gezeigt. Durch Stichproben aus Melderegistern wird dem Interviewer nicht nur genau eine Person vorgeschrieben, die zu interviewen ist, sondern es kann auch kontrolliert werden, ob exakt diese Person befragt wurde.

Da, wie zuvor erläutert, Postmaterialisten schwieriger zu erreichen sind, kann folgende gerichtete Hypothese aufgestellt werden:

> H2: *In Stichproben, die aus Einwohnermelderegistern gezogen werden, ist die Wahrscheinlichkeit, ein Interview mit einem Postmaterialisten zu realisieren höher als in Stichproben nach dem Random-Route Verfahren.*

Obwohl in den 1950er und 1960er Jahren viele Stichproben auf Basis von Melderegistern gezogen wurden, fand dies in den folgenden Jahrzehnten aufgrund rechtlicher, technischer und vor allem finanzieller Probleme nur noch äußerst selten statt (Schnell 1997: 59). Daher basieren von den hier verwendeten Studien nur einige

7 Kleinere Stimmbezirke werden vor der Begehung zusammengefasst.

8 Dieses auf Kish (1965) zurückgehende Verfahren dient der Auswahl genau einer Befragungsperson aus einem bereits selektierten Haushalt und wird bei Hoffmeyer-Zlotnik (2006: 23f.) detailliert erläutert.

ALLBUS-Studien auf Einwohnermelderegister-Stichproben. Das Ziehen einer Stichprobe aus dem Einwohnermelderegister wurde zwar schon für den ALLBUS 1980 überlegt, jedoch aus Kostengründen nicht durchgeführt (Kirschner 1984: 118). Erstmals wurde diese Praxis für den ALLBUS 1994 angewandt, musste jedoch wegen Finanzierungsproblemen 1998 vorübergehend eingestellt werden. Seit 2000 basieren alle ALLBUS-Untersuchungen auf Stichproben aus dem Melderegister.

Random-Walk-Stichproben

Wie oben erwähnt, gibt es verschiedene Random-Walk-Verfahren, mittels derer Haushalte auf der zweiten Stufe ausgewählt werden. Diese Verfahren unterscheiden sich insbesondere hinsichtlich des Spielraums, den die Interviewer haben, um die Haushalte und damit die Befragungspersonen auszusuchen. Hoffmeyer-Zlotnik (2006: 25f.) zeigt, dass diese unterschiedlichen Spielräume während verschiedener Arbeitsschritte entstehen können:

- **Erhebung:** Bei manchen Verfahren erfolgt die Identifikation des Zielhaushalts und die Befragung der Zielperson in zwei getrennten Schritten durch zwei verschiedene Personen. Dieses Verfahren wird u. a. als „Adress-Random" (Blohm 2006: 40; Koch 2002: 13) oder „Random Route mit Adress-Vorlauf" (ALLBUS-Methodenbericht 1998: 5) bezeichnet. Bei anderen Erhebungen erfolgt beides in einem Schritt. Hier hat der Interviewer einen größeren Spielraum, da nicht kontrolliert werden kann, ob der befragte Haushalt tatsächlich ein Zielhaushalt ist.
- **Vorgabe:** Bei manchen Erhebungen wird dem Interviewer lediglich eine Nettozahl vorgegeben, das heißt eine Sollzahl an zu realisierenden Interviews entlang der Random-Walk-Strecke, und bei anderen Verfahren eine Bruttozahl, also eine festgelegte Anzahl anzulaufender Haushalte. Letzteres Verfahren wird u. a. als „klassisches Random Route" bezeichnet, ersteres auch als „Standard-Random" (Koch 1997: 109) oder „vereinfachtes Random Route" (Schnell 1997: 14). Wird nur eine Nettozahl vorgegeben, werden leicht erreichbare Haushalte bevorzugt.
- **Nachbearbeitung:** Random-Walk-Erhebungen unterscheiden sich auch dahingehend, ob eine Nachbearbeitung erfolgt, wenn nicht genügend Interviews realisiert wurden. Dies war z. B. beim ALLBUS 1986 und beim Wohlfahrtssurvey 1998 der Fall, wie aus den entsprechenden Methodenberichten hervorgeht. Wird eine Nachbearbeitung durchgeführt, werden mehr Haushalte erreicht, die als schwerer erreichbar gelten.
- **Protokoll:** Wenn der Interviewer alle Kontaktversuche und kontaktierten Haushalte genau protokollieren muss, bestehen weniger Möglichkeiten den Spielraum unerlaubter Weise zu vergrößern und leichter erreichbare Haushalte zu befragen.

Diese unterschiedlich großen Ermessensspielräume des Interviewers können zu Verzerrungen führen, da die Interviewer dazu tendieren, leichter erreichbare Haushalte und Personen auszuwählen (Blohm 2006: 41; Koch 1997: 106).

H3: *Je höher der Ermessensspielraum der Interviewer beim Random Walk, desto unwahrscheinlicher werden Postmaterialisten befragt.*

Leider liegen Informationen über die genaue Ausprägung des Random-Walk-Verfahrens nur für einige wenige Studien vor. Es werden nicht nur – wie weiter oben dargestellt – viele verschiedene Bezeichnungen für gleiche Verfahren benutzt, sondern es können außerdem die gleichen Bezeichnungen für Random-Walk-Verfahren Unterschiedliches beinhalten, wie Hoffmeyer-Zlotnik (2006: 19, 25) anschaulich erläutert. Studien, für die detailliertere Informationen darüber vorliegen, was sich hinter den Bezeichnungen tatsächlich verbirgt, sind einerseits die Studien aus den Serien ALLBUS und Wohlfahrtssurvey sowie andererseits die Studien Politische Resonanz (1995) und Verhinderung von Gewalt (1989). Bei der Klassifizierung wird hier Koch (2002: 13) gefolgt, der zwischen Adress-Random (getrennter Adressenvorlauf; geringster Ermessenspielraum), Random-Route (Vorgabe einer Brutto-Anzahl) und Standard-Random (Vorgabe einer Nettozahl; höchster Ermessenspielraum) unterscheidet.

3.2 Fragebogentext

In diesem Abschnitt werden Unterschiede in der deutschen Variante des Fragebogentexts zum Postmaterialismus-Index aufgezeigt und mögliche Auswirkungen auf die postmaterialistische Einstellung der Befragten hypothetisch formuliert. Zunächst wird die einleitende Fragestellung und dann die Formulierung der vier Items verglichen.

Fragestellung

Die Formulierung der einleitenden Postmaterialismus-Fragestellung unterscheidet sich zwischen den hier untersuchten Studien. Darauf weisen sowohl Böltken und Gehring (1984) als auch Krebs und Hofrichter (1989) hin. Tabelle A2 im Anhang listet alle in den untersuchten Studien vorgefundenen Fragestellungen auf.

Bei ihrer Analyse differenzieren Böltken und Gehring (1984) zwischen Postmaterialismus-Fragestellungen, bei denen die Befragten eher als Experten angesprochen werden und Fragestellungen, bei denen eher auf die persönlichen Präferenzen der Interviewten abgezielt wird. Sie stellen insbesondere die Version 1 den Versionen 2 bis 4 gegenüber. Bei ersterer werde eine „gewisse Beliebigkeit von

Politik-Zielen unterstellt" (Böltken/Gehring 1984: 50), was sich insbesondere in den wiederkehrenden Ausdrücken *man* und *kann* äußert. Da anschließend die Frage nach dem *persönlich* wichtigsten Ziel gestellt wird, kann vermutet werden, dass hier eher persönliche Präferenzen erfragt werden (Böltken/Gehring 1984: 50). Dagegen werden bei den Versionen 2 bis 4 Zeit (*10-15 Jahre* bzw. *10 Jahre, langfristig* in Version 2) und Subjekt der Aktivität (Bundesrepublik) angesprochen (Krebs/Hofrichter 1989: 63), und es wird explizit auf die öffentliche Diskussion (*Eine Reihe von Diskussionen*) Bezug genommen (Böltken/Gehring 1984: 50). Es wird also nicht so sehr nach persönlichen Präferenzen gefragt, sondern der Befragte wird als „Experte" (Böltken/Gehring 1984: 50) angesprochen und um eine allgemeine politische Einschätzung gebeten. Böltken und Gehring (1984: 50) konstatieren, dass man sich „persönlich postmaterialistische Ziele attestieren ... kann [,] während man andererseits als Experte ... für die allgemeine Lage ... eher auf materialistische Prioritäten setzt." Es spricht allerdings nichts dagegen, dass das nicht auch umgekehrt der Fall sein kann. Daher wird hier Krebs und Hofrichter (1989: 63) gefolgt, die einen ungerichteten Effekt erwarten.

H4: *Frageformulierungen, die den Befragten als „Experten" ansprechen, führen zu anderen Schätzungen des Anteils von Postmaterialisten als Frageformulierungen, die den Befragten nach persönlichen Präferenzen befragen.*

Da Böltken und Gehring (1984) nur die Versionen 1 bis 4 untersucht haben, ihre Kategorisierung aber auf die anderen Fragestellungs-Versionen übertragbar ist, sind diese Versionen ebenfalls in die beiden Kategorien eingeordnet worden. Bei Version 5 (*Staatsgewalt*) und Version 6 (*Bundesrepublik*, Diskussion) werden die Befragten eher um eine allgemeine Einschätzung gebeten; während bei den Versionen 7 bis 9 stärker die persönlichen Präferenzen angesprochen werden und Verweise auf die öffentliche Diskussion, das Subjekt, das die politischen Ziele verfolgen soll, und die zeitliche Perspektive nicht vorhanden sind.[9]

9 Für die Studie Political Action (1980) liegt lediglich ein Fragebogen in englischer Sprache vor, aus dessen Formulierung der Postmaterialismus-Frage aber stark zu vermuten ist, dass es sich sowohl bei der Frage- als auch bei der Itemformulierung (siehe den folgenden Abschnitt) um Version 1 handelt. Außerdem wurden diese Versionen auch schon bei der Studie Political Action (1974) verwendet. Für alle anderen Studien sind deutsche Fragebögen im GESIS-Datenarchiv vorhanden.

Itemformulierung

Die deutschen Versionen der Itemformulierungen weisen ebenfalls Unterschiede auf. Allerdings scheinen sich all diese Varianten nur durch marginale Abweichungen zu unterscheiden. Auch Böltken und Gehring (1984: 48) und Krebs und Hofrichter (1989: 69) können keinen systematischen Effekt der Itemformulierung erkennen. Da zudem gleiche Itemversionen häufig mit den gleichen Fragestellungen eingeleitet werden, kann nicht entschieden werden, ob eine (mögliche) Beeinflussung des Postmaterialismus-Index durch die Item- oder die Frageformulierung entsteht. Daher wird hier nur der mögliche Effekt der Frageformulierung untersucht.

3.3 Institutseffekte

In den beiden folgenden Abschnitten werden – streng genommen – nicht direkt Erhebungsformen untersucht. Die beiden Variablen Survey-Serie und Erhebungsinstitut sind Globalvariablen, die für eine Reihe von Merkmalen stehen. Ihr Einfluss auf den Postmaterialismus-Index kann nicht direkt untersucht werden, da sie nicht oder nur unzureichend dokumentiert sind.

Institutseffekte entstehen dann, wenn „sich die Arbeitsweise eines Befragungsinstituts als Ganzes auf die Qualität der erhobenen Daten auswirkt" (Büchel 2000: 417). Dass das Erhebungsinstitut einen Einfluss auf die Randverteilungen interessierender Variablen haben kann, hat u. a. Schnell (1997: 98f.) festgestellt: Bei der Media-Analyse weisen bestimmte Institute einen überzufällig hohen Anteil an Nichterreichten auf. Auch Büchel (2000) zeigt systematische Unterschiede zwischen den Ergebnissen verschiedener Institute auf. Für den Postmaterialismus-Index werden ebenfalls Institutseffekte erwartet (Davis/Davenport 1999: 653ff.), sind aber bislang noch nicht näher untersucht worden.

Der Stand der Forschung im Bereich Institutseffekte ist nicht sehr fortgeschritten (Büchel 2000: 417f.; Hoffmeyer-Zlotnik 2006: 49; Schnell 1997: 21), was damit zusammenhängt, dass viele wichtige Erhebungsdetails nicht öffentlich zugänglich sind (Schnell 1997: 21) und es auch im GESIS-Datenarchiv nicht viele Datensätze gibt, zu denen eine genaue Dokumentation der Feldarbeit vorliegt (Schnell 1997: 57). Es ist nicht klar, welches Element bzw. welche Elemente zu einem Institutseffekt führen. Über das Agieren von Instituten ist nur wenig bekannt (Koch 2002: 10). Theoretisch kann ein Institutseffekt durch jedes Erhebungsdetail hervorgerufen werden, in dem sich die Institute unterscheiden. Neben den in den vorhergegangenen Abschnitten diskutierten Erhebungsformen werden in der Literatur weitere Elemente genannt, die einen Institutseffekt bedingen können: Die

Gewichtung, die Länge der Feldarbeit und die Wochentage der Befragung, außerdem die Vorgehensweise der Interviewer sowie die Mindestanzahl an Kontakten, bis eine Person als „nicht erreichbar" eingestuft wird. Darüber hinaus sind noch eine Vielzahl von weiteren Elementen denkbar, die einen Institutseffekt bewirken, wie z. B. der Ruf des Erhebungsinstituts, das Verfassen von Überzeugungs- und Ankündigungsbriefen, das vorherige Abstimmen eines Interviewtermins mit dem Befragten oder die Restriktionen für erneute Kontaktaufnahmeversuche, wenn die potentielle Befragungsperson nicht beim ersten Kontakt erreicht werden konnte (z. B. anderer Tag, andere Uhrzeit, Länge des Intervalls bis zur nächsten Kontaktaufnahme etc.). Auch die genaue Begehungsanweisung für alle weiter oben näher erläuterten Arten des Random Walks obliegt (mit Ausnahme der Media-Analyse) den Erhebungsinstituten (Hoffmeyer-Zlotnik 1997: 37).

Für die hier analysierten Surveys liegen keine bzw. nicht genügend Informationen über diese Erhebungsdetails vor. Daher werden diese Elemente nicht einzeln untersucht, sondern in der Globalkategorie „Institut". Um die Hypothese zu testen, muss unterstellt werden, dass die Arbeitsweisen der Institute über Zeit und Survey-Serie (weitestgehend) konstant bleiben, da über die einzelnen Arbeitsweisen nicht genügend Informationen vorliegen. Diese Annahme ist zwar für diese Ausarbeitung unvermeidlich, aber nicht ganz unproblematisch: Schnell (1997: 98ff.) macht für an der Media-Analyse beteiligte Institutionen Veränderungen über die Zeit im Feldprozedere aus.

Zwar unterscheiden sich Institute auch hinsichtlich ihrer Ausschöpfungsquote (Schnell 1997: 98f.), aber es wurde kein Zusammenhang zwischen Datenqualität und Ausschöpfungsquote festgestellt (Blohm 2006: 38; Büchel 2000: 418). Das liegt auch daran, dass die Ausschöpfungsquoten unterschiedlich definiert sind (Schnell/Hill/Esser 2005: 307f.). Obwohl die Ausschöpfungsquoten für einige der untersuchten Surveys vorliegen, wurde aufgrund der genannten Gründe eine Analyse ihrer Auswirkung auf den Postmaterialismus-Index nicht untersucht.

Es ist zu vermuten, dass von all den genannten Elementen vor allem diejenigen einen Einfluss auf den Postmaterialismus-Index haben, die die Erreichbarkeit von Befragungspersonen betreffen, da Postmaterialisten schwieriger zu erreichen sind; also insbesondere die Wochentage, an denen eine Befragung stattfinden kann, die Mindestkontaktanzahl, die Restriktionen für eine erneute Kontaktaufnahme und die vorherige Abstimmung eines Interviewtermins. Weil zu diesen Erhebungsformen jedoch fast keine Informationen bei den untersuchten Surveys vorliegen, ist nicht klar, welche Institute welche dieser Erhebungsformen verwenden, und somit kann keine Vermutung im Vorfeld darüber angestellt werden, welches Institut den Postmaterialismus-Index in welcher Weise beeinflusst. Daher ist die folgende Hypothese ungerichtet.

H5: *Die Datenerhebungsstrategie der Umfrageinstitute beeinflusst das Ergebnis der Schätzung des Anteils von Postmaterialisten.*

Für den Test dieser Hypothese werden Dummyvariablen für diejenigen Institute gebildet, die mehr als zehn (entspricht ca. 10 %) der untersuchten Studien durchgeführt haben (das sind Emnid, GFK-Getas, Infratest und Sample/INRA), damit einzelne Zellen nicht zu dünn besetzt sind und einzelne Surveys nicht ein zu großes Gewicht bekommen. Zwar haben sich viele Marktforschungsinstitute gegenseitig aufgekauft bzw. sind miteinander fusioniert (z. B. ist IPSOS aus dem Zusammenschluss von GFK, Getas und INRA hervorgegangen),[10] aber dennoch werden alle Institute einzeln betrachtet. Ausnahmen sind zum einen das Sample Institut in Mölln und INRA, die zusammengefasst werden, da INRA 1996 aus dem Sample Institut hervorgegangen ist und es sich somit nur um eine Namensänderung handelt.[11] Zum anderen werden GFK-Getas und Getas zusammengefasst, da hier nur ein Wechsel von Namen und Standort vorliegt (ALLBUS-Methodenbericht 1988: 52), sowie die verschiedenen Infratest-Namen (Infratest Sozialforschung, Infratest Wirtschaftsforschung, Infratest Burke etc.) zu Infratest.

3.4 Effekte der Survey-Serie

Als Survey-Serie werden hier Gruppen von mindestens fünf Surveys, die denselben Namen tragen und die einen Zeitraum von mindestens zehn Jahren abdecken, bezeichnet. Das trifft auf die Serien Eurobarometer, Wohlfahrtssurvey und ALLBUS zu. Ähnlich wie der Institutseffekt ist auch der Effekt der Survey-Serie eine Globalvariable, hinter der sich viele Faktoren verbergen. Ihre genaueren Ausprägungen sind nicht bekannt, weil sie in den Datendokumentationen nicht enthalten sind. Im vorhergehenden Abschnitt wird unterstellt, dass es bestimmte institutsspezifische Faktoren gibt, die einen Einfluss auf den Postmaterialismus-Index haben könnten. Hier wird angenommen, dass es solche Faktoren gibt, die „typisch" für eine bestimmte Survey-Serie sind.

Das können ähnliche Faktoren sein, die auch zum Institutseffekt führen, z. B. ist es möglich, dass die Länge der Feldarbeitsperiode oder das Senden von Ankündigungsbriefen vor dem ersten Kontaktversuch sich als spezifisch für eine Survey-Serie erweisen (und nicht unbedingt für das Institut). Zu vermuten ist, dass auch Ruf und Bekanntheit einer Survey-Serie unterschiedlich sind. Wie für

10 Siehe http://www.ipsos.de/default.asp?c=100 (08.04.2009).
11 Siehe http://www.ipsos.com/news/releases/2002/052802.aspx (08.04.2009).

die Institute gilt auch für die Survey-Serie, dass insbesondere Differenzen, die das Erreichen von potentiellen Befragungspersonen betreffen, einen Einfluss auf den Postmaterialismus-Index haben.

Falls der Preis einer Studie – unter Konstanthaltung von Fallzahl und Interviewmethode – ein guter Indikator für die Güte einer Studie ist, so sollten höhere Kosten der Survey-Serie ebenfalls einen positiven Effekt auf die Postmaterialismus-Variable haben: Es ist zu vermuten, dass bei teuren Surveys mehr für das Antreffen schwer Erreichbarer getan wird. Insbesondere ist aber zu erwarten, dass die vorgegebene Mindestkontaktzahl einen einflussreichen Faktor darstellt. Hier scheinen Unterschiede zwischen den Survey-Serien zu bestehen: Während über die Eurobarometer-Serie vermutet wird, dass ein Haushalt, der zweimal nicht erreicht wurde, als Non-Response eingestuft wird (Moschner 2008),[12] kann man aus dem (in den ALLBUS-Methodenberichten dokumentierten) recht geringen Anteil an Haushalten und Zielpersonen, die gar nicht angetroffen wurden, sowie aus der hohen Kontaktversuchsanzahl schließen, dass beim ALLBUS eine deutlich höhere Mindestkontaktzahl existiert.

Sowohl ALLBUS als auch Wohlfahrtssurvey besitzen eine stärker wissenschaftliche Ausrichtung (Koch 1997: 112) und hohe Qualitätsansprüche (Blohm 2006: 37, 39). Es kann daher vermutet werden, dass hier mehr Wert auf die Methodik gelegt wird. Für den ALLBUS ist dies auch daran zu erkennen, dass deutlich mehr Erhebungsdetails dokumentiert sind als für andere Studien: Auf den Internetseiten der GESIS sind z. B. alle Dokumentationen und Datensätze der ALLBUS-Serie für wissenschaftliche Zwecke frei verfügbar, während für die Eurobarometer nur spärliche, länderspezifische Informationen über Details der Erhebung vorhanden sind: So schreiben Saris und Kaase (1997: 21) über den von ihnen herausgegebenen Sammelband, dass er die detaillierteste Beschreibung von Erhebungsformen der Eurobarometer-Serie sei. Der Informationsgehalt dieser Aufsatzsammlung kommt jedoch bei weitem nicht an die ALLBUS-Methodenberichte heran.

H6: *Der Postmaterialismus-Index unterscheidet sich zwischen den Survey-Serien. Wenn der Survey Teil der ALLBUS- oder Wohlfahrtssurvey-Serie ist, dann ist die Wahrscheinlichkeit größer, dass Postmaterialisten befragt werden, als bei Eurobarometer-Surveys.*

Da ALLBUS und Wohlfahrtssurvey sowohl mehr Wert auf die Methodik im Allgemeinen als auch auf das mögliche Erreichen aller Befragten im Speziellen legen, steht zu vermuten, dass in beiden Serien mehr Personen aus der Gruppe der schwer

12 Das kann zumindest durch Schubert und Greil (1997: 25) für den Eurobarometer 41.0 bestätigt werden.

Erreichbaren befragt werden als bei den Eurobarometer-Untersuchungen. Weil dies jedoch eine Gruppe ist, in der ein höherer Anteil von Postmaterialisten anzutreffen ist, scheint es gerechtfertigt, die Hypothese gerichtet zu formulieren.

4 Daten

In die Analyse werden alle Surveys mit einbezogen, die unter Postmaterialismus im Datenbestandskatalog des GESIS-Datenarchivs mit den Zugangsklassen „0" und „A" (d. h. ohne Einschränkung für die akademische Forschung verfügbar) vermerkt sind, die 4-Item Version der Postmaterialismus-Frage enthalten und die mindestens die gesamte nicht institutionalisierte erwachsene Bevölkerung Westdeutschlands (ohne Berlin) als Zielpopulation haben. Wurde Postmaterialismus in mehreren Wellen eines Surveys erhoben, dann werden nur die Ergebnisse der ersten Welle berücksichtigt, um zu gewährleisten, dass alle Stichproben unabhängig voneinander sind.

Von den 256 Surveys, die im Datenbestandskatalog des GESIS-Datenarchivs unter dem Stichwort Postmaterialismus geführt wurden, trafen diese Kriterien auf 103 Surveys zu (Stand 01.08.2007). Das GESIS-Datenarchiv bildet keine vollständige Sammlung aller Datensätze der empirischen Sozialforschung, da es für Primärforscher keine Pflicht gibt, ihre Datensätze hier zu hinterlegen. Dennoch stellen die im GESIS-Datenarchiv vorhandenen Studien „fast die vollständige Grundgesamtheit aller tatsächlich Sekundäranalysen zugänglichen Datensätze" (Schnell 1997: 50) dar. Es ist somit zu erwarten, dass diejenigen Surveys in diesem Beitrag enthalten sind, welche im akademischen[13] Kontext am häufigsten für Postmaterialismus-Untersuchungen in Deutschland Verwendung finden. Ausnahme ist das nicht über das GESIS-Datenarchiv vertriebene Sozio-ökonomische Panel (SOEP), das jedoch zu Vergleichsrechnungen herangezogen wird (siehe Abschnitt 6).

Damit die ausgewählten Surveys sich im Hinblick auf die Zielpopulation nicht unterscheiden, werden nur diejenigen Personen ausgewählt, die mindestens 18 Jahre alt sind, in Westdeutschland (nicht in Berlin) leben und nicht explizit als Ausländer gekennzeichnet sind. Außerdem werden die Personen, die keine gültigen (metrischen) Angaben zum Alter gemacht haben, nicht mit in die Analyse aufgenommen.[14] Da für die meisten Surveys alle Befragungspersonen Angaben zum Alter

13 Postmaterialismus hat auch Einzug in die privatwirtschaftliche Forschung gehalten. So bilden „Postmaterielle" z. B. eines der Sinus-Milieus.

14 In vier Surveys wurde das Alter nur in Kategorien erfragt. Diese vier Surveys werden von der Analyse ausgeschlossen, da metrische Angabe für das Hypothesentesten benötigt werden.

gemacht haben und es sich bei den anderen Surveys nur um eine sehr geringe Anzahl von ausgeschlossenen Personen handelt, ist nicht davon auszugehen, dass dies einen Einfluss auf den Postmaterialismus-Index hat. Weiterhin werden nur die Personen berücksichtigt, die Angaben zur Postmaterialismus-Frage gemacht haben. Jedoch ist die Anzahl von Personen, die aus diesem Grund nicht berücksichtigt werden können, sehr gering.

5 Methodik

Die Struktur der vorliegenden Daten ist hierarchisch: Individuen (Level 1) sind eingebettet in Surveys (Level 2). Daher ist ein Mehrebenenmodell für die Datenanalyse adäquat.

Würde die hierarchische Struktur der Daten nicht berücksichtigt und somit alle Beobachtungen als unabhängig voneinander aufgefasst, würden die Standardfehler der Erhebungsform-Variablen deutlich unterschätzt werden (Huber/Kernell/Leoni 2005: 376). Die Folge wären fälschlicherweise signifikante Ergebnisse (Fehler 1. Art).

Es wird ein zweistufiges Mehrebenenmodell mit *fixed intercepts* verwendet. Auf der ersten Stufe erfolgt eine Probit-Regression:

$$\mathrm{PM}^{*}_{ij} = \beta_0 + \beta_1 \cdot \mathrm{Geburtsjahr} + R_{ij}$$

wobei β_{0j} der surveyspezifische Achsenabschnitt, β_1 der über alle Gruppen konstante Steigungsparameter des Geburtsjahres und R_{ij} der Fehlerterm der ersten Stufe mit $R_{ij} \sim \mathrm{N}(0, \sigma)$ ist. PM^{*}_{ij} ist eine unbeobachtbare, stetige Variable. Die binäre Postmaterialismus-Variable PM_{ij} nimmt den Wert 1 an, falls PM^{*}_{ij} größer oder gleich Null ist, und den Wert 0, falls PM^{*}_{ij} kleiner Null ist.

Das Geburtsjahr geht als unabhängige Variable auf der ersten Stufe in die Regression ein, denn jüngere Menschen neigen nach Ingleharts Theorie eher zum Postmaterialismus. Auf diese Weise wird dafür Rechnung getragen, dass die Surveys in verschiedenen Jahren durchgeführt wurden und die Zielpopulationen sich somit hinsichtlich der Zusammensetzung der Geburtskohorten unterscheiden. Allerdings wird so der Effekt von Erhebungsformen möglicherweise unterschätzt, weil durch bestimmte Details der Erhebung bestimmte Alterskohorten bevorzugt ausgewählt werden könnten, was den Anteil der Postmaterialisten signifikant beeinflussen könnte. Im Sinne eines konservativen Schätzens scheint es jedoch weniger angebracht, das Geburtsjahr, welches einen großen Einfluss in der Postmaterialismus-Theorie hat, völlig außen vor zu lassen.

Natürlich ist dieses Modell stark unterspezifiziert, da neben dem Geburtsjahr potenziell einflussreiche Faktoren, wie Bildung, berufliche Stellung, Einkommen etc. (vgl. dazu Kroh 2008) keine Berücksichtigung finden. Dies ist zum einen darin begründet, dass diese Faktoren in den Surveys unterschiedlich operationalisiert (z. B. Einkommen, Bildung) und nicht in allen Surveys ermittelt werden. Zum anderen verzerren ausgelassene Variablen nur dann die Schätzungen, wenn sie sowohl einen Einfluss auf die abhängige als auch auf eine unabhängige Variable haben.

Auf der zweiten Stufe werden die surveyspezifischen Achsenabschnitte regressiert:

$$\begin{aligned} \mathrm{PM}^{*}_{ij} = {} & \gamma_{00} + \gamma_{01} \cdot \mathrm{Erhebungsform}_{1j} + \ldots + \gamma_{0q} \cdot \mathrm{Erhebungsform}_{qj} \\ & + \gamma_{0r} \cdot \mathrm{Inflation}_{rj} + \gamma_{0s} \cdot \mathrm{Surveyjahr}_{sj} + \gamma_{0t} \cdot \mathrm{Surveyjahr}^{2}_{qj} + U_{0j} \end{aligned}$$

Für die Hypothesen sind vor allem die Koeffizienten γ_{01} bis γ_{0q} von Interesse. Mit U_{0j} wird der Fehlerterm auf der zweiten Stufe bezeichnet. Auch er wird als normalverteilt angenommen.

Für diese Regression wird ein FGLS-Verfahren verwendet, das auf Hanushek (1974) zurückgeht und von Lewis und Linzer (2005) auf mehrstufige Verfahren übertragen wurde. Dieses Verfahren erzielt, wenn der Anteil der Stichprobenfehler an den Residuen groß genug ist, durch die Einbeziehung der Genauigkeit (Standardfehler) der Level-1-Koeffizienten effizientere Schätzungen auf der zweiten Stufe als WLS und OLS mit robusten Standardfehlern.

Da verschiedentlich, z. B. von Clarke und Dutt (1991), kritisiert wurde, dass die Postmaterialismus-Variable stark negativ von der Inflationsrate beeinflusst wird, hat auch Inglehart die Inflationsrate mit in sein Modell aufgenommen und veranschaulicht, dass es zwar Periodeneffekte durch eine hohe Inflation gibt, aber trotzdem ein Wandel zu mehr Postmaterialismus stattgefunden hat (Inglehart/Abramson 1994: 338; Inglehart/Abramson 1999: 675). Der starke Zusammenhang mit der Inflationsrate ergibt sich daraus, dass das dritte Item äußerst sensitiv auf die Inflation reagiert, weil hier direkt nach Preissteigerungen gefragt wird (Inglehart/Abramson 1994: 341).

Zwar wurde von Clarke und Dutt (1991) noch die Arbeitslosigkeit als möglicher Einflussfaktor genannt, doch konnten Inglehart und Abramson (1994: 1999) unter Kontrolle der Inflationsrate keinen Einfluss der Arbeitslosenquote auf den Postmaterialismus-Index ausmachen. Da außerdem auch die theoretische Verbindung nicht so deutlich wie bei der Inflationsrate ist, wird hier, wie auch bei Klein/Pötschke (2000: 209), nur die Inflationsrate, nicht aber die Arbeitslosenquote berücksichtigt.

6 Ergebnisse

Zunächst wird getestet, ob überhaupt ein überzufälliger Unterschied der Postmaterialisten-Anteile zwischen verschiedenen Surveys existiert. Dazu wird der Intraklassen-Korrelationskoeffizient[15] ρ berechnet und auf einen signifikanten Unterschied von Null getestet. ρ kann einerseits als Homogenität innerhalb eines Surveys verstanden werden und andererseits als Anteil der Varianz der Postmaterialismus-Variablen, der auf die Survey-Ebene zurückgeht (Snijders/Bosker 1999: 46). Wenn ρ nicht signifikant von Null verschieden ist, bedeutet das, dass zwei Individuen aus demselben Survey sich nicht ähnlicher in Bezug auf Postmaterialismus sind als zwei Individuen aus unterschiedlichen Surveys.

Tabelle 1 Intraklassen-Korrelationskoeffizienten

	Leeres Modell	Modell A	Modell B	Modell C
ρ	0,0507	0,0409	0,0349	0,0305
$\overline{\chi}^2$	3044***	2249***	1880***	1645***

*** $p < 0{,}001$

In Tabelle 1 aufgelistet sind die Intraklassen-Korrelationskoeffizienten für verschiedene Modelle sowie die Ergebnisse der entsprechenden Likelihood-Verhältnis-Tests, ob ρ signifikant von Null verschieden ist.[16] Es ist ersichtlich, dass ρ sowohl für das leere Modell (bei dem nicht für das Geburtsjahr kontrolliert wird) als auch für Modell A (unter Kontrolle von Geburtsjahr) auf den konventionellen Niveaus signifikant ist.

Auch unter Kontrolle der Inflationsrate (Modell B) sowie der Inflationsrate und dem Jahr der Erhebung (Modell C) ist der Anteil der Varianz, die auf die Survey-Ebene

15 Bei Probit-Modellen gibt es zwei verschiedene Arten ρ zu definieren. Hier wird der von Snijders & Bosker (1999: 225) empfohlenen Version gefolgt, bei der die Residuenvarianz der latenten Variable PM* und nicht die Varianz der beobachteten Variable PM zur Berechnung von ρ verwendet wird.

16 Da ρ nicht negativ werden kann (es handelt sich um den Quotient aus Varianz zwischen den Surveys und der Gesamtvarianz, zwei nicht-negativen Termen), folgt die Teststatistik des Likelihood-Verhältnis-Tests keiner normalen χ^2-Verteilung mit einem Freiheitsgrad, sondern einer Mischung aus χ^2 mit einem und keinem Freiheitsgrad. Bei $\overline{\chi}^2$ wird dies berücksichtigt (Rabe-Hesketh/Skrondal 2008: 441).
Für die Berechnungen in Tabelle 1 wurden einstufige Mehrebenenmodelle verwendet, da damit auch die Kontrolle für Variablen auf der Survey-Ebene möglich ist. Allerdings sind die Unterschiede zum zweistufigen Verfahren für die ersten beiden Modelle, wo ein Vergleich möglich ist, sehr gering (ρ = 0,041 vs. ρ = 0,044 für Model A).

zurückgeht, deutlich von Null verschieden.[17] Somit unterscheiden sich Postmaterialisten-Anteile signifikant zwischen den Surveys: Befragte im selben Survey besitzen eine ähnlichere Ausprägung auf dem Postmaterialismus-Index und korrelieren auch unter Kontrolle von Geburtsjahr, Inflationsrate und Erhebungsjahr signifikant von Null verschieden.

Um die Größe der Intraklassen-Korrelationskoeffizienten aus Tabelle 1 besser einordnen zu können, wird ein Vergleich mit dem SOEP angestellt, in dem der Postmaterialismus-Index in den Jahren 1984 bis 1986, 1996 und 2006 abgefragt wurde.[18] Aus diesen Wellen werden nur Personen berücksichtigt, die in allen Wellen befragt wurden und valide Angaben zum Postmaterialismus-Index gemacht haben (*balanced panel design*). Anschließend wird ρ berechnet und auf signifikante Abweichung von Null getestet. Die Wellen werden dabei als separate Surveys aufgefasst. Es wird also geprüft, ob zwei Individuen aus derselben Welle sich ähnlicher in Bezug auf Postmaterialismus sind als zwei Individuen aus unterschiedlichen Wellen.

Dieses Verfahren hat den Vorteil, dass ρ unter Konstanthaltung von Geburtsjahr und Erhebungsformen (d. h. selbes Stichprobenverfahren, selber Fragebogentext, selbes Erhebungsinstitut, meist sogar selber Interviewer) berechnet werden kann. Werden alle fünf SOEP-Wellen berücksichtigt, so ergibt sich ein auf konventionellen Niveaus signifikantes ρ von 0,0208.[19] Dieser Wert lässt sich am besten mit dem Intraklassen-Korrelationskoeffizienten von Modell A vergleichen (ρ = 0,0409), da hier das Geburtsjahr, nicht aber die Erhebungsformen, konstant gehalten werden. Die Erhebungsformen scheinen einen Großteil des Varianzanteils, der auf die Survey-Ebene zurückgeht, auszumachen.

Im Folgenden wird getestet, ob bestimmte Erhebungsformen einen Einfluss auf den Postmaterialismus-Index haben. Dazu wird zunächst das im vorherigen Abschnitt erläuterte Probit-Modell der ersten Stufe geschätzt. Die β_0 -Koeffizienten dieser Regression finden sich in der letzten Spalte von Tabelle A1 im Anhang.[20] Mit dem Modell auf Stufe 2 wird anschließend getestet, ob die surveyspezifischen Achsenabschnitte durch Unterschiede in den Erhebungsformen erklärt werden können.

17 Ähnliche Ergebnisse zeigen sich bei Aggregation der Postmaterialisten-Anteile auf Survey-Ebene: Der nicht-parametrische Kruskal-Wallis-Test (Kruskal/Wallis 1952) indiziert auf dem 5 %-Signifikanzniveau, dass in 19 von den 22 Jahren, in denen mehrere Surveys durchgeführt wurden, die verwendeten Stichproben der Surveys eines Jahres nicht der selben Grundgesamtheit zu entstammen scheinen.

18 Vielen Dank an einen der beiden anonymen Gutachter für die Idee zu diesem Vergleich.

19 Dieser Wert ist mit der in Pannenberg et al. (2005: 178f.) beschriebenen Längsschnittsgewichtung berechnet. Ohne Gewichtung ergibt sich ρ = 0,0224. Werden nur die zeitlich eng aufeinander folgenden Wellen 1984-1986 verwendet, ist ρ = 0,0066 bzw. ρ = 0,0056 (ungewichtet). Für die Wellen bis einschließlich 1996 erhält man ρ = 0,0182 bzw. ρ = 0,0148 (ungewichtet).

20 Alle β_0 -Koeffizienten sowie der Koeffizient des Geburtsjahres sind in dieser Regression signifikant. Die Residuen sind ungefähr normalverteilt. Allerdings ist der gesamte Modellfit nicht sehr hoch, was vermutlich auf die Unterspezifikation des Modells zurückzuführen ist.

Tabelle 2 Die Modelle der zweiten Stufe

	Modell			
Variable	M_1	M_2	M_3	M_4
Inflationsrate	-0,0609*** (0,0107)	-0,0569*** (0,00985)	-0,0527*** (0,0104)	-0,0535*** (0,00987)
Quotenstichprobe	-0,106 (0,0537)			0,251** (0,0763)
Adress-Random	-0,161 (0,156)			
Random Route	-0,138 (0,155)			
Standard Random	0,0185 (0,0943)			
Registerstichprobe	0,0114 (0,141)			
„Experte"-Fragestellung	0,313*** (0,0895)	0,273*** (0,0766)		
Eurobarometer	-0,510*** (0,105)	-0,440*** (0,0793)	-0,183*** (0,0348)	-0,160*** (0,0327)
Wohlfahrtssurvey	0,103 (0,190)			
ALLBUS	0,284* (0,142)	0,191*** (0,0382)	0,174*** (0,0402)	0,194*** (0,0376)
Emnid	0,162* (0,0729)	0,128** (0,0403)	0,115** (0,0426)	0,180*** (0,0480)
Getas	-0,0377 (0,0470)			
Infratest	-0,0529 (0,0702)			
Sample/INRA	0,0291 (0,0707)			
Surveyjahr	6,834*** (0,943)	7,253*** (0,848)	6,847*** (0,893)	6,729*** (0,879)
Surveyjahr^2	-0,00172*** (0,000237)	-0,00183*** (0,000213)	-0,00172*** (0,000224)	-0,00169*** (0,000221)
Emnid X Quota				-0,352*** (0,0857)
Konstante	-6823*** (937,7)	-7244*** (844,2)	-6840*** (888,2)	-6723*** (874,6)
R^2	0,733	0,698	0,656	0,710
Adj. R^2	0,681	0,675	0,634	0,684
BIK	-83,5	-112,6	-104,3	-112,1

*Standardfehler in Klammern; *** p<0,001, ** p<0,01, * p<0,05.*

Modell M_1 in Tabelle 2 untersucht den Effekt aller hier diskutierten Erhebungsformen gemeinsam. Es zeigt sich, dass Quotenstichproben – ceteris paribus – einen schwach signifikant negativen Einfluss auf die abhängige Variable β_{0j} ausüben und somit auf den Anteil von Postmaterialisten in der realisierten Stichprobe. Einen negativen Effekt scheinen ebenfalls Surveys der Eurobarometer-Serie zu haben, während ALLBUS-Surveys und das Emnid-Institut einen positiven Effekt haben. Auch die Frageformulierung macht einen Unterschied: Wenn die Befragten als „Experten" angesprochen werden, steigt – ceteris paribus – die Wahrscheinlichkeit, dass sie die beiden postmaterialistischen Items auswählen.

Für die anderen Institute, die Wohlfahrtssurveys sowie die weiteren Stichprobendesigns kann kein signifikanter Einfluss festgestellt werden. Allerdings muss angemerkt werden, dass eventuelle Effekte des Stichprobendesigns möglicherweise von der ALLBUS-Dummyvariablen überdeckt werden, da außer für die ALLBUS-Studien nur für wenige andere Studien detaillierte Informationen über das (Zufalls-) Stichprobendesign vorliegen. Diese Studien ohne genauere Informationen über die Stichprobenziehung bilden gemeinsam die Referenzkategorie in M_1, obwohl diese Gruppe vermutlich sehr heterogen ist. Daher werden im Anschluss zusätzlich die Auswirkungen des Stichprobendesigns innerhalb der ALLBUS-Studien untersucht. Die Inflation hat den erwartet negativen Effekt. Der Einfluss des Erhebungsjahres ist kurvilinear (n-förmig mit 1986 als Höhepunkt).

Ein gutes Regressionsmodell sollte (als Faustregel) mindestens zehn Beobachtungen pro unabhängiger Variable haben (Heck/Thomas 2000: 23). Außerdem sind in M_1 viele nicht signifikante Variablen enthalten, die das Ergebnis nichtsdestotrotz verzerren können. Daher sind in M_2 alle Erhebungsformen herausgenommen worden, die den Wert des Bayesianischen Informationskriteriums (BIK) nicht weiter senken.[21] Es zeigt sich, dass – mit Ausnahme der Quotenstichprobe – alle signifikanten Erhebungsformen aus M_1 Signifikanz und Einflussrichtung in M_2 beibehalten.

Problematisch ist in diesem Modell jedoch, dass fast ausschließlich Surveys der Eurobarometer-Serie die „Experten"-Fragestellung verwendet haben (Korrelation zwischen den beiden Variablen ist $r > 0{,}9$), sodass hier die mit Multikollinearität verbundenen Probleme auftauchen können. Obwohl beide Variablen einen hochsignifikanten Effekt besitzen, kann der Einfluss dieser beiden Variablen kaum getrennt voneinander evaluiert werden und die Koeffizienten liefern möglicherweise ein verzerrtes Bild. Aber selbst wenn eine der beiden Erhebungsformen aus dem Modell herausgenommen wird, bleibt die verbliebene der beiden hochkorre-

21 Das BIK eignet sich hier besonders, da es sowohl zum Modellvergleich als auch als Gütekriterium der Regressionen geeignet ist (Raftery 1995).

lierten Erhebungsformen signifikant, ebenso wie alle anderen Koeffizienten (siehe Modell M_3). Würde jedoch anstelle der „Experten"-Dummy, die Eurobarometer-Variable herausgenommen, würde sich das Vorzeichen des „Experten"-Koeffizienten umdrehen. Daher sind die Ergebnisse aus M_1 und M_2 bezüglich des Vorzeichens der „Experten"-Fragestellung mit Vorsicht zu interpretieren.

Die Regressionsdiagnostik von M_3 zeigt, dass sechs von sieben Surveys, die einen kritischen Cooks-D-Wert haben (d. h. $D > 4/N = 4/99$), Quotenstichproben sind.[22] Auffallend ist weiterhin, dass von diesen sechs Surveys alle vier, die von Emnid durchgeführt wurden, ein negatives Residuum, während die anderen beiden auf Quotenstichproben basierenden Surveys ein positives Residuum haben. Dies lässt vermuten, dass die Emnid-Quotenstichproben anders als die anderen Quotenstichproben sind. In Modell M_4 wird daher ein Interaktionsterm für Emnid-Quotenstichproben mit aufgenommen. Nicht nur dieser Term ist signifikant, sondern sowohl die Quotenstichproben- als auch die Emnid-Variable. Dieses Ergebnis indiziert, dass bei Quotenstichproben im Allgemeinen mehr Postmaterialisten ausgewählt werden, während bei Emnid-Quotenstichproben weniger, aber immer noch leicht überdurchschnittlich viele Postmaterialisten befragt werden.

Eine Untersuchung der M_4-Residuen der ALLBUS-Studien lässt nicht erkennen, dass sich die Unterschiede im Stichprobendesign auf den Postmaterialismus-Index auswirken: Bei allen verwendeten Stichprobendesigns kommen positive und negative Residuen jeweils ungefähr zur Hälfte vor.[23]

7 Fazit

In diesem Beitrag wurde anhand von über 140.000 Befragten in 99 Surveys untersucht, ob Erhebungsformen einen Einfluss auf den Postmaterialismus-Index haben. Es konnte mittels eines zweistufigen Modells gezeigt werden, dass auch unter Kontrolle des Erhebungsjahres, der Inflationsrate und des Geburtsjahres des Respon-

22 Diese Analyse einflussreicher Fälle wurde auf Basis einer gewöhnlichen OLS-Regression durchgeführt, da die Berechnung von Cooks D für gewichtete Daten in Stata noch nicht möglich ist. Die Ergebnisse der OLS-Regression sind aber sehr ähnlich wie die der FGLS. Grafische Analysen für die FGLS-Regression deuten zudem auf die gleichen Ausreißer hin.

23 Wenn auf der ersten Stufe statt eines binären ein ordinales Probit-Modell berechnet wird, ändern sich die Vorzeichen und Signifikanzen der Koeffizienten aus Tabelle 2 nicht – mit der Ausnahme, dass in M_4 die Variable Quotenstichprobe auch auf dem 0,1 %-Niveau signifikant ist. Die abhängige Variable ist in diesem Fall eine diskrete Variable mit den Kategorien Materialist, materialistischer Mischtyp, postmaterialistischer Mischtyp und Postmaterialist, wobei die Mischtypen entsprechend der Prioritäten-Reihenfolge gebildet werden, in der das postmaterialistische und das materialistische Item genannt werden (siehe Abschnitt 2).

denten verschiedene Erhebungsformen einen signifikanten Einfluss auf die Wahrscheinlichkeit besitzen, dass ein Befragter als Postmaterialist eingeordnet wird.[24]

Quotenstichproben im Allgemeinen haben einen signifikant erhöhenden Effekt, der jedoch abgeschwächt wird, wenn Emnid die Erhebung durchführt. Bei Zufallsstichproben, die dieses Institut durchführt, werden jedoch überzufällig viele Befragte als Postmaterialisten eingestuft. Für die anderen Stichprobendesigns konnten keine Unterschiede festgestellt werden, was aber möglicherweise daran liegt, dass für viele Surveys keine detaillierten Informationen über die Art der Stichprobenziehung vorliegen.

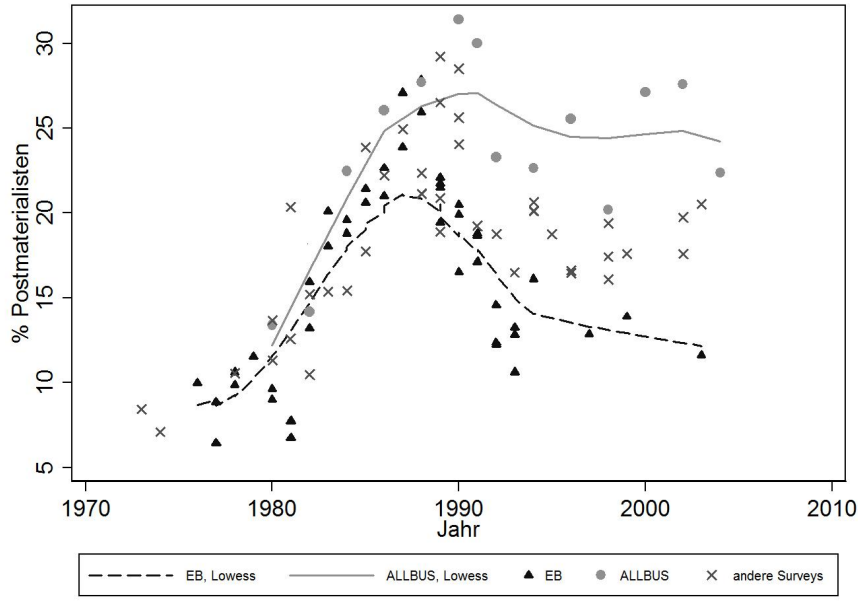
Effekte der Survey-Serie konnten hingegen ausgemacht werden. Bei der ALLBUS-Serie wurde ein positiver Effekt festgestellt, während bei den Eurobarometer-Studien ein negativer Effekt vorlag. Ob der negative Einfluss der Eurobarometer-Serie auf die Formulierung der Postmaterialismus-Frage zurückzuführen ist, konnte nicht genau festgestellt werden, da fast ausschließlich die Eurobarometer eine bestimmte Frageformulierung verwenden.

Interessant ist, dass Ingleharts Aufstieg zum „weltweit wichtigsten Theoretiker des Wertewandels und einem der bekanntesten Sozialwissenschaftler überhaupt" (Klein 2005: 265) nach Saris und Kaase (1997: 6) nur dadurch möglich wurde, dass es ihm gelang, die Fragen zum Postmaterialismus-Index im Eurobarometer zu platzieren. Nun ist es aber gerade diese Survey-Serie, die den Anteil der Postmaterialisten zu unterschätzen scheint. Dies ist jedoch weniger eine Ironie des Schicksals, als vielmehr darauf zurückzuführen, dass bei den Eurobarometern weniger Wert darauf gelegt wird, potenzielle Befragungspersonen auch zu erreichen (z. B. niedrige Mindestkontaktanzahl). Jedoch ist wenig darüber bekannt, welche genauen Erhebungsdetails der Eurobarometer diesen Einfluss auf den Postmaterialismus-Index ausüben. Das liegt insbesondere daran, dass die Datenerhebung bei den Eurobarometern (zumindest für Deutschland) eine „Black Box" ist und gerade für eine so häufig analysierte Serie äußerst schlecht dokumentiert ist.

Dass die hier festgestellten Unterschiede zwischen den Survey-Serien (und den Surveys im Allgemeinen) keine Marginalien sind, zeigt Abbildung 1, in der der Anteil der Postmaterialisten in den einzelnen Surveys im Verlauf der Zeit abgebildet ist. Für die ALLBUS- und Eurobarometer-Surveys sind jeweils Lowess-Kurven eingezeichnet, die indizieren, dass die beiden Serien unterschiedliche Verläufe des Postmaterialisten-Anteils aufzeigen.

24 Da die im GESIS-Datenarchiv vorhandenen Studien die vermutlich methodisch aufwändigeren sind (Schnell 1997: 50), ist nicht zu erwarten, dass die Auswirkungen der Erhebungsformen geringer werden, wenn Studien hinzugenommen werden, die sich nicht im GESIS-Datenarchiv befinden.

Abbildung 1 Vergleich von ALLBUS und Eurobarometer



Um das Ausmaß der Verzerrung durch verschiedene Erhebungsformen noch genauer zu bestimmen, werden ausführliche Methodenberichte benötigt, die für Sekundäranalysen offen zugänglich sind und akkurat Details der Datenerhebung (z. B. Stichprobenziehung, Mindestkontaktzahl) beschreiben. Besser wären darüber hinaus Methodenstudien, in denen den Befragungspersonen zufällig eine bestimmte Erhebungsmethode zugewiesen wird. Mit solchen experimentellen Designs könnte Wissen nicht nur über den Einfluss auf den Postmaterialismus-Index gewonnen werden. Der Postmaterialismus-Index eignet sich jedoch besonders gut für solche Methodenstudien: Dieser verbreitete Index dient nicht nur der empirischen Prüfung einer der bekanntesten sozialwissenschaftlichen Theorien, sondern diskriminiert auch zwischen schwer und leicht Erreichbaren. Daher kann die Relevanz von methodischen Fragestellungen für eher „inhaltlich" orientierte Sozialwissenschaftler gut anhand des Konzepts Postmaterialismus demonstriert werden.

Literatur

Büchel, F., 2000: Institutseffekte bei Befragungen – Auswirkungen auf Datenqualität und Analyseergebnisse. Allgemeines Statistisches Archiv 84: 417-446.

Blohm, M., 2006: Datenqualität durch Stichprobenverfahren bei der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften – ALLBUS. S. 37-54 in: F. Faulbaum und C. Wolf (Hg.): Stichprobenqualität in Bevölkerungsumfragen. Bonn: GESIS.

Böltken, F. und A. Gehring, 1984: Zur Empirie des Postmaterialismus – Quota und Random, Äpfel und Birnen, Kraut und Rüben. ZA-Information 15: 38-52. http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/archiv/za-information/ (15.06.2009).

Böltken, F. und W. Jagodzinski, 1984: Viel Lärm um nichts – Zur „Stillen Revolution" in der Bundesrepublik Deutschland 1970-1980. S. 60-72 in: A. Stiksrud (Hg.): Jugend und Werte – Aspekte einer Politischen Psychologie des Jugendalters. Weinheim: Beltz.

Clarke, H. D. und N. Dutt, 1991: Measuring value change in Western industrialized societies – The impact of unemployment. The American Political Science Review 85: 905-920.

Cuzick, J., 1985: A Wilcoxon-Type test for trend. Statistics in Medicine 4: 87-89.

Davis, D. W. und C. Davenport, 1999: Assessing the validity of the postmaterialism index. The American Political Science Review 93: 649-664.

Hanushek, E., 1974: Efficient estimator for regressing regression coefficients. American Statistician 28: 66-67.

Heck, R. H. und S. L. Thomas, 2000: An introduction to multilevel modeling techniques. Mahwah: Erlbaum.

Hoffmeyer-Zlotnik, J. H. P., 1997: Random-Route-Stichproben nach ADM. S. 33-42 in: S. Gabler und J. H.P. Hoffmeyer-Zlotnik (Hg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Hoffmeyer-Zlotnik, J. H.P., 2006: Stichprobenziehung in der Umfragepraxis – Die unterschiedlichen Ergebnisse von Zufallsstichproben in face-to-face-Umfragen. S. 19-36 in: F. Faulbaum und C. Wolf (Hg.): Stichprobenqualität in Bevölkerungsumfragen. Bonn: GESIS.

Huber, J., Kernell, G., und E. Leoni, 2005: Institutional context, cognitive resources and party attachments across democracies. Political Analysis 13: 365-386.

Inglehart, R., 1971: The silent revolution in Europe: Intergenerational change in post-industrial societies. The American Political Science Review 65: 991-1017.

Inglehart, R., 1977: The silent revolution – Changing values and political styles among Western publics. Princeton: Princeton University Press.

Inglehart, R., 1981: Post-materialism in an environment of insecurity. The American Political Science Review 75: 880-900.

Inglehart, R. und P. R. Abramson, 1994: Economic security and value change. The American Political Science Review 88: 336-354.

Inglehart, R. und P. R. Abramson, 1999: Measuring postmaterialism. The American Political Science Review 93: 665-677.

Küchler, M., 1984: Eine sozio-demographische Beschreibung der Träger postmaterialistischer Einstellungen. S. 215-232 in: K. U. Mayer und P. Schmidt (Hg.): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften – Beiträge zu methodischen Problemen des ALLBUS 1980. Frankfurt: Campus Verlag.

Kirschner, H.-P., 1984: ALLBUS 1980 – Stichprobenplan und Gewichtung. S. 114-182 in: K. U. Mayer und P. Schmidt (Hg.): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften – Beiträge zu methodischen Problemen des ALLBUS 1980. Frankfurt: Campus Verlag.

Kish, L., 1965: Survey Sampling. New York: J. Wiley.

Klein, M., 2005: Der Stellenwert von Persönlichkeitseigenschaften im Rahmen einer Theorie des Postmaterialismus. S. 265-278 in: S. Schumann (Hg.): Persönlichkeit – eine vergessene Größe der empirischen Sozialforschung. Wiesbaden: VS Verlag für Sozialwissenschaften.

Klein, M. und M. Pötschke, 2000: Gibt es einen Wertewandel hin zum „reinen" Postmaterialismus? Eine Zeitreihenanalyse der Wertorientierungen der westdeutschen Bevölkerung zwischen 1970 und 1997. Zeitschrift für Soziologie 29: 202-216.

Koch, A., 1997: ADM-Design und Einwohnermelderegister-Stichprobe-Stichprobenverfahren bei mündlichen Bevölkerungsumfragen. S. 99-116 in: S. Gabler und J. H. P. Hoffmeyer- Zlotnik (Hg.): Stichproben in der Umfragepraxis. Opladen: Westdeutscher Verlag.

Koch, A., 2002: 20 Jahre Feldarbeit im ALLBUS – Ein Blick in die Blackbox. ZUMA-Nachrichten 51: 9-37. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_51.pdf (15.07.2009).

Kohler, U., 1998: Zur Attraktivität der Grünen bei Älteren Wählern. Kölner Zeitschrift für Soziologie und Sozialpsychologie 50: 536-559.

Krebs, D. und J. Hofrichter, 1989: Materialismus-Postmaterialismus – Effekte unterschiedlicher Frageformulierungen bei der Messung des Konzeptes von Inglehart. ZUMA-Nachrichten 24: 60-72. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_24.pdf (15.07.2009).

Kroh, M., 2008: Wertewandel: Immer mehr Ost- und Westdeutsche ticken postmaterialistisch. DIW Wochenbericht 75: 480-486.

Kruskal, W. H. und W. A. Wallis, 1952: Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47: 583-621.

Lewis, J. B. und D. A. Linzer, 2005: Estimating regression models in which the dependent variable is based on estimates. Political Analysis 13(4): 345-364.

Moschner, M., 2008: Sampling and fieldwork in the standard Eurobarometer. http://www.gesis.org/en/services/data/survey-data/eurobarometer-data-service/standard-special-eb/sampling-fieldwork/ (25.06.2009).

Pannenberg, M., Pischner, R., Rendtel, U., Spiess, M. und G.G. Wagner, 2005: Sampling and weighting. S. 153-186 in: J. P. Haisken-DeNew und J. R. Frick (Hg.): Desktop companion to the German Socio-Economic Panel (SOEP) – Version 8.0, Berlin: http://www.diw.de/deutsch/sop/service/dtc/dtc.pdf (28.06.09).

Rabe-Hesketh, S. und A. Skrondal, 2008: Multilevel and longitudinal modeling using stata. College Station: Stata Press (2. Auflage).

Raftery, A., 1995: Bayesian model selection in social research. Sociological Methodology 25(1): 111-163.

Saris, W. E. und M. Kaase, 1997: The Eurobarometer – A tool for comparative survey research. S. 5-23 in W. E. Saris und M. Kaase (Hg.): Eurobarometer – Measurement instruments for opinions in Europe. ZUMA-Nachrichten Spezial 2. Mannheim: Zentrum für Umfragen, Methoden und Analysen. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial2.pdf (15.07.2009).

Schnell, R., 1993: Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität" und Gewichtungsverfahren. Zeitschrift für Soziologie 22(1): 16-32.

Schnell, R., 1997: Nonresponse in Bevölkerungsumfragen – Ausmaß, Entwicklung und Ursachen. Opladen: Leske u. Budrich.

Schnell, R., 2002: Antworten auf Nonresponse. Vortrag auf dem XXXVII. Kongress der deutschen Martktforschung. http://www.bvm.org/user/dokumente/kongress/schnell.pdf (25.06.2009).

Schnell, R., Hill, P. B., und E. Esser, 2005: Methoden der empirischen Sozialforschung. München/Wien: Oldenbourg (7. Auflage).

Schubert, P. und A. Greil, 1997: Sample design and consequences. S. 24-31 in: W. E. Saris und M. Kaase (Hg.): Eurobarometer – Measurement instruments for opinions in Europe. ZUMA-Nachrichten Spezial 2. Mannheim: Zentrum für Umfragen, Methoden und Analysen. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial2.pdf (15.07.2009).

Snijders, T. A. und R. J. Bosker, 1999: Multilevel Analysis – An introduction to basic and advanced multilevel modeling. London: Sage.

Anhang

Tabelle A1 Übersicht über die verwendeten Surveys

Survey	Jahr	Vol	ZA	Sample	Frage	Institut	N	PM	β_{0j}
ALLBUS	1980	1	4241	AR	1	Getas	2900	13,4	-40,24
ALLBUS	1982	1	4241	AR	1	Getas	2963	14,2	-40,24
ALLBUS	1984	1	4241	RR	1	Getas	2943	22,5	-39,96
ALLBUS	1986	1	4241	RR	1	Infratest	3077	26,1	-39,89
ALLBUS	1988	1	4241	RR	1	Getas	2995	27,7	-39,88
ALLBUS	1990	1	4241	AR	1	Infas	2983	31,4	-39,79
ALLBUS	1991	1	4241	SR	1	Infratest	1460	30,0	-39,87
ALLBUS	1992	1	4241	SR	1	Infratest	2294	23,3	-40,08
ALLBUS	1994	1	4241	EWM	1	Infratest	2145	22,7	-40,14
ALLBUS	1996	1	4241	EWM	1	Infratest	2160	25,6	-40,08
ALLBUS	1998	1	4241	AR	1	Getas	2030	20,2	-40,23
ALLBUS	2000	1	4241	EWM	1	Infratest	2242	27,1	-40,06
ALLBUS	2002	1	4241	EWM	1	Infas	1808	27,6	-40,11
ALLBUS	2004	1	4241	EWM	1	Infratest	1761	22,4	-40,31
Ansprüche der Bürger	1985	1	1487	ADM	1	Getas	1809	23,9	-39,94
Ansprüche der Bürger	1989	1	1487	ADM	1	Getas	1916	20,9	-40,10
Aussenpol. Einstell.	1992	1	2980	ADM	1	Basis Research	991	18,8	-40,26
Eurobarometer	1976	6.0	3521	QS	4	EMNID	830	10,0	-40,36
Eurobarometer	1977	7.0	3521	QS	4	EMNID	811	8,9	-40,43
Eurobarometer	1977	8.0	3521	QS	4	EMNID	805	6,5	-40,61
Eurobarometer	1978	9.0	3521	QS	4	EMNID	865	10,6	-40,35
Eurobarometer	1978	10.0	3521	QS	3	EMNID	820	9,9	-40,37
Eurobarometer	1979	12.0	3521	QS	3	EMNID	865	11,6	-40,30
Eurobarometer	1980	13.0	3521	QS	3	EMNID	861	9,6	-40,46
Eurobarometer	1980	14.0	3521	QS	2	EMNID	854	9,0	-40,47
Eurobarometer	1981	15.0	3521	QS	2	EMNID	839	7,7	-40,65
Eurobarometer	1981	16.0	3521	QS	2	EMNID	814	6,8	-40,69
Eurobarometer	1982	17.0	3521	QS	2	EMNID	991	15,9	-40,26
Eurobarometer	1982	18.0	3521	QS	2	EMNID	839	13,2	-40,34
Eurobarometer	1983	19.0	3521	QS	2	EMNID	881	18,0	-40,17
Eurobarometer	1983	20.0	3521	QS	2	EMNID	856	20,1	-40,06
Eurobarometer	1984	21.0	3521	QS	2	EMNID	797	18,8	-40,12
Eurobarometer	1984	22.0	3521	QS	2	EMNID	867	19,6	-40,07
Eurobarometer	1985	23.0	3521	QS	2	EMNID	835	21,4	-40,05
Eurobarometer	1985	24.0	3521	ADM	2	EMNID	854	20,6	-40,10

Survey	Jahr	Vol	ZA	Sample	Frage	Institut	N	PM	β_{0j}
Eurobarometer	1986	26.0	3521	ADM	2	EMNID	924	21,0	-40,06
Eurobarometer	1987	27.0	3521	ADM	2	EMNID	834	27,1	-39,90
Eurobarometer	1987	28.0	3521	ADM	2	EMNID	816	23,9	-40,02
Eurobarometer	1988	29.0	3521	ADM	3	EMNID	905	27,8	-39,92
Eurobarometer	1988	30.0	3521	ADM	2	EMNID	902	25,9	-39,97
Eurobarometer	1989	31.0	3521	ADM	2	EMNID	928	22,1	-40,13
Eurobarometer	1989	31.1	3521	ADM	2	EMNID	1070	21,5	-40,15
Eurobarometer	1989	32.0	3521	ADM	2	INRA	1016	21,8	-40,09
Eurobarometer	1989	32.1	3521	ADM	2	INRA	961	19,5	-40,19
Eurobarometer	1990	33.0	3521	ADM	2	INRA	969	19,9	-40,17
Eurobarometer	1990	34.0	3521	ADM	2	INRA	883	20,5	-40,15
Eurobarometer	1990	34.1	3521	ADM	2	INRA	908	16,5	-40,33
Eurobarometer	1991	35.0	3521	ADM	2	INRA	962	18,8	-40,24
Eurobarometer	1991	35.1	3521	ADM	2	INRA	921	18,7	-40,26
Eurobarometer	1991	36.0	3521	ADM	2	INRA	898	17,1	-40,31
Eurobarometer	1992	37.0	3521	ADM	2	INRA	952	14,6	-40,44
Eurobarometer	1992	37.1	3521	ADM	2	INRA	930	12,4	-40,55
Eurobarometer	1992	38.0	3521	ADM	2	INRA	912	12,3	-40,54
Eurobarometer	1993	39.0	3521	ADM	2	INRA	950	12,8	-40,55
Eurobarometer	1993	39.1	3521	ADM	2	INRA	942	13,3	-40,51
Eurobarometer	1993	40.0	3521	ADM	2	INRA	941	10,6	-40,68
Eurobarometer	1994	42.0	3521	ADM	2	INRA	911	16,1	-40,42
Eurobarometer	1997	47.1	3521	ADM	2	INRA	924	12,9	-40,60
Eurobarometer	1999	52.1	3521	ADM	2	INRA	928	13,9	-40,56
Eurobarometer	2003	59.0	3903	ADM	5	INRA	903	11,6	-40,71
European Values Study	1981	1	1838	QS	6	Allensbach	1174	20,4	-40,07
European Values Study	1990	1	4460	QS	6	Allensbach	1944	28,5	-39,90
European Values Study	1999	1	3778	ADM	1	Infas	999	17,6	-40,37
Frühjahrsstudie	1982	1	1453	ADM	7	Contest-Census	1928	15,2	-40,20
Frühjahrsstudie	1983	1	1455	ADM	7	Contest-Census	2030	15,4	-40,20
Verhinderung von Gewalt	1989	1	1813	SR	1	EMNID	1855	29,2	-39,86
Herbststudie	1981	1	1452	ADM	7	Getas	1994	12,6	-40,31
Einstell. zur Innenpol.	1988	1	1698	ADM	1	Marplan	2031	21,1	-40,06
Einstell. zur Innenpol.	1989	1	1763	ADM	1	Marplan	1996	18,9	-40,19
Einstell. zur Innenpol.	1991	1	2120	ADM	1	Marplan	1472	19,2	-40,19
Massenkommunikation	1985	1	2825	ADM	1	Infratest	1786	17,7	-40,14
Massenkommunikation	1990	1	2825	ADM	1	Infratest	3736	25,6	-39,96
Nachwahlstudie	1994	1	2601	ADM	1	Getas	945	20,1	-40,24
Wohlfahrtssurvey	1984	1	2933	AR	1	Infratest	2048	15,4	-40,22
Wohlfahrtssurvey	1988	1	2933	RR	1	Infratest	2051	21,2	-40,08

Survey	Jahr	Vol	ZA	Sample	Frage	Institut	N	PM	β_{0j}
Wohlfahrtssurvey	1993	1	2933	RR	1	Infratest	2030	16,5	-40,34
Wohlfahrtssurvey	1998	1	3398	RR	1	Infratest	1825	19,4	-40,32
Persönlichk./Wahlverh.	2003	1	4052	ADM	1	Marplan	1958	20,5	-40,34
Political Action	1974	1	757	ADM	1	Getas	2224	7,1	-40,50
Political Action	1980	1	1188	ADM	1	Getas	1959	13,7	-40,24
Pol. Einstellungen	1994	1	3065	ADM	1	Basis Research	997	20,7	-40,23
Pol. Einstellungen	1994	2	3065	ADM	1	Basis Research	918	20,2	-40,25
Pol. Einstellungen	1998	1	3066	ADM	1	Basis Research	1012	16,1	-40,45
Pol. Einstellungen	1998	2	3066	ADM	1	Basis Research	1049	17,4	-40,39
Pol. Einstellungen	2002	1	3861	ADM	1	INRA	1054	19,7	-40,32
Pol. Einstellungen	2002	2	3861	ADM	1	INRA	966	17,6	-40,37
Politbarometer	1982	1	2201	ADM	1	Marplan	1585	10,5	-40,44
Politbarometer	1988	1	1762	ADM	1	Marplan	948	22,4	-39,99
Politische Resonanz	1995	1	2820	SR	1	Getas	949	18,8	-40,29
Politische Resonanz	1996	1	2965	ADM	1	Getas	996	16,5	-40,39
Umweltbewusstsein	1996	1	2964	ADM	1	Getas	1035	16,6	-40,39
Begl.-forsch. z. VZ	1986	1	2292	QS	1	EMNID	1413	22,2	-40,05
Einstellung zur VZ	1987	1	1592	ADM	1	Getas	1704	24,9	-39,91
Wählerverhalten	1990	1	2429	ADM	1	EMNID	1875	24,1	-40,09
Wahlstudie	1989	1	1919	ADM	1	Marplan	2002	26,5	-39,89
Wohlfahrtssurvey	1978	1	2933	AR	1	Infratest	1958	10,6	-40,34
Wohlfahrtssurvey	1980	1	2933	AR	1	Infratest	2356	11,3	-40,34

ADM = Zufallsauswahl mit Random Walk ohne genauere Informationen; AR = Adress-Random; EWM = Einwohnermelderegister-Stichprobe; Frage = Version der Fragestellung (wie im Abschnitt Fragebogentext beschrieben); N = Fallzahl; QS = Quotenstichprobe; RR = Random Route; SR = Standard-Random; PM = Anteil Postmaterialisten; Vol = Volume; VZ = Volkszählung; ZA = Nummer des GESIS-Datenarchivs; β_{0j} *= Achsenabschnitte der einzelnen Surveys auf der ersten Stufe.*

A2 Übersicht über die Versionen der Fragestellung

Es sind nur inhaltliche Änderungen als unterschiedliche Versionen definiert. Formale Unterschiede sind in eckigen Klammern dargestellt.

Version 1: [Ich möchte Ihnen nun einige Fragen zu gesellschaftlichen Problemen stellen.]

[Auch] in der Politik kann man nicht alles auf einmal haben. Auf dieser Liste finden Sie einige [vier] Ziele [Ich nenne nun einige Ziele], die man in der Politik haben kann. Wenn Sie zwischen diesen verschiedenen [vier] Zielen wählen müssten, welches Ziel erschiene [erscheint] Ihnen persönlich am wichtigsten?

Version 2: Es gibt im Augenblick eine Reihe von Diskussionen, was die Ziele der Bundesrepublik in den nächsten 10-15 Jahren sein sollten. Auf dieser Liste sind einige Ziele aufgeführt, denen verschiedene Leute den Vorrang einräumen würden. Würden Sie mir bitte sagen, welches davon Sie selbst als das wichtigste auf längere Sicht halten?

Version 3: Es gibt im Augenblick eine Reihe von Diskussionen, was die Ziele der Bundesrepublik in den nächsten 10 Jahren sein sollten. Auf dieser Liste sind einige Ziele aufgeführt, denen verschiedene Leute den Vorrang einräumen würden. Würden Sie mir bitte sagen, welches davon Sie selbst für am wichtigsten halten?

Version 4: Man spricht zur Zeit sehr viel darüber, was die Ziele der Bundesrepublik in den nächsten 10 Jahren sein sollten. Welches der Ziele auf dieser Liste erscheint Ihnen am wichtigsten?

Version 5: Ich habe hier vier Aufgaben. Bitte wählen Sie davon die zwei Aufgaben aus, die Ihrer Meinung nach für die Staatsgewalt [auf Länderebene, Bundesebene oder Europäischer Ebene] die größte Bedeutung haben sollten.

Version 6: Es wird ja viel darüber gesprochen, welche Ziele die Bundesrepublik [in diesem Land] in den nächsten zehn Jahren vor allem verfolgen soll [verfolgt werden sollen]. Auf dieser Liste [hier] stehen vier [einige] Ziele, die verschiedene Leute für besonders wichtig halten. Welches davon halten Sie für das wichtigste?

Version 7: Hier habe ich vier Kärtchen, auf denen verschiedene politische Forderungen stehen. Würden Sie bitte die Kärtchen einmal so ordnen, dass diejenige Forderung, die für Sie persönlich am wichtigsten ist, ganz oben liegt?

Version 8: Hier auf dieser Liste stehen einige Forderungen. Bitte suchen Sie sich doch die zwei aus, an denen Ihnen am meisten liegt.

Version 9: In der Politik kann man nicht immer erreichen, was man will. Wenn Sie unter den folgenden Dingen wählen müssten, was wäre nach Ihrer Ansicht am erstrebenswertesten?

Anschrift des Autors

Jan Marcus
Graduate Center
DIW Berlin
Mohrenstraße 58
10117 Berlin
jan.marcus@uni-konstanz.de

Rücklauf gut, alles gut? Zu erwünschten und unerwünschten Effekten monetärer Anreize bei postalischen Befragungen

Desirable and Undesirable Effects of Monetary Incentives in Mail Surveys

Sven Stadtmüller

Zusammenfassung
Mit dem Erstversand gewährte monetäre Anreize bei postalischen Befragungen weisen, so legen es verschiedene Studien nahe, einen positiven Einfluss auf die Rücklaufquote auf. Doch aus theoretischer Perspektive gibt es auch andere Effekte, die mit ihrem Einsatz einhergehen könnten. Hierzu zählen Einflüsse des Anreizes auf die Rücklaufgeschwindigkeit und die Zusammensetzung der realisierten Stichprobe. Schließlich könnten auch die Datenqualität und die Bewertung des Interviews durch die Respondenten beeinflusst werden. Eine im Bundesland Hessen durchgeführte postalische Befragung, die sich mit dem Themenfeld des demografischen Wandels beschäftigte, bildet die Datengrundlage für die Analyse dieser, zum Teil auch unerwünschten Effekte. Die Ergebnisse sind für Forscher, die beabsichtigen, mit einem monetären Anreiz zu arbeiten, ermutigend: Belohnungen sparen Zeit und Geld, da sie nicht nur die Rücklaufquote, sondern auch die Rücklaufgeschwindigkeit erhöhen. Auf der anderen Seite lassen sich keine Hinweise erkennen, dass mit dem Einsatz von monetären Anreizen auch unerwünschte Effekte einhergehen.

Abstract
Prepaid monetary incentives in mail surveys have a positive impact by raising response rates, as several studies suggest. From a theoretical point of view, however, there are also some other effects which may be evoked by their usage. This includes impact on the response speed and on the composition of the realized sample. Furthermore, data quality and the respondents' evaluation of the interview may also be influenced. A mail survey which was conducted in the German federal state of Hesse and which focussed on the topic of demographic change is used to examine these possible, partly undesirable, effects. The results are rather encouraging for researchers who intend to work with monetary incentives in their survey: These incentives save both time and money by increasing the response rate and accelerating the response speed. On the other hand, there are no indications that any undesirable effects go along with the use of monetary incentives.

1 Einleitung

In der Vergangenheit häufig als Kompromisslösung betrachtet und mit dem Makel niedriger Rücklaufquoten behaftet, hat die postalische Befragung in den letzten Jahren an Bedeutung gewonnen. Dies liegt wesentlich daran, dass in der jüngeren Forschungsliteratur zu postalischen Befragungen immer häufiger sehr respektable Rücklaufquoten berichtet werden (z. B. Mehlkop/Becker 2007; Becker/Imhof/Mehlkop 2007). Um solche Quoten zu erreichen, haben sich gewisse Vorgehensweisen und Instrumente etabliert, zu denen mittlerweile auch der Einsatz von Anreizen bzw. Belohnungen zählt. Richtet man den Blick ausschließlich auf monetäre Belohnungen und vernachlässigt sämtliche anderen vorstellbaren Anreize (z. B. in Form von Lotterielosen, Spendenbeiträgen oder Kugelschreibern), so stellt der positive Einfluss von monetären Anreizen auf die Rücklaufquote eine inzwischen gut gesicherte Erkenntnis in der Umfrage- und Methodenforschung dar.[1]

Zwar ist auch für die vorliegende Untersuchung die Frage nach der Wirksamkeit eines Anreizes mit Blick auf die Rücklaufquote von Interesse. Allerdings liegt der Schwerpunkt des Beitrags auf der Analyse von anderen denkbaren Effekten von Belohnungen: So wird untersucht, ob diese einen Einfluss auf die Geschwindigkeit des Rücklaufs, die Zusammensetzung der realisierten Stichprobe, die Qualität der Daten und die Bewertung des Interviews durch die Befragungspersonen ausüben. Im Vergleich zu Analysen, die sich auf die Wirksamkeit eines Anreizes im Hinblick auf die Rücklaufquote beziehen, ist diesen Aspekten in der Literatur bislang nur wenig Aufmerksamkeit zuteil geworden. Sie sind aber von besonderer Bedeutung, da sich mit der Verwendung von Belohnungen auch die berechtigte Frage anschließt, ob der Forscher, der mit einem Anreiz arbeitet, neben einer erwünschten Erhöhung des Rücklaufs mit weiteren – erwünschten wie unerwünschten – Effekten zu rechnen hat.

In einem ersten Schritt wird zunächst dargelegt, welche Einflüsse ein Anreiz auf die genannten Aspekte aus theoretischer Sicht ausüben könnte. Gleichzeitig wird, sofern vorhanden, auf einige empirische Evidenzen aus anderen Untersuchungen verwiesen. Kapitel 3 stellt im Anschluss die zu prüfenden Hypothesen vor, ehe in Kapitel 4 in gebotener Kürze die Befragung vorgestellt wird, die als Grundlage der darauf folgenden Analysen dient (Kapitel 5). Die Untersuchung schließt mit einer Zusammenschau der zentralen Ergebnisse.

1 Da im Rahmen dieser Untersuchung ausschließlich auf die Wirkung monetärer Anreize abgestellt wird, wird im weiteren Verlauf auf das Adjektiv „monetär" verzichtet. Zudem werden die Begriffe „Anreize" und „Belohnungen" synonym verwendet.

2 Theoretischer Hintergrund

2.1 Die Wirkung von Anreizen auf die Rücklaufquote

Wie viele andere Entscheidungen, mit denen sich ein Individuum tagtäglich konfrontiert sieht, ist auch die Teilnahme oder Nichtteilnahme an einer Befragung für den Einzelnen mit einer Abwägung der entstehenden Kosten und dem zu erwartenden Nutzen verbunden. Nach der Social Exchange Theory (Blau 1964; Homans 1961; Thibaut/Kelly 1959) stellt die Befragung einen sozialen Austauschprozess dar. Die Teilnahme an der Befragung ergibt sich genau dann, wenn der subjektive Nutzen, den die Befragungsperson mit der Mitarbeit verbindet, die subjektiven Kosten übersteigt. Nutzen kann in diesem Zusammenhang zum Beispiel ein für die Befragungsperson interessantes Befragungsthema darstellen, aber auch das Gefühl, in einer Sache von Bedeutung um Rat gefragt zu werden oder die Möglichkeit eröffnet zu bekommen, seiner Meinung Ausdruck zu verleihen. Auf der Kostenseite spielen vielfältige Faktoren eine Rolle: Die Opportunitätskosten der Zeit, die Befürchtung, eigene Daten preis zu geben, oder auch die Gefahr, z. B. im Zuge von Wissensfragen bloß gestellt zu werden. Für den Forscher, der die Durchführung einer postalischen Befragung plant, lässt sich daraus die Aufgabe ableiten, die Kosten für die Befragungspersonen möglichst gering zu halten. Dies kann z. B. über einen nicht allzu umfangreichen Fragebogen, die glaubwürdige Zusicherung der Anonymität oder einen beigelegten, frankierten und adressierten Rückumschlag (spart Zeit und Geld) erreicht werden. Auch an der anderen Stellschraube, der Nutzenseite, kann der Forscher zu drehen versuchen, so z. B. über ein motivierendes Anschreiben, einen optisch ansprechenden Fragebogen oder eben auch über eine Belohnung.

Gerade mit Blick auf einen Anreiz stellt sich nun aber die Frage, ob ein Individuum nicht dadurch seinen Nutzen maximiert, indem es diesen zwar annimmt, sich aber dennoch nicht an der Befragung beteiligt. Dem widerspricht jedoch, dass sich in der Forschung die positive Wirkung von Belohnungen auf die Rücklaufquote immer wieder nachweisen lässt (so beispielhaft in den Metaanalysen von Church (1993) und Jobber et al. (2004)). Eine theoretische Erklärung der Befragungsteilnahme, die ausschließlich auf der Basis von Kosten-Nutzen-Erwägungen beruht, führt somit offenkundig nicht weiter (vgl. Stadtmüller/Porst 2005: 4). Berücksichtigt man zusätzlich die Theorie der kognitiven Dissonanz von Festinger (1957) und die soziale Reziprozitätsnorm (Gouldner 1960), so kann die Rücklauf erhöhende Wirkung des Anreizes theoretisch erklärt werden. Mit der Verwendung einer Belohnung initiiert der Forscher nämlich eine soziale Austauschbeziehung und erzeugt eine Verpflichtung, dem Wunsch nach Ausfüllen des Fragebogens zu entsprechen

(vgl. Arzheimer/Klein 1998: 8). Natürlich kann sich die Befragungsperson diesem Wunsch widersetzen. Allerdings führt dies nur dann nicht zu einer kognitiven Dissonanz, wenn die Befragungsperson den Anreiz wieder an den Forscher retourniert. Behält die Befragungsperson die Belohnung jedoch, ohne dem Wunsch nach einer Beantwortung des Fragebogens zu entsprechen, so verletzt sie die Reziprozitätsnorm. Ist diese vom Individuum internalisiert, so sind kognitive Dissonanzen die Folge, nach deren Vermeidung man in aller Regel strebt.

Folgt man dieser theoretischen Erklärung der Wirksamkeit von Anreizen, so ergibt sich daraus jedoch die Konsequenz, dass diese nur dann eine Rücklauf erhöhende Wirkung entfalten, wenn ihnen von der Befragungsperson ein symbolischer Charakter zugeschrieben wird. Dementsprechend sollte der Wert der Belohnung nicht zu hoch gewählt werden, da hiermit zwei Gefahren einhergehen: Zum einen mag ein finanziell wertvoller Anreiz die Befragungsperson dazu verleiten, diesen als Bezahlung oder Aufwandsentschädigung zu interpretieren. Prüft das Individuum in der Folge, ob der Anreiz eine adäquate Vergütung für die zu verrichtende „Arbeit" darstellt, wird es zumeist zu dem Schluss gelangen, unterbezahlt zu sein und nicht an der Befragung teilnehmen. Zum anderen besteht bei sehr wertvollen Belohnungen die Gefahr, dass sich die Befragungsperson in ihrer Entscheidungsfreiheit über Teilnahme oder Nichtteilnahme eingeengt fühlt. Gemäß der Theorie der psychologischen Reaktanz (Brehm/Brehm 1981) ist für diesen Fall die Entscheidung zur Nichtteilnahme erwartbar, da die Befragungsperson auf diesem Wege ihre persönliche Freiheit wieder herstellt (vgl. Stadtmüller/Porst 2005: 5). Dies erklärt, warum sich in der Forschung häufig kein annähernd linearer Zusammenhang zwischen der Höhe des Anreizes und der erzielten Rücklaufquote nachweisen lässt: In der Untersuchung von Mizes et al. (1984) zeigt sich kein Unterschied in der Rücklaufquote, je nachdem, ob als Anreiz ein Dollar oder fünf Dollar gewählt werden. Auch bei Warriner et al. (1996) erhöht sich die Ausschöpfungsquote beim Übergang von fünf zu zehn Dollar nur um einen Prozentpunkt. Folglich scheint ein effizienter Einsatz von Belohnungen vor allem dann gegeben zu sein, wenn sie einen eher symbolischen Charakter aufweisen und auch, z. B. in dem den Fragebogen begleitenden Anschreiben, als solche deklariert werden.[2]

2 Insgesamt gilt es jedoch zu bedenken, dass die Theorie der kognitiven Dissonanz, der psychologischen Reaktanz und der Reziprozitätsnorm zwar geeignet erscheinen, um die in Auszügen dargelegten empirischen Befunde zu erklären. Allerdings ist damit nicht ausgeschlossen, dass andere Einflüsse den über Experimentalstudien belegbaren Zusammenhang zwischen Anreizen und erhöhten Rücklaufquoten hervorrufen. Annähernd auszuschließen wäre dies allenfalls, sofern auch Informationen über jene Personen vorliegen, die sich nicht an der Befragung beteiligen.

2.2 Die Wirkung von Anreizen auf die Rücklaufgeschwindigkeit

Aus den bisherigen theoretischen Überlegungen kann ebenso gefolgert werden, dass ein Anreiz einen schnelleren Rücklauf nach sich ziehen dürfte. Da durch die Belohnung eine soziale Austauschbeziehung initiiert und gleichzeitig der Versuch unternommen wird, die Reziprozitätsnorm zu aktivieren, kann davon ausgegangen werden, dass die Befragungsperson zügig darauf bedacht ist, diese Beziehung einzugehen bzw. der Norm zu entsprechen. Wird die Befragungsperson dem Wunsch nach dem Ausfüllen des Fragebogens nicht mehr oder weniger unmittelbar gerecht, so sollten die kognitiven Dissonanzen kurz nach dem Erhalt der Befragungsunterlagen am stärksten ausgeprägt sein. Mit der Zeit jedoch verblassen nicht nur die kognitiven Dissonanzen, sondern ebenso die Wirksamkeit der Norm, bis schließlich die Befragung und die damit empfundenen negativen Empfindungen, welche daraus resultieren, der Norm nicht entsprochen zu haben, in Vergessenheit geraten. Aus diesen Überlegungen folgt die Erwartung, dass Probanden, denen ein Anreiz zuteil wurde, im Vergleich zur Residualgruppe im Mittel zügiger antworten. Insbesondere in den ersten Tagen nach dem Erstversand sollten im Rahmen von Methodenexperimenten vermehrt Antworten von denjenigen Befragungspersonen eingehen, die einen Anreiz erhalten haben.

In der Forschung liegen bislang nur wenige Studien vor, die sich mit der Wirkung einer Belohnung im Hinblick auf die Geschwindigkeit des Rücklaufs beschäftigen. Allerdings weisen diese in die gleiche Richtung. Brennan (1992) kann nachweisen, dass bereits ein Anreiz in Höhe von 50 Cent den Rücklauf beschleunigt und auch Arzheimer und Klein (1998), die mit Telefonkarten, also mit geldnahen Anreizen arbeiteten, zeigen dies auf: Demnach benötigten Befragungspersonen, die eine Telefonkarte erhielten, im Mittel 15,7 Tage für ihre Antwort, während sich die Residualgruppe ohne Belohnung mit ihrer Rückantwort im Schnitt etwa drei Tage länger Zeit ließ (18,6 Tage). Der Vorteil eines schnelleren Rücklaufs liegt auf der Hand: Gerade bei Befragungen mit Nachfassaktionen kann so die Zahl derer reduziert werden, die ein zweites oder gar drittes Mal angeschrieben werden muss (vgl. Stadtmüller/Porst 2005: 10).

2.3 Die Wirkung von Anreizen auf die Zusammensetzung der realisierten Stichprobe

Die Belohnung stellt, wie bereits angedeutet, eine externe motivierende Größe zur Teilnahme an der Befragung dar und gibt auf der Seite der Befragungsperson, so legen es jedenfalls die empirischen Befunde zur Rücklauf steigernden Wirkung von Anreizen nahe, nicht selten den Ausschlag für die Entscheidung zur Mitarbeit. Diese Personen, die sich aufgrund des Anreizes zur Teilnahme entschließen, werden von

Mehlkop und Becker (2007) als „Unentschlossene" bezeichnet, da bei ihnen die intrinsische Motivation nicht ausreicht, um an der Befragung teilzunehmen, sondern es eines zusätzlichen Anreizes bedarf. Daneben gibt es Personen, die schon allein aufgrund ihrer Motivation an der Befragung teilnehmen und für die die Belohnung nicht mehr als ein „nettes Zubrot" darstellt, ihre Entscheidung aber nicht weiter beeinflusst. Das Gleiche gilt, nur unter umgekehrten Vorzeichen, für die Verweigerer, die sich auch nicht durch einen Anreiz zur Teilnahme animieren lassen.

Die intrinsische Motivation einer Person, sich an einer Befragung zu beteiligen, ist mit anderen Merkmalen, wie z. B. dem Alter oder dem Bildungsgrad korreliert. Nach Goyder (1987) fördern eine hohe Bildung und ein niedriges Lebensalter die Teilnahmebereitschaft an Befragungen. Andere Studien (z. B. Pötschke/Müller 2006) suggerieren dagegen einen kurvilinearen Zusammenhang zwischen dem Lebensalter und der Teilnahmebereitschaft, der für sehr junge und für sehr alte Menschen niedrige Teilnahmequoten unterstellt. Die Vermutung einer Überrepräsentation älterer Menschen in der realisierten Teilstichprobe mit Anreiz ließe sich auch mit dem Charakter der Reziprozitätsnorm begründen: Da ein Individuum eine Norm vor allem dadurch internalisiert, dass es ihr immer wieder aufs Neue entspricht bzw. ihr gemäß handelt, dürfte sich folglich mit steigendem Lebensalter die Chance erhöhen, häufiger mit solchen Situationen konfrontiert gewesen zu sein, indem auf die Reziprozitätsnorm abgestellt wurde. Für die vorliegende Untersuchung wird mit Blick auf die Gruppe jener Befragten, die einen Anreiz erhielt, von einer Überrepräsentation von Personen mit geringer formaler Bildung ausgegangen. Für die Variable Lebensalter liegen dagegen aus theoretischer Sicht konkurrierende Wirkungszusammenhänge vor, die einer empirischen Untersuchung bedürfen. Allerdings wird vermutet, dass insbesondere ältere Menschen positiv auf den Anreiz reagieren.

Die bisherigen Studien, die sich mit der Frage nach Effekten eines Anreizes auf die Zusammensetzung der realisierten Stichprobe beschäftigten, deuten indes darauf hin, dass die Reziprozitätsnorm unabhängig von soziodemografischen Merkmalen wirkt. Dies konnten sowohl Shettle und Mooney (1999) als auch Warriner et al. (1996) sowie Arzheimer und Klein (1998) feststellen.

2.4 Die Wirkung von Anreizen auf die Qualität der Daten

Von Befragungspersonen, die vom Forscher einen Anreiz erhalten, dürfte aus theoretischer Sicht eine höhere Qualität der Daten erwartbar sein. Anders ausgedrückt: Befragte, denen ein Anreiz zuteil wurde, sollten sich aufgrund der subjektiv empfundenen erhöhten Verpflichtung beim Ausfüllen des Fragebogens besondere Mühe

geben, was sich bspw. an der Zahl der fehlenden Antworten oder dem geleisteten Aufwand (z. B. an der Zahl der Wörter bei offenen Fragen) ablesen lassen könnte. Auf der anderen Seite könnte die Belohnung, wie bereits erwähnt, vorrangig die Gruppe der Unentschlossenen animieren, die nicht genügend intrinsische Motivation aufbringt, um sich auch ohne externen Anreiz an der Befragung zu beteiligen. Demzufolge wäre bei diesen Personen eher ein „Dienst nach Vorschrift" denn ein sorgfältiges Bearbeiten des Fragebogens zu erwarten (vgl. Davern et al. 2003: 140). Die empirischen Befunde zum Zusammenhang von Datenqualität und der Gewährung eines Anreizes liefern ein uneinheitliches Bild: Während z. B. Wotruba (1966), McDaniel und Rao (1980) und James und Bolstein (1990) von einer positiven Korrelation zwischen der Höhe des Anreizes und dem von der Befragungsperson erbrachten Aufwand für den Fragebogen berichten, sind bei Davern et al. (2003) oder Shettle und Mooney (1999) solche Effekte nicht nachweisbar.

2.5 Die Wirkung von Anreizen auf die Bewertung des Interviews durch die Befragungsperson

Der Anreiz dient, wie bereits eingangs erwähnt, vorrangig dazu, den Nutzen, den die Befragungsperson aus der Teilnahme am Interview zieht, positiv zu beeinflussen. Dies ist, in Verbindung mit der Reziprozitätsnorm und der Theorie der kognitiven Dissonanz, eine mögliche Erklärung für den Befund, dass sich Personen, denen ein Anreiz beigelegt wird, eher an der Befragung beteiligen als solche, die keine Belohnung erhalten. Ein höherer Nutzen, der aus der Befragungsteilnahme resultiert, sollte sich auch in der Bewertung des Interviews durch die Befragungsperson widerspiegeln. Diese kann wiederum über die subjektiv empfundene Dauer und Interessantheit der Befragung operationalisiert werden. Folglich ist zu erwarten, dass Personen, die eine Belohnung erhalten, das Interview in dem Sinne besser bewerten, als dass sie im Vergleich zur Residualgruppe die Länge des Fragebogens als geringer und die Interessantheit der Befragung als höher einschätzen. Würden sich diese Zusammenhänge empirisch zeigen lassen, so könnte es in der Konsequenz gelingen, hohe Rücklaufquoten auch für Befragungen zu erzielen, deren Themen in der Bevölkerung auf ein vergleichsweise geringes Interesse stoßen.

Zwar finden sich in der Literatur einige Studien, die sich vorrangig unter dem Stichwort „Respondent Burden" mit Effekten des Befragungsthemas und der Länge des Fragebogens auf die Rücklaufquote beschäftigen (so z. B. Sharp/Frankel 1983; Bogen 1996; Galesic/Bosnjak 2009). Jedoch werden diese Zusammenhänge entweder nicht in Verbindung mit dem Einsatz eines Anreizes untersucht oder die interessierende Erhebungsmethode stellt nicht die postalische Befragung dar.

3 Formulierung von Hypothesen

Bevor im folgenden Abschnitt das Datenmaterial vorgestellt wird, auf dem die in Kapitel 5 durchgeführten Analysen beruhen, sollen kurz die Forschungshypothesen formuliert werden. Diese lauten:

H1: **Rücklaufquote:** *Der Anreiz wirkt sich in einer signifikant höheren Rücklaufquote in der Experimentalgruppe (Befragte, die einen Anreiz erhielten) im Vergleich zur Kontrollgruppe (ohne Anreiz) aus.*

H2: **Rücklaufgeschwindigkeit:** *Der Anreiz führt zu einer Beschleunigung des Rücklaufs in der Experimental- im Vergleich zur Kontrollgruppe.*[3] *Dies sollte sich darin zeigen, dass die Dauer des Rücklaufs (in Tagen seit dem Erstversand berechnet) in der Experimentalgruppe signifikant niedriger liegt als in der Kontrollgruppe.*

H3: **Zusammensetzung der realisierten Stichprobe:** *Der Anreiz spricht vor allem die Gruppe der „Unentschlossenen" an, die nur über eine geringe intrinsische Motivation zur Bearbeitung des Fragebogens verfügt. Da die Motivation mit dem Bildungsgrad korrespondiert, ist in der Experimentalgruppe ein im Vergleich zur Kontrollgruppe höherer Anteil an Personen mit geringer formaler Bildung zu erwarten. Zudem ist ein Zusammenhang zwischen Motivation und Alter anzunehmen. Allerdings ist die Richtung des Zusammenhangs aus theoretischer Sicht unklar. Jedoch ist zu erwarten, dass ältere Menschen die Reziprozitätsnorm stärker internalisiert haben als jüngere und daher verstärkt in der Experimentalgruppe zu finden sind.*

H4: **Qualität der Daten:** *Der Anreiz geht entweder mit einem erhöhten Aufwand der Befragungsperson beim Ausfüllen des Fragebogens einher. Denkbar ist es aber auch, dass der Anreiz vorrangig Personen mit einer geringen intrinsischen Motivation anspricht, die dann den Fragebogen lediglich pflichtgetreu ausfüllen. Daher muss die Motivation kontrolliert werden, wenn die Qualität der Daten in Experimental- und Kontrollgruppe verglichen wird.*

3 Während sich in der Hypothese zur Rücklaufquote die Begriffe Experimental- und Kontrollgruppe auf die Brutto- bzw. Nettostichprobe beziehen, sind damit in den übrigen Hypothesen der sprachlichen Einfachheit halber die realisierten Teilstichproben gemeint, die entweder vorab einen Anreiz erhielten (Experimentalgruppe) oder eben nicht (Kontrollgruppe).

H5: **Bewertung des Interviews:** *Der Anreiz führt über die Steigerung des individuellen Nutzens an der Befragungsteilnahme zu einer positiveren Bewertung des Interviews in der Experimental- im Vergleich zur Kontrollgruppe. Dies sollte sich darin zeigen, dass Personen, denen ein Anreiz zuteil wurde, die Interessantheit der Befragung höher und die Dauer des Interviews kürzer einschätzen als Personen, die keine Belohnung erhielten. Dieser Zusammenhang sollte auch dann bestehen bleiben, wenn die Motivation zur Mitarbeit kontrolliert wird.*

4 Die Studie „Zukunftswerkstatt Deutschland"

Von Oktober bis Dezember 2008 führte das Forschungszentrum Demografischer Wandel (FZDW) der Fachhochschule Frankfurt am Main eine postalische Befragung von insgesamt 2.000 Personen in drei zufällig ausgewählten hessischen Landkreisen und einer kreisfreien Stadt durch. Unter dem Titel „Zukunftswerkstatt Deutschland – Wie sehen die Hessen die Zukunft ihres Landes?" wurden die angeschriebenen Personen gebeten, Fragen zu politischen Einstellungen, Medienkonsum, Wertorientierungen und Politikpräferenzen zu beantworten. Im Zentrum der Befragung stand jedoch das Themenfeld demografischer Wandel. Insbesondere ging es um die Bekanntheit des Begriffs, die Assoziationen mit der demografischen Entwicklung und die Einschätzung von demografischen Trends, wie dem Geburtenrückgang oder dem demografischen Alterungsprozess. Da jedoch der Begriff des demografischen Wandels breiten Bevölkerungsgruppen nicht bekannt ist – was sich auch im Zuge der Befragung herausstellte – wurde es bewusst vermieden, sowohl im Anschreiben als auch im Titel der Befragung auf diesen thematischen Fokus hinzuweisen.

Die Auswahl der Befragungspersonen bzw. Haushalte erfolgte aus Kostengründen aus dem öffentlichen Telefonbuch. Der Fragebogen sollte, so wurde es dem angeschriebenen Haushalt an zentraler Stelle im Anschreiben und im Fragebogen mitgeteilt, von derjenigen Person ausgefüllt werden, die im betreffenden Haushalt zuletzt Geburtstag hatte und gleichzeitig über 18 Jahre alt ist. Dies funktionierte erwartungsgemäß nicht in allen Haushalten, was z. B. an einer im Vergleich zu den Telefonbucheinträgen zwar abgeschwächten, aber noch immer recht deutlichen Überrepräsentation von Männern abzulesen ist. Zudem erwiesen sich zahlreiche längst veraltete Einträge im Telefonbuch als problematisch.

Zwei Wochen nach dem Erstversand erhielten alle Haushalte, in Anlehnung an die Vorgaben aus Dillmans Total Design Method (1978), eine Erinnerungs- bzw. Dankespostkarte. Wiederum zwei Wochen später erfolgte die zweite und letzte

Nachfassaktion: Jene Haushalte, die zu diesem Zeitpunkt ihren Fragebogen noch nicht retourniert hatten, wurden erneut angeschrieben und erhielten nochmals ein Fragebogenexemplar. Auch die übrigen von Dillman benannten Aspekte zur Erhöhung der Rücklaufquote fanden Berücksichtigung: Die Anschreiben erhielten stets einen adressierten und frankierten Rückumschlag, das Layout des insgesamt 12 Seiten umfassenden Fragebogens und die Titelseite wurden optisch ansprechend gestaltet und das Anschreiben verdeutlichte die Bedeutung der Studie und motivierte zur Teilnahme.[4]

Gleich mit dem Erstanschreiben erhielt die Hälfte der angeschriebenen Haushalte eine Ein-Euro-Münze, die auf das Anschreiben geklebt und als „eine kleine Anerkennung für Ihre Mühe" deklariert wurde. Dabei wurde darauf geachtet, dass innerhalb der vier eingangs erwähnten Auswahlbezirke auch nur jeweils die Hälfte (N=250) der angeschriebenen Personen bzw. Haushalte den Anreiz erhielt.[5]

5 Empirische Analysen

5.1 Rücklaufquote

Von den 2.000 angeschriebenen Haushalten bzw. Personen kamen insgesamt 699 Antworten zurück, was einer Rücklaufquote von 35 % entspricht. Rechnet man die 96 stichprobenneutralen Ausfälle (z. B. Empfänger verzogen, aus gesundheitlichen Gründen nicht zur Teilnahme in der Lage) heraus, so beträgt die Nettorücklaufquote knapp 37 %.

Die Nettostichprobe enthielt 951 Haushalte/Personen, die mit dem Erstanschreiben einen Anreiz in Form einer Ein-Euro-Münze erhielten. Von diesen antworteten 406 Personen (42,7 %). Aus der Kontrollgruppe (Nettostichprobe: 953 Haushalte bzw. Personen) kamen dagegen nur 286 Fragebögen zurück (30,0 %).[6] Der Unterschied in den beiden Rücklaufquoten der Experimental- und Kontrollgruppe ist statistisch signifikant und belegt einmal mehr die Wirksamkeit eines (vorab verabreichten) Anreizes bei postalischen Befragungen.

4 Für wertvolle Hinweise sei an dieser Stelle Herrn Rolf Porst von GESIS gedankt.

5 Der Ergebnisbericht der Befragung Zukunftswerkstatt Deutschland ist unter www.fh-frankfurt.de/zukunftswerkstatt.de abrufbar und enthält im Anhang den vollständigen Fragebogen und die einzelnen Anschreiben.

6 Bei insgesamt sieben Antworten war nicht ersichtlich, ob die Person bzw. der Haushalt einen Anreiz erhielt, da die zugehörige Identifikationsnummer unkenntlich gemacht wurde.

5.2 Rücklaufgeschwindigkeit

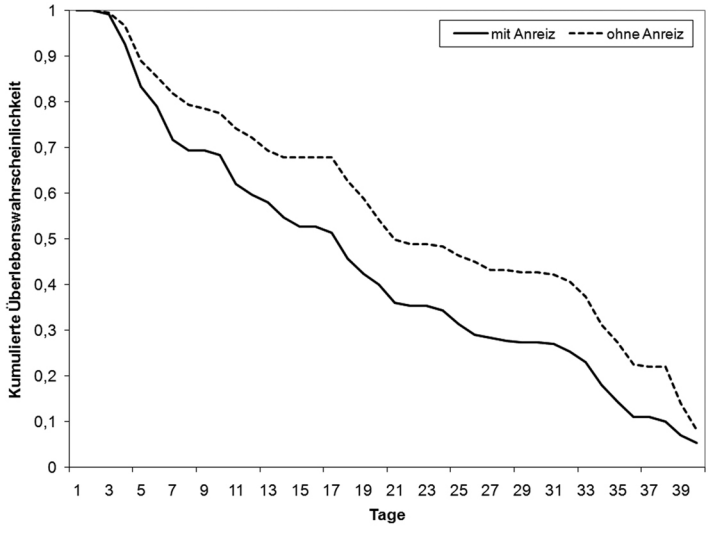
Um zu untersuchen, ob ein Anreiz den Rücklauf tatsächlich beschleunigt, ist es notwendig, über die Rückantworten exakt „Buch zu führen". Dies gestaltete sich im Rahmen der Zukunftswerkstatt Deutschland aus mehreren Gründen als problematisch: Zum einen nahm die Zustellung der Befragungsmaterialien in den verschiedenen Regionen unterschiedlich viel Zeit in Anspruch. Unerfreulich war dies nicht zuletzt deshalb, weil sich dadurch der Erhalt der Unterlagen bei vielen Personen in die (Arbeits-)Woche hinein verlegte und sich nicht, wie durch den Versand an einem Donnerstag eigentlich beabsichtigt, auf das Wochenende beschränkte. Einige Befragungspersonen erhielten gar erst fünf Tage nach dem Erstversand die entsprechenden Unterlagen. So war es einigen Teilnehmern, operationalisiert man die Rücklaufgeschwindigkeit mit der Zahl der verstrichenen Tage seit dem Erstversand, nicht möglich, vergleichsweise zügig zu antworten. Zum anderen wurden in den ersten Wochen der Feldphase die Rückantworten durch die Zusteller offenkundig gesammelt und erst, sobald sich einige Antworten angehäuft hatten, der Fachhochschule auch tatsächlich zugestellt.

Um dennoch Aussagen über die Rücklaufgeschwindigkeit und mögliche Unterschiede zwischen Experimental- und Kontrollgruppe treffen zu können, wurde eine Variable gebildet, die der Zahl der verstrichenen Tage zwischen dem Erstversand und dem Datum des Poststempels, der auf dem Rückantwortumschlag ausgewiesen ist, entspricht. Hier reduziert sich jedoch die Fallzahl auf 513, da in 186 Fällen kein Poststempel auf dem Rückantwortumschlag vorzufinden war. Implizit wird mit der Konstruktion der Variablen unterstellt, dass die Zustellgeschwindigkeit der Befragungsmaterialien innerhalb der Experimental- und Kontrollgruppe zufällig verteilt war, dass also die Dauer der Zustellung der Unterlagen nicht davon abhängig war, ob in den Unterlagen ein Anreiz enthalten war oder nicht. Zudem wird unterstellt, dass es sich auch bei den Rückumschlägen ohne Poststempel um zufällige Ereignisse handelt, die nicht in Abhängigkeit von der Verabreichung eines Anreizes an die Befragungsteilnehmer stehen.

In Abbildung 1 sind nun die Überlebenswahrscheinlichkeiten, zu verstehen als die relativen Anteile derjenigen, die zu dem jeweiligen Zeitpunkt den Fragebogen noch nicht zurückgeschickt haben, in Abhängigkeit von der Gewährung eines Anreizes abgetragen.[7]

7 Dabei ist in der Darstellung zu beachten, dass die realisierte Stichprobe und nicht die Nettostichprobe als Prozentuierungsbasis verwendet wurde (anders: Becker/Imhof/Mehlkop 2007: 147). Dies liegt in den fehlenden Werten der Variablen zur Messung der Rücklaufgeschwindigkeit begründet, die in anderen Parametern für die Nettostichprobe und den Rücklauf resultieren würden.

Abbildung 1 Überlebenswahrscheinlichkeiten für den Rücklauf in der Experimental- und Kontrollgruppe



So ist z. B. mit Blick auf den 20. Tag der Feldphase festzustellen, dass in der Gruppe der Personen, die einen Anreiz erhielt, nur noch 40 % der retournierten Fragebögen ausstanden, bzw. bis zu diesem Tag bereits 60 % aller Antworten der Experimentalgruppe bereits eingegangen waren. In der Kontrollgruppe dagegen standen zu diesem Zeitpunkt noch mehr als die Hälfte (54 %) aller letztlich retournierten Fragebögen aus. Der Verlauf der beiden Kurven zeigt eindeutig, dass der Anreiz eine Beschleunigung des Rücklaufs bewirkt. Auch die Tests für Zeitdifferenzen des Medians, so z. B. der Log-Rank oder Wilcoxon-Test, weisen allesamt statistisch signifikantes Niveau auf.

Insgesamt wird somit deutlich, dass ein Anreiz tatsächlich zu schnelleren Rückantworten von Seiten der Befragungspersonen führt. Somit können, sofern Nachfassaktionen vorgesehen sind, die Kosten, die eine Belohnung verursacht, durch geringere Kosten für die Erinnerungsaktionen, die aus einem höheren und schnelleren Rücklauf resultieren, zum Teil kompensiert werden.

5.3 Zusammensetzung der realisierten Stichprobe

Unterschiede in der sozialstrukturellen Zusammensetzung der beiden realisierten Teilstichproben (Experimental- und Kontrollgruppe) sollten sich, den theoretischen Überlegungen folgend, erkennen lassen, da der Anreiz ein ganz bestimmtes Segment der Befragungspersonen, namentlich die Gruppe der Unentschlossenen, besonders stark ansprechen dürfte. Diese, so wird unterstellt, verfügen über eine vergleichsweise geringe intrinsische Motivation zur Teilnahme an der Befragung, die wiederum mit bestimmten Merkmalen wie z. B. dem Bildungsniveau korreliert. Wird die Motivation zur Teilnahme an der Befragung über das politische Interesse operationalisiert – was angesichts der hohen gesellschaftspolitischen Bedeutung des Befragungsthemas als legitim erscheint – so zeigt sich jedoch kein signifikanter Mittelwertunterschied im subjektiven Politikinteresse, auch wenn sich die Experimentalgruppe im Mittel als etwas weniger politisiert darstellt.

In Tabelle 1 wird in der Folge die Zusammensetzung der beiden realisierten Teilstichproben mit Blick auf den formalen Bildungsgrad der Befragungsteilnehmer untersucht.

Tabelle 1 Die Zusammensetzung der Stichprobe nach Bildungsabschlüssen in Abhängigkeit von der Gewährung eines Anreizes

	Hauptschule %	mittlere Reife %	(Fach)-Abitur %	Hochschulabschluss %
kein Anreiz	31,3*	22,1	16,4	28,2
Anreiz erhalten	24,3*	25,0	15,3	33,0

*Signifikanzniveau der Mittelwertunterschiede: * $p < 0{,}05$.*

Abgesehen von einer allgemein starken Überrepräsentation von Personen mit Hochschulabschluss ist erkennbar, dass der Anteil der Absolventen der Hauptschule in der Experimentalgruppe niedriger liegt als in der Kontrollgruppe. Dieser Unterschied ist in den betrachteten Kategorien der einzige, der ein statistisch signifikantes Niveau aufweist. Gleichzeitig widerspricht er der formulierten Hypothese. Zudem zeigte sich auch, dass Personen, die ihre eigene wirtschaftliche Lage als schlecht oder sehr schlecht einstuften, dann häufiger an der Befragung teilnahmen, wenn sie keinen Anreiz erhielten. Betrachtet man diese Befunde in der Zusammenschau, ist es denkbar, intrinsische Motivation zur Teilnahme an der Befragung anders zu deuten: Sie entsteht nicht primär durch ein Interesse an Politik oder an sozialwissenschaftlichen Fragestellungen, sondern ist vielmehr bei jenen Personen stark ausgeprägt, die mit ihren Lebensbedingungen nicht einverstanden sind und dies im

Zuge der Befragung kommunizieren wollen. Um jedoch gehaltvolle Aussagen darüber treffen zu können, welche Bildungsgruppen durch eine Belohnung besonders stark angesprochen werden, wären Informationen der Nichtteilnehmer vonnöten. Daher kann mit Blick auf die bisherigen Ergebnisse nur gefolgert werden, dass ein Anreiz allenfalls geringe Effekte auf die Zusammensetzung der Stichprobe ausübt.

Gleiches gilt auch für das Lebensalter der Befragungspersonen: Das arithmetische Mittel des Lebensalters unterscheidet sich zwischen Experimental- und Kontrollgruppe um 1,7 Jahre, wobei es in der Experimentalgruppe niedriger liegt, die Differenz aber kein statistisch signifikantes Niveau aufweist. Bildet man vier gleich stark besetzte Altersgruppen (18 bis 40 Jahre, 41 bis 53 Jahre, 54 bis 65 Jahre und 66 bis 89 Jahre) und betrachtet die gemeinsame Verteilung mit dem Merkmal „Anreiz erhalten/keinen Anreiz erhalten", so fallen die Anteile in derjenigen Gruppe mit Belohnung unter den jüngeren Altersgruppen höher aus als unter den älteren Befragten. Diese Tendenz bleibt auch bestehen, wenn das politische Interesse kontrolliert wird.

Während die Hypothese postulierte, ältere Menschen würden stärker auf den Anreiz reagieren, da sie die Reziprozitätsnorm eher verinnerlicht haben dürften, so scheint dies folglich eher auf die Jüngeren zuzutreffen, wenngleich dieser Zusammenhang kein statistisch signifikantes Niveau erreicht.

Insgesamt ist somit nur bedingt erkennbar, dass sich die beiden Subgruppen im Hinblick auf ihre sozialstrukturelle Zusammensetzung voneinander unterscheiden. Zudem ist aus den Daten nicht abzulesen, dass ein bestimmtes Geschlecht besonders stark auf den Anreiz reagiert.

5.4 Qualität der Daten

Im Folgenden soll die Frage geklärt werden, ob eine Befragungsperson, der eine Belohnung zuteil wurde, beim Ausfüllen des Fragebogens eine größere Sorgfalt an den Tag legt. Die Qualität der Daten bzw. das Ausmaß an geleistetem Aufwand durch die Befragungsperson soll hierfür zum einen an der Zahl des Item-Nonresponse und zum anderen an der Anzahl der Worte gemessen werden, die die Befragungsperson bei offenen Abfragen verwendete. Zunächst richtet sich hier der Blick auf den Item-Nonresponse, also auf die Frage, ob es zwischen der Experimental- und der Kontrollgruppe dahingehend einen Unterschied gibt, dass Personen, die keinen Anreiz erhielten, im Mittel weniger Fragen beantworteten. Insgesamt wurden 68 Items identifiziert, die im Grunde von jeder Befragungsperson problemlos hätten beantwortet werden können. Tatsächlich zeigt sich hier ein größeres Engagement in der Experimentalgruppe: Im Mittel gaben die Probanden mit Belohnung nur auf 1,60 Fragen keine Antwort, während der Wert in der Kontrollgruppe mit 1,85 Items

etwas höher liegt. Allerdings handelt es sich hierbei eher um eine zufällig auftretende Differenz, da der Mittelwertunterschied nicht statistisch signifikant ist. Kontrolliert man zusätzlich das politische Interesse, das ebenfalls einen starken Einfluss auf Item-Nonresponse haben sollte, so ergibt sich das folgende Bild:

In der Gruppe der politisch eher Desinteressierten begünstigt ein Anreiz den Item-Nonresponse. Allerdings sind die Fallzahlen hier recht gering und der Mittelwertunterschied weist kein statistisch signifikantes Niveau auf. Dennoch würde dieser Zusammenhang eher dafür sprechen, dass die Belohnung gerade deshalb *nicht* zu einer höheren Datenqualität führt, da sie vornehmlich die Gruppe der Unentschlossenen anspricht, die sich zwar durch den Anreiz zur Teilnahme animieren lässt, dann aber nicht mehr als „Dienst nach Vorschrift" verrichtet. Dem entspricht, dass sich der Zusammenhang in der Gruppe der Befragungspersonen, die sich zumindest partiell für Politik interessieren, umkehrt: Hier füllen jene den Fragebogen besonders sorgfältig aus, die einen Anreiz erhalten haben. Der gleiche Zusammenhang zeigt sich, wenn auch in abgeschwächter Form, in der Gruppe der politisch Interessierten. Dennoch erreicht keine Mittelwertdifferenz statistisch signifikantes Niveau, weshalb es nahe liegt, dass die Verabreichung eines Anreizes keinen Einfluss auf den Item-Nonresponse ausübt.

Wie ist es weiterhin mit der Bereitschaft bestellt, auf offene Fragen zu antworten? Der Fragebogen eröffnete den Befragungsteilnehmern an einigen Stellen die Möglichkeit, eigene Vorschläge zu machen oder Ergänzungen vorzunehmen. Zunächst wurden die Befragungspersonen nach ihren Assoziationen mit dem Begriff des demografischen Wandels gefragt. Gegen Ende des Fragebogens zielten zwei weitere Fragen auf Reformen im System der gesetzlichen Rentenversicherung ab. Hier konnte der Befragte eigene Vorschläge formulieren. Schließlich blieb auf der letzten Seite des Fragebogens viel Platz für die Probanden, sowohl Anregungen als auch Kritik an der Befragung zu formulieren.

Bei allen betrachteten Antworten konnte kein Zusammenhang zwischen dem Erhalt eines Anreizes und der erbrachten Sorgfalt durch die Befragungsperson festgestellt werden. Richtet man den Blick auf die Anzahl der Worte, die die Teilnehmer gebrauchten, um ihre Assoziationen mit dem demografischen Wandel zu formulieren, so liegt der Mittelwert in der Experimentalgruppe (nur Befragte, die angaben, den Begriff demografischer Wandel zu kennen und gleichzeitig einen Anreiz erhielten) mit 15,2 zwar leicht höher als in der Kontrollgruppe (14,5). Der Mittelwertunterschied erweist sich jedoch nicht als statistisch signifikant. Zudem ergeben sich auch dann keine bedeutsamen Unterschiede, wenn die intrinsische Motivation bzw. das politische Interesse berücksichtigt werden. Das gleiche Bild zeigt sich bei den Antworten zu möglichen Reformoptionen in der gesetzlichen

Rentenversicherung: Dort liegt die mittlere Zahl der verwendeten Wörter in der Kontrollgruppe sogar höher als in der Experimentalgruppe (8,2 gegenüber 6,3), weist aber ebenfalls kein statistisch signifikantes Niveau auf. Auch die Gruppenvergleiche, die zusätzlich die intrinsische Motivation der Probanden kontrollieren sollen, ändern nichts an diesem Bild. Lediglich in jener Gruppe der Befragten, die angaben, sich auf mittlerem Niveau für Politik zu interessieren, liegt ein signifikanter Mittelwertunterschied vor: Diejenigen unter ihnen, die keine Belohnung erhielten, gebrauchen hier mehr Worte als die vergleichende Gruppe, der ein Anreiz beigelegt wurde. Keine Differenzen in der Zahl der gebrauchten Wörter finden sich schließlich auch im letzten Teil des Fragebogens, der für Anmerkungen und Kritik reserviert war: Hier liegt der Mittelwert in der Kontrollgruppe ein wenig höher (6,8) als in der Experimentalgruppe (6,4). Jedoch zeigt sich hier wiederum kein signifikanter Zusammenhang, der sich auch dann nicht einstellt, wenn zusätzlich das politische Interesse kontrolliert wird. Insgesamt offenbaren die Daten somit keinerlei Hinweise dafür, dass ein Anreiz einen Einfluss auf die Sorgfalt bei der Beantwortung der Fragen ausübt.

5.5 Bewertung des Interviews

Schließlich richtet sich der Fokus der Analyse auf die Frage, ob Personen, denen ein Anreiz zuteil wurde, das Interview positiver bewerten als die Residualgruppe. Diese Vermutung speist sich aus der Nutzen stiftenden Funktion des Anreizes: ein höherer Nutzen bei der Befragungsteilnahme könne demnach mit einer besseren Bewertung des Interviews einhergehen. Im Rahmen der vorliegenden Untersuchung kann die Bewertung des Interviews anhand zweier die Befragung abschließender Items untersucht werden, die eine Einschätzung der Befragten im Hinblick auf die Befragung selbst (von „sehr interessant" bis „sehr langweilig") und hinsichtlich des Umfangs des Fragebogens (von „viel zu lang" bis zu „hätte ruhig auch länger sein können") erbaten.

In Tabelle 2 sind die Mittelwerte in der Experimental- und Kontrollgruppe für diese beiden Items wiedergegeben. Die Ergebnisse lassen darauf schließen, dass sich die Bewertung des Interviews völlig unabhängig von der Vergabe einer Belohnung vollzieht. Auch unter Berücksichtigung der intrinsischen Motivation ändert sich nichts an diesen Ergebnissen.

Tabelle 2 Die Bewertung des Interviews durch die Respondenten in Abhängigkeit von der Gewährung eines Anreizes

	kein Anreiz	Anreiz erhalten
Wie interessant fanden Sie unsere Befragung? (1 „sehr langweilig" bis 5 „sehr interessant")	3,80	3,79
Fanden Sie den Fragebogen zu lang? (1 „ja, viel zu lang" bis 4 „nein, er hätte ruhig auch etwas länger sein können")	2,95	2,93

6 Schlussbetrachtung

Die Analysen zu erwünschten und unerwünschten Effekten eines Anreizes in postalischen Befragungen dürften den Forscher, der für seine Untersuchung den Einsatz einer Belohnung plant, durchweg erfreuen. Einmal mehr bestätigte sich die Rücklauf erhöhende Wirkung eines Anreizes. Darüber hinaus führt die Belohnung auch zu einer Erhöhung der Rücklaufgeschwindigkeit. Diese Kombination aus erhöhtem und beschleunigtem Rücklauf spart Zeit und vor allem Geld, da sie die Zahl derjenigen Befragungspersonen reduziert, die ein weiteres Mal angeschrieben werden müssen.

In weiten Teilen ist es ebenso erfreulich, dass sich einige weitere vermutete Zusammenhänge *nicht* bestätigen lassen. Dies stellt im Ergebnis ein positives Resultat für den Einsatz von Anreizen bei postalischen Befragungen dar. An vorderster Stelle ist hier der vermutete Einfluss der Belohnung auf die Zusammensetzung der realisierten Stichprobe zu nennen: So ist allenfalls bedingt zu erkennen, dass ein Anreiz bestimmte sozialstrukturelle Gruppen stärker anspricht als andere. Zudem führt der Einsatz einer Belohnung nicht zu einer größeren Sorgfalt der Befragungsperson beim Ausfüllen des Fragebogens. Schließlich unterscheiden sich Personen, denen eine Belohnung beigelegt wurde, auch nicht hinsichtlich ihrer Bewertung des Interviews von der Kontrollgruppe. Dies bedeutet im Umkehrschluss, dass man bei vergleichsweise wenig ansprechenden Themen und Fragebögen nicht darauf setzen sollte, dass der Einsatz eines Anreizes diese negativen Aspekte aufwiegt.

Gerade jene Ergebnisse, die sich auf die Rücklaufquote und Rücklaufgeschwindigkeit unter Verwendung eines Anreizes beziehen, reihen sich in einschlägige Forschungsresultate ein und untermauern somit die Wirksamkeit einer Belohnung mit Blick auf diese beiden Variablen. Was die übrigen Aspekte der Analyse anbelangt, kann jedoch längst noch nicht von gesicherten Erkenntnissen gespro-

chen werden. Die Zusammensetzung der realisierten Stichprobe, die Sorgfalt beim Ausfüllen des Fragebogens und die Bewertung des Interviews könnten Kategorien sein, die besonders stark mit dem Wert des Anreizes, der im Rahmen dieser Befragung treffend als symbolisch bezeichnet werden kann, korrelieren. Anders ausgedrückt: Es ist nicht unwahrscheinlich, dass ein im Vergleich zur Ein-Euro-Münze wertvollerer Anreiz durchaus die eben genannten Effekte hervorrufen kann. Somit stellen auch diese Resultate ein Plädoyer für Anreize dar, die nicht mehr als einen rein symbolischen Wert besitzen.

Literatur

Arzheimer, K. und M. Klein, 1998: Die Wirkung materieller Incentives auf den Rücklauf einer schriftlichen Panelbefragung. ZA-Information 43: 6-43. http://www.za.uni-koeln.de/publications/pdf/za_info/ZA-Info-43.pdf (9.9.2009).

Becker, R., R. Imhof und G. Mehlkop, 2007: Die Wirkung monetärer Anreize auf den Rücklauf bei einer postalischen Befragung und die Antworten auf Fragen zur Delinquenz. Empirische Befunde eines Methodenexperiments. Methoden – Daten – Analysen 1: 131-159. http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/jg-1-2007-heft-2/ (9.9.2009).

Blau, P., 1964: Exchange and power in social life. New York: Wiley.

Bogen, K., 1996: The effect of questionnaire length on response rates – a review of the literature. Proceedings of the Survey Research Methods Section of the American Statistical Association: 1020-1025.

Brehm, S. S. und J. W. Brehm, 1981: Psychological reactance: a theory of freedom and control. New York: Academic Press.

Brennan, M., 1992: The effect of a monetary incentive on mail survey response rates: new data. Journal of the Market Research Society 34: 173-177.

Church, A. H., 1993: Estimating the effect of incentives on mail survey response rates: a meta-analysis. Public Opinion Quarterly 57: 62-79.

Davern, M., T. H. Rockwood, R. Sherrod und S. Campbell, 2003: Prepaid monetary incentives and data quality in face-to-face interviews. Data from the 1996 Survey of Income and Program Participation Incentive Experiment. Public Opinion Quarterly 67: 139-147.

Dillman, D. A., 1978: Mail and telephone surveys. New York: Wiley.

Festinger, L., 1957: A theory of cognitive dissonance. Stanford: Stanford University Press.

Galesic, M. und M. Bosnjak, 2009: Effect of questionnaire length on participation and indicators of response quality in a web survey. Public Opinion Quarterly 73: 349-360.

Gouldner, A. W., 1960: The norm of reciprocity: A preliminary statement. American Sociological Review 25: 161-178.

Goyder, J., 1987: The silent minority: Nonrespondents on sample surveys. Cambridge: Polity Press.

Homans, G. C., 1961: Social behavior: its elementary forms. New York: Harcourt.

James, J. M. und R. Bolstein, 1990: The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. Public Opinion Quarterly 54: 346-361.

Jobber, D., J. Saunders und M. Vince-Wayne, 2004: Prepaid monetary incentive effects on mail survey response. Journal of Business Research 57: 21-25.

McDaniel, S. W. und C. P. Rao, 1980: The effect of monetary inducement on mailed questionnaire response quality. Journal of Marketing Research 17: 265-268.

Mehlkop, G. und R. Becker, 2007: Zur Wirkung monetärer Anreize auf die Rücklaufquote in postalischen Befragungen zu kriminellen Handlungen. Theoretische Überlegungen und empirische Befunde eines Methodenexperiments. Methoden – Daten – Analysen 1: 5-24. http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/mda/jg-1-2007-heft-1/ (9.9.2009).
Mizes, S. J., L. E. Fleece und C. Roos, 1984: Incentives for increasing return rates: Magnitude levels, response bias and format. Public Opinion Quarterly 48: 794-800.
Pötschke, M. und C. Müller, 2006: Erreichbarkeit und Teilnahmebereitschaft in Telefoninterviews: Versuch einer mehrebenenanalytischen Erklärung. ZA-Information 59: 83-99. http://www.za.uni-koeln.de/publications/pdf/za_info/ZA-Info-59.pdf (9.9.2009).
Sharp, L. M. und J. Frankel, 1983: Respondent burden: A test of some common assumptions. Public Opinion Quarterly 47: 36-53.
Shettle, C. und G. Mooney, 1999: Monetary incentives in U.S. Government Surveys. Journal of Official Statistics 15: 231-250.
Stadtmüller S. und R. Porst, 2005: Zum Einsatz von Incentives bei postalischen Befragungen. ZUMA How-to-Reihe Nr. 14. http://www.gesis.org/forschung-lehre/gesis-publikationen/gesis-reihen/how-to/ (9.9.2009).
Thibaut, J. W. und H. H. Kelly, 1959: The social psychology of groups. New York: Wiley.
Warriner, K., J. Goyder, H. Gjertsen, P. Hohner und K. McSpurren, 1996: Charities, no; lotteries, no; cash, yes. Public Opinion Quarterly 60: 542-562.
Wotruba, T. R., 1966: Monetary inducements and mail questionnaire research. Journal of Marketing Research 3: 393-400.

Anschrift des Autors

Sven Stadtmüller
Forschungszentrum Demografischer Wandel (FZDW)
Fachhochschule Frankfurt am Main
Nibelungenplatz 1
60318 Frankfurt am Main
svenstad@fzdw.fh-frankfurt.de

Zur Analyse von Wahlergebnissen in Parteihochburgen unter Berücksichtigung von Regressionsphänomenen

Methodological Contribution to the Analysis of Election Results in Party Strongholds

Thomas Ostermann und Rainer Lüdtke

Zusammenfassung
Regelmäßige Wahlen bilden ein Kernstück jeder demokratischen Verfassungsordnung. Insbesondere in den letzten drei Jahrzehnten hat sich die Wahlforschung im Bereich der politischen Wissenschaften mehr und mehr etabliert und spielt bei den Berichterstattungen sowohl im Vorfeld von Wahlen als auch in der Analyse von Wahlergebnissen eine entscheidende Rolle. Bei der Wahlanalyse sind vor allem die Änderungen in den sogenannten „Partei-Hochburgen" von besonderem Interesse. Dieser Artikel soll anhand von Daten der Landtagswahl in Hessen 2008 sowie der Bundestagswahl 2005 klären helfen, ob es sich bei den Verlusten, die alle großen demokratischen Parteien in ihren Hochburgen erlitten haben, tatsächlich um beunruhigende Entwicklungen oder um ein Regressionsphänomen zur Mitte handelt. Entsprechende statistische Verfahren und Modelle werden in diesem Zusammenhang vorgestellt, angewandt und diskutiert.

Abstract
Regular elections are at the heart of every democratic constitutional order. Particularly in the last three decades electoral studies have increasingly established themselves within the field of political sciences and now play an important role in media and press coverage prior to elections and in the analysis of election results. In this analysis the results in party strongholds are of particular interest to researchers. Based on data of the election to the German State Parliament of Hesse in 2008 and of the election to the German Bundestag in 2005 this article aims at answering the question whether the losses of all big democratic parties in their strongholds are a serious and alarming matter or whether to a certain extend these losses can be explained by regression-to-the-mean. Relevant statistical methods and models are introduced, applied and discussed.

1 Einleitung

Wahlen stellen in der demokratischen Ordnung ein wichtiges und öffentlich hoch diskutiertes Instrument der politischen Willensbekundung der Bevölkerung dar. Freie Wahlen gelten als Qualitätsmerkmal demokratischer Verfahren in der Politik, mit denen dem Bürger ein Recht auf Mitbestimmung und Partizipation zur Bestimmung der Repräsentanten des Volkes geboten wird (Derichs et al. 2006). Regelmäßige Wahlen bilden daher ein Kernstück jeder demokratischen Verfassungsordnung. In der Bundesrepublik Deutschland wird dies in Artikel 28, Absatz 1 und Artikel 38, Absatz 1 des Grundgesetzes für die Wahlen zum Bundestag und zu den Vertretungen in Ländern, Kreisen und Gemeinden nach dem Prinzip der allgemeinen, gleichen, unmittelbaren, geheimen und freien Wahl geregelt, welche in regelmäßigen Abständen stattfinden.

Insbesondere in den letzten drei Jahrzehnten hat sich die Wahlforschung im Bereich der demoskopischen Wissenschaften mehr und mehr etabliert und spielt bei den Berichterstattungen sowohl im Vorfeld von Wahlen als auch in der Analyse von Wahlergebnissen eine entscheidende Rolle (Brehm 1999). Neben den obligatorischen Umfrage- und Hochrechnungsszenarien, die mittlerweile bereits bei kommunalen Wahlen eingesetzt werden, sind Wahlanalysen immer dann von besonderem parteiinternen, aber auch öffentlichen Interesse, wenn sich Ergebnisse der vorhergegangenen Wahl nicht wiederholen. Neben globalen Wählerwanderungen wird in diesem Zusammenhang auch analysiert, in welchen Bezirken die deutlichsten Verluste aufgetreten sind.

Hier sind die sogenannten „Partei-Hochburgen" von besonderem Interesse, also diejenigen Gebiete, in denen eine Partei über einen längeren Zeitraum besonders hohe Stimmenanteile erhält. Nach Kirschey (2006) verfolgt die Analyse solcher Hochburgen das primäre Ziel aufzuzeigen, „wie das aktuelle Wahlergebnis in diesen Gebieten ausgefallen ist und ob bzw. welche Abweichungen zum Landesergebnis eingetreten sind." Dieses spielt auch in der öffentlichen Diskussion eine große Rolle. So hört man immer wieder die Behauptung, dass die eine oder andere Partei besonders in ihren Hochburgen „herbe Verluste" hinnehmen musste, zuletzt z. B. in einem Bericht der Südthüringischen Zeitung (2008) zur Landtagswahl 2008 in Hessen.

Auf den ersten Blick erscheint der Sinn einer solchen Analyse unmittelbar evident. Wählerverluste in Gebieten, in denen eine Partei bisher besonders viele Stimmen sammeln konnte, müssen per se beunruhigend sein. Aus wissenschaftlich-statistischer Sicht gilt dies nicht uneingeschränkt, es ist durchaus möglich, dass es sich bei den besonders herben Verlusten in Hochburgen um ein rein technisches Phänomen handelt, das als „Regression zur Mitte" bekannt ist.

Dieser Artikel soll anhand von Daten der Landtagswahlen in Hessen 2008 sowie der Bundestagswahl 2005 klären helfen, ob es sich bei den Verlusten, die alle großen demokratischen Parteien in ihren Hochburgen erlitten haben, tatsächlich um beunruhigende Entwicklungen oder um ein Regressionsphänomen handelt.

Ziel ist es dabei nicht, sich fokussierend auf diese beiden Wahlen zu beschränken. Vielmehr soll exemplarisch dargelegt werden, wie eine Analyse von Wahlhochburgen adäquat unter Berücksichtigung statistischer Sondereffekte erfolgen kann.

2 Regression zur Mitte

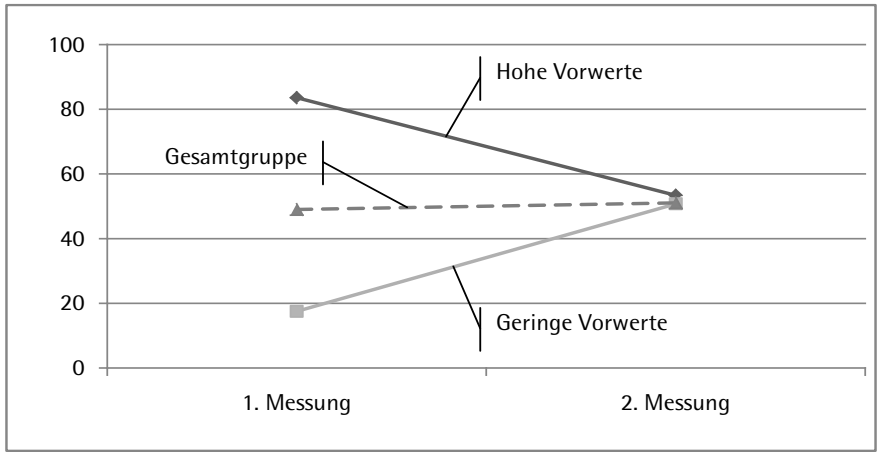
2.1 Beschreibung des Phänomens

Regression-zur-Mitte (regression-to-the-mean, RTM) ist ein rein statistisches Phänomen, das bei Wiederholungsmessungen auftritt, die an selektierten Populationen vorgenommen werden. Das Phänomen wurde erstmals von Galton (1887) bei der Analyse von Körpergrößen von Nachkommen auffällig großer bzw. kleiner Väter beschrieben. Demzufolge haben besonders große Väter tendenziell kleinere Söhne, während besonders kleine Väter Söhne bekommen, die größer sind als sie selbst.

Wie man heutzutage weiß, liegt dieser Beobachtung ein einfaches Gesetz zugrunde, das völlig unabhängig vom inhaltlichen Zusammenhang auftritt. Entscheidend ist der selektive Blick, mit dem die Daten analysiert werden. Man behält nicht mehr das Ganze im Auge sondern konzentriert sich ausschließlich auf eine Gruppe von Beobachtungen, die in ihrer Ausprägung besonders extrem sind, also besonders klein oder besonders groß. Ist man nun an der zeitlichen Fortentwicklung dieser extremen Gruppe interessiert (und blendet daher die „durchschnittlichen" Beobachtungen aus), so werden die Folgemessungen in dieser Subgruppe im Durchschnitt näher am Mittelwert der Gesamtgruppe liegen.

Dieses lässt sich einfach mithilfe von zwei Reihen von Zufallszahlen zwischen 0 und 100 erzeugen. Abbildung 1 zeigt bspw. ein Szenario, in dem auf zwei Subgruppen selektiert wurde, zum einen auf die Gruppe von Beobachtungen, bei denen bei der ersten Messung Werte über 70 gemessen wurden (extrem hohe Vorwerte), zum anderen die Gruppe mit Erstmessung unter 30 (extrem geringe Vorwerte). Dargestellt sind die Mittelwerte der beiden Zufallszahlenreihen für diese Subgruppen sowie die Mittelwerte der Gesamtgruppe aller Zahlen. Es ist deutlich zu sehen, dass die Mittelwerte der Wiederholungsmessung näher am Mittelwert der Gesamtpopulation liegen als die der Erstmessung (Lüdtke/Ostermann 2005).

Abbildung 1 Regression zur Mitte am Beispiel zweier Zufallszahlenreihen zwischen 0 und 100



2.2 Mathematischer Hintergrund

Mathematisch lässt sich dieser Sachverhalt über eine bivariate Verteilung beschreiben. Im obigen, künstlichen Beispiel wäre eine bivariate Gleichverteilung angebracht, aus mathematischen Gründen wird üblicherweise aber eine bivariate Normalverteilung angenommen. Bezeichnet man mit X die Erstmessung eines Merkmals, so lässt sich die Zweitmessung Y über eine lineare Regressionsgleichung auf die Erstmessung zurückführen:

$$Y - \mu = \rho(X - \mu) + \varepsilon, \qquad \varepsilon \approx N\left(0, \sigma^2(1-\rho^2)\right) \tag{1}$$

Dabei bezeichnet μ den Mittelwert der Verteilung (der bei beiden Messungen ja gleich ist), σ^2 deren Varianz und ρ die Korrelation zwischen beiden Messungen, die üblicherweise zwischen 0 und 1 liegt; ε ist ein normalverteilter Fehler mit Erwartungswert 0 und Varianz . Damit ergibt sich für den bedingten Erwartungswert von Y bei gegebenem $X = x$

$$Y - \mu = \rho(X - \mu) + \varepsilon, \qquad \varepsilon \approx N\left(0, \sigma^2(1-\rho^2)\right) \tag{2}$$

Dieser unterscheidet sich von x um

$$\Delta = x - E(Y \mid X = x) = x - [(1-\rho)\mu + \rho x] = (1-\rho)(x-\mu) \tag{3}$$

woraus sich unmittelbar der RTM-Effekt ablesen lässt: Ist x größer als μ, so ist Δ größer als 0, also x größer als der zu erwartende Wert bei der Zweitmessung; falls $x < \mu$ folgt genau das Gegenteil. Dieser Effekt ist um so größer, erstens, je größer der tatsächlich gemessene Wert x ist und zweitens, je geringer die Korrelation ρ zwischen X und Y ist.

Dieses gilt auch dann noch, wenn man die Voraussetzung aufgibt, dass die Verteilung zu beiden Messzeitpunkten stabil ist, d. h. der Mittelwert μ bei Erst- und Zweitmessung gleich ist. Bezeichnet man mit τ die Veränderung, die allen Messwerten gleichartig widerfährt (also etwa landesweite Verluste bzw. Gewinne einer Partei), so folgt aus

$$Y - \mu = \tau + \rho(X - \mu) + \varepsilon, \qquad \varepsilon \approx N\left(0, \sigma^2(1-\rho^2)\right) \tag{4}$$

$$E(Y \mid X = x) = (1-\rho)\mu + \tau + \rho x \tag{5}$$

dass für besonders große x der erwartete Gewinn kleiner ist als der durchschnittliche Gewinn, bzw. der erwartete Verlust größer ausfällt. Umgekehrt wird für kleine x ein größerer Gewinn bzw. kleinerer Verlust erwartet.

Gleichung (4) kann dazu benutzt werden, einen statistischen Test zu konstruieren, der den Effekt einer Intervention bewertet, die nur in einer selektierten Gruppe von besonders auffälligen Merkmalsträgern (z. B. Wahlhochburgen) durchgeführt wird. Dazu ist es lediglich notwendig, den wahren Wert von μ zu kennen. Dieses ist letztendlich der t-Test innerhalb des linearen Regressionsmodells von Y-μ auf X-μ, ob der Achsenabschnitt (hier τ) von 0 verschieden ist oder nicht. Die Teststatistik t ist gegeben durch

$$t = \frac{\hat{\tau}}{\sqrt{s^2\left(1/n + \overline{X}^2 \Big/ \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)}} \tag{6}$$

wobei $\hat{\tau}$ den Schätzer des Achsenabschnitts und s^2 den Mean Squared Error (MSE) aus der Regressionsanalyse bezeichnet. X_i sind die Werte der Erstmessung des i-ten Merkmalsträgers aus der selektierten Gruppe und $\overline{X}$ ihr Mittelwert, $i=1,\ldots,n$. t ist gemäß einer t-Verteilung mit (n-2) Freiheitsgraden verteilt.

Eine ausführliche Darstellung und Herleitung des Tests findet sich bei Mee und Chua (1991). Dort findet sich auch ein illustratives Beispiel zu acht Studenten, die bei einer Prüfung für einen Highschool-Abschluss mit einem durchschnittlichen Punktewert von 57,4 durchfielen. Nach einem speziell ausgerichteten Auffrischungskurs stieg die Prüfungsleistung auf 60,4 Punkte an. Auf den ersten Blick ist dies ein klares Indiz dafür, dass der Auffrischungskurs erfolgreich war; eine Interpretation, die vor allem auch durch das Ergebnis eines verbundenen t-Tests mit einem einseitigen p-Wert von p=0,0428 gestützt wird. Diese Sichtweise berücksichtigt aber nicht, dass es sich bei den Prüflingen ausschließlich um diejenigen Studenten handelte, die in der ersten Prüfung besonders schlecht abgeschnitten hatten und damit einer selektierten Gruppe angehörten. RTM-Effekte, und damit verbesserte Prüfungsergebnisse in der Wiederholungsprüfung, waren daher zu erwarten. Statt eines verbundenen t-Tests wäre daher ein korrigierter t-Test gemäß Gleichung (6) angebracht gewesen. Tabelle 1 zeigt die Einzeldaten der acht Studenten. Nimmt man nun wie Mee und Chua an, dass das mittlere Prüfungsergebnis aller Studenten bei der Erstprüfung bei μ=75 Punkten liegt, dann lässt sich der korrigierte t-Test mit jedem statistischen Programmpaket leicht berechnen, indem man zunächst von allen Einzelwerten (bei Erst- und Wiederholungsprüfung) 75 abzieht und dann eine einfache lineare Regressionsanalyse durchführt. Im Falle der obigen acht Studenten ergibt sich ein t-Wert von t=1,08 und demzufolge ein (einseitiger) p-Wert von p=0,16, also kein statistisch signifikantes Ergebnis, und damit kein Indiz dafür, dass der Auffrischungskurs den gewünschten Effekt hatte.

Tabelle 1 Prüfungsleistungen von acht Studenten, die bei einer Prüfung durchgefallen waren, und nach einem Ergänzungskurs eine Wiederholungsprüfung absolvierten

Student	Leistung bei Erstprüfung (Punkte)	Leistung bei Wiederholungsprüfung (Punkte)
1	45	49
2	52	50
3	63	70
4	68	71
5	57	53
6	55	61
7	60	62
8	59	67

2.3 Regressionseffekte im sozialwissenschaftlichen Kontext

Wie das in Kapitel 2.1 beschriebene Beispiel der Zufallszahlen zeigt, bedarf es keinerlei inhaltlich begründeter Mechanismen, um einen Regressionseffekt auszulösen. Entscheidend ist lediglich der selektive Blick auf die Daten. Damit ist das Phänomen universell und natürlich auf jede andere Datenerhebung im sozial- oder politikwissenschaftlichen Kontext anwendbar. Nachtigall und Suhl (2002) geben hierzu ein eindrucksvolles Beispiel: Nach Furby (1973) findet sich bei Kindern von besonders intelligenten Eltern eine durchschnittlich geringere Intelligenz. Da die Intelligenz eines Menschen positiv mit seinem sozioökonomischen Status korreliert, wurden diese Daten verschiedentlich dahingehend interpretiert, dass ein höher sozioökonomischer Status kausal eine (relative) Minderbegabung in der Folgegeneration bedingt; eine Interpretation, die den dahinter liegenden Regressionseffekt vollständig vernachlässigt und daher als nicht zwingend bewertet werden muss.

Bei der Analyse von Wahlergebnissen stellt eine Fokussierung auf Wahlhochburgen, zumindest wenn man diese nur auf Basis der Ergebnisse der vorhergehenden Wahl definiert, ebenfalls einen Selektionsprozess im obigen Sinne dar: Betrachtet werden dabei (ausschließlich) Messungen am oberen Rand der Verteilung, Messwerte aus der Mitte oder gar dem unteren Rand der Verteilung werden ausgeblendet oder zumindest weniger stark berücksichtigt. Dementsprechend kommt es zwangsläufig zu Regressionseffekten und damit möglicherweise zu Fehlinterpretationen der Daten.

3 Daten

Nach diesen Vorbemerkungen ist also zu erwarten, dass eine Partei in einem Gebiet, in dem sie ein besonders gutes Wahlergebnis erzielt hat, bei der Folgewahl überdurchschnittlich verliert (bzw. falls sie gewonnen hat, dass sie unterdurchschnittlich gewinnt). Nur wenn diese überdurchschnittlichen Verluste (bzw. unterdurchschnittlichen Gewinne) nicht durch Regressionseffekte zu erklären sind, hat dieses eine Bedeutung, die einer detaillierteren Ursachenanalyse bedarf.

Wir haben daher für CDU, SPD, FDP und Bündnis 90/die Grünen die Wahlergebnisse der hessischen Landtagswahlen vom 27.1.2008 mit denen vom 2.2.2003 verglichen. Datengrundlage war jeweils der Anteil der Landesstimmen in allen

55 hessischen Wahlkreisen, wie sie auf den entsprechenden Webseiten des Hessischen Statistischen Landesamtes veröffentlicht wurden.[1]

Analoge Analysen können für die Bundestagswahlen vom 22.9.2002 und 18.9.2005 durchgeführt werden. Hier beziehen wir uns auf die Anteile der Zweitstimmen je Wahlkreis, entsprechend den Veröffentlichungen des Bundeswahlleiters.[2]

3.1 Analyse der hessischen Landtagswahlen

Abbildung 2 zeigt die Verluste bzw. Gewinne der vier großen Parteien je Wahlkreis bei den Landtagswahlen in Hessen 2008 im Vergleich zu den Stimmanteilen bei den Wahlen 2003. Für alle vier Parteien wird der RTM-Effekt durch die eingezeichnete Regressionsgerade eindrucksvoll belegt: je größer der Stimmanteil 2003 war desto geringer waren die Gewinne 2008 (SPD, FDP) bzw. desto größer die Verluste (CDU, Grüne).

Besonders deutlich wird das bei den Grünen. Sie hatten 2003 im Wahlkreis Frankfurt/Main V einen Stimmanteil von 26,8 % und verloren hier mit 9,8 Prozentpunkten überdurchschnittlich; landesweit betrug der Verlust nur 2,6 Prozentpunkte. Abbildung 2 (c) belegt nun, dass es sich dabei nicht etwa um (im statistischen Sinn) unerwartete Verluste handelte: genau dieser Wert war aufgrund des RTM-Effekts prognostizierbar. In ähnlicher Weise ist der unterdurchschnittliche Gewinn der FDP von 1,4 Prozentpunkten in ihrer Hochburg Hochtaunus II zu interpretieren (landesweite Gewinne 1,5 Prozentpunkte): Nicht der geringere Gewinn ist überraschend, sondern eher die Tatsache, dass der tatsächliche Gewinn nur so wenig unterhalb der landesweiten Ergebnisse lag (Abbildung 2 (d)).

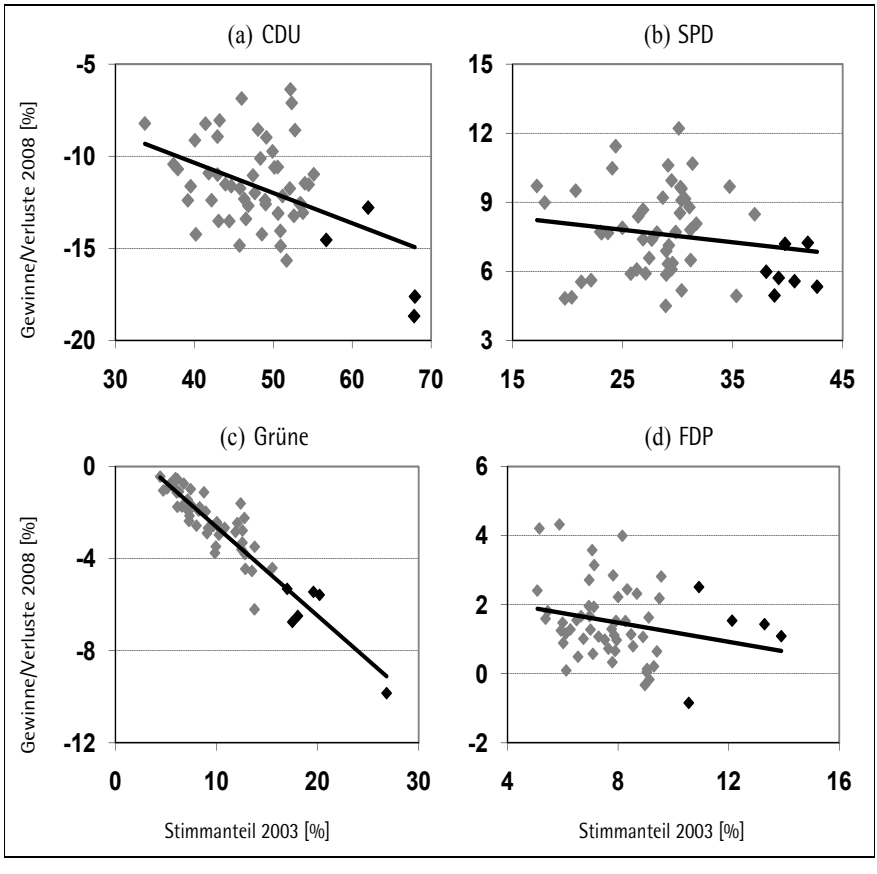
Anders bei der CDU, die in Hessen landesweit durchschnittlich zwölf Prozentpunkte verlor. Ihre größten Verluste (18,7 % und 17,6 %) hatte sie genau in den Wahlkreisen Fulda I und Fulda II zu verzeichnen, bei denen sie 2003 die höchsten Stimmanteile hatte (jeweils 67,9 %). Abbildung 2 (a) bestätigt den RTM-Effekt zwar grundsätzlich, zeigt aber auch, dass aufgrund dieses Effekts in beiden Kreisen lediglich ein Verlust von etwa 15 Prozentpunkten zu erwarten war.

Nimmt man nun an, dass Parteien in ihren Wahlhochburgen einen anderen Wahlkampf führen als in den restlichen Wahlkreisen (z. B. weil man die dortigen Wähler besonders motivieren möchte oder man sich umgekehrt deren Stimmen sehr sicher ist und daher nur wenig Mittel einsetzt), so stellt sich unmittelbar die Frage, ob die gewählte Strategie erfolgreich war oder nicht. Mit anderen Worten: es ist zu über-

1 http://www.statistik-hessen.de/subweb/ltw2003/endg/tabz999001.htm;
http://www.statistik-hessen.de/subweb/ltw2008/index.htm.
2 http://www.bundeswahlleiter.de/de/bundestagswahlen/BTW_BUND_05/downloads/.

prüfen, ob sich das Ergebnis in den Wahlhochburgen signifikant vom Gesamttrend der Partei unterscheidet. Statistisch lässt sich das durch das Einfügen eines Interventionseffekts τ lösen, wie dieses in Gleichung (4) geschehen ist. Mit Gleichung (6) steht dann ein Test zur Verfügung, der genau diese Frage überprüft. Völlig falsch wäre es dagegen, einen einfachen verbundenen t-Test zu verwenden, da dieser den Selektionsmechanismus, nur extreme Werte zu analysieren, nicht berücksichtigt und daher Veränderungen zur Mitte ausschließlich auf den Interventionseffekt zurückführt.

Abbildung 2 Gewinne und Verluste der einzelnen Parteien je Wahlkreis bei der Landtagswahl in Hessen 2008 im Vergleich zur Wahl 2003



„Hochburgen" schwarz gekennzeichnet. Zur Illustration haben wir alle diejenigen Wahlkreise als Hochburgen definiert, in denen die Partei 2003 deutlich über dem mittleren Wahlergebnis lag (in Abbildung 2 hervorgehoben), wobei die Definition von „deutlich" allein auf der Basis des grafischen Eindrucks getroffen wurde.

Tabelle 2 Tests auf überdurchschnittliche Verluste (bzw. unterdurchschnittliche Gewinne) in auf der Basis von Stimmanteilen im Wahljahr 2003 willkürlich definierten Hochburgen im Vergleich der hessischen Landtagswahlen von 2003 und 2008

Partei	Definition Hochburg %	N	Wahl 2003 %	Wahl 2008 %	Test auf Interventionseffekt *
CDU	> 56	4	63,6	47,7	t_{verb}: p = 0,027 $t_{korr.}$: p = 0,345
SPD	> 38	7	40,1	46,1	t_{verb}: p = 0,002 $t_{korr.}$: p = 0,812
FDP	> 10	5	12,2	13,3	t_{verb}: p = 0,282 $t_{korr.}$: p = 0,697
Grüne	> 16	6	19,9	13,3	t_{verb}: p = 0,001 $t_{korr.}$: p = 0,975

** p-Werte sind einseitig.*
t_{verb} *– verbundener t-Test,* $t_{korr.}$ *– nach Mee & Chua korrigierter t-Test gemäß Gleichung (6).*

Tabelle 2 zeigt, dass für keine Partei überdurchschnittliche Verluste in ihren jeweiligen Hochburgen reklamiert werden können, zumindest wenn man den um den RTM-Effekt korrigierten t-Test (Gleichung (6)) zugrunde legt. Mit anderen Worten: ein eventuell in diesen Hochburgen anders geführter Wahlkampf hätte keinen nachweisbaren Effekt gehabt. Ein Urteil, das völlig anders ausgesehen hätte, wenn der einfache verbundene t-Test als Entscheidungsgrundlage verwendet worden wäre. Hier hätten CDU, SPD und Grüne die in ihren Hochburgen erzielten Ergebnisse als auffällig beurteilt.

3.2 Analyse der Bundestagswahl 2005

Bei der Bundestagswahl 2005 zeigt sich ein ähnliches Ergebnis (Abbildung 3), was nach dem bisher Gesagten auch kaum anders zu erwarten war: In ihren jeweiligen Hochburgen gewinnen alle Parteien unterdurchschnittlich hinzu oder verlieren überdurchschnittlich. Aufgrund eines Sondereffekts muss man bei der Analyse allerdings eine Einschränkung machen: sie gilt für Westdeutschland mit Ausnahme des Saarlands. Dieser Sondereffekt liegt im Wahlergebnis der PDS begründet, die 2005 besonders zugewinnen konnte, ihre Gewinne aber überwiegend in Ostdeutschland und dem Saarland erzielte (hier trat der ehemalige SPD-Ministerpräsident Oskar Lafontaine für die PDS an). Diese Gewinne gingen fast ausschließlich zu Lasten der SPD, so dass deren Verluste hier besonders groß waren und aus dem Gesamtbild herausfallen.

Abbildung 3 Gewinne und Verluste der einzelnen Parteien je Wahlkreis in Westdeutschland (außer Saarland) bei der Bundestagswahl 2005 im Vergleich zur Wahl 2002

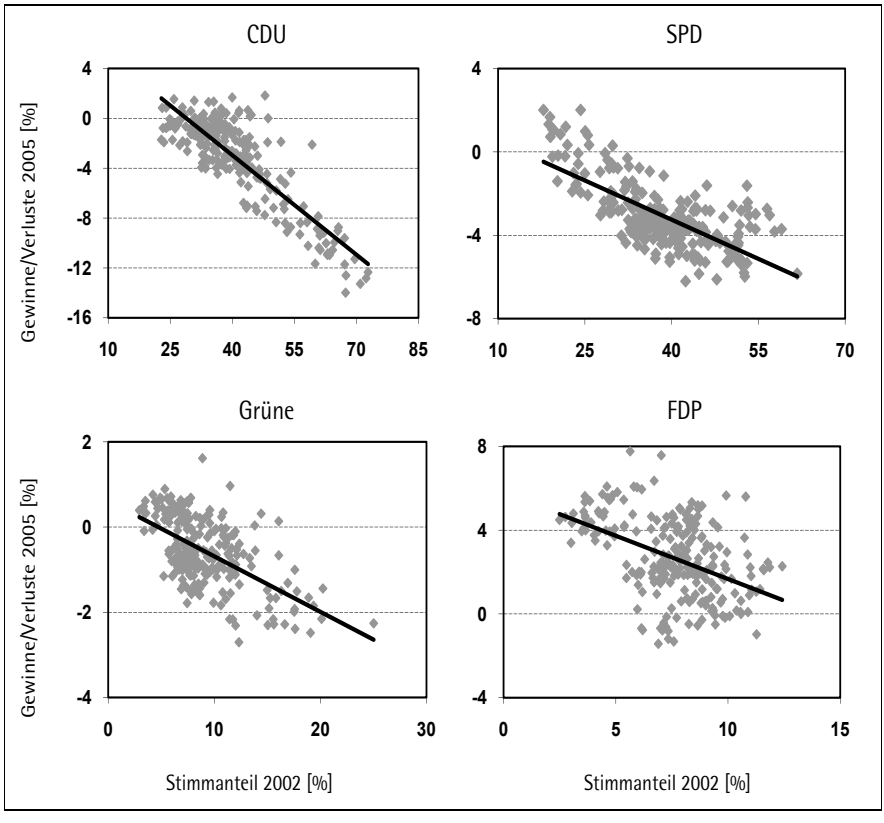


Tabelle 3 zeigt zudem, dass weder bei der CDU noch bei FDP oder Grünen Zusatzeffekte in den Hochburgen nachweisbar sind. Bei der SPD ist ein solcher allerdings mit dem korrigierten t-Test auch statistisch nachweisbar: die Partei verlor in ihren Hochburgen zwar leicht überdurchschnittlich, aber deutlich weniger als aufgrund des RTM-Effekts zu erwarten gewesen wäre.

Tabelle 3 Tests auf überdurchschnittliche Verluste (bzw. unterdurchschnittliche Gewinne) in willkürlich definierten Hochburgen im Vergleich der Bundestagswahlen von 2005 und 2002

Partei	Definition Hochburg %	N	Wahl 2002 %	Wahl 2005 %	Test auf Interventionseffekt*
CDU	> 67,0	8	69,4	57,1	$t_{korr.}$: p = 0,244
SPD[1]	> 55,0	5	58,4	54,6	$t_{korr.}$: p = 0,004
FDP	> 11,5	5	11,8	14,1	$t_{korr.}$: p = 0,767
Grüne	> 20,0	5	22,2	20,2	$t_{korr.}$: p = 0,593

** p-Werte sind einseitig, 1 nur Westdeutschland außer Saarland.*
$t_{korr.}$ *– nach Mee & Chua korrigierter t-Test.*

4 Diskussion

In der Analyse von Wahlergebnissen spielen viele Effekte eine Rolle, bspw. der von Lazarsfeld et al. (1969) beschriebene Bandwagon-Effekt oder Mitläufer-Effekt (Anschluss an die Mehrheitsmeinung, z. B. um schließlich als Sieger dazustehen) und der Underdog-Effekt (Anschluss an die Minderheitsmeinung, z. B. aus Trotz), die speziell für Wahlhochburgen von Interesse sind (Schoen 2002). Aber auch Interventionen auf Seiten der Partei sind zu berücksichtigen: So ist z. B. denkbar, dass Parteien in Wahlkreisen, in denen sie in der Wahl zuvor sehr gut abgeschnitten haben, einen weniger intensiven Wahlkampf betreiben, bzw. der politische Gegner – auch wegen des Anreizes der Überhangmandate – besondere Anstrengungen unternimmt. Andererseits ist es auch möglich, dass Anhänger der Parteien, die zuvor sehr gut abgeschnitten haben, bei der kommenden Wahl aufgrund einer eher moderaten Werbung der Partei weniger mobilisiert werden.

Überlagert werden diese Effekte aber in jedem Fall vom reinen Regressionseffekt, auf den wir uns in der hier dargestellten Analyse beschränkt haben. Dieser Regressionseffekt ist ein wenig beachtetes statistisches Artefakt, das allein auf Selektionsbedingungen zurückzuführen ist und zunächst keine inhaltliche Entsprechung hat. Erst Abweichungen vom erwarteten Regressionseffekt können die Grundlage für eine inhaltliche Diskussion und Interpretation von Wahlergebnissen aus Wahlhochburgen sein. Auf diesem Wege ist es durchaus möglich, die oben genannten Effekte als „Interventionseffekte" zu modellieren (obwohl ja im engeren Sinne keine Intervention vorliegt) und die von uns vorgeschlagene Technik, den korrigierten t-Test, anzuwenden.

Dabei ist zu beachten, dass es sich in unseren Beispielen um reine prä-post Szenarien ohne Baseline-Absicherung handelt. Bezogen auf die Parteihochburgen ist es daher unwahrscheinlich, dass der korrigierte Test auch dann die entsprechenden Resultate ergibt, wenn er auf Hochburgen angewandt wird, die konstant über mehrere Jahre von einer Partei dominiert wurden und plötzlich einbrechen. Für eine solche Entwicklung sind die Gründe eher inhaltlich, bspw. im Wechsel/Rücktritt eines Spitzenkandidaten zu suchen.

Weiterhin ist in der Interpretation unserer Ergebnisse zu berücksichtigen, dass es sich bei den verwandten Daten bereits um Aggregatdaten handelte. Diese sind sicher nicht ausreichend, um ein individuelles Wahlmodell zu testen, was aber auch nicht Intention der vorliegenden Arbeit war. Trotzdem könnte mit entsprechenden Daten möglicherweise noch deutlicher gezeigt werden, dass die größeren Verluste der Parteien in den Hochburgen der Wahl 2005 auf das Phänomen der Regression zur Mitte zurückzuführen sind und nicht auf tatsächlich andere Kalküle der Wähler. Ein Individualmodell könnte dann Aufschluss über die verbleibenden Effekte geben, wie diese in der Bundestagswahl 2005 bei der SPD auch bei Anwendung des korrigierten Tests deutlich hervortreten (siehe Tabelle 3).

Da der Status einer Wahlhochburg multifaktoriell begründet ist und je nach Partei unterschiedlich ausfällt, ist sowohl der zugrunde liegende Selektionsmechanismus für den Regressionseffekt als auch der Einfluss anderer Parameter nur schwer modellierbar. Dementsprechend und vor dem Hintergrund des Vorliegens von Daten aus der Grundgesamtheit der Wähler (einschließlich eines tatsächlichen Mittelwerts) bietet sich für dieses Szenario das Verfahren von Mee und Chua in der hier vorgestellten Form optimalerweise an.

Am Beispiel von Wahlergebnissen in Hochburgen konnte hier mit dem letztgenannten Verfahren eine solche Analyse dargestellt werden. Auch wenn in den vorgestellten Beispielen die Art der Wahlkampfstrategie in Hochburgen nicht explizit bekannt ist, so sprechen Arbeiten zur Wahlforschung dafür, dass Hochburgen von Parteien sowohl in der Phase des Wahlkampfs (Hoecker 2005) als auch in der Analyse des Wählerverhaltens (Schoon 2006; Probst 2007) eine besondere Bedeutung für die Parteien besitzen. Gerade in diesen Regionen sollen Stammwähler der Partei mobilisiert werden. Umgekehrt wird es in den entsprechenden Analysen immer wieder als bedeutsam hervorgehoben, wenn andere Parteien eine Hochburg für sich erobern.

Insgesamt hat die Auseinandersetzung mit diesem Phänomen der Regression zur Mitte und dessen Modellierung seit seiner ersten Formulierung durch Galton zu einer umfangreichen Bibliographie in allen Bereichen der quantitativen Wissenschaften geführt. In der Medizin z. B. stellt sich die zentrale Frage, wie groß ein solcher Regressionseffekt ist und wie er in epidemiologischen Längsschnitter-

hebungen von einem Effekt einer gesundheitsrelevanten Intervention abzugrenzen ist. So wird z. B. die Frage nach dem Erfolg einer rehabilitationsmedizinischen Behandlung in der Reduktion von Arbeitsunfähigkeitstagen (AU-Tagen) gemessen. Seit den Längsschnittuntersuchungen von Wagner (1977), in denen sich Personen mit vielen AU-Tagen in den nächsten Jahren deutlich in ihrem Gesundheitszustand verbesserten wird dieser Effekt unter dem Stichwort „AU-Trend" kontrovers diskutiert (Zwingmann/Wirtz 2005).

Aber auch bei anderen Szenarien der empirischen Sozialforschung taucht dieses Phänomen auf: So berichten Tversky und Kahnemann (1974), dass sich Piloten nach einer guten (besonders sanften) Landung auf einem Flugzeugträger in den Folgelandungen deutlich verschlechtern. In einer anderen Untersuchung stellte Secrist (1933) fest, dass große Firmen mit der Zeit an Bedeutung (gemessen an deren Umsatz) verlieren. Lee und Smith (2002) konnten in einem ähnlichen gelagerten Szenario zeigen, dass die Perfomance von erfolgreichen Football-Vereinen sich in den Folgejahren verschlechtert und daß diese Beobachtung auf einen Regressionseffekt zurückzuführen ist. Tatsächlich lassen sich solche Szenarien mit den mittlerweile zur Verfügung stehenden statistischen Methoden recht gut voraussagen, wenn entweder der Selektionsprozess modelliert werden kann (Varghese 1997) oder der tatsächliche Mittelwert der Grundpopulation bekannt ist (Mee/Chua 1991). Aber auch wenn der tatsächliche Mittelwert in einem bestimmten Intervall vermutet wird, so kann mit einer Erweiterung des Algorithmus von Mee und Chua eine Aussage über die Wahrscheinlichkeit des Vorliegens eines Regressionseffekts gemacht werden (Ostermann et al. 2008).

Literatur

Brehm, J., 1999: Alternative corrections for sample truncation: Applications to the 1988, 1990, and 1992 senate election studies. Political Analysis 8(2): 183-199.

Derichs, C., T. Heberer und J. Hippler, 2006: Wahlen und Regierbarkeit im globalen Rahmen. S. 11-24 in: T. Heberer und C. Derichs (Hg.): Wahlsysteme und Wahltypen – Politische Systeme und regionale Kontexte im Vergleich, Wiesbaden: VS Verlag für Sozialwissenschaften.

Furby, L., 1973: Interpreting regression toward the mean in developmental research. Developmental Psychology 8(2): 172-179.

Galton, F., 1886: Regression towards mediocrity in hereditary stature. Journal of the Anthropological Institute 15: 246-263.

Hoecker, M., 2005. Die Oberbürgermeisterwahl in Stuttgart 1996 – Parteipolitik und Wahlkampfstrategie: die kommunale Persönlichkeitswahl im Spannungsfeld der modernen Parteiendemokratie. Eine Einzelfallstudie. Dissertation Universität Stuttgart.

Kirschey, T., 2005: Bundestagswahl 2005 – regionale Parteihochburgen und ihre Strukturen. Statistische Monatshefte Rheinland-Pfalz 10: 593-604.

Lazarsfeld, P. F., B. R. Berelson und H. Gaudet, 1969: Wahlen und Wähler. Neuwied und Berlin: Verlag Luchterhand.

Lee, M. und G. Smith, 2002: Regression to the mean and football wagers. Journal of Behavioral Decision Making 15(4): 329-342.
Lüdtke, R. und T. Ostermann, 2005: Regression zur Mitte – ein Thema in der Krebsforschung? Deutsche Zeitschrift für Onkologie 37(3): 169-175.
Mee, R. und T. Chua, 1991: Regression Toward the mean and the paired sample t-test. American Statistician 45(1): 39-42.
Nachtigall, C. und U. Suhl, 2002: Der Regressionseffekt – Mythos und Wirklichkeit. Methevalreport 4(2). Jena: Lehrstuhl für Psychologische Methodenlehre und Evaluationsforschung am Institut für Psychologie der Friedrich-Schiller-Universität Jena: http://www.metheval.uni-jena.de/materialien/reports/report_2002_02.pdf (24.7.2009).
Ostermann, T., S. N. Willich und R. Lüdtke, 2008: Regression toward the mean – a detection method for unknown population mean based on Mee and Chua's algorithm. BMC Medical Research Methodology 8: 52.
Probst, L., 2007: Vorwahlanalyse zur Bürgerschaftswahl 2007. Institut für Politikwissenschaft Universität Bremen.
Schoen, H., 2002: Wirkungen von Wahlprognosen auf Wahlen. S. 171-191 in: T. Berg (Hg.): Moderner Wahlkampf, Opladen: Leske und Budrich.
Schoon, S., 2006: Wählerverhalten und Strukturmuster des Parteienwettbewerbs in Mecklenburg-Vorpommern nach der Landtagswahl 2006. Rostocker Informationen zu Politik und Verwaltung 27: 9-20.
Secrist, H., 1933: Triumph of mediocrity in business. Chicago: Bureau of Business Research, Northwestern University.
Südthüringische Zeitung 29.01.2008: Wahl in Hessen: Herbe Verluste in CDU-Hochburgen. Rhön im Landestrend.
Tversky, A. und D. Kahneman, 1974: Judgment under uncertainty: Heuristics and Biases. Science 185: 1124 – 1131.
Varghese, G., W. D. Johnson, A. Shahane und T. G. Nick, 1997: Testing for treatment effect in the presence of regression toward the mean. Biometrics 53 (1): 49-59.
Wagner, H,, 1977: Fehlerquellen bei Kurerfolgsbeurteilungen mittels Arbeitsausfallzeiten wegen Krankheit. Zeitschrift für Physiotherapie 29: 313-338.
Zwingmann, C. und M. Wirtz, 2005: Regression zur Mitte. Die Rehabilitation 44(4): 244-251.

Anschrift der Autoren

PD Dr. Thomas Ostermann
Lehrstuhl für Medizintheorie, Integrative und Anthroposophische Medizin
Universität Witten/Herdecke
Gerhard-Kienle-Weg 4
58313 Herdecke
thomaso@uni-wh.de

Rainer Lüdtke
Karl and Veronica Carstens Stiftung
Am Deimelsberg 36
45276 Essen
r.luedtke@carstens-stiftung.de

Entwicklung einer neuen fehlertoleranten Methode bei der Verknüpfung von personenbezogenen Datenbanken unter Gewährleistung des Datenschutzes

Development of a New Method for Privacy-Preserving Record Linkage Allowing for Errors in Identifiers

Rainer Schnell, Tobias Bachteler und Jörg Reiher

Zusammenfassung
Die Verknüpfung der Angaben mehrerer Datenbanken über dieselbe Person wird immer häufiger für Forschungszwecke genutzt. Aus Datenschutzgründen müssen die Identifikatoren in vielen Fällen vor der Zusammenführung verschlüsselt werden. Bisher verwendete Techniken sind hierbei ineffizient, da Fälle mit Fehlern in den Identifikatoren fast immer vollständig verloren gehen. Die Autoren haben ein neues Verfahren entwickelt, das trotz starker Verschlüsselung Fehler in den Identifikatoren toleriert. Testergebnisse anhand simulierter und echter Datenbestände zeigen, dass das Verfahren ähnlich gute Ergebnisse erbringt wie unverschlüsselte Identifikatoren. Das Verfahren kann für viele Probleme in der Forschungspraxis der empirischen Sozialforschung verwendet werden.

Abstract
Combining multiple databases with additional information on the same person is increasingly occurring throughout research. In many applications, identifiers have to be encrypted due to privacy concerns. Existing protocols are inefficient in actual research practice since cases with errors in identifiers are almost always in their entirety lost. Therefore, a new protocol for privacy-preserving record linkage with encrypted identifiers allowing for errors in identifiers has been developed by the authors. The results from tests on simulated and actual databases are comparable to non-encrypted identifiers. This new technique will have many practical applications in social research.

1 Problemstellung

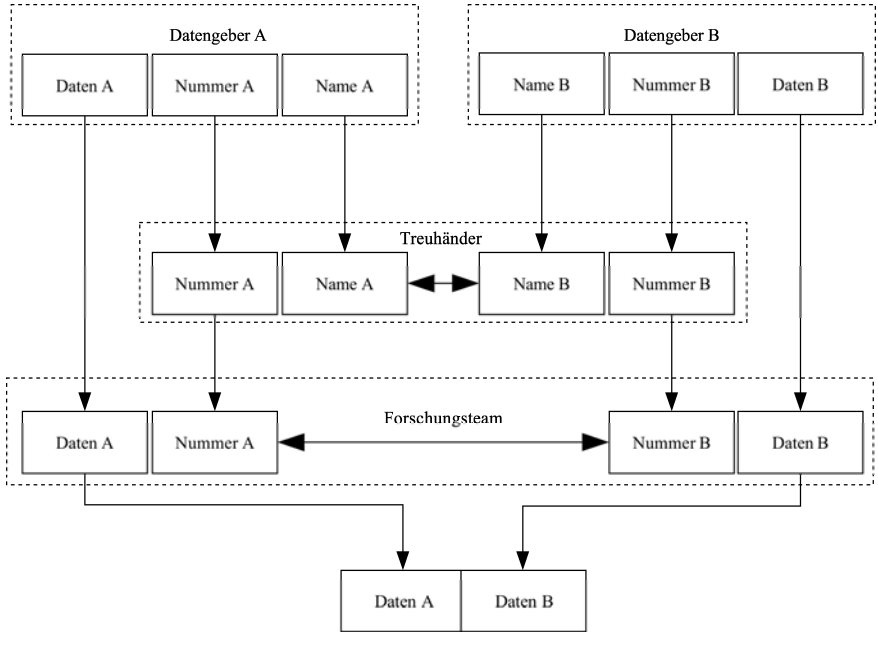
Weltweit werden in zunehmendem Maße bei Forschungsprojekten bereits existierende Datenbestände zu neuen Datenbeständen zusammengeführt (Bethlehem 2008; Schnell 2009). Dies gilt nicht nur für die hier führende Medizin, sondern auch für die Sozialwissenschaften. International üblich ist z. B. die gemeinsame Verwendung von Befragungsdaten mit Registerdaten, so bei Studien über Erkrankungshäufigkeiten, kriminologischen Fragestellungen oder Arbeitsmarktstudien. Die Zusammenführung von Datenbanken ist dann technisch unproblematisch, wenn in allen beteiligten Datenbanken eine gemeinsame Identifikationsnummer vorhanden ist, etwa eine Sozialversicherungsnummer oder – wie in den skandinavischen Ländern – eine über den Lebenslauf unveränderliche Personenkennziffer. Die meisten dieser Identifikationsnummern enthalten interne Prüfziffern, so dass Fehler in diesen Identifikatoren meist schnell entdeckt und korrigiert werden können. Probleme entstehen dann, wenn in den Datenbanken lediglich potenziell fehlerbehaftete alphanumerische Variablen wie Name, Vorname, Adresse und Geburtsort als primäre Identifikatoren vorhanden sind. Für die Zusammenführung solcher Datenbanken werden spezielle Verfahren benötigt, die zusammenfassend in der statistischen Literatur als „Record-Linkage"-Verfahren bezeichnet werden (Herzog/Scheuren/Winkler 2007). In der sozialwissenschaftlichen Forschung besteht ein zentrales Problem des Record-Linkage in der Lösung der Datenschutzprobleme: Wie können Datenbanken mithilfe von alphanumerischen Identifikatoren zusammengeführt werden, wenn die Identifikatoren Fehler aufweisen und die Anonymität der Personen in den Datenbanken vollständig gewahrt werden soll? Die Lösung dieses Problems ist Gegenstand unseres von der DFG geförderten Projekts „Spezifizierung und Implementierung eines datenschutzrechtlich unbedenklichen Verfahrens zur Verknüpfung sozialwissenschaftlicher Mikrodaten".

2 Bisherige Lösungsansätze

Um die Vertraulichkeit der Angaben in den Datenbanken zu erhalten, werden Datenbestände der beschriebenen Art meist mit Hilfe eines Datentreuhändermodells zusammengeführt. Bei einem einfachen Treuhändermodell (vgl. Abbildung 1) übermitteln die beiden Datengeber einem Datentreuhänder lediglich die zur Verknüpfung benötigten Merkmale, etwa Name, Geburtsdatum und Adresse, zusammen mit einer beliebigen, aber für den jeweiligen Datensatz ein-eindeutigen laufenden Nummer. Der Datentreuhänder verknüpft die Datenzeilen anhand der Merkmale

und erhält so Paare laufender Nummern von zusammengehörigen Records. Danach löscht der Treuhänder die Verknüpfungsmerkmale und übermittelt der Forschergruppe lediglich die Paare laufender Nummern. Die Forschungsgruppe kann so die von den Datengebern zuvor übermittelten Sachdaten zusammenführen, ohne dass die identifizierenden Merkmale bekannt sind.

Abbildung 1 Datentreuhändermodell



Quelle: Schnell, Hill & Esser (2008: 256).

In der Bundesrepublik existieren keine zentralen Datentreuhänderstellen, daher müssen die Treuhänderlösungen praktisch für jedes neue Projekt neu eingerichtet und mit Datenschützern verhandelt werden. Faktisch verhindert diese infrastrukturelle Hürde viele Projekte, da solche Genehmigungsprozesse Jahre in Anspruch nehmen können (Schnell/Bachteler 2006). Die in der Medizin in der BRD gebräuchlichen und mit beachtlichem finanziellen und personellen Aufwand betriebenen Pseudonymisierungsstrategien (Eichelberg/Aden/Thoben 2005) sind aufwendig und fehleranfällig. Aufgrund ähnlicher Probleme in anderen Ländern gibt es international eine Reihe von Forschergruppen, die Problemlösungen vorgeschlagen haben. Das bekannteste Verfahren in diesem Zusammenhang wurde von Churches und

Christen (2004) vorgestellt.[1] Aber auch dieses Verfahren besitzt zahlreiche Effizienzprobleme und Angriffsmöglichkeiten, so dass bislang kein praktisch verwendbares Verfahren existiert.[2] Es galt also, ein neues Verfahren zu entwickeln.

3 Entwicklungsprozess des neuen Verfahrens

Die Arbeitsgruppe der Autoren hat in einer Reihe von DFG-Projekten ein Programm zum Record-Linkage für die empirische Sozialforschung entwickelt: Die sogenannte „Merge Toolbox" (MTB) (Schnell/Bachteler/Bender 2004; Schnell/Bachteler/Reiher 2005). Mit MTB lassen sich Datenbanken mit probabilistischem Record-Linkage zusammenführen. MTB ist für akademische Zwecke frei verfügbar und wurde in zahlreichen empirischen Projekten erfolgreich eingesetzt. Zu diesen Anwendungen gehört der Einsatz in mehreren Krebsregistern sowie in zahlreichen Studien des Instituts für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit. Aus diesen Anwendungen wurde der Bedarf nach einem effizienten Verfahren für die Zusammenführung verschlüsselter Identifikatoren deutlich. Über zahlreiche Zwischenstufen (Schnell/Bachteler/Reiher 2007) wurde von den Autoren dann das neue Verfahren entwickelt, implementiert und getestet (Schnell/Bachteler/Reiher 2009).

4 Technische Details des SAFELINK-Verfahrens

Das Problem der Verknüpfung von Datenbanken mit fehlerbehafteten Identifikatoren wie z. B. Namen lässt sich auf das Problem der Berechnung der Ähnlichkeit dieser Identifikatoren reduzieren. Da aus Gründen des Datenschutzes auf keinen Fall Identifikatoren übermittelt werden sollen, die einen Rückschluss auf Personen zulassen, müssen die Identifikatoren so verschlüsselt werden, dass sie anschließend nicht mehr entschlüsselt werden können. Das Problem besteht also in der Berechnung

1 Für das Verständnis des Verfahrens von Churches & Christen sind einige technische Begriffe, wie z. B. N-Gramme, erforderlich, die wir erst später in diesem Aufsatz einführen. Die Kenntnis dieser Begriffe vorausgesetzt lässt sich das Verfahren folgendermaßen zusammenfassen: Beide Datengeber bilden für jeden Namen die Potenzmenge seiner N-Gramm Menge und verschlüsseln jede Teilmenge der Potenzmenge gesondert durch einen HMAC Algorithmus. Die resultierenden Hashwerte werden zusammen mit der Zahl der in der Teilmenge enthaltenen N-Gramme an eine Drittpartei übermittelt. Für jedes Paar an Namen kann die Drittpartei anhand der größten übereinstimmenden Teilmenge der beiden Namen die N-Gramm Ähnlichkeit bestimmen. Für Einzelheiten muss auf die Publikation von Churches & Christen verwiesen werden. Eine Kritik der Effizienz des Verfahrens findet sich bei (Verykios/Karakasidis/Mitrogiannis 2009).

2 Literaturübersichten finden sich bei Trepetin (2008) und Schnell, Bachteler & Reiher (2009).

von Ähnlichkeiten zwischen unentschlüsselbaren Identifikatoren. Exakt dies ist die Aufgabenstellung für das von uns neu entwickelte Verfahren „SAFELINK". Um das Verfahren zu erläutern, ist die Einführung einiger technischer Begriffe erforderlich.

Viele Algorithmen der Informatik für Zeichenfolgen basieren auf N-Grammen. Unter einem N-Gramm versteht man eine Folge aus N Zeichen; üblich in diesem Kontext sind Bigramme (N=2) und Trigramme (N=3). Häufig wird eine Zeichenkette (z. B. „Reisebus") vor der Zerlegung in N-Gramme an beiden Enden mit N-1 Leerzeichen ergänzt, so dass sich im Beispiel die Bigramm-Menge {_R; RE; EI; IS; SE; EB; BU; US; S_} ergibt. Die Ähnlichkeit von zwei N-Gramm-Mengen wird häufig mit dem Dice-Koeffizienten

$$\frac{2c}{a+b} \tag{1}$$

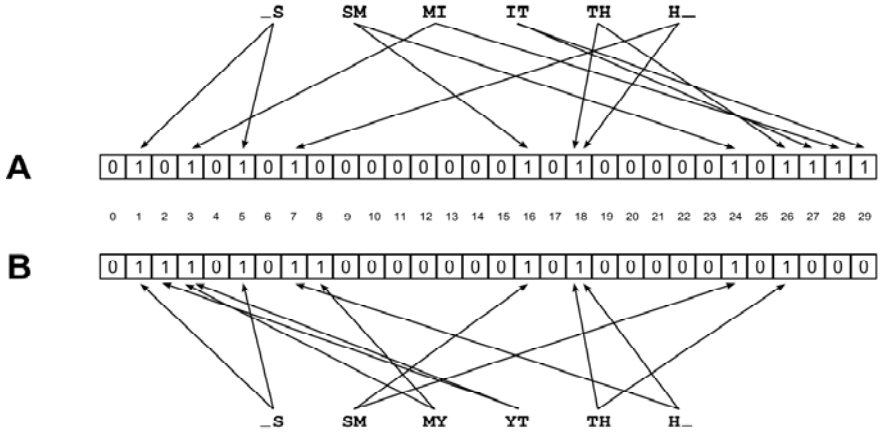
angegeben, wobei a die Zahl der N-Gramme in der Zeichenkette A, b die Zahl der N-Gramme in der Zeichenkette B und c die Zahl der N-Gramme ist, die in beiden Zeichenketten vorkommen. Viele Verfahren zur Zusammenführung von Datenbanken basieren darauf, dass diejenigen Records zusammengeführt werden, deren Identifikatoren sich am stärksten ähneln. Eine einfache Möglichkeit besteht in der Wahl derjenigen Paare, deren Dice-Koeffizienten ihrer N-Gramme am höchsten sind.

Ein Vektor der Länge l aus Nullen und Einsen wird in der Informatik als Bit-Vektor bezeichnet. Bildet man einen numerischen Wert n aus der Menge der natürlichen Zahlen mit einer Funktion auf das Intervall $0 \leq n \leq l$ ab, so wird die Funktion als „Hash-Funktion" bezeichnet. Die Abbildung mehrerer Hash-Funktionen auf einen Bit-Vektor ist ein sogenannter „Bloom-Filter" (Bloom 1970). Bloom-Filter werden meist zum Test der Mitgliedschaft eines Objekts in einer Menge verwendet (Brass 2008). Von besonderem Interesse ist die Verwendung von kryptografischen Hash-Funktionen. Hierbei handelt es sich um Einwegfunktionen, bei denen vom Ergebnis nicht mehr auf die Eingabewerte der Funktionen geschlossen werden kann. Das klassische Beispiel für eine solche Funktion ist der häufig für Prüfsummen eingesetzte Algorithmus MD-5, ein moderneres Beispiel SHA-1 (Swoboda/Spitz/Pramateftakis 2008).

Um die Ähnlichkeit zwischen Identifikatoren zu berechnen, ohne die Identifikatoren offenzulegen, speichern wir im Safelink-Verfahren die N-Gramme jedes Namens in einem eigenen Bloom-Filter. Für die Speicherung werden immer mehrere unabhängige Hash-Funktionen verwendet. Für die Zuordnung der Namen aus verschiedenen Datenbanken werden dann nicht mehr die Namen verglichen, sondern nur noch bitweise die Bloom-Filter.

Abbildung 2 illustriert das Verfahren für die beiden Namen SMITH und SMYTH mit Bigrammen, einem (unrealistisch kurzen) Bloom-Filter mit 30 Bits und zwei Hash-Funktionen. Die Namen werden in Bigramme zerlegt und jedes Bigramm der beiden Namen in den Bloom-Filtern A und B gespeichert. So erbringt z. B. das gemeinsame Bigramm „_S" den Hash-Funktionswert 1 für die erste Hash-Funktion und den Wert 5 für die zweite Hash-Funktion: Entsprechend werden die Bits an den Positionen 1 und 5 in beiden Bloom-Filtern auf den Wert 1 gesetzt. Im Gegensatz dazu sind die Bigramme „YT" (Hash-Werte 2 und 3) und „IT" (Hash-Werte 27 und 29) nur in einem Namen vorhanden, daher werden in den beiden Bloom-Filtern unterschiedliche Bits auf den Wert 1 gesetzt.

Abbildung 2 Beispiel für die Verwendung zweier Bloom-Filter für die Berechnung von Stringähnlichkeiten



Nach der Speicherung aller Bigramme in die Bloom-Filter sind 8 identische Bit-Positionen in beiden Bloom-Filtern auf 1 gesetzt. Insgesamt sind 11 Bits im Filter A und 10 Bits in Filter B gleich 1 gesetzt. Der Dice-Koeffizient ergibt sich als $\frac{2 \cdot 8}{(10+11)} \approx 0{,}762$. Die Ähnlichkeit der Namen kann also allein durch die Ähnlichkeit der Bloom-Filter approximiert werden. Da aber für die Speicherung kryptografische Einwegfunktionen genutzt wurden, können aus den Bloom-Filtern die Eingabenamen nicht mehr rekonstruiert werden. Dadurch wird eine fehlertolerante Verknüpfung zweier Datenbanken durch eine Forschungsgruppe oder eine Drittpartei bei vollständiger Anonymität der Identifikatoren möglich.

Für den Einsatz des Verfahrens sind eine Reihe von Entscheidungen notwendig. Weitgehend unkritisch ist die Wahl, ob Bi- oder Trigramme eingesetzt

werden: Unsere Simulationen zeigen kaum Unterschiede zwischen den Ergebnissen. Ebenso ist die Wahl der Länge der Bloom-Filter innerhalb eines Intervalls von 500-1.000 Bits eher unproblematisch. Derzeit neigen wir eher zu längeren Bloom-Filtern.[3] Die Wahl der eigentlichen Hash-Funktion erscheint uns auf der Grundlage der Arbeiten von Kirsch und Mitzenmacher (2006) ebenfalls weitgehend unproblematisch. Sie schlagen vor, mit zwei unabhängigen Hash-Funktionen k Hash-Funktionen zu realisieren. Die k Hash-Werte ergeben sich durch

$$g_i(x) = h_1(x) + ih(x) \bmod l \tag{2}$$

wobei i von 0 bis k-1 läuft und l die Länge des Bit-Arrays ist. Wir verwenden die erwähnten kryptografischen Funktionen SHA1 (h_1) und MD5 (h_2). Die Wahl der Zahl k der Hash-Funktionen beeinflusst das Verhalten des Verfahrens deutlich. Daher stellen wir dies weiter unten ausführlich dar.

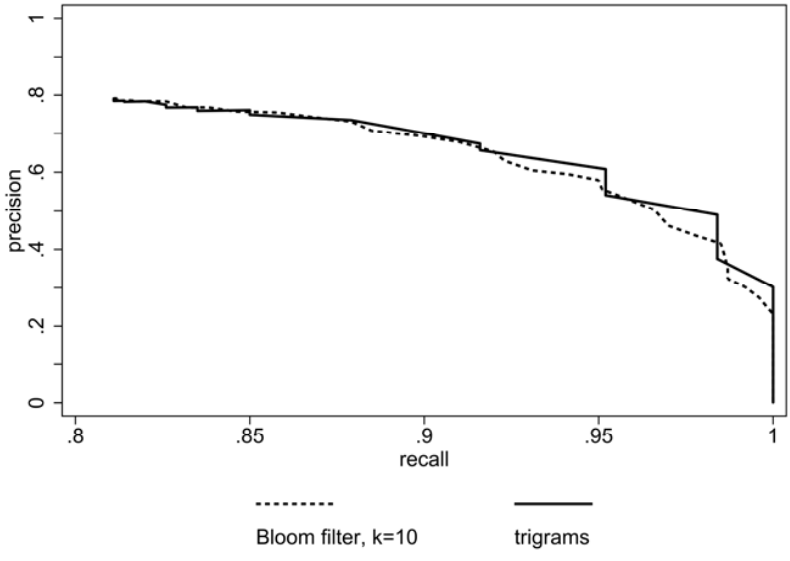
5 Empirische Tests des Verfahrens

Wir haben das Verfahren sowohl mit simulierten als auch mit echten Daten getestet. Um den Effekt der Anzahl der Hash-Funktionen auf die Leistung des Verfahrens zu testen, haben wir in einem Experiment mit simulierten Daten die Anzahl der Hash-Funktionen variiert (k = 1, 5, 10, 25, 50, 100). Die Länge der Filter war dabei konstant 1.000 Bits. Verglichen wurde mit den Ergebnissen unverschlüsselter Trigramme. Die Ausgangsdaten waren dabei 1.000 verschiedene aus einer Telefon-CD zufällig ausgewählte Nachnamen. Aus dem Datenbestand wurden alle nicht-alphabetischen Zeichen entfernt, die Umlaute umgewandelt und häufige Namensbestandteile wie akademische Titel und ehemalige Adelsprädikate gelöscht. Diese bereinigte Liste wurde kopiert und in der Kopie in jedem Namen mit der Wahrscheinlichkeit 0,25 an einer jeweils neu ermittelten zufälligen Stelle exakt ein Buchstabe durch einen anderen Buchstaben ersetzt. Auf diese Weise erhielten wir zwei Listen von jeweils 1.000 Namen, bei denen 250 Namen exakt eine Differenz aufwiesen.

3 Die Länge der Bloom-Filter beeinflusst auf recht komplizierte Weise die Sicherheit des Verfahrens: Längere Schlüssel sind für eine bestimmte Form des Angriffs (einen Wörterbuchangriff) weniger sicher als kurze. Der geringe Verlust an Sicherheit bei einem langen Schlüssel wird aber durch den Gewinn an Präzision deutlich aufgewogen. Eine genaue Analyse dieses Problems ist der Gegenstand einer laufenden Studie der Arbeitsgruppe.

Ein Beispiel für die Ergebnisse (hier mit 10 Hash-Funktionen) zeigt Abbildung 3 mit einem sogenannten „precision versus recall plot". In einem PR-Plot wird die Genauigkeit eines Abrufs aus einer Datenbank („precision") gegen die Empfindlichkeit des Abrufs („recall") geplottet. PR-Plots sind die in der Informatik übliche Variante der in der medizinischen Literatur gebräuchlichen Receiver-Operating-Characteristic-Kurven, bei denen Sensitivität eines Verfahrens gegen die Spezifität des Verfahrens abgebildet wird (Davis/Goadrich 2006).

Abbildung 3 Precision-Recall-Kurven von unverschlüsselten Trigrammen und Bloom-Filtern der Länge 1.000 mit 10 Hash-Funktionen



Exakter: Für ein gegebenes Maß an Ähnlichkeit zweier Schlüssel wird ein Record-Paar als „match" bezeichnet, wenn die Records tatsächlich zusammengehören. Alle anderen Paare sind daher „non matches" (Winkler 1995). Entsprechend ergibt sich dann die übliche Klassifikation in korrekt übereinstimmende Paare („true positive", TP), falsch positive Paare (FP), falsch negative Paare (FN) und korrekt nicht übereinstimmende Paare („true negative", TN). Die Vergleichskriterien ergeben sich dann als

$$\text{recall} = \frac{\sum TP}{\sum TP + \sum FN} \tag{3}$$

$$\text{precision} = \frac{\sum TP}{\sum TP + \sum FP} \tag{4}$$

In Abbildung 3 ist die Precision-Recall-Kurve für die unverschlüsselten Trigramme der Precision-Recall-Kurve für die Bloom-Filter sehr ähnlich. Wenn man nur wenige Hash-Funktionen in einem Bloom-Filter verwendet, dann ist die Wahrscheinlichkeit, dass verschiedene Trigramme auf identische Bit-Positionen abgebildet werden, sehr klein. Je mehr Hash-Funktionen verwendet werden, desto eher werden verschiedene Trigramme auf die gleiche Position abgebildet. Dies ist für die Leistung von SAFELINK von zentraler Bedeutung: Mit steigender Zahl der Hash-Funktionen k wird ein Wörterbuch-Angriff ohne Kenntnis der Parameter der Verschlüsselung immer schwieriger.[4] Andererseits wird durch ein Ansteigen der Zahl der Hash-Funktionen die Wahrscheinlichkeit einer falsch-positiven Identifikation eines Record-Paares immer höher. Es handelt sich also um eine Abwägung zwischen erhöhter Sicherheit für einen sehr unwahrscheinlichen (und natürlich illegalen) Angriff und Verringerung der Präzision der Verknüpfung. Die Zahl der Hash-Funktionen sollte daher nicht zu hoch gewählt werden.

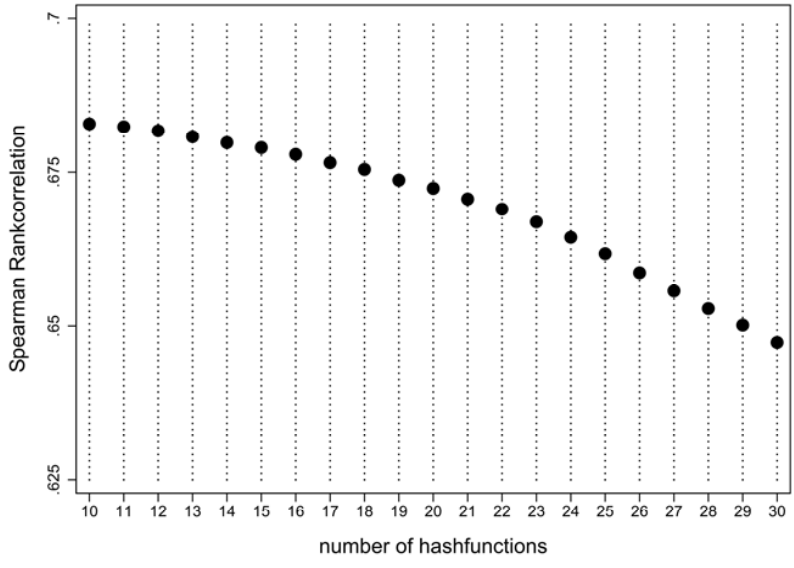
Nach unseren ersten Versuchen schien ein vernünftiges Intervall für die Zahl der Hash-Funktionen zwischen 10 und 30 zu liegen. Für dieses Intervall haben wir das Experiment für jede Zahl k zwischen 10 und 30 wiederholt. Für jedes Record in den beiden Datenbanken wurde die Ähnlichkeit anhand der unverschlüsselten Trigramme und der Bloom-Filter berechnet. Anschließend wurde die Rang-Korrelation zwischen diesen beiden Ähnlichkeiten berechnet. Das Ergebnis zeigt Abbildung 4.

Die Rangkorrelation zwischen den Ähnlichkeiten auf der Basis der unverschlüsselten Trigramme und der Bloom-Filter sinkt monoton mit steigender Zahl k der Hash-Funktionen. Bei 30 Hash-Funktionen liegt der Rangkorrelationskoeffizient nur noch bei 0,647. Bis weitere Ergebnisse auf der Basis realer Datensätze vorliegen, halten wir daher 15 Hash-Funktionen für einen akzeptablen Kompromiss.

In einem zweiten Test wurde SAFELINK mit den Ergebnissen unverschlüsselter Bigramme und der sogenannten „Kölner Phonetik" (Postel 1969) verglichen. Phonetiken sind Sammlungen von Regeln, die eine Zeichenkette in einen phonetischen Code übersetzen. Falls die phonetischen Codes übereinstimmen, wird dem Paar als Ähnlichkeitswert die Zahl 1 zugewiesen, sonst die Zahl 0.

4 In der Kryptografie bezeichnet man den Vergleich der verschlüsselten Werte mit den Verschlüsselungen bekannter Eingabewerte in den Verschlüsselungsalgorithmus als Wörterbuchangriff.

Abbildung 4 Rangkorrelationen zwischen den Dice-Koeffizienten von unverschlüsselten Trigrammen und Bloom-Filtern der Länge 1.000 nach Zahl der Hash-Funktionen



Die Kölner Phonetik wurde speziell für den deutschen Sprachraum entwickelt und wird häufig verwendet, so z. B. von den deutschen Krebsregistern (Eichelberg/Aden/Thoben 2005). Für diesen zweiten Test des Verfahrens wurden ausschließlich die Namen (ohne jede inhaltliche Information) aus zwei Verwaltungsdatenbanken mit jeweils ca. 15.000 Einträgen verwendet. Die Aufgabe bestand darin, zusammengehörende Namen in beiden Datenbanken zu finden. Bei diesem Test haben wir Bigramme, 15 Hash-Funktionen und Bloom-Filter mit 500 Bits verwendet. Die Leistung der unverschlüsselten Bigramme, der Kölner Phonetik und SAFELINK wurde dadurch verglichen, dass für jedes der drei Verfahren ein unabhängiges Record-Linkage mit exakt denselben Parametern durchgeführt wurde. Für das Record-Linkage wurde unser Programm „MTB" (Schnell/Bachteler/Reiher 2005) verwendet. Abbildung 5 zeigt den PR-Plot von SAFELINK im Vergleich zu den unverschlüsselten Bigrammen, Abbildung 6 im Vergleich zur Kölner Phonetik. SAFELINK erzielt nahezu die gleiche Leistung wie die unverschlüsselten Bigramme und bessere Ergebnisse als die Kölner Phonetik. Dies gilt vor allem für Recall-Level über 0,75, da die Phonetik dann besonders viele falsch-positive Ergebnisse erbringt.

Abbildung 5 Precision-Recall-Kurven von unverschlüsselten Bigrammen und Bloom-Filtern der Länge 500 mit 15 Hash-Funktionen

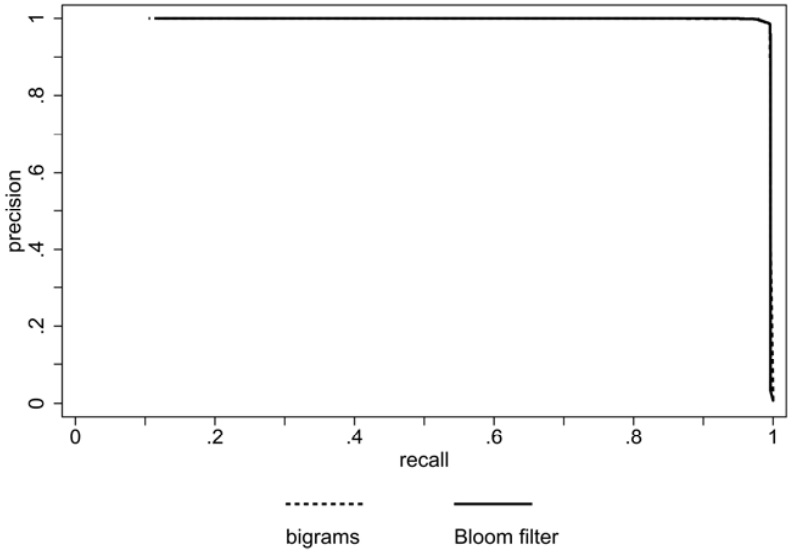
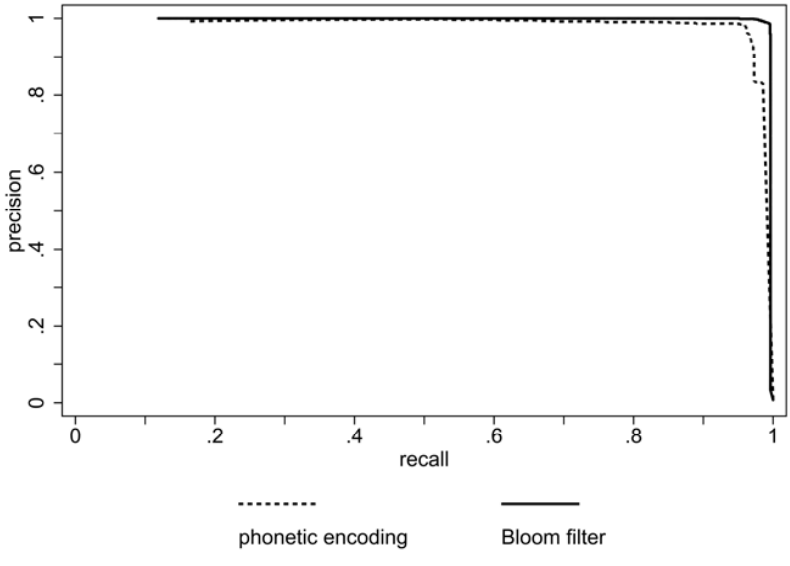


Abbildung 6 Precision-Recall-Kurven der Kölner Phonetik und Bloom-Filtern der Länge 500 mit 15 Hash-Funktionen



6 Ein Protokoll zur Datenverknüpfung mit SAFELINK

Die zuvor beschriebenen erfolgreichen Tests legen es nahe, die vorgeschlagene Methode innerhalb eines Protokolls für das datenschutzrechtlich unbedenkliche Record-Linkage zu verwenden. Um dem Protokoll noch eine weitere Sicherheitsschicht hinzuzufügen, schlagen wir vor, die Hash-Funktionen SHA1 und MD5 durch deren kryptografische Varianten mit einem Schlüssel K zu ersetzen.[5] Bei diesen sogenannten „keyed hash message authentication codes" (HMAC) wird ein zusätzlicher Schlüssel K verwendet. Bei einem HMAC ist es auch bei Kenntnis des Schlüssels K nicht möglich, aus einem gegebenen HMAC-Wert eine dazu passende Nachricht zu rekonstruieren (Swoboda/Spitz/Pramateftakis 2008: 108).

Basierend auf der so erweiterten Bloom-Filter-Methode ist die Implementation eines Record-Linkage-Protokolls recht einfach. Unser Protokoll verwendet eine halb vertrauenswürdige Drittpartei[6], da ansonsten einer der beider Datengeber A oder B versuchen könnte, mit Hilfe einer Liste möglicher Einträge die Bloom-Filter des anderen Datengebers zu identifizieren.[7] Neben den Datengebern A und B mit den Datenbeständen S_a und S_b werden die Drittpartei C und der Empfänger D der zusammengeführten Datensätze im Protokoll benötigt. Das Protokoll läuft folgendermaßen ab:

1. Die Datengeber A und B einigen sich über die Länge l der Bloom-Filter, die Anzahl k der Hash-Funktionen sowie über den vertraulichen Schlüssel K.
2. Für jeden Namen i in seinem Datenbestand führt jeder der Datengeber die folgenden Schritte durch:
   a) Umwandlung des Namens in seine N-Gramme.
   b) Speichern der N-Grammmenge mit k Funktionen und dem Schlüssel K in einem Bloom-Filter der Länge l.
   c) Speichern des Bloom-Filters und einer systemfreien, eindeutigen und zufällig generierten Identifikationsnummer id in eine Liste BF. Hinzufügen der Identifikationsnummer zum Datenbestand.

5 Diese Varianten werden als HMAC-MD5 und HMAC-SHA1 bezeichnet und wurden von Krawczyk, Bellare & Canetti (1997) vorgeschlagen.

6 In der englischsprachigen Literatur wird zwischen „trusted third parties" (TTPs) und „semi-trusted third parties" (STTPs) unterschieden. Für eine STTP gelten weniger anspruchsvolle Annahmen bezüglich ihrer Vertrauenswürdikeit als für eine TTP. Gefordert wird, dass eine STTP ein vereinbartes Protokoll einhält und nicht böswillig mit einer anderen Partei kooperiert. Im Gegensatz zu einer TTP wird aber nicht angenommen, dass eine STTP nicht versucht, einen kryptanalytischen Angriff auf verschlüsselte Informationen zu unternehmen. Für das Safelink-Verfahren ist die Annahme einer STTP ausreichend.

7 Ein solcher Wörterbuchangriff wäre möglich, falls die Datengeber die Zahl der Hash-Funktionen k, den Schlüssel K und die Länge des Bloom-Filters l kennen. Weiterhin wüssten die Datengeber welche Einträge in beiden Datenbanken vorhanden wären. Beide Probleme werden durch den Einsatz einer Drittpartei vermieden.

3. Beide Datengeber entfernen aus ihren Datenbeständen alle Namen und sonstige Identifikatoren bis auf id.
4. Beide Datengeber übersenden ihre Datenbestände (ohne Identifikatoren bis auf id) an D.
5. Beide Datengeber übersenden ihre Bloom-Filter-Listen samt id an die Drittpartei C.
6. C vergleicht alle Paare von Bloom-Filtern und berechnet die Ähnlichkeit der Filter mit dem Dice-Koeffizienten.
7. Paare mit den höchsten Dice-Koeffizienten werden zu Tupeln (id in BF_a, id in BF_b, Dice-Koeffizient) zusammengefasst.
8. C übersendet die Liste der besten Paare mit ihren Tupeln an D.
9. D führt die Datenbestände unter Verwendung der Tupel zusammen.

Das Protokoll ist in Hinsicht auf die Datengeber A und B sicher, da keiner der beiden Zugang zu den Bloom-Filtern des anderen hat. Die Drittpartei sieht nur die Bloom-Filter, kennt aber weder k, noch K, noch l. Selbst wenn er alle diese Informationen hätte, müsste die Drittpartei zusätzlich eine Liste potenzieller Namen abfragen. So lange die Drittpartei also nicht mit einem Datengeber konspiriert, ist das Protokoll sicher.[8]

7 Implementierung des Protokolls

Einer der Autoren hat das Verfahren zunächst innerhalb unseres Record-Linkage-Programms MTB in Java implementiert. Um potenzielle Fehler auszuschalten, hat ein anderer der Autoren vollkommen unabhängig davon das Protokoll erneut mit Python implementiert.[9] Beide Implementierungen umfassen ca. 100 Zeilen Programm-Code, wobei aber zahlreiche (jeweils andere) Standardbibliotheken eingebunden werden. Die Ergebnisse beider Programme sind identisch. MTB führt die Berechnungen dabei deutlich schneller aus, ca. 1.000 Records können dabei in 5 Minuten vollständig paarweise verglichen werden. Dies entspricht bei Anwendungen in der Praxis der maximalen Größe einer Teilmenge einer Datenbank, die tatsächlich vollständig

8 Sollte die Drittpartei mit einem Datengeber konspirieren, ist der direkte Austausch der Datenbestände ohnehin einfacher.

9 Da beide Sprachen plattformunabhängig sind, sollten die Programme unter jedem modernen Betriebssystem (Linux, Mac-OS, Windows) funktionieren.

verglichen wird.[10] Der Geschwindigkeitsverlust gegenüber der Verwendung unverschlüsselter Identifikatoren liegt bei ca. 20 %.

8 Zukünftige Weiterentwicklungen und Anwendungen

Das Kernproblem der fehlertoleranten Verknüpfung von stark kryptografierten Namen wurde von uns mit SAFELINK gelöst. Trotz der Einsatzbereitschaft der Programme und des Protokolls bleiben einige Probleme offen. Bei der Arbeit an diesem Protokoll wurde deutlich, dass die Art der Aufbereitung der Namen vor der Verschlüsselung noch erheblichen Forschungsbedarf besitzt. Weiterhin ist in der Literatur das Problem der kryptografischen Verschlüsselung numerischer Daten mit kardinalen Eigenschaften derart, dass Abstandsberechnungen zwischen den verschlüsselten Daten möglich bleiben, bislang nicht befriedigend gelöst. Dies ist z. B. für die Berücksichtigung des Geburtsdatums beim Record-Linkage von zentraler Bedeutung. Zusammen mit SAFELINK würden dadurch Personenidentifikatoren möglich, die dezentral immer neu generiert werden können, ohne dass sie zentral gespeichert werden müssen. Damit wären Verknüpfungen von personenbezogenen Daten im Längsschnitt möglich, die bisher aus rechtlichen Gründen kaum realisierbar waren. Neben den offensichtlichen Möglichkeiten von medizinischen Längsschnittstudien ergeben sich so z. B. technische Realisierungsmöglichkeiten für kriminologische Panels oder ein Bildungspanel mit Individualdaten im Längsschnitt. Die Lösung dieser Probleme ist Gegenstand unserer laufenden Bemühungen.

Literatur

Bethlehem, J., 2008: Surveys without questions. S. 500-511 in: E. D. de Leeuw, J. D. Hox und D. A. Dillman (Hg.): International handbook of survey methodology, New York: Erlbaum.

Bloom, B. H., 1970: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM 13: 422-426.

Brass, P., 2008: Advanced data structures. Cambridge: Cambridge University Press.

Churches, T. und P. Christen, 2004: Some methods for blindfolded record linkage. BMC Medical Informatics and Decision Making 4: 1-17.

Davis, J. und M. Goadrich, 2006: The relationship between precision-recall and ROC curves. S. 233-240 in: Proceedings of the 23rd International Conference on Machine Learning, New York.

10 Beim Record-Linkage vermeidet man aus Rechenzeitgründen den vollständigen Paarvergleich zweier Datenbanken. Bei großen Datenbeständen wird immer nur innerhalb von Teilmengen („blocks“) verglichen. Die Bildung dieser Blöcke ist ein eigenes Forschungsgebiet innerhalb des Record-Linkage.

Eichelberg, M., T. Aden und W. Thoben, 2005: A distributed patient identification protocol based on control numbers with semantic annotation. International Journal on Semantic Web and Information Systems 1: 24-43.

Herzog, T. N., F. J. Scheuren und W. E. Winkler, 2007: Data quality and record linkage techniques. New York/Berlin: Springer.

Kirsch, A. und M. Mitzenmacher, 2006: Less hashing, same performance: building a better Bloom filter. S. 456-467 in: Y. Azar und T. Erlebach (Hg.): Algorithms-ESA 2006. Proceedings of the 14th Annual European Symposium, 11-13 September 2006, Zürich. Berlin: Springer.

Krawczyk, H., M. Bellare und R. Canetti, 1997: HMAC: Keyed-hashing for message authentication. Internet RFC 2104. http://tools.ietf.org/html/rfc2104 (9.9.2009).

Postel, H. J., 1969: Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. IBM-Nachrichten 19: 925-931.

Schnell, R. und T. Bachteler, 2006: Der Bedarf nach einer Treuhänderlösung für die Verknüpfung von Mikrodaten in der Bundesrepublik. Diskussionspapier, Zentrum für Quantitative Methoden und Surveyforschung, Universität Konstanz. http://www.uni-due.de/methods/documents/SchnellDatenTreuhandRatWSD.pdf (9.9.2009).

Schnell, R., T. Bachteler und J. Reiher, 2009: Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics and Decision Making 9: 41.

Schnell, R., T. Bachteler und J. Reiher, 2007: Die sichere Berechnung von Stringähnlichkeiten mit Bloom-Filtern, Diskussionspapier, Universität Konstanz, September 2007.

Schnell, R., T. Bachteler und S. Bender, 2004: A toolbox for record linkage. Austrian Journal of Statistics 33: 125-133.

Schnell, R., T. Bachteler und J. Reiher, 2005: MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. Zentralarchiv-Informationen 56: 93-103. (http://www.za.uni-koeln.de/publications/pdf/za_info/ZA-Info-56.pdf (9.9.2009).

Schnell, R., P. Hill und E. Esser, 2008: Methoden der empirischen Sozialforschung. 8. Auflage, München: Oldenbourg.

Schnell, R., 2009: Record linkage from a technical point of view, Expertise für den Rat für Wirtschafts- und Sozialdaten, Projekt: Developing the Research Infrastructure for the Social and Behavioral Sciences in Germany and Beyond: Progress since 2001, Current Situation and Future Demands, Februar 2009.

Swoboda, J., S. Spitz und M. Pramateftakis, 2008: Kryptographie und IT-Sicherheit: Grundlagen und Anwendungen. Wiesbaden: Vieweg+Teubner.

Trepetin, S., 2008: Privacy-preserving string comparisons in record linkage systems: a review. Information Security Journal: A Global Perspective 17: 253-266.

Verykios, V. S, A. Karakasidis und V. K. Mitrogiannis, 2009: Privacy preserving record linkage approaches. International Journal of Data Mining, Modelling and Management 1: 206-221.

Winkler, W. E., 1995: Matching and record linkage. S. 355-384 in: B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge und P. S. Kott (Hg.): Business survey methods. New York: Wiley.

Anschrift der Autoren

Prof. Dr. Rainer Schnell
Tobias Bachteler
Jörg Reiher
Universität Duisburg-Essen
Lotharstraße 65
47057 Duisburg
rainer.schnell@uni-due.de
www.methodenzentrum.de

Data Collection Quality Assurance in Cross-National Surveys: The Example of the ESS

Qualitätssicherung bei der Datenerhebung von international vergleichenden Umfragen am Beispiel des ESS

Achim Koch, Annelies G. Blom, Ineke Stoop and Joost Kappelhof

Abstract

The significance of cross-national surveys for the social sciences has increased over the past decades and with it the number of cross-national datasets that researchers have access to. Cross-national surveys are typically large enterprises that demand dedicated efforts to coordinate the process of data collection in the participating countries. While cross-national surveys have addressed many important methodological problems, such as translation and the cultural applicability of concepts, the management of the data collection process has yet had little place in cross-national survey methodology. This paper describes the quality standards for data collection and their monitoring in the European Social Survey (ESS). In the ESS data are collected via face-to-face interviewing. In each country a different survey organisation carries out the data collection. Assuring the quality across the large number of survey organisations is a complex but indispensable task to achieve valid and comparable data.

Zusammenfassung

International vergleichende Umfragen haben in den vergangenen Jahrzehnten zunehmende Bedeutung in den Sozialwissenschaften erlangt. Diese Umfragen sind für gewöhnlich große Unterfangen, die gezielte Anstrengungen zur Koordinierung der Datenerhebung in den teilnehmenden Ländern erfordern. Probleme des Managements der Datenerhebung bei international vergleichenden Umfragen haben bislang jedoch nur wenig Aufmerksamkeit gefunden, im Unterschied etwa zu anderen methodischen Herausforderungen wie Fragen der Übersetzung oder der interkulturellen Übertragbarkeit von theoretischen Konzepten. Der vorliegende Beitrag beschreibt die Qualitätsstandards für die Datenerhebung und deren Überwachung im European Social Survey (ESS). Im ESS werden Daten in persönlich-mündlichen Interviews erhoben; in jedem Teilnehmerland ist ein anderes Umfrageinstitut mit der Feldarbeit betraut. Um valide und vergleichbare Daten zu erzielen, sind Maßnahmen zur Sicherung der Qualität der Datenerhebung über die große Zahl von Umfrageinstituten hinweg unverzichtbar.

1 Introduction

With growing globalisation the importance of cross-national data has increased and with it also the number of cross-national surveys (Kish 1994; Jowell 1998; Heath et al. 2005; Lynn et al. 2006). Cross-national surveys are large enterprises that demand considerable financial, human and infrastructural resources, at the country-level and the cross-national level. To assure reliable and valid measurement within each country as well as cross-country data comparability survey standards are specified and their implementation is monitored centrally. The participating countries are to adhere to the survey standards and provide proof of correct implementation. Cross-national surveys differ in the level of standardisation that they pursue. Whereas some surveys only specify a very limited set of survey standards (such as question wording and a minimum sample size), other surveys cover all aspects of the survey life-cycle. The present paper describes the quality standards for data collection and their monitoring in the European Social Survey (ESS).

It is uncontested that social measurements like quantitative surveys need some kind of standardisation of methods and processes to provide reliable and valid data (Jowell 1998). This holds both for national and for cross-national surveys except for one important difference. National surveys usually have a single design (Lynn et al. 2006). This means that there is one sample design and one questionnaire is administered in a standard way by interviewers who have received the same training and instructions. In cross-national surveys, designs differ across countries due to differences in financial resources, legislation regarding the survey business, available sampling frames, the geographical dispersion of the population, languages, the experience and capability of survey organisations and survey practices (like the typical methods and content of interviewer training or the prevailing mode of interviewing) (Park/Jowell 1997; Smith 2007). Consequently, even in highly standardised cross-national surveys some aspects of the survey design will be implemented differently across countries.

When differences in methods affect survey outcomes, comparisons across countries can be jeopardised, because observed cross-country differences may be mere methodological artefacts. If standardisation in methods leads to equivalent outcomes, cross-national surveys should therefore strive for perfect standardisation. However, for reasons mentioned above perfect standardisation is impossible. Moreover, occasionally the *effect* of methods can differ across countries. Skjåk and Harkness (2003) for instance argue that optimal modes of interview administration (face-to-face, telephone, self-completion, etc.) in one country may be quite

problematic in others. Therefore, sometimes comparability of results may best be achieved by a deliberate variation in design. Such considerations seem quite plausible also with regard to response and nonresponse. For instance, in order to achieve similar response rates between countries it can be prudent to allow for the use of different types of respondent incentives across countries. Or, given differences between countries in at-home-patterns of their population, it may be advisable to accept different call schedules to achieve similarly low noncontact rates in all countries.

> '[T]he challenge is to identify which aspects of design need to be identical, which should be allowed (encouraged) to vary – and within what parameters – and which may be less important, in the sense that relevant characteristics of the survey data may be insensitive to variations in design.' (Lynn et al. 2006: 14f.)

With regards to equivalence in probability sampling Lynn et al. (2006) argue that different sampling strategies may be the best way to achieve equivalent samples (see also Kish 1994; Häder/Lynn 2007); while equivalence of measurement may be best achieved by standardising question wording and mode of interview. For other aspects of survey design, such as data collection practices, the authors note that little is known yet about the effects of different design options.

We look at quality assurance for data collection in cross-national surveys using the example of the ESS. The ESS is a biennial cross-national survey of social and political attitudes in Europe. Data are collected via face-to-face interviewing.[1] In the ESS standards for data collection are set by a Central Coordinating Team (CCT), which also produces guidelines, assists countries in preparing fieldwork, monitors the progress of fieldwork in all countries and evaluates the implementation processes. In each country a different survey organisation carries out the data collection.[2] Assuring the quality across such a large number of survey organisations is a complex but indispensable task to achieve valid and comparable data. We describe how the CCT of the ESS coordinates data collection in the more than 30 participating countries and how it tries to find a viable balance between standardisation and national adaptation.

1 Cross-national surveys often rely on face-to-face interviewing. Apart from the ESS, for instance also the Adult Literacy and Life Skills Survey, the Eurobarometer, the European and World Values Surveys, the Survey on Health, Ageing and Retirement in Europe and the World Mental Health Survey are conducted face-to-face.

2 A few ESS countries appoint local branches of globally acting groups like Ipsos, TNS or Gallup to carry out the ESS. In such a case, the national branches of global survey organisations usually act quite independently from each other. Some cross-national surveys (e. g. the Eurobarometer) subcontract the entire cross-national data collection to one global survey organisation. Here the cross-national coordination of the data collection is the task of the central office of the global survey organisation.

We first provide basic background information on the ESS and describe how standards for data collection are set and monitored. Subsequently, outcomes of this approach for key data collection features in the first three rounds of the ESS are presented. We describe to what extent countries adhered to data collection standards and discuss reasons for deviations from these standards. The conclusion provides some final considerations.

2 The ESS: Basic Features, Aims and Organisation

The ESS is an academically-driven social survey designed to chart and explain the attitudes, beliefs and behaviour patterns of its diverse populations (Jowell et al. 2007). In addition to monitoring and interpreting social change, the ESS also seeks to consolidate and improve cross-national quantitative measurements within Europe and beyond (O'Shea et al. 2003). Since 2002 the survey has been fielded every two years and now, in its fourth round, it covers more than 30 countries. Each of the participating countries conducts approximately 2000 face-to-face interviews in each round, either as paper-and-pencil interviews (PAPI) or as computer assisted personal interviews (CAPI). The ESS questionnaire includes two main sections: a 'core' module which remains relatively constant from round to round plus two or more 'rotating' modules repeated at intervals. The core module monitors change and continuity in a wide range of social variables. The rotating modules focus on particular academic or policy concerns, like 'immigration' or 'family, work and well-being'. The average interview length is about 70 minutes.

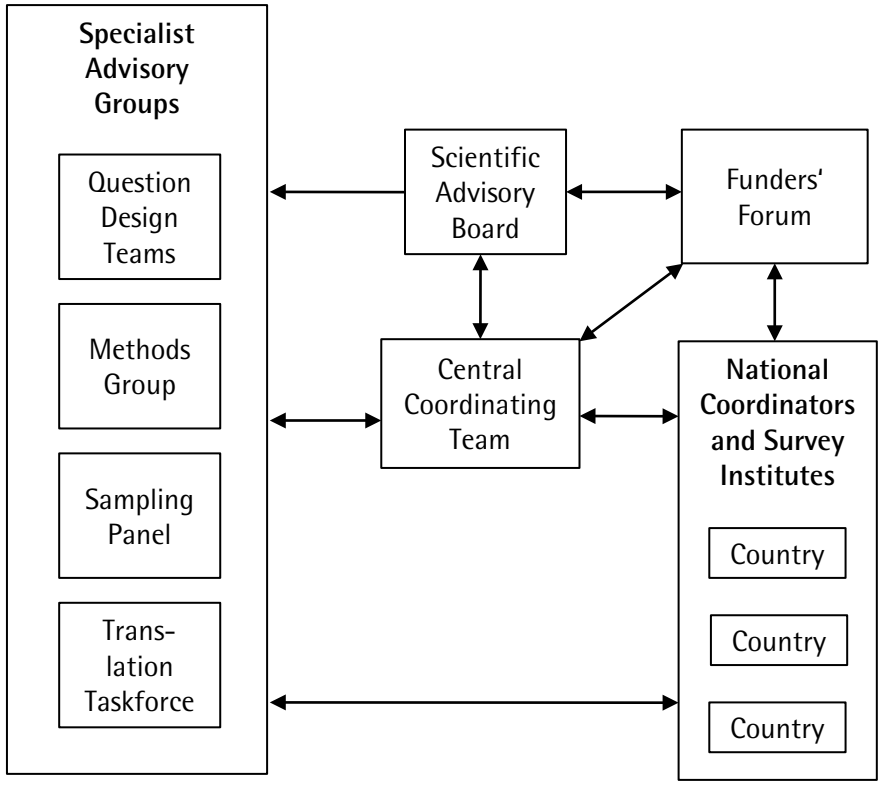
The ESS project is directed by the CCT, led by the Centre for Comparative Social Surveys at City University London (see Figure 1). The CCT is responsible for the design and coordination of the project. Its work is primarily funded by the European Commission. A multi-disciplinary team of researchers from seven European research institutes cooperates in the CCT.[3] Each partner institute has pre-specified and self-contained responsibilities, some of which continue throughout the project's life, others for shorter periods. The work comprises more than ten workpackages, including the general coordination and implementation of the project, sampling design, translation, fieldwork commissioning, piloting, archiving

3 Apart from City University, these institutes include NSD in Norway, GESIS in Germany, SCP in the Netherlands, the Universitat Pompeu Fabra in Barcelona, Spain, the University of Leuven in Belgium, and the University of Ljubljana in Slovenia.

and dissemination. The seven institutes are also jointly responsible for overall quality control and quality assessment.

Data collection and other national costs in each country are borne by national funding bodies. In each participating country the national funding agency appoints a National Coordinator (NC) and a survey organisation to implement the survey according to common ESS specifications. The NC and the survey organisation are responsible for the national implementation, including the sampling, translation, data collection, data editing and survey documentation.

Figure 1 Organisational Structure of the ESS



The two key actors – the CCT and NCs – are supplemented by a network of overseeing and supporting groups: the Scientific Advisory Board, the Question Design Teams, the Sampling Panel, the Translation Taskforce, a methodological advisory board (the Methods Group) and a group representing the national funding bodies and the European Commission (the Funders' Forum).

3 Survey Standards in the ESS

3.1 Specification of Standards for Data Collection in the ESS

Cross-national surveys vary in the balance of responsibilities at the cross-national and national levels. Given the large number of participating countries and aspired methodological rigour, the ESS needs a strong cross-national organ (the CCT) that stipulates the survey design and monitors quality. A standard specification designed by the CCT establishes the methods and procedures to be followed in all participating countries. Regarding the data collection process these *'Specifications for Participating Countries'* (European Social Survey 2001; 2003; 2005) cover three core areas: (1) the selection of a survey organisation, (2) data collection outcomes and (3) data collection procedures. In the following we summarise the ESS standards and describe the rationale behind them. Tables A1 and A2 in the appendix list the ESS data collection specifications distinguishing between required and recommended procedures. The tables also indicate the leeway for national adaptation for both groups of procedures.

Selection of a Survey Organisation

The ESS urges participating countries to contract the best European fieldwork organisations to ensure that its regular rounds of data collection are carried out to the same exacting standards (O'Shea et al. 2003). Survey organisations to be appointed for the ESS must be capable of conducting national probability-based face-to-face surveys to the most rigorous standards. Furthermore, the specifications stipulate that, if necessary, the survey organisations should be willing to change their routine procedures and methods to ensure cross-national comparability. Accordingly, the ESS requires some flexibility on the part of survey organisations intending to field the ESS. The advantage of the country-wise selection of survey organisations is that NCs are best aware of the quality that organisations in their country can produce. However, it can be a challenge to get survey organisations to replace their traditional approach with ESS standards. This is an important task for the NCs. Section 3.3 demonstrates how this aim was pursued.

Adherence to the ESS specifications is a prerequisite for each participating country and for each survey organisation selected to field the ESS. At times this may require a higher budget than is necessary for fielding a survey according to the usual standards in a country. Examples of fundamental changes to typical fieldwork practices come from France and Switzerland. In France the major challenge was to replace the traditional quota sampling by probability sampling, and in Switzerland the prevailing telephone mode had to be substituted by face-to-face interviewing.

Data Collection Outcomes

The ESS standards for data collection outcomes concern the sample size and the response and noncontact rate. The ESS specifications require a minimum *effective* sample size of 1500 interviews for each participating country based on a probability sample (Häder/Lynn 2007). Countries may use different sampling designs which may have a different effect on standard errors (independent from the size of the sample). To standardise the level of precision of results across countries the ESS prescribes an effective sample size, which takes account of the design effects associated with a country's sample design. The concept of an effective sample size operated in the ESS requires countries with geographically clustered samples to provide a higher number of completed interviews than countries using a simple random sample.

Nonresponse is a major threat to sample surveys, since it decreases the net sample size and can lead to biased survey results (Groves/Couper 1998; Groves et al. 2002). In most Western countries response rates have been declining during the past decades (de Leeuw/de Heer 2002). The ESS specifies a minimum target response rate of 70 percent. When setting this target the CCT was aware that some countries would reach the target, while others would struggle. The CCT felt that specifying a target outcome rate to competing survey organisations would make the target a contractual obligation that the selected survey organisation must strive and budget for (Jowell et al. 2007).[4] The rationale was to maximise response rates in each country and to reduce variation in response rates across countries in order to optimise comparability. In addition to setting a target response rate the ESS limits the noncontact rate to three percent of the eligible sample. The reason for specifying a maximum noncontact rate was that this source of nonresponse can be easier controlled by insisting on certain design features (especially the number and timing of contact attempts) than the other major source of nonresponse, i. e. refusals (Groves/Couper 1998).

Obviously, as regards nonresponse the ultimate goal should be to minimise nonresponse bias. However, minimising bias is even more difficult than enhancing

4 Of course, a certain response rate cannot be enforced. Individual target persons always have the right to refuse, may not be at home for a prolonged time or may not be able to participate in the survey because of illness, mental incapacities or language problems. If a survey organisation does not achieve the agreed upon rate, it has to be discussed whether additional fieldwork efforts and measures might be helpful when fielding the survey again in the future. Also a change of the survey agency might be considered. We should note that in the ESS only few countries included payment sanctions for not achieving the response rate target in their contract with the survey organisation.

response rates. Nonresponse bias is estimate-specific and can vary substantially across variables within the same survey (Groves/Peytcheva 2008). Estimating nonresponse bias requires comparative auxiliary information for both respondents and nonrespondents, which cross-national surveys have trouble providing (Blom et al., forthcoming). Furthermore, a target nonresponse bias is extremely difficult to budget for in fieldwork reality. Nonresponse bias targets are demanding in national studies (for an interesting attempt see Schouten et al. 2009), and nearly impossible to use – at least for the time being – in cross-national multi-topic surveys like the ESS.

Data Collection Procedures

In order to achieve the specified data collection outcomes and to improve comparability across countries the ESS defines data collection procedures that each participating country needs to follow. These procedures include the mode of interview, maximum interviewer workloads and interviewer briefings, a set fieldwork period, interviewer calling schedules, the collection of contact data and quality control back-checks.

Research has shown that the mode of data collection can affect survey results (Biemer/Lyberg 2003; Groves et al. 2004). Even within a country differential coverage, nonresponse and measurement errors across modes can cause mode effects; across countries the scope for differential errors are magnified. Consequently, the ESS collects its data in the same mode across all countries, namely by means of face-to-face interviews. For cross-national surveys face-to-face fieldwork offers several advantages over other modes including the best possible coverage of the target population and higher response rates (in most European countries). Furthermore, it is generally thought that the duration of a face-to-face interview can be longer than interviews in other modes.

In face-to-face surveys interviewers have a great potential to affect data quality.

> 'The task of the interviewer is more comprehensive and complex than merely asking questions and recording the respondent's answer. Interviewers implement the contact procedure, persuade the respondents to participate, clarify the respondent's role during the interview and collect information about the respondent.' (Loosveldt 2008: 202)

This can lead to interviewer effects in the resulting survey data, which the ESS tries to minimise via two main strategies: by training all interviewers in personal briefings (administered by NCs and/or researchers from the survey organisation) and by restricting the interviewer workload (at a maximum of 48 sample units per interviewer).

In each country interviewers should collect the ESS data in a fieldwork period of at least one month within a four-month period between 1st September and 31st December of the survey year. This serves to guarantee that the reference period of the ESS data is kept comparable, which is particularly important for an attitudinal survey like the ESS. At the same time the prescribed fieldwork period allows sufficient time for collecting data from difficult (in terms of contacting or gaining cooperation) sample units. Also practical considerations provide arguments in favor of a standardisation of the fieldwork period across countries. Lengthy and/or non-concurrent fieldwork periods make the coordination of the survey more time-consuming and can lead to delays in the data release. The risk of perpetuating delays from one round to another is another concern for repeated cross-sectional surveys like the ESS.

The probability of contacting a sample unit depends on the interplay of the sample unit's available at-home-pattern and the interviewer's number and timing of contact attempts (Groves/Couper 1998). Therefore, the ESS carefully specifies interviewer calling schedules that include sufficient calls spread over two weeks, on different days of the week and different times of day (a minimum of 4 contact attempts, of which at least one on a weekday evening and one at the weekend). To achieve a standardised measurement of the response process and allow cross-country comparisons thereof, the ESS countries have to collect contact data by means of ESS model contact forms.[5] These contact data include information on the timing, mode and outcome of each contact attempt and reasons for refusal, as well as information on the housing and neighbourhood of the sample unit. The ESS does not allow any substitution of difficult to reach or reluctant target persons and survey organisations are required to carry out checks of noncontacted, refusing and interviewed sample units.

The survey climate, that is the societal conditions that facilitate or mitigate survey participation, may vary between countries (Groves/Couper 1998), and in some countries the mandatory fieldwork specifications of the ESS may be insufficient to reach the response rate target. Therefore, ESS specifications suggest several additional measures (such as the selection of experienced interviewers, the use of respondent incentives or refusal conversion attempts, see Table A2 in the appendix). Each country is requested to consider these suggestions, but there is no general obligation to implement them.

5 The implementation of the ESS model forms is optional, provided countries deliver all mandatory variables described in the data protocol. ESS contact forms and contact data are available from http://ess.nsd.uib.no/.

With these standards and recommendations the ESS aims at a good balance between standardisation on the one hand and the provision of (some) leeway for national adaptation and customisation on the other hand. The same targets are set for all countries and minimum data collection standards are defined to help countries achieve these targets. These minimum standards may be complemented by optional measures. For example, countries can increase the number of call attempts interviewers make and use respondent and interviewer incentives of different types and values.

3.2 Support Documents Provided by the CCT

In addition to specifying details of survey design and implementation, the central coordination of the ESS also provides guidance and support documents[6], and personal assistance in tailoring the ESS procedures to the national situation. The ESS documents are quality assurance tools (Lyberg/Biemer 2008) designed to help planning and implementing national data collection. They are updated each round to reflect experiences from previous rounds and the latest scholarly insights into process quality. In the following we provide a short overview of the most important documents.

The *'Project Instructions'* guide the national teams in producing interviewer instructions. They cover information on the general background of the ESS, sampling, contact procedures and the use of contact forms, data protection, general interviewing principles and more specific aspects of individual survey questions. As circumstances and fieldwork traditions vary across countries, NCs are not supposed to produce *verbatim* translations of the ESS project instructions. Instead, NCs are advised to base their interviewer briefing agenda on these instructions and to ensure that all topics are covered.

A document on *'Field Procedures in the European Social Survey: Enhancing Response Rates'* assists countries in deciding on fieldwork strategies when aiming for the high ESS target response rates. It summarises advice on interviewer recruitment, training and organisation, as well as specific measures for reducing the two main sources of nonresponse: noncontact and refusal. Some of the procedures discussed in the paper are mandatory measures specified in the *'Specifications for Participating Countries'*. Other measures are additional recommendations or suggestions for maximising response.

6 These documents can be found at the ESS website: www.europeansocialsurvey.org.

The ESS specifications recommend the use of an advance letter to announce the upcoming interviewer call to the target persons. A model *'Advance Letter'* is provided as a guide on how countries might draft such a letter. Again a verbatim translation of this letter is not recommended. Rather the model letter outlines all the issues to be included in a national advance letter.

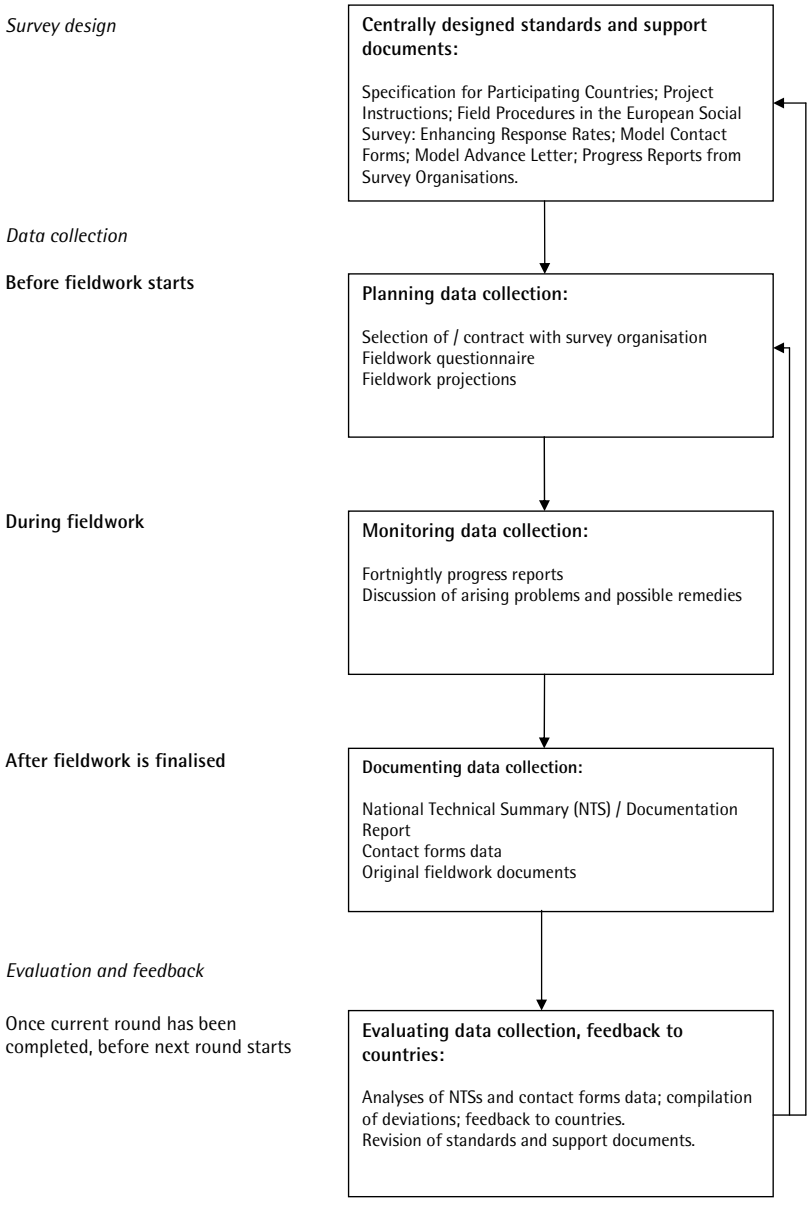
To calculate consistent response rates, evaluate fieldwork procedures and analyse possible nonresponse bias the CCT requires countries to deliver detailed data on the contacting and cooperation process for both respondents and non-respondents. To standardise the collection of these data the CCT designs and provides countries with model *'Contact Forms'*. The provision of model contact forms is accompanied by instructions and examples of how to complete them, and by an algorithm demonstrating how to arrive at final outcome codes from the contact form data.

The document *'Progress Reports from Survey Organisations'* instructs NCs on monitoring the implementation of fieldwork in their country. The survey organisations appointed for ESS have to provide regular feedback (at least every fortnight) regarding fieldwork progress during the data collection period. Such a close monitoring of the fieldwork progress allows for the early identification of difficulties and enables timely solving of problems.

3.3 Data Collection: Planning, Monitoring, and Documenting

Cross-national data collection management in the ESS broadly distinguishes three phases: before, during and after fieldwork (see Figure 2). First the CCT formulates the ESS specifications and supporting documents. Before the start of fieldwork the CCT also aids NCs in planning national data collection in accordance to the ESS standards and recommendations. During fieldwork NCs monitor the fieldwork progress in their country and pass information on to the CCT for cross-national progress monitoring. If necessary, problems and corrective actions are discussed between the CCT and NCs and put into practice. After fieldwork completion NCs document the data collection process and the CCT collects documentation from each participating country in the ESS *Documentation Report* (http://ess.nsd.uib.no/). For single cross-sectional surveys the documentation phase completes the survey life-cycle. The repeated cross-sections of the ESS, however, contain an additional phase of analysis, evaluation and feedback regarding the data collection process (as suggested by Lynn 2003). This aims at transferring experiences and improvements from one survey round to the next.

Figure 2 Phases of Survey Implementation in the ESS



Before the Start of Fieldwork: Planning Data Collection

The ESS data collection life-cycle starts with the receipt of funding from the national research council, the subsequent tender invitation and selection of the national survey organisation. Since the ESS fieldwork costs have to be borne by national funding agencies, the main work in this phase is carried out by the NCs; the CCT only *oversees and supports* the process of fieldwork commissioning.[7] By means of a short fieldwork questionnaire covering the major parameters of fieldwork a CCT workgroup supports NCs in ensuring that the contracts with the survey organisations comply closely with the ESS specifications. The specifications provide the general framework for data collection in the ESS. However, in preparing national fieldwork NCs have to make a multitude of specific planning decisions, for example on the concrete target sample size and response rate, on fieldwork start and end dates, on the number of interviewers, on dates and contents of interviewer briefings, on the use of advance letters and incentives, etc. Each participating country has to fill in the fieldwork questionnaire before the contract with the survey organisation is signed. This will usually require the NC to consult with the national survey organisation. The fieldwork plan has to be discussed, any envisaged problems solved and an agreement on a final fieldwork strategy reached between the NC and the CCT. Once the fieldwork plan has been signed-off by the CCT, ensuring that at least the design of the survey is according to the rules, the respective country can start fieldwork and the NC needs to make sure that all agreed procedures are actually implemented.

In many ESS countries the process of fieldwork planning takes place without facing major problems. However, usually there are also some countries where the concrete fieldwork planning constitutes a compromise between conflicting targets or procedures. These conflicts arise against the background of national particularities, such as the available budget, personnel resources (including interviewers), or upcoming events like a national election. Typical issues in the fieldwork questionnaire discussions between the CCT and NCs include target response rates (if lower than 70 percent), the planned number of interviewers and their average workload, the timing of fieldwork, or the number and timing of contact attempts. In round 2, for example, one country was signed-off with a maximum interviewer workload of more than 48 sample units, to enable them to work with a small, but highly expe-

7 As a result of these national selection processes, fieldwork in the ESS is carried out by a somewhat eclectic mixture of survey organisations, including commercial survey agencies, national statistical institutes, non-profit organisations and university institutes (Billiet et al. 2007).

rienced and well-trained interviewer corps. In another country the fieldwork start was postponed to prevent fieldwork from coinciding with national elections.

Most countries participate in multiple ESS rounds. For these countries the CCT provides feedback on difficulties encountered in previous rounds of data collection at an early stage of planning. Countries are asked to explicitly address these deviations in the planning phase of the current round and demonstrate ideas for improvement.

The process of discussing and signing-off fieldwork plans of approximately 25 countries in each round of ESS is time-consuming. It often covers a period of more than 12 months due to differences in the timing of funding decisions and differences in data collection schedules across countries. The CCT's involvement in the data collection planning phase is pivotal in coordinating ESS implementation in 25 different survey organisations across Europe; it contributes to the development of country-specific fieldwork plans and to preventing deviations from ESS survey standards from the outset.

The final tool for planning fieldwork is the fieldwork projections, which each NC sends to their designated CCT contact person one month prior to the start of fieldwork. At a minimum these projections comprise the expected number of interviews per fortnight. The projections are used by the CCT (and the NC) as a standard to evaluate actual fieldwork progress against.

During Fieldwork: Monitoring Data Collection

Collecting data via face-to-face interviews in a cross-national survey usually consists of decentralised operations of thousands of (mostly) free-lance interviewers. Implementing strict quality standards can be demanding and close monitoring of the fieldwork progress is crucial. Only a close supervision allows for an early identification of difficulties, and makes it easier to diagnose and remedy problems within the fieldwork period. For this purpose all ESS countries are assigned a CCT contact person who monitors and discusses fieldwork progress with their country. During fieldwork the NCs are required to regularly check the fieldwork progress in their country. Survey organisations and NCs have to produce a progress report at least fortnightly and discuss this report with their designated CCT contact person. At a minimum the progress report includes the number of completed interviews conducted each week. The CCT contact person and the NC can compare this information to the fieldwork projections to identify possible problems and a need for action. Further essential information includes a breakdown of the issued sample into major outcome codes (like 'noncontact', 'refusal', 'language barrier', etc.) and an assessment of the overall response rate.

The progress reports give a broad overview of how fieldwork develops. In addition, NCs are required to ask the survey organisation to provide or have accessible more detailed information that can be consulted for national fieldwork monitoring. This detailed information is particularly relevant for providing advice on trouble shooting. Important additional information might include response rates for regions or individual interviewers, response rates for demographic subgroups of target persons, data on the number and timing of contact attempts or information about re-issues of reluctant target persons.

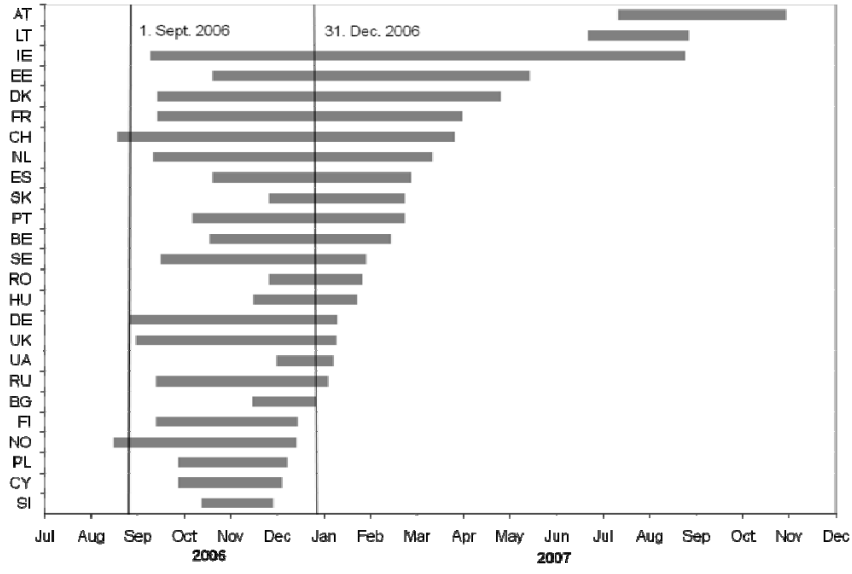
Typical problems encountered during fieldwork monitoring include delayed start dates for data collection, too few available interviewers or lower response rates than expected. This phase often requires decisions between conflicting targets: for example extending the fieldwork period to increase the response rate or allowing successful interviewers to work on additional sample units to decrease refusal rates. Budget constraints frequently complicate finding viable solutions, for example when discussing the training of additional interviewers or when considering respondent incentives.

The cross-national monitoring of data collection in the ESS covers a long time period. All national fieldwork periods taken together ESS fieldwork is considerably longer than the core four months between September and December. Both delayed start dates of fieldwork (especially in countries with problems in receiving funding on time), and (in some countries) fieldwork extensions well into the next year contribute to this. Figure 3 (see next page) shows the timing and duration of fieldwork in ESS 3.

After Completing Fieldwork: Documenting Data Collection

Meticulous documentation of procedures is an imperative for each survey aiming at high quality. For a survey like the ESS detailed documentation is even more important. The reason for this is twofold. First, in cross-national surveys the 'distance' between data producers and data users is larger than in national surveys (Lynn et al. 2006). Considering the multinational character of data collection with diverse survey organisations all over Europe, it is more challenging for a researcher to gain insight into the relevant aspects of the survey design and implementation. Second, this is even more relevant when the cross-national data are not primarily analysed by the researchers involved in designing and producing the survey, but constitute a public good available for interested researchers all over the world. Such datasets require an especially comprehensive documentation, so that secondary data analysts can evaluate the quality of the data.

Figure 3 Start and End Dates of the National ESS Fieldwork Periods in Round 3



In the ESS each country has to deliver its data to the ESS data archive (at NSD in Norway) once fieldwork, data entry and editing is completed. In addition to the interview data further documents (e. g. the original questionnaires, showcards, contact forms, interviewer instructions and advance letters) and data sets (e. g. the contact data) have to be delivered. Especially relevant for the documentation of fieldwork are the National Technical Summary (and the Documentation Report based thereon) and the contact data. The Documentation Report provides standardised information on the survey implementation and fieldwork procedures in each country: about the length of the fieldwork period, the selection, payment and briefing of interviewers, the call schedule, the use of advance letters, respondent incentives, refusal conversion strategies, the distribution of response outcomes, the use of quality control back-checks, etc. In addition to this aggregate-level process information the ESS makes micro-level process information available in the contact data. This includes information on the timing, mode and outcome of each contact attempt and reasons for refusal, as well as information on the housing and neighbourhood of the sample unit. The CCT uses these data, for example, to calculate response rates in all ESS countries in a consistent way and it publishes detailed reports on cross-national fieldwork processes (Symons et al. 2009). The

contact data are made publicly available via the ESS archive, and anyone interested can analyse them.

From One ESS Round to the Next – Aiming for Continuous Improvement

The ESS aims for charting and explaining social change. In order to achieve this, the study is based on cross-sectional surveys repeated at regular intervals. In contrast to single cross-sections the systematic and planned replication every two years allows for learning from round to round. Striving for continuous improvement is a key feature of the ESS as a long-term project. This requires a dedicated effort to analyse fieldwork parameters in the ESS and to implement feedback processes to convey information about successes and shortcomings from round to round.

A team within the CCT is concerned with the analysis and evaluation of fieldwork procedures. Researchers review and assess the fieldwork and interviewing procedures in all ESS countries by means of the contact data. The CCT particularly studies interviewer calling patterns and refusal conversion efforts (Symons et al. 2009). In addition, the CCT documents and seeks to improve adherence to the ESS specifications. Information on data collection is assembled for each country and compared to the ESS specifications. These activities are mainly based on the Documentation Report, but also draw on analyses of the ESS interview and contact data sets.

The ESS documentation and evaluation activities form part of a round-to-round improvement process, whose basic features are described below.

Transparency: The ESS aims at making explicit all survey standards and at documenting all departures from these standards that occur during survey implementation. For this the ESS website publishes rich information: the original fieldwork documents for each country are provided; the Documentation Report is generated for every survey round; contact forms data are made publicly available; and reports on methodological analyses with the data are produced in a timely fashion.

Feedback: As mentioned above, at the beginning of each round all countries in the ESS receive individualised feedback about their performance in previous rounds. Problems and deficiencies such as low response rates, deficient calling patterns or high interviewer workloads are raised and individual strategies for future improvement are discussed.

Revision of standards and protocols: When planning the next ESS round all specifications and protocols are evaluated in the light of experiences and results in previous rounds. As a consequence, several – mostly slight – revisions and additions have been implemented in the past.

It is obvious that improvements in such a large-scale enterprise as the ESS can only take place gradually. The ESS involves multiple players that need to coordinate their actions for sustained success. In the following section we provide an overview of adherence to fieldwork standards in the first three rounds of the ESS.

4 Adherence to Data Collection Standards in ESS Rounds 1, 2 and 3

Participation in the ESS was high from the very beginning. Already in the first round in 2002/2003 22 countries took part (see Table A3 in the appendix). In rounds 2 and 3 the participation was even higher with 26 and 25 countries, respectively. 17 countries can be described as perennial ESS participants, having taken part in each of the first three biennial rounds. A total of 32 countries have fielded at least one round.

Table 1 provides an overview of compliance with a number of data collection targets and procedures for all countries in ESS rounds 1 to 3. The selection of criteria is guided by the idea that the more ESS countries adhere to each of these, the better the quality and comparability of the resulting data will be. For example, we expect that achieving the same (minimum) effective sample size across countries assures a similar (minimum) level of precision of results, independently from differences in the sampling design which may exist between countries. Similarly, low noncontact rates and high response rates contribute to low or comparable nonresponse biases across countries. Insisting on the same mode of interviewing (face-to-face) in all countries is a basic requirement to foster comparability. The duration of fieldwork is expected to be related in particular to the noncontact rate: the longer the fieldwork period, the higher the probability of finding someone at home. A joint fieldwork period (i. e. completion by the end of the year) minimises the chance of major events (like the credit crunch) impacting on survey results differentially across countries. In-person briefings of interviewers ensure that interviewers understand the meaning of the questions, promote a higher response rate and prevent substitution and other interviewer misconduct. The maximum interviewer assignment size is a means to minimise the chance of large interviewer effects. Taken together these targets and procedures give some idea of compliance and deviations in the ESS. In this overview we do not focus on individual countries but just present general patterns. Detailed information on individual countries (e. g. response rates, length of fieldwork, etc.) is available on the ESS data website (http://ess.nsd.uib.no/).

Table 1 Adherence to Targets and Procedures in ESS Rounds 1, 2 and 3 (All Countries)

Target / Procedure	Round 1 (# of countries)	Round 2 (# of countries)	Round 3 (# of countries)
Effective sample size of at least 1500 (800) interviews*	8	9	11
Response rate 70 percent or higher	5	6	5
Response rate 65 percent or higher	11	10	12
Noncontact rate 3 percent or less	8	7	11
Noncontact rate 5 percent or less	15	13	17
Interview mode: face-to-face	22	26	25
Fieldwork period at least one month (30 days)	21	26	25
Fieldwork period 4 months (122 days) at maximum	11	15	14
Fieldwork completed by the end of the survey year	5	5	5
Fieldwork completed by the end of January of the following year	8	13	12
All interviewers briefed in person	18	21	20
No interviewer with more than 48 realised interviews	12	15	16
Total number of countries	22	26	25

** Data on effective sample sizes provided by M. Ganninger. A description of how the effective sample sizes were estimated can be found in Ganninger (2006). No information on effective sample size available for one country in round 1, and three countries each in rounds 2 and 3.*

Table 1 shows that only a minority of countries in rounds 1 to 3 managed to reach the ambitious ESS targets concerning the effective sample size, the response rate and noncontact rate. Between eight and eleven countries achieved an effective sample size of at least 1500 interviews. In several countries budget constraints limited the possibilities for increasing the (effective) sample size. The proportion of countries obtaining a response rate of 70 percent or higher is lower. Five to six countries in each round achieved a response rate of 70 percent or higher. However, if we relax the criteria slightly, we find that ten to twelve countries in each round achieved a response rate of 65 percent or higher. Of course, a high response rate cannot be legislated for. The response rate of a country is the result of the at-home-patterns and the willingness to participate of its population on the one hand, and the efforts of the survey organisation and the interviewers on the other hand (Groves/Couper 1998). Between seven and eleven countries achieved the

noncontact rate target of three percent or less. Therefore, more countries reached the ESS noncontact rate target than the ESS response rate target. Apparently, it is easier to control the number of noncontacted sample units than the number of sample units refusing to participate.

Turning to some of the ESS fieldwork procedure requirements one finds that in each round all countries complied with fielding the ESS face-to-face. Similarly, with the exception of one country in ESS 1[8], all countries had a fieldwork period which lasted at least one month. A reasonably long fieldwork period is one prerequisite for achieving a low noncontact rate. On the other hand, the number of countries adhering to the maximum duration of fieldwork of four months is lower. In rounds 1 to 3 between eleven and fifteen countries completed fieldwork within four months. In several countries difficulties in achieving high response rates led to prolonged fieldwork periods.

Prolonging the fieldwork period beyond the limit of four months is one reason for a delayed end of fieldwork. Another reason are delayed start dates (see also Figure 3), often caused by difficulties in receiving funding in time. Both processes contribute to the finding that only five countries completed fieldwork in time (i. e. by the end of the year) across rounds 1 to 3. Fortunately, especially in rounds 2 and 3, several of the remaining countries managed to finish fieldwork by the end of January of the following year. Therefore, in ESS rounds 2 and 3 approximately half the countries completed fieldwork by the end of January at the latest.

The ESS specifications require that all interviewers are briefed during in-person briefing sessions before carrying out an assignment. In each ESS round, four-fifth of the countries complied with this requirement. One-fifth of the countries, however, did not brief all their interviewers personally (including in each round one country that did not conduct any interviewer briefings). In addition the interviewer workload is limited in the ESS: no interviewer should work on more than 48 issued sampling units. A somewhat weaker criterion for adherence to the interviewer workload limit is that no interviewer may complete more than 48 interviews.[9] According to this criterion between twelve and sixteen countries complied with the ESS requirement in each round.

8 Even in this country the fieldwork period took nearly one month, namely 29 days.

9 This is a somewhat weaker criterion, since usually interviewers that have completed 48 interviews worked on more than 48 issued sampling units. However, using this criterion simplifies our analysis, because calculating the number of interviews conducted is easier than calculating number of sample units worked on.

Table 2 Adherence to Targets and Procedures in ESS Rounds 1, 2 and 3 of Perennial Countries

Target / Procedure	Round 1 (# of countries)	Round 2 (# of countries)	Round 3 (# of countries)
Effective sample size of at least 1500 (800) interviews*	7	6	8
Response rate 70 percent or higher	3	4	2
Response rate 65 percent or higher	9	7	5
Noncontact rate 3 percent or less	5	7	9
Noncontact rate 5 percent or less	12	9	13
Interview mode: face-to-face	17	17	17
Fieldwork period at least one month (30 days)	16	17	17
Fieldwork period 4 months (122 days) at maximum	8	9	7
Fieldwork completed by the end of the survey year	5	3	4
Fieldwork completed by the end of January of the following year	7	9	7
All interviewers briefed in person	14	14	14
No interviewer with more than 48 realised interviews	10	9	9
Total number of countries	17	17	17

** Information not available for one country in rounds 1, 2 and 3.*

This is a rather cursory overview of adherence to fieldwork procedures in the ESS. We conclude this section by comparing the results across rounds to see whether compliance with procedures and targets has improved. For this we restrict our analyses to those 17 countries that participated in each of the first three ESS rounds (Table 2).

The overarching finding of Table 2 is stability. At this general level there is no clear indication of improved compliance with the ESS data collection rules and targets across the first three rounds. Regarding the response rate target (70 percent) and the maximum length of the fieldwork period (four months) even a slight deterioration may be observed in round 3 compared to rounds 1 and 2. Only the maximum noncontact rate (three percent) describes a modest improvement in round 3 compared to rounds 1 and 2.

This result is not surprising. Improvements in cross-national surveys can only be achieved incrementally, given the nature and complexity of such enter-

prises. They require the successful interplay of numerous survey organisations, dozens of researchers, hundreds of interviewers and thousands of respondents. For several survey specifications that we examined external constraints limit the possibilities for improvement. If, for example, funding decisions in a country are made too late, there is no way to resolve a delayed start of fieldwork (unless one excludes the country from taking part in that round). Similarly, a sampling design causing lower design effects may simply not be available in a country, where no up-to-date register of residents exists. Increasing the sample size or the number of primary sampling units might be alternatives, but usually these options are associated with higher costs. Therefore, some deviations will inevitably occur, despite all parties' dedication to improvement. In addition one has to note that some non-adherence to survey specifications is the result of deliberate trade-off decisions: in order to (better) comply with one specification, another specification has to be sacrificed. Examples for this are lengthy fieldwork periods to enhance response rates or too large workloads for the most experienced, well-trained interviewers. However there are also other deviations, resulting from errors, misunderstandings or deliberate non-adherence to the specifications because of local/national traditions or procedural habits of survey organisations. Such issues need attention, because in these cases improvements are essential and possible.

5 Final Considerations and Conclusion

The present paper investigated data collection quality standards and their monitoring in the ESS. We gave an overview of the different steps which can be distinguished when implementing a cross-national survey, and described the various documents and tools designed to provide support during the ESS data collection process. The first three rounds of the ESS showed that adherence to data collection targets and procedures was not perfect and that despite dedicated efforts no clear evidence of improvement from round to round is found. It should be noted, however, that this only holds for the examination at a rather general level as we did.

It does not mean that no improvements in more specific areas and/or in individual countries took place.[10]

However, to what extent does non-adherence to the data collection standards set in the ESS matter? Of course, one can easily imagine that some deviations have negative practical consequences. For instance, a notable delay in the time schedule because of a prolonged fieldwork period endangers the timely delivery of data to the users. However, the more basic question is: Does better compliance with the ESS data collection targets and procedures also come along with better quality and comparability of the data? Or vice versa: Does increasing noncompliance mean a worsening of data quality and comparability? It is obvious that this question cannot be answered satisfactorily within the scope of the present paper since that would require conducting various experiments and detailed data analyses. Nevertheless, it should be mentioned that there are some, admittedly scattered pieces of evidence that adherence to fieldwork standards really matters – from experiences within the ESS and beyond. Vehovar and Zupanič (2007), for example, report that ESS countries with a lower response rate are characterised by a larger nonresponse bias, when bias is measured as the difference between unweighted and weighted survey outcomes (using a post-stratification weight for sex, age and education). Billiet and Pleysier (2007) find that the number of interviewer visits makes a difference for the response rate achieved. According to their analyses the average response rate in ESS 2 would have been 7.5 percentage points lower, if all countries had stopped contacting target persons after four visits. Finally, Heath et al. (2009) show with data from the International Social Survey Programme (ISSP) that differences between countries in response rates and the mode of interview affect the substantive outcomes and can be a threat for the validity of cross-national comparisons.

10 An example is interviewers' contacting habits. Detailed analyses of the ESS contact data (Billiet/Pleysier 2007; Symons et al. 2009) show that adherence to the ESS call schedule (at least four visits to noncontacts, including at least one visit in the evening and at least one at the weekend) is far from perfect. For 14 countries comparable information on the number and timing of contact attempts to noncontacts is available for ESS rounds 2 and 3. In round 2 on average 58 percent of the noncontacted cases were attempted at least four times (unweighted mean across all 14 countries). And on average 73 percent of the noncontacted cases were visited at least once in the evening, and 44 percent at least once at the weekend. The CCT fed back this information to NCs on an individual basis and in round 3 compliance with the call schedule was improved. On average 72 percent of all noncontacted cases in round 3 were visited at least four times and 83 percent of the noncontacted cases received at least one visit in the evening and 68 percent at least one visit at the weekend. These additional investments in contact efforts came along with a (slight) increase in the number of countries achieving a low noncontact rate between ESS 2 and 3 (see Table 2).

We conclude our paper with three further considerations regarding the evaluation of standards and improvements in ESS fieldwork over time. First, it is worth mentioning that the ESS aimed for high standards from the very beginning. In a way, already continuing to pursue these standards can be seen as a success. Retaining the response rate target of 70 percent, for instance, is valuable in an environment where the general trend is towards decreasing response rates (de Leeuw/de Heer 2002). In the first three rounds of the ESS we find that especially countries with above-average response rates face difficulties maintaining their response rate level.

Second, the paper focussed on data collection issues. However, these are only one aspect of the survey process, which unfortunately the researcher has limited control over. Designing reliable and valid survey questions and instruments, translating them adequately into all survey languages and selecting efficient probability samples are other aspects which also need careful planning and implementation to assure high survey quality. Kohler (2008), for example, showed that the sampling quality of the ESS is higher than in other cross-national surveys.

Finally, a perennial question in cross-national surveys aiming at high quality is how to deal with countries not complying with the standards. This will, of course, depend on the type, the severity and the reasons for non-compliance, as well as on the relative importance of a specific quality aspect. The strictest reaction is excluding these countries' data from the integrated dataset. A milder reaction is flagging such breaches alongside the data. In the ESS, for example, deviations from question formulations resulted in the removal of the question from the combined dataset for the respective country. Deviations from random sampling were not accepted at all. Low response rates were accepted, as were deviations from the fieldwork period.

When considering reactions to non-compliance one needs to take into account what effect these may have on a country's ability (and sometimes also its willingness) to participate in subsequent rounds. Excluding a country's data from the integrated dataset can impact on the country's eligibility for future funding (and thus denying it the chance to improve survey quality). However, including data that deviate too far from the cross-national standards endangers comparability. Furthermore, accepting deviations once can easily lead to institutionalising their acceptance overall. This is one of the trade-off decisions faced in cross-national surveys like the ESS.

References

Biemer, P. B. and L. E. Lyberg, 2003: Introduction to survey quality. Hoboken: Wiley.

Billiet, J., A. Koch, and M. Philippens, 2007: Understanding and improving response rates. Pp. 113-137 in: R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva (eds.): Measuring attitudes cross-nationally - Lessons from the European Social Survey. London: Sage.

Billiet, J. and S. Pleysier, 2007: Response based quality assessment in the ESS – Round 2: An update for 26 countries. Version May 5, 2007. Center for Sociological Research, K. U. Leuven.

Blom, A. G., A. Jäckle, and P. Lynn, (forthcoming): The use of contact data in understanding cross-national differences in unit non-response. In: J. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, and P. Ph. Mohler (eds.): Survey methods in multinational, multiregional, and multicultural contexts. New York: Wiley.

de Leeuw, E. and W. de Heer, 2002: Trends in household survey nonresponse: A longitudinal and international comparison. Pp. 41-54 in: R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (eds.): Survey nonresponse. New York: Wiley.

European Social Survey, 2001: Round 1 specification for participating countries. London: Centre for Comparative Social Surveys, City University London.

European Social Survey, 2003: Round 2 specification for participating countries. London: Centre for Comparative Social Surveys, City University London.

European Social Survey, 2005: Round 3 specification for participating countries. London: Centre for Comparative Social Surveys, City University London.

Ganninger, M., 2006: Estimation of design effects for ESS round II. Unpublished manuscript. Mannheim: GESIS.

Groves, R. M. and M. P. Couper, 1998: Nonresponse in household interview surveys. New York: Wiley.

Groves, R. M., D. A. Dillman, J. L. Eltinge, and R. J. A. Little (eds.), 2002: Survey nonresponse. New York: Wiley.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau, 2004: Survey methodology. Hoboken: Wiley.

Groves, R. M. and E. Peytcheva, 2008: The impact of nonresponse rates on nonresponse bias. A meta-analysis. Public Opinion Quarterly, 72: 167-189.

Häder, S. and P. Lynn, 2007: How representative can a multi-nation survey be? Pp. 33-52 in: R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva (eds.): Measuring attitudes cross-nationally - Lessons from the European Social Survey. London: Sage.

Heath, A., S. Fisher, and S. Smith, 2005: The globalization of public opinion research. Annual Review of Political Science, 8: 297-333.

Heath, A., J. Martin, and Th. Spreckelsen, 2009: Cross-national comparability of survey attitude measures. International Journal of Public Opinion Research, 21: 293-315.

Jowell, R. (1998): How comparative is comparative research? American Behavioural Scientist, 42: 168-177.

Jowell, R., M. Kaase, R. Fitzgerald, and G. Eva, 2007: The European Social Survey as a measurement model. Pp. 1-31 in: R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva (eds.): Measuring attitudes cross-nationally - Lessons from the European Social Survey. London: Sage.

Kish, L., 1994: Multipopulation survey designs: Five types with seven shared aspects. International Statistical Review, 62: 167-186.

Kohler, U., 2008: Quality assessment of European surveys. Towards an open method of coordination for survey data. Pp. 405-424 in: J. Alber, T. Fahey, and Ch. Saraceno (eds.): Handbook of Quality of Life in the Enlarged European Union. London: Routledge.

Loosveldt, G., 2008: Face-to-face interviews. Pp. 201-220 in: E. D. de Leeuw, J. J. Hox, and D. A. Dillman (eds.): International Handbook of Survey Methodology. New York: Psychology Press.

Lyberg, L. and P. B. Biemer, 2008: Quality assurance and quality control in surveys. Pp. 421-441 in: E. D. de Leeuw, J. J. Hox, and D. A. Dillman (eds.): International Handbook of Survey Methodology. New York: Psychology Press.
Lynn, P., 2003: Developing quality standards for cross-national survey research: Five approaches. International Journal of Social Research Methodology, 6: 323-336.
Lynn, P., L. Japec, and L. Lyberg, 2006: What's so special about cross-national surveys? ZUMA-Nachrichten Spezial, 12: 7-20. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten_spezial/znspezial12.pdf (11/04/2009).
O'Shea, R., C. Bryson, and R. Jowell, 2003: Comparative attitudinal research in Europe (ESS deliverable number 1). London: Centre for Comparative Social Surveys, City University.
Park, A. and R. Jowell, 1997: Consistencies and differences in a cross-national survey. The International Social Survey Programme (1995). ISSP 1995 National Identity ZA No. 2880. Codebook Central Archive, Cologne.
Schouten, B., F. Cobben, and J. Bethlehem, 2009: Indicators for the representativeness of survey response. Survey Methodology, 35: 101-113.
Skjåk, K. K. and J. Harkness, 2003: Data collection methods. Pp. 179-193 in: J. Harkness, F. J. R. van de Vijver, and P. Ph. Mohler (eds.): Cross-cultural survey methods. Hoboken: Wiley.
Smith, T. W., 2007: Survey nonresponse procedures in cross-national perspective: The 2005 ISSP non-response survey. Survey Research Methods, 1, 45-54.
Symons, K., H. Matsuo, K. Beullens, and J. Billiet, 2009: Response based quality assessment in the ESS – Round 3: An update for 23 countries. Version 5 February 2009. Center for Sociological Research, K. U. Leuven.
Vehovar, V. and T. Zupanič, 2007: Weighting in the ESS round 2. University of Ljubljana, Faculty of Social Sciences.

Addresses of the Authors

Achim Koch
Annelies G. Blom
GESIS - Leibniz-Institute for the Social Sciences
PO Box 12 21 55
68072 Mannheim
Germany
Achim.Koch@gesis.org
Annelies.Blom@gesis.org

Ineke Stoop
Joost Kappelhof
SCP - The Netherlands Institute for Social Research
PO Box 16164
2500 BD, The Hague
The Netherlands
i.stoop@scp.nl
j.kappelhof@scp.nl

Appendix

Table A1 Data Collection Standards in the ESS

Target/Procedure	International Standard	Leeway for National Adaptation[a)]
Data collection outcomes		
Effective sample size	1500 (800) interviews[b)]	
Response rate	Target 70 % (minimum)	
Noncontact rate	Target 3 % (maximum)	
Data collection procedures		
Mode of data collection	Face-to-face	Countries to decide whether PAPI or CAPI.
Fieldwork period	1 - 4 months between September and December	Only in substantiated circumstances may a country deviate from this timetable (after discussion with the CCT).
Briefing of interviewers	In-person briefing sessions (administered by NCs and/or researchers from the survey organisation). Briefings should cover respondent selection procedures (if applicable) and recording of the fieldwork process using the standard contact forms. Sections of the questionnaire that require special attention should be pointed out and explained carefully to interviewers. A practice interview should be conducted.	Countries free to decide on the specific content and the length of briefing sessions. Countries may for example provide training in response-maximisation techniques and doorstep interactions.
Interviewer workload	Maximum of 48 issued sampling units per interviewer	
Number and timing of contact attempts	At least four personal visits, including one visit on a weekday evening and one visit at the weekend, spread over at least two weeks.	Additional contact attempts possible.
Mode of first contact	In person	In countries using a sample of named individuals with telephone numbers, first contact may be made by phone (in order to make an appointment to visit the respondent).
Contact forms	Each country to provide a dataset with the timing, mode and outcome of each contact attempt, reasons for refusals and a number of specified observable area, dwelling and household characteristics.	Use of model contact forms recommended; but countries free to use own forms as long as all required data are delivered in required format.
Quality control back-checks	To be carried out and documented on at least 5 % of respondents, 10 % of refusals and 10 % of noncontacts.	Some discretion in the way the back-checks are carried out.

a) It should be noted that in addition to the leeway described in the table countries that have trouble achieving certain standards (e. g. the effective sample size or the target response rate) may discuss alternatives with the CCT.
b) Minimum of 800 interviews, if the target population of a country is less than 2 million people.

Table A2 Further Recommendations Concerning Data Collection in the ESS

Procedure	Recommendation and Leeway for National Choice
Selection of interviewers	Selection of experienced interviewers recommended.
Payment of interviewers	Recommended to discuss interviewer pay arrangements with the survey organisation. Consider implementing a bonus system. The pay rates for ESS should be attractive for interviewers, both with respect to the study difficulty and with respect to the pay on other studies.
Advance letter	Use of advance letters (personalised, if possible) recommended; model advance letter provided; recommended to include letter in interviewer workpackages and instructing them to post the letter a few days before they intend to call at the address.
Respondent incentive	Use of incentives recommended; type and handling of incentives to be decided by individual countries.
Refusal conversion	Use of refusal conversion procedures recommended for all countries. Countries ultimately to decide whether they re-issue refusals. If possible, experienced interviewers should carry out the conversion attempts. Interviewers should be familiar with refusal avoidance techniques.
Other response enhancing measures	All potential survey organisations should be invited to suggest a range of techniques that they believe would enhance the final response rate.

Table A3 Participating Countries in ESS Rounds 1 to 3

Country	R1 Participant	R2 Participant	R3 Participant
Austria	✓	✓	✓
Belgium	✓	✓	✓
Bulgaria			✓
Cyprus			✓
Czech Republic	✓	✓	
Denmark	✓	✓	✓
Estonia		✓	✓
Finland	✓	✓	✓
France	✓	✓	✓
Germany	✓	✓	✓
Greece	✓	✓	
Hungary	✓	✓	✓
Iceland		✓	
Ireland	✓	✓	✓
Israel	✓		
Italy	✓	✓	
Latvia			✓
Luxembourg	✓	✓	
Netherlands	✓	✓	✓
Norway	✓	✓	✓
Poland	✓	✓	✓
Portugal	✓	✓	✓
Romania			✓
Russia			✓
Slovakia		✓	✓
Slovenia	✓	✓	✓
Spain	✓	✓	✓
Sweden	✓	✓	✓
Switzerland	✓	✓	✓
Turkey		✓	
UK	✓	✓	✓
Ukraine		✓	✓
Total	22	26	25

Kommentar zu Anna Schnauber und Gregor Daschmann:

„States oder Traits? Was beeinflusst die Teilnahmebereitschaft an telefonischen Interviews"
(MDA 2008, 2: 97–123)

Comment to Anna Schnauber and Gregor Daschmann:

„States or Traits? Factors Influencing the Willingness to Participate in Telephone Surveys"
(MDA 2008, 2: 97–123)

Olaf Bock und Kai-Uwe Schnapp

1 Einleitung

Das zentrale Ergebnis der Non-Response-Studie von Schnauber und Daschmann (im Folgenden S/D) lautet:

> „Die Umfrageeinstellung [...] ist ein stabiler Einflussfaktor. Da sich aber nur wenige und schwache Zusammenhänge mit grundlegenden Persönlichkeitseigenschaften und soziodemografischen Merkmalen zeigen, spricht dies zwar dafür, dass es bestimmte Personen gibt, die Befragungen gegenüber grundsätzlich abgeneigt sind, diese sich aber nicht grundlegend von den Teilnehmern einer Befragung unterscheiden. Somit kann davon ausgegangen werden, dass Verweigerungen *nicht* zu systematischen Verzerrungen der Ergebnisse von Umfragen führen." (Schnauber/Daschmann 2008: 97; Hervorhebungen: O.B./K.-U.S.)

Diese Schlussfolgerung hat, wird sie in der Praxis ernst genommen, weitreichende Konsequenzen für die Qualitätssicherung bei (telefonischen) Umfragen. Sie erscheint uns allerdings nicht so gut begründet, als dass sie unbefragt Bestand haben sollte. S/D's Konklusion basiert *erstens* vorrangig auf Aussagen zum fehlenden Zusammenhang zwischen „grundlegenden Persönlichkeitsmerkmalen" auf der einen und sozialstrukturellen Variablen sowie dauerhaften Einstellungen zu Umfragen auf der anderen Seite. Dieser Nichtzusammenhang wird nach unserer Lesart a) nicht nachgewiesen und wäre b) selbst bei einem Nachweis keine hinreichende Absicherung für die oben getroffene Feststellung. Diese Position werden wir begründen.

Zweitens wird die Frage der Nichterreichbarkeit im Aufsatz gar nicht thematisiert. S/D begründen dies im ersten Absatz damit, dass die Gruppe der Verweigerer „besonders interessant" für die Nonresponseforschung wäre (S/D 2008: 98). Wir werden erläutern, warum wir die Exklusion der Nichterreichbaren für unangemessen halten. *Drittens* erschwert das Fehlen einer Darstellung zum Verhältnis zwischen Panel- und Erstbefragten in der Studie deutlich die Interpretation der Ausschöpfungsquoten. Darauf möchten wir vor allem deshalb hinweisen, weil S/D die hohe Ausschöpfungsquote ihrer Verweigererstudie im Vergleich zu „Studien in der Vergangenheit" besonders positiv hervorheben (S/D 2008: 111). Zunächst soll aber der für die Diskussion zentrale Begriff der „Nonresponse" formal definiert werden.

2 Definition von Non-Response

Fixpunkt der von S/D aufgegriffenen Debatte ist der sogenannte *Nonresponse-Bias.* Er setzt sich zusammen aus dem Nonresponse-*Error*, dem Unterschied zwischen Respondenten R und Nichtrespondenten N, sowie der Nonresponse-*Rate*, dem Anteil der Nonresponse an der Gesamtstichprobe (Schnell 2008: 12). Für ein arithmetisches Mittel kann der Nonresponse-Error dann z. B. wie folgt formuliert werden:

$$Bias_{\bar{x}} = (\bar{x}_R - \bar{x}_N) * \frac{n_N}{n_N + n_R}$$

Träfe S/D's Studienergebnis zu – Respondenten (R) und Nichtrespondenten (N) unterscheiden sich in Umfragen *nicht* systematisch (S/D 2008: 97, 120) – so wäre stets von einem Nonresponse-*Error* von Null auszugehen. Das würde das alarmierte Eingangsplädoyer der Autoren (vgl.: 98) für eine dringende Erhöhung der Stichprobenausschöpfungen zur Verringerung der Nonresponse-*Rate* überflüssig machen, denn eine Verringerung der Nonresponse-*Rate* verändert die Untersuchungsdaten und -ergebnisse nicht, wenn ein Nonresponse-*Error* gar nicht vorhanden, der Unterschied zwischen Respondenten und Nichtrespondenten also Null ist. Lässlich würden auch neuere elaborierte Methoden zur *Korrektur von* Nonresponse-Bias (siehe dazu Schnell/Hill/Esser 2005: 314ff.; Schnell 2008). Wir halten es jedoch weder für plausibel, von einem über unterschiedliche Umfragestudien einheitlichen noch gar von einem Nonresponse-Fehler von Null auszugehen. Diese Position begründen wir im Folgenden.

3 Zusammenhang zwischen „grundlegenden Persönlichkeitsmerkmalen" und anderen Variablen

S/D stellen zunächst auf der Basis ihrer Daten fest, dass „sowohl bestimmte Traits als auch bestimmte States einen Einfluss auf die Teilnahmebereitschaft" (119) haben. Weiterhin beobachten sie, dass weder Traits noch States „mit grundlegenden Persönlichkeitsmerkmalen zusammenhängen" (119). Diese „grundlegenden Persönlichkeitsmerkmale" werden in der Analyse von S/D über die Variablen „Vertrauen in Mitmenschen", „Extrovertiertheit" und „Kommunikation/Kompetenz" (114f., 118f.) operationalisiert. Die Analyse mündet in der zunächst auf ihre Studie *(Exba 2007)* bezogenen Feststellung, dass aufgrund des fehlenden Zusammenhanges zwischen grundlegenden Persönlichkeitsmerkmalen auf der einen, sozialstrukturellen Variablen und dauerhaften Umfrageeinstellungen auf der anderen Seite, keine grundlegenden Repräsentativitätsprobleme zu befürchten sind.

Um die Aussage eines fehlenden Zusammenhanges zwischen diesen Variablen und z. B. den sozialstrukturellen Variablen zu belegen, reicht es u. E. aber nicht, eine logistische Regressionsanalyse durchzuführen, wie sie S/D in diesem Fall präsentieren, da hier lediglich die Effekte der erklärenden Variablen auf die Befragungsteilnahme geprüft werden. Vielmehr ist die Korrelation innerhalb der Gruppe der erklärenden Variablen explizit zu prüfen.[1] Eine solche Analyse wird in dem Beitrag jedoch nicht vorgelegt. Neben diesem formalen Problem ist aber vor allem auf den substanziellen Punkt hinzuweisen, dass das Vorliegen oder Fehlen von Korrelationen zwischen zwei Variablengruppen (hier „grundlegende Persönlichkeitsmerkmale" und „sozialstrukturelle Variablen") keine Auskunft darüber liefert, ob und wie diese Variablengruppen mit anderen Variablengruppen korreliert sind, wie es also um die Eignung der Daten für Inferenzschlüsse in nicht betrachteten Variablensegmenten bestellt ist.

Wenn es zutreffen soll, dass Verweigerungen nicht zu systematischen Verzerrungen bei Umfragen führen, muss man unterstellen, dass S/D von einem vollständig zufälligen Ausfallprozess ausgehen, die Ausfälle wären also „missing completely at random" (MCAR). Diese Annahme liegt nahe, weil S/D unkonditional von einem Fehlen systematischer Verzerrungen sprechen. Sie müssen also davon ausgehen, dass die Ausfälle weder von den beobachteten noch von den unbeobachteten

1 Die Notwendigkeit der Überprüfung von Korrelationen zwischen den erklärenden Variablen ergibt sich natürlich auch aus der Notwendigkeit, das Fehlen von Multikollinearität als einer Anwendungsvoraussetzung der Regressionsanalyse nachzuweisen. Über einen entsprechenden Test wird in dem Beitrag jedoch nicht berichtet.

Fällen abhängen. Die Ausfälle wären dann komplett ignorierbar, ihr Einschluss in die Analyse würde die Analyseergebnisse nicht ändern (vgl. Weins 2006: 207f.). Wie Weins jedoch weiter feststellt, ist der Ausfallmechanismus in der Regel unbekannt und die Annahme ignorierbarer Ausfälle nicht prüfbar (ebd. 208). Den Beweis der Ignorierbarkeit der Ausfälle treten S/D folgerichtig auch nicht an. Vielmehr rekurrieren sie auf die Unterstellung, dass eine Homomorphie der Verteilung bestimmter (sozialstruktureller) Variablen zwischen den Teildatensätzen der Teilnehmer und der Verweigerer auf eine Homomorphie der Verteilungen in allen analytisch interessierenden Variablen schließen lässt. Anders ausgedrückt ist die Unterstellung von S/D, dass die Repräsentativität der Stichprobe und damit ihre Eignung für Inferenzschlüsse an keiner Stelle beeinträchtigt ist. Diese Form eines korrelationsbezogenen „Analogieschlusses" ist jedoch nicht zulässig (Schnell 1993: 29; Gabler et al. 1994). Es muss vielmehr davon ausgegangen werden, dass eine strukturelle Übereinstimmung von Datensätzen in einem Variablenbereich nichts Ausreichendes über eine äquivalente Übereinstimmung auf anderen Variablen aussagt. Darüber hinaus zeigt die Nonresponseforschung, wie oben erläutert, dass diese Übereinstimmung in der Regel auch gar nicht gegeben ist. Nimmt man diese Einwände ernst, dann kann also allenfalls von einem Ausfallmechanismus ausgegangen werden, bei dem die Ausfälle „missing at random" (MAR) sind. Hier dürfen die Ausfälle zwar von den beobachteten, nicht aber von den unbeobachteten Werten abhängen (Weins 2006: 207f.). Auch den Beweis für die Richtigkeit dieser Annahme treten S/D aber nicht an und können ihn auch nicht antreten (vgl. erneut Weins 2006: 208). Ist ein Nachweis also nicht möglich, dass die Ausfälle MCAR oder MAR sind, müsste man im Sinne einer konservativen Strategie der Qualitätssicherung von einem systematischen Ausfallprozess ausgehen („missing not at random" – MNAR). Ein solcher Ausfallprozess erlaubt jedoch keine inferenzstatistische Nutzung der Daten. Zusammengefasst lautet unser Argument also: S/D unterstellen implizit einen MCAR-Ausfallmechanismus, dessen Existenz sie aber nicht nachweisen (können). Dieser fehlende Nachweis macht ihre allgemeine Schlussfolgerung, dass die Ausfälle in ihrer speziellen Studie und darüber hinaus in Umfragen generell ignorierbar seien, unglaubwürdig. Deskriptive Aussagen für die erhobenen Daten bleiben damit natürlich möglich. Die Zulässigkeit von Inferenzschlüssen auf der Basis der vorliegenden Daten muss jedoch angezweifelt werden.

4 States und Traits

Neben der Diskussion über die Repräsentativität der Daten ist es nach unserer Ansicht geboten, die Frage zu stellen, welche Aussagen mit den Daten von S/D auf der deskriptiven Ebene über die Rolle von States und Traits, also Eigenschaften der Situation und Eigenschaften der Personen aufgrund der Analyseergebnisse von S/D gemacht werden können.

In der deutschen (Schnell 1997, 2008) wie internationalen (Groves/Couper 1998) Forschung zur Umfragebeteiligung wird seit geraumer Zeit betont, dass die Beteiligung vor allem eine Funktion der Situation, weniger aber fester Überzeugungen für oder gegen die Teilnahme an Umfragen sei:

> „We believe that few householders have strongly preformed decisions about survey requests. Rather, these decisions are made largely at the time of the request for participation."

Nach den Ergebnissen einer Sekundäranalyse des Teilnahmeverhaltens in mehreren Bevölkerungsumfragen sind die Einflüsse auf die Teilnahmeentscheidung „higly susceptible to situational factors." (Groves/Couper 1998: 321) Im Sinne der Nomenklatur von S/D sind es also eher States (situative Faktoren) denn Traits (situationsübergreifend invariante Faktoren), die Einfluss auf die Teilnahmeentscheidung an Umfragen haben. Auch andere Arbeiten dokumentieren die Nichtangemessenheit der Vorstellung von der Existenz eines fixen Bevölkerungsstratums so genannter „harter" Verweigerer (von der Heyde 2002: 39; Niemann/Abel 2000: 119f.).

Diese Ergebnisse der empirischen Beteiligungsforschung werden von S/D nur zum Teil bestätigt. So erklären die situationsübergreifend stabilen Umfrageeinstellungen (vgl. die Systematisierung von S/D: 101) im Regressionsmodell für die Teilnahmeentscheidung einen sehr großen Teil der Teilnahmeentscheidung (vgl. die in der Regressionstabelle auf S. 118 berichteten Pseudo-R^2-Werte). Gleichzeitig hat nur ein Teil der situativen Faktoren, nämlich die hier als Stress- und Zeitfaktor deklarierten Variablen einen hohen Erklärungsanteil, während Interviewermerkmale sowie Dauer und Thema des Interviews als weitere Bündel situativer Faktoren nur wenig Erklärungskraft beitragen. Diesen Umstand versuchen S/D vor allem in Bezug auf den geringen gemessenen Einfluss der Interviewer argumentativ abzuschwächen. Das mutet aus zwei Gründen eigenartig an. Zum *einen* könnte man jedes Ergebnis einer konkreten Datenanalyse mit der Begründung nicht ernst nehmen, dass man aus anderer Forschung wisse, dass dieser Faktor einflussreich sei. Zum *anderen* hat die Variablengruppe der sozialstrukturellen Indikatoren eine im Vergleich viel höhere Erklärungskraft, wird aber in der Interpretation der Daten eher abschwächend behandelt. Berücksichtigt man das von Kozak (2009) vorge-

brachte Argument, dass man einen Varianzaufklärungsanteil nicht vor absoluten, sondern vor an den Hypothesen orientierten Schwellwerten interpretieren sollte, dann ist dieses Vorgehen von S/D um so weniger plausibel. Vor dem Hintergrund des Forschungsstandes würde man für die stabilen Merkmale eine sehr geringe, für die situativen Merkmale eine sehr hohe Varianzaufklärung erwarten. Die vorgelegten Ergebnisse weisen aber eine entgegengesetzte Struktur auf.

Diese Abweichung der eigenen Ergebnisse vom Forschungsstand wird von den Autoren nach unserer Einschätzung nicht ausreichend thematisiert. Vielmehr mündet die Analyse etwas voreilig einerseits in einer Entwarnung (keine Zusammenhänge zwischen diesen Umfrageeinstellungen und grundlegenden Persönlichkeitsmerkmalen, daher auch kein Zusammenhang des Ausfalls mit jedweden anderen interessierenden Merkmalen) und andererseits in der aktiven Forderung, dass es vor allem darum gehen müsse, das Image der Umfrageforschung zu verbessern, um höhere Ausschöpfungsquoten und damit eine höhere Repräsentativität von Umfragen zu erreichen (119f.). Letztere Forderung erscheint zwar insofern plausibel, als sie eine Kur für das Problem der ausfallbedingten Verzerrungen empfiehlt, die tatsächlich in der Hand derjenigen liegt, die Umfragen durchführen. Vor dem Hintergrund der oben geführten Argumentation überzeugt uns dieser Lösungsansatz allerdings nicht.

5 Nichterreichbarkeit

Ein weiteres, aus unserer Sicht bedeutsames Manko der Arbeit von S/D ist die grundsätzliche Nichtbefassung mit den nicht erreichbaren Mitgliedern der Bruttostichprobe. Nach einem ersten generellen Bezug auf Ausschöpfungsquoten beschränken sich S/D auf die „besonders interessante Gruppe der Verweigerer" (98). Diese Einschränkung bei der Untersuchung von Unit-Nonresponse erscheint uns als nicht plausibel, denn für einen Teil der Fachdiskussion ist das Problem der Nichterreichbarkeit inzwischen wesentlich gravierender, als das Problem der Verweigerung (vgl. u. a. Atrostic et al. 2001; Steeh et al. 2000; Baur 2006: 164ff.; Niemann/Abel 2000: 120; Dethlefsen 2000: 63; Tuckel/O'Neill 2002). Die schwindende Teilnahme an Telefonumfragen veranlasst einige Autoren sogar, von einer Transitionsphase zu sprechen, „in which the telephone survey is losing its status as the most popular mode of data gathering." (Tuckel/O'Neill 2002: 34) Indem „States oder Traits?" das Problem der Nichterreichbarkeit völlig unbeachtet lässt, bleibt ein Teil der Unit-Nonresponse in dieser Studie vollkommen unaufgeklärt. Das sehen wir als problematisch an, weil der Ausfall nichterreichbarer Personen nicht zufällig (also mindes-

tens MAR), sondern abhängig von spezifischen Attributen dieser Personengruppe ist (Schnell et al. 2008: 311). Frühere Nonresponsestudien haben gezeigt, dass sich leicht und schwer Erreichbare deutlich unterscheiden, so dass der Ausfall von schwer Erreichbaren eine Stichprobe erheblich verzerren kann (Baur 2006: 169). Unklar bleibt bei den von S/D gemachten Angaben zum Stichprobendesign (107, Anm. 5), ob im Rahmen der Hauptstudie überhaupt mehrere Kontaktversuche bei allen ausgewählten Stichprobenelementen unternommen wurden oder nicht vielmehr Nichtkontakte durch „neue" Stichprobenelemente ersetzt wurden. Im ersten Fall wäre die unzureichende Dokumentation der Ausfälle zu kritisieren, im zweiten läge einerseits ein problematisches, dem Redressment nahes Auswahlverfahren der Studie zugrunde, das Inferenzschlüsse problematisch macht. Mit redressmentähnlichen Ziehungsverfahren werden aber Nichterreichbarkeitsquoten über dem Erwartungswert erzeugt, weil vor allem leicht erreichbare Personen in die Stichprobe gelangen. Schwer erreichbare Personen haben eine deutlich unter dem Erwartungswert von Studien mit mehrfachen Kontaktversuchen liegende Wahrscheinlichkeit, überhaupt in die Stichprobe zu gelangen. Die oben zitierten strukturellen Unterschiede zwischen leicht und schwer erreichbaren Personen führen dann zu einer gegenüber der Grundgesamtheit verzerrten Stichprobe. Da S/D die Leserschaft nicht ausreichend über das Stichprobendesign aufklären[2], bleibt zunächst ungeklärt, ob neben dem Verweigerungsbias in ihrer Studie nicht auch ein erheblicher Nichterreichbarkeitsbias vorliegt.

Die Nichtbeachtung der Nichterreichbaren führt gleichzeitig zu einem weiteren analytischen Problem. Wird nämlich die Unterscheidung zwischen Unit-Nonresponse wegen Verweigerung und Unit-Nonresponse wegen Nichterreichbarkeit durch Nichtbeachtung einer der beiden Gruppen ignoriert, so kann ein wichtiger Wandel bei der Unit-Nonresponse in telefonischen Umfragen, wie ihn Steeh et al. (2001: 242) etwa für die USA aufgezeigt haben, gar nicht erkannt werden:

> „[…] refusal rates are not decreasing or moderating because potential respondents are more willing to talk to survey interviewers over the telephone. Rather what we are seeing […] may be a change in the character of nonresponse. Evidence of this is […] clearly apparent in the metropolitan area where refusals significantly declined and noncontacts significantly increased."

(vgl. auch: Collier/Bienstock 2007: 179) Den Verzicht auf eine Diskussion der Teilnahmeausfälle, die nach Ausfallarten differenziert, halten wir angesichts dieses Forschungsstandes für unplausibel und problematisch.

2 Vgl. die rudimentären Hinweise zur Stichprobenziehung von S/D auf S. 107 (Anm. 5); insbesondere der Hinweis auf das so genannte „Rösch-Design".

6 Verhältnis von Panel- und Erstbefragten

S/D belegen die Qualität ihrer Verweigererstudie unter anderem auch mit dem Verweis auf die sehr hohe Ausschöpfungsquote (59 %, S/D: 111f.), die sie bei Ihrer Verweigererbefragung erreicht haben. Vor dem Hintergrund der Ausgangsfrage des Beitrages *(States or Traits?)* verliert diese Ausschöpfungsquote aber dadurch an Überzeugungskraft, dass sowohl Erst- als auch Panelbefragte in der Studie befragt wurden, ohne dass S/D die Leserschaft darüber aufklären, wie hoch der Anteil der Panel- und Erstbefragten in der Stichprobe ist. Das ist insofern von Bedeutung, als man bei den Panelbefragten davon ausgehen muss, dass sie weder allgemein noch bezogen auf diesen Studiengegenstand strukturelle Verweigerer sind, denn als Panelbefragte haben sie ja mindestens bereits einmal an genau dieser Umfrage teilgenommen. Gibt es also stabile „Traits", – eine Frage, die mit der hier kritisierten Studie untersucht werden soll – so ist bei diesen Befragten davon auszugehen, dass tatsächlich vor allem situative Faktoren die Verweigerung beim Erstanruf erklären, während grundsätzlich eine Teilnahmebereitschaft vorhanden ist. Gleichzeitig ist aber von einer Konvertierbarkeit dieser Befragten auszugehen, die über der einer Zufallsstichprobe der Bevölkerung liegt, wie die Erstbefragten sie darstellen. Die leichtere Konvertierbarkeit der Panelbefragten mindert daher den Eindruck, den die hohe Ausschöpfung von 59 % hinterlassen soll. Die aufkommenden Zweifel hätten gemindert werden können, wenn S/D, wie oben bereits angedeutet, über den Anteil von Erst- und Panelbefragten berichtet hätten. Dies tun sie jedoch nur an einer Stelle, wenn nämlich berichtet wird, dass von den 139 Personen, die die Nachbefragung durch die Autoren verweigerten, 24 Panel- und 115 Erstbefragte waren. Diese Information ist insofern wertlos, als man ohne eine Information über das Ausgangsverhältnis nichts über unterschiedliche Teilnahmebereitschaften erfährt.

Den bereits gefallenen Begriff der Konversion aufgreifend möchten wir außerdem darauf verweisen, dass es sinnvoll gewesen wäre, die Nachbefragung als Wiederholungsbefragung mit Konversionscharakter zu interpretieren. Dies tut explizit Schräpler (2000) mit einer Verweigereranalyse der SOEP-Befragten. Die von Schräpler berichtete Konversionsquote durch Nachfrage bei Erstverweigerern lag bei 15 %. Interpretiert man die Verweigererbefragung als Konversionsversuch, dann verändert dies notwendigerweise auch die Erwartungen an die Befragten. Man versteht die Befragten bei dieser Herangehensweise als Personen mit unterschiedlicher Bereitschaft zur Befragungsteilnahme und eher nicht als eine Gruppe von Befragungsteilnehmern und eine Gruppe von Verweigerern. Mit Blick auf die S/D leitende Frage nach dem Stellenwert von States und Traits ist dies eine relevante Verschiebung der Perspektive. Außerdem ließen sich an dieses Verständnis kontinu-

ierlicher Unterschiede in der Teilnahmebereitschaft weitere Fragen anschließen, die leider in dem Artikel von S/D gar nicht diskutiert werden. Zu diesen Fragen gehört u. a., wie Befragte trotz einer ersten negativen Reaktion zu einer Umfrageteilnahme bewegt (konvertiert) werden können (vgl. dazu nochmals Schräpler 2000).

Es sei an dieser Stelle erwähnt, dass die Autoren bei allen Analysen des Abschnittes 5 regelmäßig darauf verweisen, dass sich Panel- und Erstbefragte in ihrem Antwortverhalten nicht signifikant unterscheiden (S/D: 107). Damit wird zwar deutlich, dass S/D sich mit den Unterschieden beider Gruppen befasst haben, unverständlich bleibt aber, warum der Leserschaft die Details dieser Analyse wie der Verteilung von Panel- und Erstbefragten vorenthalten werden. Beides hätte den Ergebnissen höhere Glaubwürdigkeit verleihen können.

7 Fazit

Schnauber und Daschmann stellen in ihrem Beitrag die Frage, ob es stabile Verweigererdispositionen gebe, die systematisch Ausfälle bei (Telefon)surveys erzeugen und damit die Eignung solcher Umfragedaten für repräsentative Aussagen über Grundgesamtheiten in Frage stellen. Ihre Antwort lautet nein. Wie wir zeigen konnten, weist die Argumentation von S/D eine Reihe von Problemen auf. Zuvörderst sind es Probleme der Interpretation ihrer statistischen Ergebnisse, die die sehr weitreichende Schlussfolgerung aus der Zusammenfassung in Frage stellen. Im Weiteren ist es die Nichtbefassung mit dem Problem der Nichterreichbarkeit, die die weitreichenden Schlussfolgerungen als diskussionsbedürftig erscheinen lässt.

Ein genereller Nachweis auch nur der Nichtexistenz eines durch Verweigerung verursachten Non-Response-Bias liegt nach unserer Einschätzung mit dem Beitrag von Schnauber und Daschmann nicht vor. Unit-Nonresponse, sowohl durch Verweigerung als auch durch Nichterreichbarkeit, bleibt also bis auf Weiteres ein Problem, mit dem sich die Umfrageforschung jeweils studienspezifisch auseinandersetzen muss. Dies macht es u. a. erforderlich, dass Umfragedaten Datensätze für Nichtrespondenten aufweisen, die so viel Information über die Befragungssituation und den Interviewer enthalten wie möglich, um eine sinnvolle Analyse der Nonresponse zu ermöglichen (Schnell 2008).

Literatur

Atrostic, B. K. , N. Bates, G. Burt und A. Silberstein, 2001: Nonresponse in U.S. Government household surveys: consistent measures, recent trends, and new insights. Journal of Official Statistics 17: 209-226.

Backhaus, K. , B. Erichson, W. Plinke und R. Weiber, 2008: Multivariate Analysemethoden. Eine anwendungsorientierte Einführung. Berlin/Heidelberg: Springer.

Baur, N., 2006: Ausfallgründe bei zufallsgenerierten Telefonstichproben am Beispiel des Gabler-Häder-Designs. S. 159–183 in: F. Faulbaum und C. Wolf (Hg.): Stichprobenqualität in Bevölkerungsumfragen. Bonn: Informationszentrum Sozialwissenschaften.

Behnke, J., N. Baur und N. Behnke, 2006: Empirische Methoden der Politikwissenschaft. Paderborn/München/Wien/Zürich: Schöningh.

Collier, J. E. und C. C. Bienstock, 2007: An analysis of how nonresponse error is assessed in academic marketing research. Marketing Theory 7: 163-183.

Dethlefsen, H.-A., 2000: Qualitätsmamagement in der CATI-Forschung. S. 49–64 in: V. Hüfken (Hg.): Methoden in Telefonumfragen. Opladen: Westdeutscher Verlag.

Gabler, S., J. H.P. Hoffmeyer-Zlotnik und D. Krebs, 1994: Gewichtung in der Umfragepraxis. Opladen: VS Verlag für Sozialwissenschaften.

Groves, R. M. und M. P. Couper 1998: Nonresponse in household interview surveys. New York: Wiley.

Heyde, C. v. d., 2002: Das ADM-Telefonstichproben-Modell. S. 32–45 in: S. Gabler, S. Häder (Hg.): Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland. Münster/New York/München/Berlin: Waxmann.

Kozak, M., 2009: What is strong correlation? Teaching Statistics 31: 85–86.

Niemann , S. und T. Abel, 2000: Stichprobenselektion in einer telefonischen Bevölkerungsbefragung: ein Vergleich von Teilnehmern und Nichtteilnehmern im Berner Lebensstil-Panel. S. 105-122 in: V. Hüfken (Hg.): Methoden in Telefonumfragen. Opladen: Westdeutscher Verlag.

Reuband, K.-H. und J. Blasius, 2000: Situative Bedingungen des Interviews, Kooperationsverhalten und Sozialprofil konvertierter Verweigerer. Ein Vergleich von telefonischen und face-to-face-Befragungen. S. 139-169 in: V. Hüfken (Hg.): Methoden in Telefonumfragen. Opladen: Westdeutscher Verlag.

Schendera, C. F.G., 2008: Regressionsanalyse mit SPSS. München: Oldenbourg.

Schnauber, A. und G. Daschmann, 2008: States oder Traits? Was beeinflusst die Teilnahmebereitschaft an telefonischen Interviews? Methoden — Daten — Analysen 2: 97-123.

Schnell, R., 2008: Antworten auf Nonresponse. Sozialwissenschaftlicher Informationsdienst. Methoden und Instrumente der Sozialwissenschaften 2008/1-1: 11-23. http://www.gesis.org/fileadmin/upload/dienstleistung/fachinformationen/servicepublikationen/sofid/Fachbeitraege/Methoden_2008-1-1.pdf (9.9.2009).

Schnell, R., P. B. Hill und E. Esser, 2005: Methoden der empirischen Sozialforschung. München/Wien: Oldenbourg.

Schnell, R., 1997: Nonresponse in Bevölkerungsumfragen. Opladen: Leske Budrich.

Schnell, R., 1993: Die Homogenität sozialer Kategorien als Voraussetzung für „Repräsentativität" und Gewichtungsverfahren. Zeitschrift für Soziologie 22: 16-32.

Schräpler, J.-P., 2000: Was kann man am Beispiel des SOEP bezüglich Nonresponse lernen? ZUMA-Nachrichten 46: 117-149. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_46.pdf (9.9.2009).

Steeh, C., N. Kirgis, B. Cannon und J. DeWitt, 2001: Are they really as bad as they seem? Nonresponse rates at the end of the Twentieth Century. Journal of Official Statistics 17: 227-247.

Tuckel, P. und H. O'Neill, 2002: The vanishing respondent in telephone surveys. Journal of Advertising Research (September - October): 26-48.

Weins, C., 2006: Multiple Imputation. S. 205–216 in: J. Behnke, T. Gschwend, D. Schindler und K.-U. Schnapp (Hg.): Methoden der Politikwissenschaft: Neuere qualitative und quantitative Analyseverfahren. Baden-Baden: Nomos.

Anschrift der Autoren

Prof. Dr. Kai-Uwe Schnapp
Olaf Bock
Universität Hamburg
Fakultät Wirtschafts- und Sozialwissenschaften
Fachbereich Sozialwissenschaften
Institut für Politikwissenschaft
Allende-Platz 1
20146 Hamburg
kai-uwe.schnapp@wiso.uni-hamburg.de
olaf.bock@wiso.uni-hamburg.de

Replik zum Kommentar von Olaf Bock und Kai-Uwe Schnapp zu Anna Schnauber und Gregor Daschmann:

„States oder Traits? Was beeinflusst die Teilnahmebereitschaft an telefonischen Interviews“
(MDA 2008, 2: 97–123)

Reply to Olaf Bock's and Kai-Uwe Schnapp's Comment to Anna Schnauber and Gregor Daschmann:

„States or Traits? Factors Influencing the Willingness to Participate in Telephone Surveys“
(MDA 2008, 2: 97–123)

Gregor Daschmann und Anna Schnauber

1 Einleitung

Bock und Schnapp (im Folgenden: B/S) haben sich intensiv mit unserer Studie zu Verweigerungsursachen bei Telefonumfragen auseinandergesetzt und ihre Überlegungen auf den vorausgehenden Seiten eingehend dargestellt. Wir schätzen das Interesse der Kollegen an unserer Arbeit und danken für die zahlreichen und kenntnisreichen Anmerkungen. Ein Teil ihrer Ausführungen hat uns dazu veranlasst, einige zusätzliche Informationen, deren Darstellung wir in der Erstpublikation versäumten, in dieser Replik nachzureichen. Allerdings müssen wir im Folgenden auch einen Teil der Kritik zurückweisen – dort wo sie uns unangemessen, unbegründet oder unsachgemäß erscheint.

Anlass des Kommentars von B/S sind deren grundlegende Zweifel am zentralen Befund unserer Studie. Als zentralen Befund referieren B/S eingangs ihres Kommentars folgende Passage aus dem Abstract unseres Aufsatzes: „Die Umfrageeinstellung [...] ist ein stabiler Einflussfaktor. Da sich aber nur wenige und schwache Zusammenhänge mit grundlegenden Persönlichkeitseigenschaften und soziodemografischen Merkma-

len zeigen, spricht dies zwar dafür, dass es bestimmte Personen gibt, die Befragungen gegenüber grundsätzlich abgeneigt sind, diese sich aber nicht grundlegend von den Teilnehmern einer Befragung unterscheiden. Somit kann davon ausgegangen werden, dass Verweigerungen *nicht* zu systematischen Verzerrungen der Ergebnisse von Umfragen führen." (Schnauber/Daschmann 2008: 97 nach B/S 2009: 249) Das zentrale Fazit von B/S lautet hierzu: „Ein genereller Nachweis auch nur der Nichtexistenz eines durch Verweigerung verursachten Non-Response-Bias liegt nach unserer Einschätzung mit dem Beitrag von Schnauber und Daschmann nicht vor." (B/S 2009: 257)

Dem stimmen wir voll und ganz zu. Denn dieser Beweis ist erkenntnistheoretisch unmöglich zu erbringen. Die Behauptung, Verweigerer und Teilnehmer von Befragungen zeigten keine systematischen Unterschiede, postuliert die Nullhypothese. Diese kann stets nur falsifiziert, nie jedoch verifiziert werden. Dies liegt schon in der Vorläufigkeit wissenschaftlicher Erkenntnis begründet. Ergebnisse, die für die Nullhypothese sprechen, schließen nicht aus, dass morgen eine Studie erscheint, die gegenteilige Befunde produziert. Eine wissenschaftliche Antwort auf unsere Fragestellung kann sich daher nur auf kumulative Befunde stützen. Jede Studie, die keine Unterschiede zwischen Verweigerern und Teilnehmern nachweisen kann, spricht für die Nullhypothese, ohne jedoch deren Richtigkeit zu beweisen, und stellt somit einen kleinen Mosaikstein im gesamten Forschungsfundus dar. So ist auch unsere Untersuchung zu verstehen – nicht mehr, jedoch auch nicht weniger. Mit ihrer Forderung nach einem endgültigen Beweis stellen B/S somit Anforderungen an unsere Studie, die Wissenschaft schlichtweg nicht einlösen kann.

Aufgrund dieser erkenntnistheoretischen Prämisse, die wir als selbstverständlich bekannt vorausgesetzt haben, sind wir auch weit davon entfernt, unsere Befunde zu generalisieren. Passim unterstellen B/S uns dies jedoch in ihrem Kommentar. So bezeichnen sie es z. B. als unsere „… allgemeine Schlussfolgerung, dass die Ausfälle in [unserer] … speziellen Studie und darüber hinaus in Umfragen generell ignorierbar seien … ." (B/S 2009: 252) In Zusammenhang mit diesem Generalisierungsvorwurf ist es allerdings verwunderlich, dass B/S ihre Kritik ausschließlich auf die Passage in unserem Abstract stützen, der – einem Abstract entsprechend – Befunde und Folgerungen natürlich verdichtet und verkürzt wiedergibt. Nicht zur Kenntnis nehmen sie hingegen folgenden Passus aus unserem Fazit: „Im Fall der hier beispielhaft untersuchten Befragungsstudie sind somit inhaltliche Verzerrungen der Umfrageergebnisse durch Teilnahmeverweigerung unwahrscheinlich. Natürlich kann dieser Befund derzeit noch nicht auf andere Befragungen oder Umfragen allgemein generalisiert werden. Aus praktischen wie finanziellen Gründen konnte die Untersuchung nur im Rahmen einer einzelnen Marktforschungsstudie erfolgen; der Einfluss von Thema und Länge der konkreten Befragung oder des Interview-

ers konnte somit nicht durch echte Variation überprüft, sondern nur anhand der Selbsteinschätzung des Befragten als Größe einbezogen werden. Der Einfluss der Erhebungsart und des Auftraggebers konnte gar nicht untersucht werden. Es ist somit nicht auszuschließen, dass es hier zu Wechselwirkungen von Befragungs- und Erklärungsvariablen kommt … ." (S/D 2008: 120) Diese Passage dürfte bereits klarstellen, dass wir unsere Befunde nicht auf andere Befragungen oder Umfragen allgemein übertragen, da dies – wie bereits ausgeführt – erkenntnistheoretisch nicht statthaft wäre. B/S gliedern ihren weiteren Kommentar in mehrere Abschnitte. Wir haben uns an diesen Aufbau gehalten, und gehen im Folgenden auf diese einzelnen Abschnitte und die darin enthaltenen Argumente ein.

2 Definition von Non-Response

Die von B/S angeführte Definition des Nonresponse-Error ist unstrittig. Daran anknüpfend allerdings folgern sie: „Träfe S/D's Studienergebnis zu – Respondenten (R) und Nichtrespondenten (N) unterscheiden sich in Umfragen *nicht* systematisch, … so wäre stets von einem Nonresponse-*Error* von Null auszugehen." (B/S 2009: 250) Unser Studienergebnis ist nicht, dass sich Respondenten von Nichtrespondenten generell nicht systematisch unterscheiden – wir generalisieren diese Aussage nicht, wie schon ausgeführt, auf die Gesamtheit aller anderen Umfragen und treffen, wie die Verfasser selbst feststellen, keine Aussage über Nicht-Erreichbare, die in unserer Studie ausgeklammert sind (vgl. unten, Abschnitt 5). Einen Nonresponse-Error = 0 haben wir somit nie postuliert.

Weiter führen B/S aus, unser „alarmiertes Eingangsplädoyer … für eine dringende Erhöhung der Stichprobenausschöpfungen" würde durch unsere Befunde „überflüssig" (B/S 2009: 250). Das widerspricht der Logik und Rhetorik einer wissenschaftlichen Publikation. Das „Eingangsplädoyer", wie B/S es nennen, stellt die Relevanz unserer Fragestellung dar und begründet somit, warum wir welchen Sachverhalt untersuchen. Wie und warum sollte es schon die Befunde der empirischen Untersuchung vorwegnehmen? Allerdings müssten wir natürlich, folgten wir der Argumentation von B/S, zumindest im Fazit Bemühungen um die Erhöhung von Stichprobenausschöpfungen als überflüssig darstellen. Denn B/S argumentieren, dass, wenn ein Nonresponse-Error nicht vorhanden ist, eine höhere Ausschöpfung, also eine Verringerung der Nonresponse-Rate, auch die Untersuchungsdaten und -ergebnisse nicht verändere und somit nutzlos sei. Aus zwei Gründen folgen wir jedoch dieser Argumentation in unserem Fazit nicht. Zum einen, weil wir die Prämisse nicht teilen: Wie oben bereits ausgeführt, gelangen wir nicht zu der Schlussfol-

gerung, dass der Nonresponse-Error = 0 sei. Zum anderen, weil wir, selbst wenn die Prämisse zuträfe, die Folgerung nicht teilen. Eine Verringerung der Ausschöpfung durch unsystematische, also zufällige Ausfälle (was einem Nonresponse-Error = 0 entspricht) ändert zwar nichts an den Ergebnissen, wohl aber an den Fallzahlen und damit am Vertrauensbereich beim Rückschluss auf die tatsächlichen Verhältnisse in der Grundgesamtheit. Je geringer die Ausschöpfung, desto größer der ökonomische Aufwand, um bei einer Befragung durch Nachkontaktieren oder Ersetzen von Stichprobenelementen die gewünschten Fallzahlen und somit die gewünschte Validität zu erhalten. Selbst bei einem Nonresponse-Error = 0 wäre es also kaum ratsam, Ausschöpfungsquoten zu ignorieren.

3 Zusammenhang zwischen „grundlegenden Persönlichkeitsmerkmalen" und anderen Variablen

Zentral befassen sich B/S in diesem Abschnitt mit unseren Inferenzschlüssen von erfassten auf nicht erfasste Variablensegmente. Sie weisen völlig zu Recht darauf hin, dass wir quasi unterstellen, dass „eine Homomorphie der Verteilung bestimmter (sozialstruktureller) Variablen zwischen den Teildatensätzen der Teilnehmer und der Verweigerer auf eine Homomorphie der Verteilungen in allen analytisch interessierenden Variablen schließen lässt. (...) Diese Form eines korrelationsbezogenen Analogieschlusses ist jedoch nicht zulässig (...). Es muss vielmehr davon ausgegangen werden, dass eine strukturelle Übereinstimmung von Datensätzen in einem Variablenbereich eben nichts Ausreichendes über eine äquivalente Übereinstimmung auf anderen Variablen aussagt." (B/S 2009: 252) Theoretisch ist dies völlig richtig dargestellt – es gibt keinen Nachweis für die Annahme, dass analytisch interessierende Variablen homomorph zu den von uns erfassten soziostrukturellen Merkmalen verteilt sind. Ein solcher Nachweis wäre, wie B/S ebenso richtig feststellen, auch gar nicht möglich, da diese Variablen und ihre Verteilungsparameter unter Umständen nicht einmal bekannt sind. Denn rein theoretisch ist eine unendliche bzw. endliche, aber sehr große Grundgesamtheit an Einflussgrößen denkbar. Dies hat zur Folge, dass erkenntnistheoretisch der Rückschluss auf nicht untersuchte Merkmale stets nur spekulativ sein kann. Dies wäre allerdings auch durch einen erheblichen Untersuchungsaufwand nicht zu ändern. Selbst wenn wir in unserer Studie zahlreiche weitere interessierende Variablen mit einbezogen hätten, hätte dies uns keine Sicherheit gegeben für den Rückschluss auf nicht erfasste Merkmale. Die Frage ist allerdings, welche Konsequenz man daraus zieht. Soll man deshalb, weil verlässliche Erkenntnis nicht möglich ist, auf die Untersuchung des Problems verzichten? Oder

das, was sich mit vertretbarem Aufwand realisieren lässt, realisieren, um zumindest Anhaltspunkte für Plausibilitätsüberlegungen zu erhalten? Wir haben uns für Letzteres entschieden. Unsere Befunde beweisen also nicht, auch nicht innerhalb unseres Befragten- bzw. Verweigerersamples, dass Verweigerer und Teilnehmer sich nicht unterscheiden. Aber sie sprechen eher für diese Annahme – und geben vor allem zu gegenteiligen Annahmen keinen Anlass.

4 States und Traits

Zentral ist in diesem Abschnitt, dass B/S uns vorhalten, wir würden den Befund, dass Interviewermerkmale sowie Dauer und Thema des Interviews als situative Faktoren nur wenig Erklärungskraft zeigen, „argumentativ abschwächen" (B/S 2009: 253) – was bedeuten würde, dass wir nicht ergebnisoffen mit unseren Daten umgehen, sondern unliebsame Befunde quasi wegdiskutieren. Dies weisen wir entscheiden zurück. Der Grund für diese defensive Argumentation ist nicht etwa willkürlicher Umgang mit unliebsamen Befunden, sondern vielmehr die Qualität unseres Materials. Wir weisen mehrfach in unserem Text darauf hin, dass wir bei den Merkmalen wie Interviewereigenschaften oder Dauer und Thema der Befragung nur sehr weiche Indikatoren in unserem Variablenset haben, da aus pragmatischen Gründen eine (quasi-)experimentelle Variation dieser Erklärungsgrößen nicht möglich war. So konnte z. B. die entscheidende Dimension „Sprache" (Tonlage, Sprachmelodie, Tempo, Dialektfärbung etc.) in der vorliegenden Studie nicht berücksichtigt werden. Die Sprache ist aber das einzige Instrument, das der Interviewer überhaupt hat. Ebenso wurden Thema und Länge der Befragung nicht variiert. Aufgrund dieser weichen Indikatoren halten wir es für nicht statthaft, diese Befunde als besonders aussagekräftig zu bewerten – im Gegensatz z. B. zu den soziodemografischen Faktoren, bei denen wie nahezu alle gängigen Variablen einbezogen haben. Uns irritiert allerdings die Asymmetrie der Kritik: B/S werfen uns passim vor, unsere Befunde voreilig oder unangemessen zu generalisieren – was wir, wie bereits dargelegt, in dieser Form nicht tun. Dort hingegen, wo wir aufgrund unzureichender Belastbarkeit unserer Daten von Generalisierungen absehen, wird uns hingegen vorgeworfen, eben dies nicht zu tun.

5 Nichterreichbarkeit

B/S vertreten die Ansicht, dass zur Klärung von Nonresponse weniger das Problem der Verweigerungen als vielmehr der Nichterreichbarkeit relevant sei und verweisen hierzu auf einige Studien. Wir könnten wir hier nun entgegnen, dass es durchaus auch Befunde gibt, die diese von B/S hervorgehobene Problematik der Nichterreichbarkeit nicht stützen, da sie keine Unterschiede im Antwortverhalten von leicht und schwer erreichbaren Zielpersonen feststellten (z. B. Blasius/Reuband 1995). Die Aussage, ob Verweigerung oder Nichterreichbarkeit ein bedeutsameres Problem darstellt, ist also durchaus strittig. Zudem haben wir – wie weiter unten aufgezeigt wird – deutliche Anstrengungen unternommen, das Problem der Nichterreichbarkeit einzuschränken. Allerdings liegt hier nicht der zentrale Streitpunkt. Problematischer ist vielmehr, dass B/S aus dieser Einschätzung einen Schwachpunkt unserer Studie konstruieren. Sie schreiben: „Ein weiteres, aus unserer Sicht bedeutsames Manko der Arbeit von S/D ist die grundsätzliche Nichtbefassung mit den nicht erreichbaren Mitgliedern der Bruttostichprobe. (...) Diese Einschränkung bei der Untersuchung von Unit-Nonresponse erscheint uns als nicht plausibel ... (...) Indem ‚States oder Traits?' das Problem der Nichterreichbarkeit völlig unbeachtet lässt, bleibt ein Teil der Unit-Nonresponse in dieser Studie vollkommen unaufgeklärt. Das sehen wir als problematisch an." (B/S 2009: 254) Wir können diesen Vorwurf des „Mankos" nicht nachvollziehen. Wir sind nicht angetreten, um das gesamte Problem der Unit-Nonresponse zu untersuchen. Unser Aufsatz heißt nicht: „Was beeinflusst Unit-Nonresponse bei telefonischen Interviews?", sondern: „Was beeinflusst die Teilnahmebereitschaft an telefonischen Interviews?". Wir haben unsere Fragestellung also auf das Problem der Verweigerung beschränkt, was wissenschaftlich selbstverständlich legitim ist. Da wir eben nicht – wie uns B/S im ersten Kapitel ihres Kommentars vorwerfen – aus unseren Befunden auf das Gesamtproblem der Nonresponse (also Verweigerer und Nichterreichbare) generalisieren, ist diese Beschränkung unproblematisch, selbst wenn – wie B/S zu Recht anmerken – ein Teil der Nonresponse in unserer Studie hierdurch unaufgeklärt bleibt. Es bleibt B/S selbstverständlich unbenommen, das, womit wir uns befasst haben, für weniger interessant oder relevant zu halten als andere Fragestellungen. Ihr Interesse an unserer Fragestellung ist allerdings nicht der Maßstab der wissenschaftlichen Qualität unserer Arbeit.

Weiter kritisieren B/S, wir würden die Leserschaft nicht ausreichend über das Stichprobendesign aufklären. Dies trifft unseres Erachtens nicht zu. Wir schildern explizit, dass es sich um eine zufallsgenerierte Telefonstichprobe nach dem in Deutschland anerkanntesten Ziehungsverfahren nach Häder und Gabler handelt,

die nach Bundesländern geschichtet ist. Dieses Verfahren der Stichprobenziehung für telefonische Stichproben ist dem Fachpublikum hinreichend bekannt und kann schon aus Raumgründen nicht in allen Details im Rahmen eines solchen Aufsatzes thematisiert und erneut beschrieben werden. Erläuternd haben wir auf die Ausführungen von Rösch (1998) verwiesen, der aufzeigt, welche Maßnahmen zur Reduktion von Stichprobenfehlern bei telefonischen Bevölkerungsumfragen möglich sind. Dieser Hinweis wird von B/S offenbar mit dem Ausdruck „Rösch-Design" gleichgesetzt, von dem in unserem Text keine Rede ist.

Schließlich – so B/S – bleibe unklar, „ob im Rahmen der Hauptstudie überhaupt mehrere Kontaktversuche bei allen ausgewählten Stichprobenelementen unternommen wurden oder nicht vielmehr Nichtkontakte durch neue Stichprobenelemente ersetzt wurden. Im ersten Fall wäre die unzureichende Dokumentation der Ausfälle zu kritisieren, im zweiten läge … ein problematisches, dem Redressment nahes Auswahlverfahren der Studie zugrunde, das Inferenzschlüsse problematisch macht." (B/S 2009: 255) Es ist in der Tat ein redaktionelles Versäumnis unsererseits, die Zahl der Kontaktversuche nicht erwähnt zu haben. Wir reichen diese Information gerne nach: Im Rahmen unserer Befragung wurde jedes Stichprobenelement bis zu fünf Mal zu unterschiedlichen Tageszeiten an unterschiedlichen Tagen kontaktiert, teilweise sogar öfter. Somit wurde das möglichste getan, um Nichterreichbarkeit zu beschränken. Wir haben also weder – wie B/S unterstellen – Nichterreichbarkeit ignoriert, noch impliziert unser Vorgehen einen erheblichen Nichterreichbarkeitsbias. Die weiteren Ausführungen von B/S zur Problematik von redressmentähnlichen Ziehungsverfahren können unkommentiert bleiben, weil sie – wie durch die Zahl der Nachkontaktierungen verdeutlicht wird – für unsere Studie schlichtweg nicht zutreffen.

6 Verhältnis von Panel- und Erstbefragten

Da Panelteilnehmer, wie B/S vollkommen zutreffend ausführen, vermutlich eine geringere Verweigerungsneigung aufweisen als andere Erstkontakte, spekulieren B/S im sechsten Abschnitt darüber, ob unsere „sehr hohe Ausschöpfungsquote" in der Verweigererbefragung von 59 % nicht dadurch zustande kommt, dass auch Panelteilnehmer – also Personen, die mindestens in einem der Vorjahre bereits an der Telefonbefragung teilgenommen hatten – befragt wurden. Diese Spekulation ist nicht zutreffend. Allerdings müssen wir einräumen, dass wir hierzu in gewissem Sinne eingeladen haben, da wir versäumt haben, das Verhältnis von Panel- und Erstkontakten entsprechend in unserem Text zu erwähnen. Es sei hiermit nachgereicht: Von allen Verweigerern waren 14 % Panel-Kontakte und dementsprechend

86 % Erst-Kontakte, von allen geführten Verweigererinterviews waren 17 % Panel-Kontakte und 83 % Erst-Kontakte. Somit ist die Ausschöpfung nicht auf die Panel-Kontakte zurückzuführen, sondern auf das Kernvorgehen unsrer Studie, in der es ja eben darum ging, durch gezielte Bemühungen mehr über die Verweigerungsgründe von den Verweigerern zu erfahren.

B/S schlagen vor, die Nachbefragung als Wiederholungsbefragung mit Konversionscharakter zu interpretieren. Dies ist aus zwei Gründen hier nicht möglich. Zum Einen haben wir faktisch nicht versucht, Konversion herzustellen: Beim Nachkontakt wurde nicht versucht, die Verweigerer doch noch zur Beantwortung der Eingangsbefragung zu bewegen. Es wäre somit nicht statthaft, dies dennoch als Konversionsbefragung zu interpretieren. Zum Anderen war es nicht unserer Ziel, das Konversionspotential bestimmter persuasiver Strategien für die Eingangsbefragung zu prüfen, sondern die Ursachen für Verweigerung. B/S kritisieren, dass wir nicht die Frage diskutieren, „... wie Befragte trotz einer ersten negativen Reaktion zu einer Umfrageteilnahme bewegt (konvertiert) werden können." (B/S 2009: 257) Dies ist mit Sicherheit ein spannendes und relevantes Problem. Es war jedoch nicht unsere Fragestellung. Dass wir uns mit der Fragestellung unserer Studie befassen und nicht mit anderen Fragestellungen, halten wir, wie schon im vorigen Abschnitt ausgeführt, für eine unsachgemäße Kritik.

Dass wir im Übrigen, wie B/S weiter bemängeln – die Details der Unterschiede von Panel- und Erstbefragten nicht eingehend darstellen, hat ausschließlich redaktionelle Gründe. Die Verdichtung der Studie auf eine zehnseitige Aufsatzpublikation erzwingt notwendigerweise die Ausklammerung zahlreicher Aspekte. Da es kaum Unterschiede zwischen Panel- und Erstkontakten gab, so dass der Nutzwert für die Leserschaft uns hier relativ gering schien, entschieden wir uns, diese nur zu erwähnen, wo sie auftraten, und ansonsten diesen Vergleich auszuklammern. Auf Anfrage stellen wir B/S wie auch anderen interessierten Lesern die ausführliche Arbeit gerne als PDF zur weiteren Einsichtnahme zur Verfügung.

7 Fazit

Insgesamt kommen wir somit zu dem Schluss, dass wir die zentralen von B/S geäußerten Kritikpunkte an unserer Untersuchung zurückweisen, da sie entweder Beweisqualität einfordern, die Wissenschaft nicht erbringen kann, Fragestellungen anmahnen, zu deren Beantwortung wir nicht angetreten sind oder Mängel im Untersuchungsdesign oder in der Datenanalyse unterstellen, die nicht zutreffend sind. Dennoch war der Kommentar für uns äußerst hilfreich, da er uns verdeutlicht

hat, dass wir vereinzelt in unserem Originaltext das ein oder andere Detail missverständlich oder unvollständig dargestellt haben, so dass wir hier die Möglichkeit hatten, diese Informationen nachzureichen. Sollten wir durch solche Darstellungsmängel im Originaltext Unklarheiten oder Fehlinterpretationen hervorgerufen haben, ist dies selbstverständlich ein Fehler unsererseits, und wir bitten die geneigten Leser, dies zu entschuldigen. Allerdings schließen B/S unseres Erachtens voreilig aus solchen fehlenden Informationen (z. B. zum Anteil der Panelbefragten oder der Zahl der Nachkontaktierungen) auf fehlerhaftes Verhalten unsererseits. Mit dieser Gleichsetzung ignorieren sie eine wesentliche Gepflogenheit des wissenschaftlichen Diskurses, wonach – bis zum Beweis des Gegenteils – anderen Forschern intellektuelle Redlichkeit und Lauterkeit im methodischen Umgang zu unterstellen ist.

Literatur

Blasius, J. und K. Reuband, 1995: Telefoninterviews in der empirischen Sozialforschung: Ausschöpfungsquoten und Antwortqualität. ZA-Information 37: 64-87. http://www.za.uni-koeln.de/publications/pdf/za_info/ZA-Info-37.pdf (18.11.2009).

Rösch, G., 1998: Maßnahmen zur Reduktion von Stichprobenfehlern bei telefonischen Bevölkerungsumfragen. S. 9–18 in: S. Gabler, S. Häder, J. H. P. Hoffmeyer-Zlotnik (Hg.): Telefonstichproben in Deutschland. Opladen: Westdeutscher Verlag.

Anschrift der Autoren

Prof. Dr. Gregor Daschmann
Institut für Publizistik
Colonel-Kleinmann-Weg 2
55099 Mainz
gregor.daschmann@uni-mainz.de

Anna Schnauber
Kellerstraße 22
65183 Wiesbaden
anna.schnauber@gmx.de

Rezensionen

Wolfgang Donsbach, & Michael W. Traugott (Hg.), 2008: The SAGE Handbook of Public Opinion Research. London, Thousand Oaks, CA: SAGE. ISBN-10: 141291177X , ISBN-13: 978-1412911771, 618 Seiten mit Index, 112,99 EUR.

Das Handbuch von Donsbach und Traugott macht es sich zur Aufgabe, „to give an overview of the most important concepts included in and surrounding the term public opinion and its application in modern social research" (S. 4). Das Vorhaben, dem Leser einen Überblick über das gesamte Forschungsfeld zu präsentieren, mag zwar für ein Handbuch zunächst selbstverständlich erscheinen, ist jedoch in Anbetracht der begrifflichen Breite von „öffentlicher Meinung", „Meinungsforschung" oder „public opinion (research)" ein hoch gestecktes Ziel. So schränken sich auch die Herausgeber ein, und stellen fest, mit dem Handbuch weder eine Kanonisierung des Feldes anzustreben noch in der Lage zu sein, eine abschließende Aussage über die korrekte Verwendung des Begriffs „public opinion" machen zu können.

Mit über 600 Seiten und 55 thematischen Artikeln ist der Band recht umfangreich und inhaltlich ausgesprochen breit angelegt. Das Handbuch ist in fünf Kapitel zu unterschiedlichen Themenbereichen gegliedert. Im ersten Teil, „History, Philosophy of Public Opinion and Public Opinion Research" wird zunächst ein genereller Überblick über „Public Opinion" gegeben. Sieben Aufsätze stellen die Entstehung, Kommunikation und Bedeutung öffentlicher Meinung dar. Der Schwerpunkt liegt zunächst auf Kommunikation und Ausdruck der öffentlichen Meinung in der Medienberichterstattung, im Internet, in der Kulturproduktion und in bürgerschaftlichem Engagement. Anschließend wird die Geschichte und Entwicklung der Meinungsforschung in vier Beiträgen beschrieben. Der zweite Teil des Buchs behandelt in insgesamt acht Essays „Theories of Public Opinion Formation and Change". Hier wird beispielsweise der Zusammenhang zwischen Wissen und Einstellungen thematisiert, es werden Perzeptionstheorien, Einflüsse der Medien auf die öffentliche Meinung oder der Effekt der Schweigespirale diskutiert und allgemein die Frage nach der Bedeutung von Einstellungen gestellt. „Methodology", genauer Forschungsdesign und Messverfahren, ist Thema des dritten und mit 19 Beiträgen umfangreichsten Teils des Handbuchs. Nach zwei grundsätzlichen Aufsätzen zum Nutzen und zu Problemen der Fragebogenforschung folgen kurze Überblicksessays zu verschiedensten quantitativen und qualitativen Erhebungsverfahren: persönlich-mündliche Befragung, Telefoninterviews, postalische Befragung, Internetsurveys, Gruppendiskussionen und Inhaltsanalyse. Weitere Beiträge behandeln speziellere Themen wie Panelbefragungen, Fragebogensplits, Nonresponse oder Stichprobenziehung. Der vierte Teil sammelt Beiträge zu „The Social and Political Environment of Public Opinion Research". Gegenstand der Aufsätze sind hier rechtliche und ethische Fragen, Einstellungen zur Meinungsforschung sowie die Verwertung der Ergebnisse durch Medien und Politik. Den Abschluss des Handbuchs bietet schließlich ein heterogenes Residualkapitel mit acht Beiträgen zu „Special Fields of Application", in dem Aspekte wie Marktforschung, die Messung langfristigen Wertewandels, Wahlforschung und andere Themen diskutiert werden.

Bei der Lektüre des Handbuchs wird deutlich, wieso die Herausgeber in ihrer Einlei-

tung einen eher defensiven Tonfall anschlagen und betonen, sie könnten das Feld nicht kanonisieren. In der Tat ist es so, dass der Leser sich ob der Fülle der Beiträge und verschiedenen Themen mitunter verloren fühlen und den Eindruck gewinnen kann, er lerne zwar viele Aspekte der Meinungsforschung, bzw. des „Public Opinion Research" kennen, erfahre aber letztlich nicht, was öffentliche Meinung, public opinion, überhaupt bedeute. Dies ist jedoch nur zu einem gewissen Teil den Herausgebern anzulasten, sondern erscheint vielmehr als ein Problem des gesamten Fachbereichs. Allenfalls hätte man erwarten können, dass in der Einleitung das Thema stärker vorstrukturiert wird und das Buch dadurch stärker einen roten Faden bekommt. Ob ein roter Faden bei einem Handbuch jedoch notwendig bzw. überhaupt möglich ist, sei dahingestellt.

Der überwiegende Teil der Beiträge ist sorgfältig geschrieben und bietet einen leicht verständlichen Einstieg in das jeweilige Thema. Sehr lesenswert ist beispielsweise der Beitrag von Visser, Holbrook und Krosnick zum Zusammenhang zwischen politischem Wissen, Einstellungen und Verhalten („Knowledge and Attitudes"). Es ist jedoch zu beachten, dass ein Teil der Autoren in der kommerziellen Markt- und Meinungsforschung beschäftigt ist, und ihre Darstellung durch die akademische Brille betrachtet irritierend sein kann. So etwa im Beitrag zu Panel Surveys: hier wird nicht auf die unschätzbaren Vorteile von Paneldaten in der Kausalanalyse oder bei der Kontrolle unbeobachteter Heterogenität eingegangen – vielmehr arbeitet der Autor überwiegend mit Prozentuierungen und weist darauf hin, dass die Ergebnisse moderner statistischer Verfahren zur Panelanalyse für die Kunden schwer verständlich sind. Auch inhaltlich kann es mitunter Überraschungen geben. Beispielsweise entpuppt sich der Aufsatz „The Start of Public Opinion Research" als ein gut lesbarer Essay über George Gallup; dies ist natürlich nicht zu beanstanden, zumal das Handbuch zwei weitere Beiträge über die Geschichte der Meinungsforschung beinhaltet. Dennoch wäre es wünschenswert gewesen, wenn der Titel den Fokus des Beitrages widergespiegelt hätte.

Insgesamt bietet das SAGE Handbook of Public Opinion Research tatsächlich einen sehr guten Überblick über viele Aspekte der Meinungsforschung und ist als Werkzeug für den niederschwelligen Einstieg hervorragend geeignet. Auch ein Einsatz in der Lehre (B.A.) bietet sich an. So könnte man beispielsweise in einem Proseminar zur Meinungsforschung für jede Sitzung einen oder zwei der Aufsätze als Grundlagentext vorsehen und die Themen mit Referaten zu aktuellen oder klassischen Studien ergänzen. Insgesamt halte ich das Handbuch für gelungen und kann dem interessierten Leser eine Anschaffung empfehlen. An Universitäten mit Studiengängen in Publizistik, Markt- oder Meinungsforschung sollte das Handbuch nicht fehlen.

Henning Best, Mannheim

* * * * *

Nikolaus Jackob, Harald Schoen & Thomas Zerback (Hg.), 2009: Sozialforschung im Internet. Methodologie und Praxis der Online-Befragung. Wiesbaden: VS-Verlag. ISBN: 978-3-531-16071-9, 377 Seiten, 34,90 EUR.

Der von Nikolaus Jackob, Harald Schoen und Thomas Zerback herausgegebene Sammelband enthält eine Anzahl von Studien mit dem Schwerpunkt der Methodologie und der Praxis der Online-Befragung von Auto-

ren unterschiedlichster wissenschaftlicher Herkunft und zu verschiedenen Problemperspektiven. Thematisch teilt sich der Band in drei Bereiche: In der Einführung zeigen die Autoren, wie die Online-Befragung als Methode in den Sozialwissenschaften etabliert ist. Die methodologisch orientierten Aufsätze im zweiten Teil befassen sich mit den Möglichkeiten und Grenzen der Online-Befragung. Stichprobenprobleme sowie der Mangel an Repräsentativität und Validität werden kritisch diskutiert. Im dritten Teil zeigen eine Reihe von Fallstudien die breiten Anwendungsmöglichkeiten von Online-Befragungen bzw. Online-Experimenten und deren Umsetzung in der Praxis.

Zerback, Schoen, Jackob und Schlereth machen im ersten Teil darauf aufmerksam, wie wichtig die methodische Qualität von Online-Befragungen ist. Sie arbeiten bei der Inhaltsanalyse von 40 sozialwissenschaftlichen Fachzeitschriften heraus, dass in einem nicht unerheblichen Teil der Beiträge methodische Defizite hinsichtlich Definition der Grundgesamtheit, Stichprobenauswahl und Repräsentationsschluss bestehen. Welker und Matzat berichten über die historische Entwicklung der Online-Forschung als interdisziplinäres Forschungsfeld. Das Internet ist nicht nur Forschungsinstrument, z.B. im Sinne von Online-Umfragen, sondern auch Gegenstand der Forschung selbst, wenn es z. B. um die sozialen und psychologischen Auswirkungen der Internet-Nutzung geht. Hier werden auch andere Datenerhebungsmethoden (wie Logfile-Analyse, Experimente, Inhaltsanalyse) angesprochen. Thomas Roessing erklärt auch für den Nicht-Informatiker verständlich die verschiedenen Kommunikationsmöglichkeiten des Internets und deren Nutzen für Online-Befragungen. Maurer und Jandura zeigen, wie wichtig es ist, den Weg der Entstehung der Daten zu beachten und schlagen Kriterien vor, wie eine Online-Umfrage gestaltet werden sollte, um qualitativ hochwertige Daten zu erhalten. Baur und Florian geben sehr gut umfassende und konkrete Handlungsanweisungen, wie man mit Stichprobenproblemen bei Online-Umfragen umzugehen hat.

Ungünstig erscheinen die inhaltlichen Wiederholungen über die Bedeutung und Beliebtheit, über Vor- und Nachteile, resp. Methodenprobleme der Online-Befragung in den Einleitungen der einzelnen Aufsätze. Es entsteht so der Eindruck, dass es sich hier nicht um ein Lehrbuch, sondern um eine Ansammlung von Aufsätzen handelt, die sich teilweise inhaltlich überschneiden.

In den weiteren Aufsätzen werden verschiedene Anwendungsbeispiele gezeigt, die sich mit der Lösung methodischer Probleme befassen. Die darauf folgenden Fallstudien im dritten Teil geben einen Einblick in die unterschiedlichsten Möglichkeiten der Online-Forschung. Hier kann man sich durchaus Anregungen für die Durchführung eigener Online-Umfragen holen.

Insgesamt erscheint mir der Band mit ca. 370 Seiten zum Thema „Sozialforschung im Internet" mit dem Schwerpunkt „Online-Befragung" zu lang. Für eine Lehrbuch-Ausgabe müssten zumindestens die inhaltlichen Wiederholungen herausgestrichen werden. Als Sammelband gibt er jedoch einen guten Einblick „hinter die Kulissen" der Online-Forschung.

Claudia Buchheister, Potsdam

* * * * *

Jochen Mayerl & Dieter Urban, 2008: Antwortreaktionszeiten in Survey-Analysen. Messung, Auswertung und Anwendungen. Wiesbaden: VS-Verlag. ISBN: 978-3-531-16175-4. 136 Seiten, mit 11 Abbildungen, 19,90 EUR.

Die Messung und Analyse von Antwortreaktionszeiten (Response Latency) ist in der methodischen Forschung kein neuer Ansatz, jedoch kommen entsprechende Verfahren in der Praxis verhältnismäßig selten zur Anwendung. Dabei hält die Computerunterstützung für telefonische und persönlich-mündliche Befragungen längst das technische Equipment zur aktiven und passiven Messung des Zeitintervalls zwischen der Fragepräsentation durch den Interviewer und der Initialisierung einer Antwort durch den Befragten bereit. Die aktuelle Weiterentwicklung dieser Methodik vorzustellen haben sich die Autoren zur Aufgabe gemacht und es gelingt ihnen dabei auf überzeugende Weise, sowohl den Nutzen solcher Messungen für die Umfrageforschung wie auch adäquate Analyseverfahren zu demonstrieren.

Der inhaltliche Teil zwischen Einführung und Resümee besteht aus drei Abschnitten: Im Kapitel „Antwortreaktionszeitmessung in Survey-Studien" werden zunächst die Techniken aktiver und passiver Reaktionszeitmessung mit verschiedenen Erhebungsverfahren dargestellt sowie Anwendungen und Ergebnisse früherer Studien berichtet. Dabei wird die Bedeutung von Reaktionszeiten etwa als Indikator für mentale Zugänglichkeit für Einstellungen in der sozialpsychologischen Forschung, zur Analyse von Prozessen der Informationsverarbeitung oder als moderierende Variable bei der Identifikation von Responseeffekten in der Methodenforschung anhand der jeweils relevanten Studien aufgezeigt. Das Kapitel endet mit einer Darstellung der zahlreichen Einflussfaktoren auf Antwortreaktionszeiten wie z. B. Übungs- und Lerneffekte, Motivationsverlust, Alter, Intelligenz oder Gesundheitszustand, die je nach Forschungsfrage als zu kontrollierende Störeffekte auftreten können. Spätestens an dieser Stelle offenbaren sich dem Leser die Komplexität einer sachgerechten Messung und Analyse dieser Zeitintervalle sowie einer adäquaten Zuschreibung der ermittelten Werte zu den jeweils in Frage kommenden Ursachen.

Den Möglichkeiten der methodischen und der statistischen Kontrolle solcher Störfaktoren ist das sich anschließende Kapitel „Kontrolle von Störeffekten und Datenbehandlung" gewidmet. Neben einer intensiven Interviewerschulung, der Einhaltung genereller Regeln für ein optimiertes Befragungsdesign sowie einer geeigneten Messwertvalidierung ist vor allem die Ermittlung und statistische Berücksichtigung einer individuellen Basisgeschwindigkeit zur Korrektur der rohen Reaktionszeit von zentraler Bedeutung. Hierfür werden jeweils die gängigen Methoden und Indizes sowie eigene Weiterentwicklungen der Autoren vorgestellt. Empfohlen wird zur Bereinigung der Messwerte die Verwendung des Residual-Index, der sich aus einer Regression der rohen Reaktionszeiten auf die Basisgeschwindigkeit errechnet. Im Falle von Latenzzeitanalysen auf Itemebene ist dieses Maß jedoch, wie die Autoren richtig bemerken, konstruktionsbedingt völlig ungeeignet. Hier sollte auf den Difference Score Index oder den Z-Score Index zurückgegriffen werden.

Wie sich die beschriebenen Mess- und Analyseverfahren in der Praxis einsetzen lassen und mit welcher Art von Ergebnissen dabei zu rechnen ist, wird im Kapitel „Empirische Anwendungen" anhand von drei Beispieluntersuchungen, die im Rahmen einer CATI-Studie mit 2.000 Interviews innerhalb eines DFG-geförderten Forschungsprojekts entstanden sind, veran-

schaulicht. Als geeignete Moderatorvariable für Response Bias erweist sich die Latenzzeit bei der Aufdeckung von Reihenfolgeeffekten mit Items zur Verhaltensintention. So korrelieren Assimilierungseffekte, die durch Voranstellung spezieller Fragen zu konkretem Verhalten (hier: gesundheitsbewusste Ernährung) vor einer allgemeinen Frage zu beabsichtigtem Verhalten (hier: intendierte Umstellung der Ernährungsgewohnheiten) negativ mit der gemessenen Antwortreaktionszeit. Der gleiche Zusammenhang wird bei der Messung von Akquieszenz mittels eines Split-Ballot-Verfahrens sichtbar. Mit Hilfe eines Strukturgleichungsmodells zur Erklärung von Verhalten und Verhaltensintention durch soziale Normen und Verhaltenseinstellungen weisen die Autoren nachfolgend signifikant stärkere Effekte bei spontan geäußerten sozialen Urteilen als bei Antworten mit langen Latenzzeiten nach. Ebenso ist ein Zusammenhang zwischen Reaktionszeit und temporaler Stabilität von Einstellungen auch unter statistischer Kontrolle von Störfaktoren wie Alter, Bildung oder Zustimmungstendenz erkennbar.

Mit ihren Ergebnissen zeigen die Autoren, dass Latenzzeiten für die Analyse von Umfragedaten einen nicht unbedeutenden Mehrwert an Informationen liefern. Dies gilt insbesondere für den Bereich der Methodenforschung etwa im Rahmen von Fragebogenevaluationen oder zur Qualitätssicherung. Für Umfrageinstitute gibt das Buch Anregungen, wie man mit verhältnismäßig geringem Programmieraufwand wertvolle Zusatzdaten zur Optimierung von Erhebungsinstrumenten gewinnen kann. Die Forderung allerdings, Reaktionszeitmessungen grundsätzlich in allen Bevölkerungsbefragungen per CATI durchzuführen, erscheint am Ende doch etwas zu hoch gegriffen. Denn will man sich dabei nicht auf kleine Pretestfallzahlen oder rein passive Messungen beschränken, dann bedeutet die zusätzliche Aufgabe einer korrekten Zeitmessung und Validierung auch für geübte Interviewer zumindest eine mentale Ablenkung und stellt damit selbst eine potentielle Fehlerquelle dar. Besonders empfohlen sei dieses Buch Lehrenden im Methodenbereich der Sozialwissenschaften für den Einsatz in Vorlesungen und Praxisseminaren. Viele Lehrstühle verfügen heute über kleinere CATI-Labore, die sich für Lehrforschungsprojekte zur Messung und Analyse von Antwortreaktionszeiten bestens eignen.

Marc Deutschmann, Offenbach am Main

* * * * *

Peter Kriwy & Christiane Gross (Hg.), 2009: Klein aber fein. Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS-Verlag, ISBN: 978-3-531-16526-4, 414 Seiten, 39,90 EUR.

Peter Kriwy und Christiane Gross haben im VS-Verlag einen Sammelband herausgegeben, dessen (Unter-)Titel wohl bei vielen Sozialforschern auf Interesse stoßen wird: "Quantitative empirische Sozialforschung mit kleinen Fallzahlen". Nicht immer kann man mit großen Samples arbeiten, in der Forschungspraxis ist man oft mit kleinen Populationen oder Stichproben konfrontiert, bei denen man mit den gängigen (insbesondere inferenzstatistischen) Methoden schnell an Grenzen stößt.

Ein erster Blick ins Inhaltsverzeichnis verspricht Interessantes, werden hier doch unterschiedlichste Konzepte, Verfahren und Methoden genannt. Die Tatsache, dass (neben einer ausführlichen Einleitung) 15 Beiträge ohne thematische Untergliederung anein-

andergereiht werden, befremdet zunächst etwas. Aus der Einleitung erfährt man, dass die ersten sieben Beiträge methodische Grundlagen darstellen sollen, während sich die restlichen acht Praxisbeiträgen widmen. Wie sich bei der folgenden Lektüre zeigt, ist diese Trennung nicht ganz stringent (und der Verzicht auf eine Gliederung der Beiträge daher nachvollziehbar), werden doch sinnvollerweise methodische Verfahren mit inhaltlichen Beispielen illustriert bzw. bei berichteten Forschungsergebnissen auch die methodischen Grundlagen erläutert.

Obwohl manche Beiträge inhaltlich oder methodisch miteinander verbunden sind, ist der Band nicht systematisch aufgebaut, sondern er umfasst vielmehr unterschiedlichste Zugänge, die sich einer durchgängigen Gliederung widersetzen. Eventuell könnte man Erhebungsdesigns von Auswertungsdesigns unterscheiden, was sich aber auch nicht bei allen Beiträgen durchhalten lässt, an dieser Stelle aber zur Orientierung dienen soll. Bei den Erhebungsdesigns zeigt sich bereits eine breite Palette bekannter und auch weniger verbreiteter Methoden:

Nicole Saam führt in die Computersimulation ein, die auch für kleine und kleinste Fallzahlen geeignet ist, weil in die Analyse eingehende Daten und Fälle im Rahmen der Simulation erst erzeugt werden. Geeignete Modellierungsstrategien, die die interne Validität garantieren, vorausgesetzt, lassen sich auf diese Weise Daten erzeugen, die im Extremfall von einem einzigen empirischen Fall ausgehen. Dies demonstriert Frank Arndt im hinteren Teil des Buchs mit der Simulation der Verhandlungen im Rahmen der Amsterdamer EU-Regierungskonferenz 1996 anhand eines formalen tausch- und verhandlungstheoretischen Modells.

Einem in der Forschungspraxis immer wieder vorkommenden Problem widmet sich Sabine Wagner. Sie beschäftigt sich mit der Datenerhebung bei Spezialpopulationen, bei denen häufig Stichprobenprobleme (etwa aufgrund geringer Grundgesamtheit bei nicht eindeutiger und nicht dokumentierter Zugehörigkeit) mit Schwierigkeiten im Feldzugang und geringer Teilnahmebereitschaft der Zielgruppe zusammentreffen. Für eine erfolgreiche Forschung ist hier eine abgestimmte Vorgangsweise bei den einzelnen Untersuchungsschritten erforderlich, wie sie in ihrer Analyse von lokalen Austauschnetzwerken zeigt.

Eine gewisse Konjunktur scheinen derzeit faktorielle Surveys (Vignetten) zu erleben, mit denen sich gleich vier Beiträge im Band beschäftigen. Gerade bei hypothetischen Entscheidungssituationen lassen sich durch das gezielte Abfragen mit gezielter Variation der Rahmenbedingungen (Vignetten) auch bei relativ kleinen Stichproben Aussagen treffen, die bei konventionellem Survey-Design mit komplexer Drittvariablenkontrolle große Stichproben erfordern würden. Jochen Groß und Christina Börensen gehen zunächst der Frage der Validität faktorieller Surveys nach und vergleichen diese mit Ergebnissen aus einer Verhaltensbeobachtung. Dabei – es geht konkret um abweichendes Verhalten im Straßenverkehr – zeigt sich zwar eine Diskrepanz zwischen (Verhaltensabsichten erfragenden) Vignetten und Beobachtung (von tatsächlichem Verhalten), doch zumindest die Effektrichtungen stimmten überein. Die anderen Beiträge zu dieser Thematik verwenden Daten aus einer Vignetten-Untersuchung zur Umzugsneigung von Ehepaaren, wenn ein Partner ein berufliches Angebot in einer fremden Stadt erhält. Katrin Auspurg u. a. untersuchen dabei die familialen Verhandlungs- und Entscheidungsprozesse in Form von Paarbefragungen, d. h. es wurden jeweils beide Partner mittels Vignetten befragt. Natascha Nisic und Katrin Auspurg vergleichen faktoriellen Survey und klassische Umfragetechnik zur gleichen Thematik. Sie ziehen dazu Daten aus dem sozioökonomischen Panel zu tatsächlich erfolgten Wohnortwechseln heran und kommen zum Schluss, dass in beiden Verfahren dieselben Faktoren für die Realisierung eines berufsbedingten Umzugs

eruiert werden können. Schließlich verwenden Martin Abraham und Thess Schönholzer die Daten aus dem faktoriellen Survey dafür, die Mobilitätsentscheidung anhand spieltheoretischer Überlegungen in Form einer Dilemmasituation zu modellieren.

Ein weiteres, in den letzten Jahren ebenfalls zunehmend öfter eingesetztes Verfahren präsentiert Andreas Techen: Er führt eine Netzwerkanalyse unter Freundschafts- und Ratgebernetzwerken durch und überprüft mit den Daten unterschiedliche Theorien aus dem Bereich Freundschafts- bzw. interpersonelle Beziehungen und Sozialkapital in Netzwerken.

Schließlich werden noch experimentelle Designs (außerhalb des faktoriellen Surveys) vorgestellt. Heiko Rauhut u. a. stellen Forschungen zur Neutralisationstheorie in der Kriminalsoziologie vor, für die sie die Fragenreihenfolge (Rechtfertigung von und Bereitschaft zur Begehung von Bagatelldelikten) variiert haben um damit Reihenfolge- bzw. Ausstrahlungseffekte im Sinn der Fragestellung messen zu können. Für das vorgestellte Forschungsdesign ist eine kleine Stichprobe zwar nicht zentrales Merkmal, es ist aber doch auch dafür geeignet. Ben Jann berichtet von einem Feldexperiment zu Hup-Verhalten von Fahrzeuglenkern, deren Auto bei grüner Ampel durch ein anderes blockiert wurde, wobei hier Unterschiede auf den Statusunterschied der beiden Fahrzeuglenker (abgelesen an den gefahrenen Autos) zurückgeführt wurden.

Für den Bereich der Auswertung liegen insgesamt weniger Beiträge vor, diese sind aber durchwegs lesenswert. Antje Buche und Johann Carstensen stellen ein interessantes Analyseverfahren vor, das weder qualitativen noch quantitativen Analysetechniken eindeutig zugeordnet werden kann: Bei der „Qualitative Comparative Analysis" (QCA) werden die Bedingungen für ein bestimmtes Ereignis in Wahrheitstafeln aufgelistet und mit Hilfe Boolescher Operatoren formallogisch dargestellt; daneben existiert eine Form von QCA, die auf Fuzzy-Logic basiert und auf diese Weise versucht, die in der Praxis oft inadäquate Reduktion von Bedingungen und Ergebnis auf die Dichotomie gegeben/nicht gegeben aufzuweichen. Nach Angabe der beiden Autoren eignet sich QCA vor allem für mittlere Fallzahlen.

Andreas Broscheid stellt die Vorteile Bayesianischer Statistik für die Analyse kleiner Fallzahlen vor. Derselbe Autor wendet Verfahren der Bayes-Statistik in einem anderen Aufsatz an, in welchem er der Frage nachgeht, ob die Entscheidungen eines bestimmten, als besonders liberal geltenden amerikanischen Berufungsgerichts sich tatsächlich von denen anderer Kreise unterscheiden.

Die beiden restlichen Beiträge behandeln statistische Verfahren. Benn Jann beschäftigt sich mit der Diagnostik von Regressionsschätzungen bei kleinen Stichproben, wobei er einen Schwerpunkt auf die Diagnose von Ausreißern legt, die gerade bei kleinen Stichproben verheerende Auswirkungen haben können. Dabei geht er auch kurz auf die logistische Regression ein. Werner Georg u. a. demonstrieren an einem Survey über studentische Fachkulturen und Lebensstile die Anwendung von Faktoren- und Latent Class Clusteranalysen, wobei angesichts einer Stichprobengröße von $n > 500$ die Eignung dieser Verfahren für „kleine" Stichproben fraglich bleibt und auch nicht näher diskutiert wird.

Insgesamt wäre eine etwas ausführlichere Beschäftigung mit statistischen Verfahren bei kleinen Fallzahlen sicher eine gute Abrundung des Sammelbandes gewesen, etwa zu den Möglichkeiten und Grenzen verteilungsfreier Verfahren oder Bootstrapping-Techniken. Der Gesamteindruck ist aber, dass dieses Buch jedenfalls wert ist, es in die Hand nehmen und darin zu schmökern. Es ist ein Buch, das Anregungen bringt, vielleicht einmal etwas Neues auszuprobieren oder einfach über Alternativen nachzudenken. Man kann und sollte von diesem Band aber kein Lehrbuch erwarten – diesen An-

spruch stellen die Herausgeber auch nicht. Es ist nicht möglich, in Form von einzelnen Beiträgen die gesamte Methodik eines Verfahrens darzustellen. Dazu muss man auf andere Literatur zurückgreifen.

Eines wird bei der Lektüre der unterschiedlichen Beiträge aber deutlich: Auch wenn die verwendeten Daten von geringer Quantität sein mögen, ihre Qualität darf nicht im Zweifel stehen. Gerade bei einer kleinen Fallzahl kommt es auf die Validität jedes einzelnen Datums an, schlampig erhobene Daten können gerade hier fatale Auswirkungen haben. In diesem Sinne ist der Titel des Buches auch als Forderung zu verstehen: Klein aber fein!

Martin Weichbold, Salzburg

* * * * *

Thomas Sauerbier, 2009: Statistiken verstehen und richtig präsentieren. München: Oldenbourg. ISBN 978-3-486-59060-9, XIX, 254 Seiten, 29,80 EUR.

Thomas Sauerbier (FH Gießen-Friedberg) bedient mit seinem Lehrbuch „Statistiken verstehen und richtig präsentieren" den deutschsprachigen Markt für Statistiklehrbücher, die die Darstellung statistischer Informationen nicht nebenher oder am Rande verhandeln, sondern Fragen der grafischen Darstellung in den Mittelpunkt rücken. Er erörtert systematisch und für den Einsteiger erschöpfend die Möglichkeiten und Regeln der anwendungsorientierten Darstellung von statistischem Zahlenmaterial. Sauerbier legt eine kompakte und dennoch präzise Einführung vor, die sowohl an Leser ohne statistische Vorbildung als auch an Leser mit Vorkenntnissen gerichtet ist, die ein Nachschlagewerk für praktische Problemstellungen suchen (S. 4f.). Eine hervorzuhebende Besonderheit des Buches ist, dass der Autor durch das gesamte Buch praktische Tipps für die Verwendung von Microsoft Excel zur grafischen Realisierung der besprochenen Diagrammarten gibt. Dabei warnt Sauerbier vor üblichen Anwendungsfehlern im Umgang mit diesem und verwandten Tabellenkalkulationsprogrammen und gibt auch konkrete Anleitungen, wie Diagrammtypen, die im Programm nicht zum standardmäßigen Repertoire gehören, aber mit einigen Kniffen realisiert werden können. Direkt am Ende des ersten Kapitels platziert Sauerbier einen Tipp zum Kopieren von Diagrammen aus Excel in Word, der alleine für viele Leser den Kauf des Buches wert ist (S. 7f.).

Sauerbier beschreibt den Weg von den Daten zum Diagramm in vier Schritten als gründliche Datenanalyse zur Identifikation der Informationen (1), als Festlegung des Teils der Daten, die präsentiert werden sollen (2) und als Auswahl der am besten geeigneten Darstellungsform (3) bei Beachtung der Regeln der Realisierung von Diagrammen (4) (S. 1f.). Der Autor verfolgt dabei „zwar einen wissenschaftlich fundierten, im Ergebnis aber eher pragmatischen Ansatz" (S. 4). Dies schließt genau die praktischen Empfehlungen ein, vor denen viele Autoren statistischer Lehrbücher oftmals zurückschrecken. Diese pragmatische Haltung ist deswegen kein Fehler, weil Sauerbier es genau versteht, den Erläuterungen stets Warnungen zur Seite zu stellen, wenn dies erforderlich ist: so etwa im Falle der Verwendung von 3D-Diagrammen, deren Schwächen so auch für den überambitionierten Anwender sofort sichtbar werden (S.82, 109f.). Für jeden Diagrammtyp werden (inkl. englischsprachiger Bezeichnung) in einem Steckbrief am Ende eines Unterkapitels zudem die wichtigsten Merkmale in Form allgemeiner Kennzeichen,

der Eignung und der Realisierung zusammengefasst.

Das Buch ist in 11 gut durchstrukturierte und für den Anfänger und Erstleser sinnvoll aufeinander aufbauende Kapitel gegliedert. Der Gebrauchsanweisung in der Einleitung folgen ausgewählte Grundlagen der Statistik (Kap. 2), die Skalenniveaus, Merkmalstypen, Verteilungsformen und die Unterschiede von Bestands- und Ereignismassen in klarer Sprache darstellen und an Positiv- und Negativbeispielen illustrieren. Das umfassende Kapitel 3 (S. 19-116) beginnt mit eigentlich herauszuhebenden Unterkapiteln zu Zahlenangaben und Tabellen: hier werden auch die Fragen der Formatierung von Zahlen, der gebotenen Genauigkeit von Angaben sowie der Darstellung und praktischen Gestaltung von Tabellen diskutieren. Diese Darstellung könnte durch zusätzliche gut formatierte Beispiele aus der Praxis verbessert werden, ist aber trotz ihrer Kürze auch so für den Studienanfänger der Wirtschafts- und Sozialwissenschaften bzw. Einsteiger in die empirische Sozial- und Wirtschaftsforschung eine Einführung von hohem praktischen Nutzen.

Das Kapitel 3 stellt in sechs weiteren Unterkapiteln die wichtigsten Diagrammtypen dar: Säulen-, Kreis- (und Ring-) sowie Linien- (und Flächen-)diagramm für eindimensionale Verteilungen, außerdem die Diagrammvarianten Gruppensäulen und -balken sowie Stapelsäulen und -balken. Für jeden Diagrammtyp wird ausgehend von der Einführung in Kapitel 3 angegeben, für welches Skalenniveau (z. B. keine metrischen Merkmale) und welchen Merkmalstypus (z. B. häufbar), sowie für welchen Zahlenbereich (z. B. keine negativen Werte) die Darstellung geeignet ist. Zudem wird genau dargestellt und abschließend zudem stets im „Steckbrief" zusammengefasst, wie der Diagrammtyp korrekt realisiert wird. Auch statistisch problematische, in Massenmedien und der Marktforschung aber gebräuchliche Diagrammtypen wie die Piktogrammmenge werden kritisch besprochen.

Am Beispiel des Stabdiagramms (46ff.) und des Paarbalkendiagramms (84ff.) wird eine besondere Stärke des Buches deutlich: Der Autor erläutert in einem Kasten, wie dieser in Excel nicht vorgegebene Diagrammtyp mit dem Programm dennoch optisch ansprechend realisiert werden kann. Freilich verzichtet das Buch auf eine Darstellung von Screenshots u. ä., so dass beim Nutzer grundlegende Kenntnisse im Umgang mit menügesteuerten Programmen vorausgesetzt werden. Das Kapitel 3 erläutert zudem die Nutzung und richtige Erstellung von Streudiagrammen und Blasendiagrammen und gibt geeignete Tipps zur grafischen Darstellung. Hierbei erweist sich die vom Autor selbst eingeräumte und aus didaktischen Gründen gewählte Redundanz als besonders nützlich: Das Unterkapitel 3.8 kann jederzeit als „How to…" für allgemeine Fragen der richtigen Beschriftung und Skalierung gelesen werden. Dabei geht Sauerbier auch auf die Darstellungsfragen und typografischen Aspekte ein, die scheinbar selbstverständlich sind, aber insbesondere von Studierenden stets fehlerhaft realisiert werden: so etwa, wenn Beschriftungen über der ersten selbst erstellten Abbildung besonders groß und zentriert ausgerichtet, aber inhaltlich diffus formuliert werden, auf einmal für eine Balkenbeschriftung nur Großbuchstaben auftauchen oder der Hintergrund des Diagramms grau ist. Solchen Fehlern beugt Sauerbier wirksam vor (113f.).

Vertieft wird der anwendungsorientierte Anspruch des Buches in den Folgekapiteln 4 bis 7, die weitere Details der zweckgerichteten Verwendung von Abbildungen (Kap. 4), ein- und zweidimensionaler Verteilungen (Kap. 5 und 6) sowie von Zeitreihen (Kap. 7) darstellen. In Kapitel 4 widmet sich Sauerbier der Frage „Welches Diagramm für welche Art von Aussage?" und synthetisiert und vertieft die Befunde der vorangegangenen Kapitel. Ebenso wie das Kapitel 3 gilt dieses dem Autor als „Pflichtlektüre", da erst „die umfassende Kenntnis der verschiedenen Diagrammarten mit ihren Voraus-

setzungen, Vor- und Nachteilen sowie formalen Details eine optimale Auswahl und Gestaltung ermöglicht" (S. 5). So wird für Einzelwerte, Rangfolgen, zeitliche Entwicklungen, den Zusammenhang von Merkmalen und den Vergleich von Verteilungen wie mit Hilfe eines Werkzeugkoffers dargestellt, welche der zuvor eingeführten Diagrammtypen verwendet werden sollten. Auf etwas zu knappen sechs Seiten wird so eine Einordnung vorgenommen.

Ausführlicher widmet sich der Autor anschließend in Kapitel 5 den eindimensionalen Verteilungen, um wiederum an Beispielen die richtigen Darstellungsformen entlang der Skalenniveaus nominal, ordinal und metrisch vorzustellen. Hierbei treten nun zwangsläufig starke Redundanzen zum bereits Dargestellten auf, da im Wesentlichen die Sortierung entlang der Skalenniveaus und darzustellenden Dimensionen das zuvor in Kapitel 3 gewählte Ordnungsprinzip des Diagrammtyps ersetzt. Positiv anzumerken ist aber, dass durch die statistisch orientierte Sortierung entlang der Skalenniveaus nun die häufig virulenten Grenzfälle zwischen ordinalen und metrischen Merkmalen (132f.) sowie Fragen der geeigneten Klassierung von Merkmalen (136f.) genauer diskutiert werden. Die Regeln zur Klassenbildung sind für den Anfänger eine wertvolle Handreichung.

Diese Zusatzinformationen gegenüber den bereits in Kapitel 3 dargestellten Inhalten bereichern auch Kapitel 6 zu zweidimensionalen Verteilungen. Aufbauend auf den Informationen zu klassierten Merkmalen stellt der Autor nun für zweidimensionale metrische Verteilungen den Weg zum richtigen Diagramm dar und illustriert dies an Beispielen. Auch gemischte Merkmalsskalen werden in einem eigenen Unterkapitel entsprechend an Beispielen illustriert und geeignete Typen wie das Paarbalkendiagramm und das Paarhistogramm werden vorgestellt (S. 159). In Kapitel 7 liefert der Autor eine ausführliche Einführung der Darstellung von Zeitreihen und vergleicht die Vor- und Nachteile von Säulen-, Linien- und Flächendarstellung (161f.). Am Beispiel von Arbeitslosenzahlen werden die Probleme saisonaler Schwankungen und ihrer Möglichkeiten ihrer Darstellung etwa durch Glättung (Varianten gleitender Durchschnitte) diskutiert (170f.). Zudem zeigen Negativbeispiele, wie mit Zeitreihendarstellungen falsche Eindrücke erweckt werden können und wie mehrere Zeitreihen mit verschiedenen Diagrammtypen gleichzeitig dargestellt werden können. Kapitel 7 ist eine eigenständige Vertiefung und wertvolle Ergänzung, die über das übliche Spektrum anderer Lehrbücher hinausreicht und durch die Orientierung an gängigen Darstellungsformen in Excel und vergleichbaren Programmen, wie im Falle des Stapelsäulendiagramms mit Zeitreihen (S. 182), für den Leser von praktischem Nutzwert ist. So werden etwa auch Vor- und Nachteile der Darstellung entlang zweier Achsen in einem Diagramm im Vergleich zu der alternativen Realisierung einer Darstellung in zwei Diagrammen dargestellt.

Das Buch schließt mit vier Kapiteln, die sich der Statistik im numerischen Sinne widmen und die praktisch bedeutsamen Themen Prozentwerte (Kap. 8), Durchschnittswerte (Kap. 9), Streuung (Kap. 10, inkl. Boxplot-Darstellung und Schwebebalken) und Stichproben (Kap. 11) behandeln. Es liegt in der Ausrichtung und der anvisierten Leserschaft begründet, dass dieser Teil – verglichen mit anderen Statistiklehrbüchern – nicht nur vom Umfang knapper ausfällt (S. 199-248), sondern am Ende steht und gewissermaßen einen Appendix für den Leser darstellt, der sich kurz und knapp seiner Kenntnisse vergewissern möchte. Trotz der Kürze werden aber neben Modus, Median und arithmetischem Mittel auch geometrisches und harmonisches Mittel nicht ausgelassen. Dies erweist sich wiederum als besonders nützlich, da dem Excelnutzer ganz praktische Hinweise zur korrekten Berechnung gegeben werden.

Ergänzt werden könnte das Buch um Hinweise, wie aus den gängigen Statistikpaketen SPSS und STATA Daten importiert und

Tabellen in Excel und z. B. Open Office eingelesen werden können. Aber der Autor hat hier verständlicherweise im Sinne des Leseflusses eine Grenze gezogen. Etwas irreführend wirkt der Frageclaim auf der Coverrückseite „Wie liest man eine Statistik?", da dies am Wert des Buches vorbeizielt. Der Leser lernt nicht bloß das Lesen statistischer Daten, sondern vor allen Dingen die richtige Erstellung statistischer Abbildungen. „Wie übertrage ich statistische Daten regelkonform in eine Abbildung?" lautete also die geeignete Leitfrage.

Ein kritischer Leser, der Nutzer von TeX-Software zum Textsatz ist und professionelle Grafikprogramme zur Diagrammdarstellung verwendet, wird vermutlich den gesamten praktischen Ansatz des Buches samt der erkennbar der Microsoftfamilie entstammenden Schriften und Grafiken ablehnen. Und in der Tat kann man über die Qualität der Abbildungen bisweilen geteilter Meinung sein. So stört, dass der Autor alle Abbildungen generell einrahmt, da er dies für Geschmacksache hält. In der Praxis wissenschaftlicher Publikationen sollten die meisten Abbildungstypen aber in aller Regel freistehen. Auch fehlen Hinweise auf die durchgehende Einheitlichkeit von Abbildungen in einem Textdokument bzw. einer Publikation. Die Wahl der Parameter der Abbildung (wie Schriftgröße, Farb- oder Graustufenspektrum sowie Anordnung) sollte nicht für eine Abbildung alleine gelten, sondern sich wenn möglich im Sinne eines einheitlichen Designs durch eine Publikation ziehen.

Leser, die für komplexere Analyseverfahren und Diagrammtypen als die oben dargestellten Hinweise zu grafischen Darstellungen suchen, werden lediglich die Grundregeln des Buches als nützlich erachten und ansonsten auf die Spezialliteratur ausweichen müssen. Den in den Literaturhinweisen gelisteten Werken früheren Datums sowie der englischsprachigen Literatur wird mit Thomas Sauerbiers „Statistiken verstehen und richtig präsentieren" eine kompakte und gut lesbare Einführung zur Seite gestellt. Auf dem deutschsprachigen Markt für einführende Lehrbücher in die Statistik liegt ein Lehrbuch vor, das für Studienanfänger und Studierende von großem praktischen Nutzen ist und für fortgeschrittene Anwender und Lehrende eine Vielzahl wichtiger Hinweise zum Nachschlagen bereithält.

Tilo Beckers, Düsseldorf

Ankündigungen

Ausschreibung

ALLBUS-Nachwuchspreis 2010

ALLBUS-Nachwuchspreis 2010

Die seit 1980 alle zwei Jahre durchgeführte Allgemeine Bevölkerungsumfrage der Sozialwissenschaften - ALLBUS - ist ein zentraler Bestandteil der sozialwissenschaftlichen Infrastruktur in Deutschland. Die Bereitstellung einer qualitativ hochwertigen Datenbasis für Sekundäranalysen ist gerade für Nachwuchswissenschaftler wichtig, die nicht auf eigene Erhebungen zurückgreifen können.

Der *ALLBUS-Nachwuchspreis* unterstreicht die besondere Bedeutung des ALLBUS für die Ausbildung des wissenschaftlichen Nachwuchses. Er wird in Zukunft alle zwei Jahre verliehen, im Wechsel mit einem zweiten, allen Wissenschaftlern offenstehenden *ALLBUS-Award*, der erstmals 2011 und danach ebenfalls im zweijährigen Turnus vergeben werden wird.

Mit dem ALLBUS-Nachwuchspreis sollen herausragende Qualifikationsarbeiten prämiert werden, in denen ALLBUS-Daten eine zentrale Rolle spielen. Über die Verleihung des Preises, der mit 1.000 € dotiert ist, entscheidet eine Jury, die aus den Mitgliedern des wissenschaftlichen Beirats des ALLBUS besteht. Die Originalität und Bedeutung der in der Arbeit behandelten Fragestellung sowie das Niveau der statistischen Analyse sind maßgebliche Kriterien der Bewertung.

Eingereicht werden können alle Studienabschlussarbeiten (Diplomarbeit, Magisterarbeit, B. A.-Abschlussarbeit oder M. A.-Abschlussarbeit) und Dissertationen in deutscher oder englischer Sprache, die in den letzten vier Jahren fertiggestellt wurden. Absolventen können sich selbst bewerben oder von Hochschullehrern vorgeschlagen werden.

Einzureichen sind:

- die Qualifikationsarbeit, in Papierform sowie als pdf-Datei
- eine maximal zweiseitige Kurzfassung der Arbeit
- ein Lebenslauf der/des Autorin/s
- eine Bestätigung des betreuenden Hochschullehrers/der betreuenden Hochschullehrerin oder der Hochschule,

dass es sich bei der eingereichten Arbeit um eine Abschlussarbeit eines akademischen Studiums oder um eine Dissertation handelt, die *nach dem 31.12.2005 eingereicht* wurde. Diese Bestätigung soll auch Angaben zu den für entsprechende Abschlussarbeiten geltenden Vorgaben enthalten, insbesondere solche zu Umfang und Bearbeitungsdauer.

Einsendungen bitte bis spätestens zum **28.02.2010** an:

GESIS – Leibniz-Institut für Sozialwissenschaften
ALLBUS-Nachwuchspreis
Postfach 12 21 55
68072 Mannheim

Autorinnen und Autoren Jahrgang 3 (2009)

- Katrin Auspurg, Konstanz
- Tobias Bachteler, Duisburg
- Mandy Beuer-Krüssel, Leipzig
- Annelies G. Blom, Mannheim
- Olaf Bock, Hamburg
- Gregor Daschmann, Mainz
- Karsten Hank, Mannheim
- Thomas Hinz, Konstanz
- Hendrik Jürges, Mannheim
- Joost Kappelhof, Den Haag
- Achim Koch, Mannheim
- Lars Eric Kroll, Berlin
- Ivar Krumpal, Leipzig
- Thomas Lampert, Berlin
- Stefan Liebig, Bielefeld
- Rainer Lüdtke, Essen
- Jan Marcus, Konstanz
- Thomas Ostermann, Witten
- Jörg Reiher, Duisburg
- Barbara Schaan, Mannheim
- Kai-Uwe Schnapp, Hamburg
- Anna Schnauber, Mainz
- Rainer Schnell, Duisburg
- Sven Stadtmüller, Frankfurt
- Ineke Stoop, Den Haag

Rezensentinnen und Rezensenten Jahrgang 3 (2009)

- Tilo Beckers, Düsseldorf
- Henning Best, Mannheim
- Claudia Buchheister, Potsdam
- Marc Deutschmann, Offenbach am Main
- Siegfried Gabler, Mannheim
- Michael Häder, Dresden
- Sabine Häder, Mannheim
- Sigrid Haunberger, Bern
- Gerd Meier, Lüneburg
- Susanne Rippl, Chemnitz
- Christian Seipl, Hildesheim
- Martin Weichbold, Salzburg

Gutachterinnen und Gutachter Jahrgang 3 (2009)

Wir danken folgenden Kolleginnen und Kollegen, die für den Jahrgang 3 (2009) der Methoden — Daten — Analysen Manuskripte begutachtet haben:

- Rolf Becker, Bern
- Christian Geiser, Berlin
- Joachim Gerich, Linz
- Josef Hartmann, München
- Jürgen Hoffmeyer-Zlotnik, Mannheim
- Sascha Huber, Mannheim
- Markus Klein, Hannover
- Thomas Klein, Heidelberg
- Michael Knigge, Berlin
- Ulrich Kohler, Berlin
- Frauke Kreuter, Maryland
- Oliver Lüdtke, Tübingen
- Andreas Mielck, München
- Karsten Müller, Mannheim
- Dieter Ohr, Berlin
- Andreas Pöge, Bielefeld
- Rolf Porst, Mannheim
- Peter Preisendöfer, Mainz
- Andreas Quatember, Linz
- Uli Rendtel, Berlin
- Jürgen Schupp, Berlin
- Nobert Schwarz, Michigan
- Tom W. Smith, Chicago
- Stefan Spitz, München
- Peter M. Steiner, Wien
- Oliver Walter, Kiel
- Cornelia Weins, Trier

Hinweise für unsere Autorinnen und Autoren

Methoden – Daten – Analysen (MDA) veröffentlicht Beiträge aus dem Bereich der Empirischen Sozialforschung, insbesondere aus dem Bereich der Umfragemethodik. Im Vordergrund stehen Artikel, welche die methodischen und/oder statistischen Kenntnisse der Profession erweitern, sowie Beiträge, die sich mit der Anwendung der Methoden der Empirischen Sozialforschung in der Forschungspraxis beschäftigen, oder solche, in denen ein statistisches Verfahren exemplarisch angewandt wird. Obwohl der Schwerpunkt auf Umfragemethoden liegt, sind Beiträge zu anderen methodischen Bereichen willkommen. Die Artikel sollen für eine breite Leserschaft von Wissenschaftlern und Praktikern im Bereich der Empirischen Sozialforschung verständlich sein.

Manuskripte, die bereits an anderer Stelle veröffentlicht sind oder gleichzeitig anderen Publikationsorganen zur Veröffentlichung angeboten worden sind, werden grundsätzlich nicht berücksichtigt. Eine spätere Veröffentlichung eines in der MDA erschienenen Beitrages ist möglich, sofern an exponierter Stelle auf die Ersterscheinung des Beitrages in der MDA hingewiesen wird.

Jeder Beitrag, der zur Veröffentlichung in MDA eingereicht wird, wird zunächst von den Herausgebern danach bewertet, ob er für eine Veröffentlichung grundsätzlich in Frage kommt.

Falls die Herausgeber einer Veröffentlichung grundsätzlich ablehnend gegenüber stehen, werden die Autoren unter Angabe von Gründen für diese Entscheidung informiert.

Falls die Herausgeber zur Ansicht gelangen, dass der Beitrag grundsätzlich zur Veröffentlichung in Frage kommt, wird er anonymisiert an mindestens zwei unabhängige Gutachter verschickt, die um eine Stellungnahme gebeten werden. Im Zweifelsfalle wird ein drittes Gutachten eingeholt.

Wird ein Beitrag nach Beschluss der Herausgeber in das Begutachtungsverfahren gegeben, erfolgt die abschließende Entscheidung über ein Manuskript auf der Basis der Gutachten durch die Herausgeber. Im Falle einer Ablehnung erhalten die Autoren eine ausführliche Begründung für die Ablehnung. Wird eine Überarbeitung eines Beitrages für erforderlich gehalten, erhalten die Autoren detaillierte Überarbeitungshinweise.

Unabhängig vom Ergebnis des Begutachtungsverfahrens werden die Autoren von der Entscheidung durch die Redaktion per E-Mail informiert.

Die folgenden Regeln sind bei der Abfassung von Manuskripten zu beachten:

Manuskripte müssen per E-Mail (mda@gesis.org) eingereicht werden. Der Umfang der Manuskripte soll inklusive Leerzeichen alles in allem nicht mehr als 70.000 Zeichen betragen.

Den Beiträgen sind Abstracts in Deutsch und Englisch (jeweils ca. 15 Zeilen) voranzustellen. Auch der Titel des Beitrages ist in Deutsch und Englisch einzureichen.

Um die Anonymität der Beiträge zu wahren, darf in einem Manuskript nur der Titel des Beitrages enthalten sein, nicht aber Namen oder Anschriften der Autoren; Name und Anschrift der Autoren müssen, gemeinsam mit dem Titel des Beitrages, auf einer separaten Seite eingereicht werden.

Beiträge sind mit dem Dezimalklassifikationssystem zu untergliedern (1 - 2 - 2.1 - 2.2 - 3 usw.). Die Gliederungstiefe geht dabei höchstens auf *eine* Stelle nach dem Punkt.

Tabellen enthalten Tabellennummer und Titel im Tabellenkopf, Abbildungen werden analog behandelt.

Grafiken sind mittels gängiger Grafiksoftware zu erstellen. Ist eine spezielle Grafiksoftware erforderlich, übernimmt der Autor/die Autorin die endgültige Formatierung der Grafiken in eigener Regie.

Bei der Erstellung von Tabellen und Grafiken ist zu berücksichtigen, dass der Satzspiegel 11,5 cm (Breite) x 18,5 cm (Höhe) beträgt. Die Grafiken sind als jpeg- oder tif-Dateien in *Graustufen (CMYK)* mit einer Auflösung von mindestens 300 dpi zu liefern.

Die Beiträge sind unter Wahrung der gültigen Rechtschreiberegelungen (neue Rechtschreibung) zu erstellen.

Werden in einem Beitrag empirische Daten verwandt, muss die Möglichkeit der Replikation bestehen. Im Falle einer Veröffentlichung in der MDA erklären sich die Autoren daher schriftlich bereit, Dritten auf deren Anfrage hin die Daten und Programmroutinen zur Verfügung zu stellen.

Anmerkungen und Fußnoten sind mit der Fußnotenfunktion des Schreibprogrammes (im Normalfalle Word) zu erstellen; bitte nicht gesondert formatieren. Fußnoten sind nur für inhaltliche Kommentare vorzusehen, nicht für bibliographische Hinweise.

Literaturhinweise im Text sind nach den folgenden Mustern aufzuführen: Müller (2002) – Schulze und Mayer (2003) – Müller, Mayer und Schulze (2004) – Müller et al. (2005) – Müller (2006: 75) – (vgl. Müller 2007: 75) – (Müller 2008; Mayer/Müller/Schulze 2009).

Das Literaturverzeichnis ist wie folgt zu gestalten:

Buchveröffentlichungen:

Strobl, R. und W. Kühnel, 2000: Dazugehörig und ausgegrenzt. Analysen zu Integrationschancen junger Aussiedler. Weinheim/München: Juventa.

Zeitschriftenbeiträge:

Becker, R., R. Imhof und G. Mehlkop, 2007: Die Wirkung monetärer Anreize auf den Rücklauf bei einer postalischen Befragung und die Antworten auf Fragen zur Delinquenz. Empirische Befunde eines Methodenexperiments. Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung 1 (2): 131-159.

Beiträge in Büchern:

Braun, M. und I. Borg, 2004: Berufswerte im zeitlichen und im Ost-West-Vergleich. S. 179-199 in: R. Schmitt-Beck, M. Wasmer und A. Koch (Hg.): Sozialer und politischer Wandel in Deutschland. Analysen mit ALLBUS-Daten aus zwei Jahrzehnten. Wiesbaden: VS-Verlag für Sozialwissenschaften.

Internetquellen:

Stadtmüller, S. und R. Porst, 2005: Zum Einsatz von Incentives bei postalischen Befragungen. GESIS How-to-Reihe, Nr. 14. Mannheim: GESIS. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/how-to14rp.pdf (1.12.2008).

ISSN 1864-6956

3. Jahrgang 2009