methoden daten analysen
ZEITSCHRIFT FÜR EMPIRISCHE SOZIALFORSCHUNG

mda

2009, Jahrgang 3, Heft 1

Lars Eric Kroll und Thomas Lampert — Soziale Unterschiede in der Lebenserwartung. Datenquellen in Deutschland und Analysemöglichkeiten des SOEP

Mandy Beuer-Krüssel und Ivar Krumpal — Der Einfluss von Häufigkeitsformaten auf die Messung von subjektiven Wahrscheinlichkeiten

Katrin Auspurg, Thomas Hinz und Stefan Liebig — Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey

Karsten Hank, Hendrik Jürges und Barbara Schaan — Die Erhebung biometrischer Daten im Survey of Health, Ageing and Retirement in Europe

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

FORSCHUNGSBERICHTE

3 Soziale Unterschiede in der Lebenserwartung. Datenquellen in Deutschland und Analysemöglichkeiten des SOEP
Lars Eric Kroll und Thomas Lampert

31 Der Einfluss von Häufigkeitsformaten auf die Messung von subjektiven Wahrscheinlichkeiten
Mandy Beuer-Krüssel und Ivar Krumpal

59 Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey
Katrin Auspurg, Thomas Hinz und Stefan Liebig

PRAXISBERICHTE

97 Die Erhebung biometrischer Daten im Survey of Health, Ageing and Retirement in Europe. Befunde und Perspektiven
Karsten Hank, Hendrik Jürges und Barbara Schaan

REZENSIONEN

109 International Handbook of Survey Methodology. E. D. de Leeuw, J. J. Hox und D. A. Dillman, 2008
Sigrid Haunberger

112 Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. Frank Faulbaum, Peter Prüfer und Margrit Rexroth, 2009
Michael Häder

114 Advances in Telephone Survey Methodology. J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link und R. L. Sangster, 2007
Sabine Häder und Siegfried Gabler

116 Telefonbefragungen über das Mobilfunknetz: Konzept, Design und Umsetzung einer Strategie zur Datenerhebung. Michael Häder und Sabine Häder, 2009
Gerd Meier

120 International vergleichende Sozialforschung. Ansätze und Messkonzepte unter den Bedingungen der Globalisierung. Birgit Pfau-Effinger, Slađana Sakač Magdalenić, Christof Wolf, 2009
Susanne Rippl und Christian Seipel

ANKÜNDIGUNGEN

123 6. Nutzerkonferenz ‚Forschung mit dem Mikrozensus' Analysen zur Sozialstruktur und zum sozialen Wandel. 15.-16. Oktober 2009 - Programm

126 XVII ISA World Congress of Sociology. Research Committee on Logic and Methodology RC33, 11.-17. July 2010, Gothenburg, Sweden - Call for Papers

129 Hinweise für unsere Autorinnen und Autoren

Soziale Unterschiede in der Lebenserwartung

Datenquellen in Deutschland und Analysemöglichkeiten des SOEP

Socio-Economic Differences in Life Expectancy

Data Sources in Germany and the Potential of the German Socio-Economic Panel Study

Lars Eric Kroll und Thomas Lampert

Zusammenfassung
In diesem Beitrag werden Möglichkeiten zur Analyse sozialer Unterschiede in der Lebenserwartung auf Basis des Sozio-oekonomischen Panels (SOEP) dargestellt. Einleitend wird ein Überblick über verschiedene Datenquellen zur Analyse sozial differenzieller Sterblichkeit gegeben und der Forschungsstand für Deutschland zusammengefasst. Anschließend wird auf methodische Besonderheiten und Probleme des SOEP hingewiesen, die sich insbesondere an einer Überschätzung der mittleren Lebenserwartung festmachen lassen. Abschließend wird ein Verfahren vorgestellt, anhand dessen sich diese Überschätzung unter Verwendung der amtlichen Sterbetafeln ausgleichen lässt.

Abstract
This paper describes the potential for analysing socio-economic inequalities in life expectancy with the German Socio-Economic Panel Study (SOEP). It includes a short review of available data sources and recent evidence on inequalities in life expectancy for Germany. Subsequently the characteristics and issues with analysing mortality with the SOEP are discussed. The main problem for providing undistorted estimates of life expectancy across social strata with SOEP data is an overestimation of life expectancy in the study population. We will present a method that allows overcoming this problem with the utilization of official life tables for Germany.

1 Einleitung

Obwohl die Lebenserwartung weltweit sukzessive steigt[1], zeigen verfügbare Daten für nahezu alle Länder soziale Unterschiede zwischen Bildungs-, Einkommens- oder Berufsgruppen (Kunst et al. 1998; Mackenbach et al. 1997). Sie treten bei allen vorherrschenden Todesursachen zutage und sind bei der Sterblichkeit infolge von Herz-Kreislauf-Erkrankungen und Lungenkrebs besonders stark ausgeprägt (Mackenbach 2006). Für viele Länder, wie z. B. Schweden, England oder Italien, lässt sich darüber hinaus zeigen, dass sich die sozialen Unterschiede im Verlauf der 1980er und 1990er Jahre ausgeweitet haben (Mackenbach et al. 2003).

Deutschland fehlt zumeist in international vergleichenden Studien zum Ausmaß sozialer Unterschiede in der Lebenserwartung. Die verfügbare Datenbasis ist hierzulande deutlich eingeschränkter als z. B. in Schweden, England oder Finnland. So lassen sich auf Basis der amtlichen Statistik keine Aussagen zu sozialen Unterschieden in der Lebenserwartung treffen. Eine wichtige Datenquelle, die regelmäßig aktualisierte Analysen zu sozialen Unterschieden ermöglicht, ist das Sozio-oekonomische Panel (SOEP)[2]. Die Eignung des SOEP für die Berechnung der Lebenserwartung wurde allerdings zuletzt in Frage gestellt (Schnell/Trappmann 2006). Als Problem wird vor allem angesehen, dass die Studienteilnehmer im Vergleich zur Grundgesamtheit eine deutlich erhöhte Lebenserwartung aufweisen.

In der vorliegenden Studie wird ein Verfahren vorgestellt, anhand dessen sich die Überschätzung der Lebenserwartung im SOEP ausgleichen lässt. Anhand aktueller Daten für den Zeitraum 1995 bis 2005 wird das Ausmaß der Überschätzung der Lebenserwartung beschrieben und deren Einfluss auf die ausgewiesenen sozialen Unterschiede quantifiziert. Auf diese Weise soll sichergestellt werden, dass künftig auch für Deutschland Ergebnisse zum Ausmaß von Differenzen in der Lebenserwartung ausgewiesen werden können, die sich mit anderen europäischen Ländern vergleichen lassen.

1 Für Deutschland – wie für die anderen Industrieländer – ist seit der zweiten Hälfte des 19. Jahrhunderts ein sukzessiver Anstieg der mittleren Lebenserwartung festzustellen. In den Jahren 1871/80 erreichte nur etwa ein Drittel der Bevölkerung das 60. Lebensjahr, um 1950 traf dies bereits auf über 75 % und am Ende des Jahrtausends sogar auf annähernd 90 % zu (Destatis 2006). Die Entwicklung in den letzten Jahren deutet darauf hin, dass auch künftig mit einem Zugewinn an Lebenszeit zu rechnen ist (Riley 2001; Oeppen/Vaupel 2002).

2 Das SOEP steht interessierten Wissenschaftlern als Scientific-Use-File zur Verfügung (Wagner et al. 2007).

2 Datenquellen zur Analyse von Mortalität in Deutschland

In vielen Ländern werden Daten zur Analyse sozial differenzieller Mortalität durch die Verknüpfung von Informationen aus einem nationalen Zensus[3] mit Bevölkerungsregistern[4] gewonnen (Mackenbach et al. 2003). So wird in Dänemark, Finnland, Norwegen, Schweden und England regelmäßig ein Zensus durchgeführt. In den skandinavischen Ländern werden dabei Informationen aus staatlichen Registern für Sozialleistungen oder Steuererhebungen zusammengetragen und mit den Melderegistern abgeglichen, ohne dass eine gesonderte Erhebung nötig ist. In anderen Ländern wie England, in denen die vorhandenen Datenbestände weniger Informationen enthalten, werden große repräsentative Bevölkerungsstichproben befragt. Die Befragten werden anschließend im Zuge eines Mortalitäts-follow-up (Dauer in der Regel fünf Jahre) anhand von Bevölkerungsregistern weiter verfolgt. Die Verknüpfung von sehr großen Stichproben bzw. Vollerhebungen mit nationalen Bevölkerungsregistern stellt valide Informationen zu sozialen Unterschieden in der Sterblichkeit bereit und minimiert die statistischen Unsicherheiten. Dadurch wird auch die Analyse von sozialen Determinanten auf der Ebene einzelner Todesursachen ermöglicht.[5]

In Deutschland erschwert die kommunale Organisation des Meldewesens bisher ein Mortalitäts-follow-up auf Basis amtlicher oder wissenschaftlicher Erhebungen. So gibt es heute über 5.000 verschiedene kommunale Melderegister. Der Verbleib von Untersuchungspersonen muss aufwendig im Zuge einzelner Anfragen an die jeweils zuständigen Meldebehörden geklärt werden.[6] Die Einrichtung eines einheitlichen Bundesmelderegisters ist erst bis zum Jahr 2010 geplant. Damit wird sich der Zugang zu Meldedaten verbessern und der logistische Aufwand entsprechender Recherchen zukünftig verringern. Eine Alternative zu einem Mortalitäts-follow-up wären Angaben zum sozioökonomischen Status, die direkt auf den

3 Ein nationaler Zensus ist eine große Erhebung, bei der Angaben zu sozialstatistisch relevanten Merkmalen erhoben werden. Dabei kann entweder die gesamte Bevölkerung erfasst (Volkszählung), oder nur eine hinreichend große Stichprobe gezogen werden (Mikrozensus als 1 %-Stichprobe der Bevölkerung).

4 In zentralen Bevölkerungsregistern werden die Geburt, der aktuelle Wohnort bzw. der Todeszeitpunkt der Bürger erfasst.

5 Die Differenzierung nach Todesursachen ermöglicht ein tieferes Verständnis von kausalen Mechanismen. So konnte beispielsweise eine in Finnland beobachtete Ausweitung sozialer Unterschiede in der Lebenserwartung auf Todesfälle bei Herz-Kreislauf-Erkrankungen, sowie auf durch Alkohol bedingte Todesfälle und Suizide zurückgeführt werden (Martikainen et al. 2001).

6 In der letzten Verbleibstudie für ehemalige Teilnehmer des SOEP mussten bis zu fünf verschiedene Ämter kontaktiert werden, bis der Wohnort und Vitalstatus eines Befragten endgültig ermittelt werden konnte (Infratest 2002).

Totenscheinen vermerkt werden.[7] Das ‚Gesetz über die Statistik der Bevölkerungsbewegung und die Fortschreibung des Bevölkerungsstandes' (Bevölkerungsstatistikgesetz) bildet die Grundlage für die auf Totenscheinen erhobenen Merkmale. Bis zur Neufassung des Gesetzes im Jahr 1971 wurde das Merkmal ‚Beruf des Verstorbenen' auf den Totenscheinen erfasst. Für den Zeitraum 1937 bis 1971 standen damit Angaben zur berufsspezifischen Mortalität zur Verfügung (Mielck 2000). Aktuell werden auf den Todesscheinen aber lediglich folgende Merkmale erfasst (BevStatG §2 Abs. 1):

- Sterbetag[8], Geschlecht, Alter, Familienstand – bei Kindern Angabe über Ehelichkeit oder Nichtehelichkeit – und Wohngemeinde,
- Erwerbstätigkeit, rechtliche Zugehörigkeit oder Nichtzugehörigkeit zu einer Kirche, Religionsgesellschaft oder Weltanschauungsgemeinschaft und Staatsangehörigkeit,
- Bei Verheirateten: Alter des überlebenden Ehegatten,
- Todesursache.

Nach heutigem Rechtsstand ist in Deutschland weder ein Mortalitäts-follow-up auf Basis amtlicher Erhebungen möglich, noch die Erfassung sozioökonomischer Merkmale auf den Totenscheinen vorgesehen. Für die Zukunft steht damit – vorbehaltlich gesetzlicher Änderungen – nicht zu erwarten, dass die amtliche Statistik Informationen zur sozial differenziellen Mortalität in Deutschland zur Verfügung stellen kann. Die Forschung ist somit auf alternative Datenquellen angewiesen. Möglichkeiten zur Analyse sozialer Unterschiede im Mortalitätsrisiko bestehen anhand von Daten der Krankenkassen, der Rentenversicherung und wissenschaftlicher Surveys. Wissenschaftliche Surveys, die bereits häufiger für Analysen zur Lebenserwartung herangezogen wurden, sind das Sozio-oekonomische Panel (SOEP), die MONICA Kohortenstudie und der Lebenserwartungssurvey (LES). Nachfolgend werden die Eigenschaften der alternativen Datenquellen beschrieben, bevor in den nächsten Abschnitten auf das SOEP eingegangen wird.

7 Eine solche Praxis ist in England schon seit 1851 etabliert (Pamuk 1985), die Qualität dieser Informationen muss aber kritisch beurteilt werden, weil unklar ist aus welcher Quelle die vermerkten Informationen stammen.

8 Bei Sterbefällen innerhalb der ersten vierundzwanzig Lebensstunden Sterbedatum und Lebensdauer.

2.1 Daten der gesetzlichen Sozialversicherungen

Sowohl die gesetzlichen Renten- als auch die Krankenversicherungsträger verfügen über Daten zu Überlebensraten und zum Einkommen ihrer Versicherten, weil diese für die Berechnung von Beiträgen bzw. Anwartschaften eine wichtige Voraussetzung sind. Daneben erheben sie routinemäßig weitere Hintergrundinformationen zu Qualifikation und Beschäftigung (Himmelreicher et al. 2006; Voges et al. 2004; Badura et al. 2007). Durch die Einrichtung von Forschungsdatenzentren werden diese Daten nun auch der wissenschaftlichen Forschung zur Verfügung gestellt. Allerdings bestehen angesichts der Sensibilität dieser Informationen zumeist größere bürokratische oder technische Hürden, so dass nur wenige Forscher oder Institutionen tatsächlich Zugriff auf die Daten haben. Für die Daten der Sozialversicherungen sprechen aber insbesondere ihr großer Stichprobenumfang und die Verlässlichkeit der Informationen.

Für die Rentnerinnen und Rentner stellt das Forschungsdatenzentrum der gesetzlichen Rentenversicherung (FDZ-RV) verschiedene Datensätze auf Basis der Versicherten bereit, die Analysen zur Mortalität ermöglichen.[9] Bei den Daten der Rentenversicherung handelt es sich um eine Vollerhebung aller ehemals abhängig beschäftigten Rentnerinnen und Rentner. Es können aber erst Todesfälle nach dem Renteneintritt untersucht werden. Informationen zur vorzeitigen Mortalität, die bei sozial Benachteiligten besonders stark ausgeprägt ist, stehen nicht zur Verfügung. Weil die erhobenen Hintergrundmerkmale nicht zur Durchführung der Routineaufgaben benötigt werden, weisen diese häufig Lücken auf. Anhand der Daten können daher vor allem Unterschiede in der Lebenserwartung nach Rentenanwartschaften (Entgeltpunkte[10]) untersucht werden. Diese sollten allerdings nur für männliche Versicherte ausgewertet werden. Die weiblichen Erwerbsbiographien der heutigen Rentnerinnen sind aufgrund der vorherrschenden Orientierung am Modell des männlichen Alleinverdieners bisher zu lückenhaft, um Rückschlüsse von ihren Rentenanwartschaften auf ihre Lebenssituation zu erlauben. Bis zur Aufhebung der

9 Informationen zum FDZ-RV unter http://fdz.deutsche-rentenversicherung.de.

10 Erwerbseinkommensbezogener Faktor zur Ermittlung der Rentenhöhe in den je nach aktuellen rentenrechtlichen Regelungen auch weitere Faktoren wie Erziehungs- und Ausbildungszeiten eingehen. Er kann als Maß für das Lebenseinkommen angesehen werden, ist jedoch an den Rändern der Verteilung nur bedingt aussagekräftig (Himmelreicher et al. 2006: 4): „Wegen der Beitragsbemessungsgrenze unterliegen die beobachteten Entgeltpunkte und Rentenzahlbeträge einer Rechtszensierung, weil über dieser liegende Arbeitseinkommen die Ansprüche an die Rentenversicherung nicht erhöhen. Zudem besteht eine Linkszensierung des Lebensarbeitseinkommens, weil die Entgeltpunkte durch umverteilende Maßnahmen in der Rentenversicherung zum Teil erhöht wurden, wie z. B. im Fall einer Rentenerhöhung wegen ‚Rente nach Mindesteinkommen', die bis 1992 bezogen werden konnte [...]".

institutionellen Trennung von Arbeiter- und Angestelltenrentenversicherung waren außerdem Vergleiche zwischen beiden Versichertengruppen möglich.

Die Daten der gesetzlichen Rentenversicherung zeigen deutliche Unterschiede in der Lebenserwartung von Angestellten und Arbeitern sowie nach Entgeltpunkten (Hoffmann et al. 2006; Himmelreicher et al. 2006). Für den Zeitraum 2002/04 betrug die fernere Lebenserwartung im Alter von 65 Jahren im Vergleich von Arbeitern und Angestellten 15,4 und 17,5 Jahre bei Männern bzw. 19,5 und 20,8 Jahre bei Frauen (Hoffmann et al. 2006). Unterschiede nach Entgeltpunkten können im Rahmen des Arbeitsfiles ‚Differentielle Sterblichkeit 2003'[11] für fast 4 Millionen Versicherte des Jahres 2003 analysiert werden. Die fernere Lebenserwartung von 65-jährigen Männern unterscheidet sich demnach um etwa fünf Jahre (14 bzw. 19 Jahre) zwischen der niedrigsten und der höchsten von zehn Einkommenskategorien (Himmelreicher et al. 2006).

Auch auf Basis der Daten der Krankenkassen sind Analysen zur Mortalität möglich. In Deutschland gibt es mehr als 200 gesetzliche Krankenkassen. Bisher konnten nur die Daten weniger Krankenkassen für die wissenschaftliche Forschung verwendet werden, z. B. die der Allgemeinen Ortskrankenkasse Mettmann und der Gmünder Ersatzkasse (Helmert 2000; Geyer/Peter 2000). Ein gemeinsamer Pool der Daten aller gesetzlichen (und privaten) Krankenversicherungen existiert bisher nicht. Probleme bei bevölkerungsbezogenen Analysen von Daten der gesetzlichen Krankenversicherungsdaten gibt es im Hinblick auf die Repräsentativität der jeweiligen Versichertengemeinschaft sowie die Qualität und Verfügbarkeit sozialepidemiologisch relevanter Hintergrundmerkmale. Seit der Liberalisierung des Marktes der Krankenversicherungen wechseln die Bürger immer häufiger ihre Versicherungen, um möglichst günstige Beiträge zu zahlen.[12] Eine längsschnittliche Weiterverfolgung der Versicherten wird dadurch erschwert. Für die Analyse sozialer Unterschiede stehen zumeist das beitragspflichtige Einkommen und berufsbezogene Merkmale wie die Stellung im Beruf oder der Bildungsabschluss zur Verfügung. Informationen, die nicht zur Berechnung der Beiträge herangezogen werden, können allerdings fehlende Werte aufweisen oder veraltet sein.

11 Diese Daten können nur bei Aufenthalten im Forschungsdatenzentrum der Rentenversicherung analysiert werden.

12 Im Zuge des Gesundheitsstrukturgesetzes (Gesetz zur Sicherung und Strukturverbesserung der gesetzlichen Krankenversicherung) vom 21. Dezember 1992 wurde die Einführung der freien Kassenwahl zum 1. Januar 1996 beschlossen. Seither können grundsätzlich alle Versicherten der gesetzlichen Krankenversicherung ihre Krankenkasse frei wählen. Im Zuge der sukzessiven Erhöhung der Beiträge seit Ende der 1990er machen die Versicherten zunehmend von dieser Möglichkeit Gebrauch (Greß 2002).

Für Mitglieder der Allgemeinen Ortskrankenkasse (AOK) liegen seit 1989 aus dem Kreis Mettmann Daten zur sozial differenziellen Sterblichkeit vor, die bereits mehrfach für wissenschaftliche Studien genutzt wurden (Geyer/Peter 1999, 2000; Gässler et al. 2005). Im Vergleich zur deutschen Bevölkerung sind die unteren Statusgruppen darin überrepräsentiert. So hatten im Jahr 1995 etwa 48 % der erwerbstätigen Versicherten, aber lediglich 27 % der Erwerbsbevölkerung die berufliche Stellung Arbeiter. Der Anteil von Versicherten mit sehr hohem Berufsstatus liegt mit 0,2 % dagegen deutlich unter ihrem Bevölkerungsanteil von 0,5 %. Zudem gibt es einen hohen Anteil fehlender Werte bei den Statusindikatoren Bildung, Beruf und Einkommen (Geyer/Peter 2000). Veränderungen im Berufsstatus werden nur nach einem Wechsel des Arbeitgebers an die Krankenkasse gemeldet und können dadurch veraltet sein (Geyer/Peter 1999). Von Januar 1987 bis Dezember 1996 wurden 112.338 abhängig beschäftigte Männer und Frauen im Alter zwischen 30 und 70 Jahren für insgesamt 743.288 Personenjahre beobachtet. Die Ergebnisse zeigen für Männer und Frauen mit niedrigem beruflichen Status ein gegenüber Versicherten mit hohem Status um das 4,3- bzw. 3,8-Fache erhöhtes Mortalitätsrisiko. Es wurden auch weiterführende Analysen zum Vergleich der Bedeutung von Bildung, Beruf und Einkommen durchgeführt (Geyer/Peter 2000). Alle Indikatoren hatten einen starken Einfluss, in der multivariaten Betrachtung erwies sich aber nur der Einkommenseffekt als statistisch signifikant.

Die Gmünder Ersatzkasse (GEK) stellt seit 1987 anonymisierte Daten für die wissenschaftliche Forschung bereit. Die Versicherten der Gmünder Ersatzkasse haben im Vergleich zur Bevölkerung ebenfalls einen unterdurchschnittlichen Sozialstatus. Im Jahr 1998 hatten nur etwa 9,9 % der GEK Versicherten einen höheren beruflichen Status, während der Anteil in der Bevölkerung mit 20,4 % mehr als doppelt so hoch war. Männer und Frauen mit gering qualifizierten manuellen Berufen waren in der GEK dagegen mit 20,8 % bzw. 13,1 % gegenüber 12,6 % und 6,0 % in der Bevölkerung deutlich überrepräsentiert. Bis April 2004 wurden über 2,8 Millionen Versicherte erfasst (Helmert et al. 2002; Voges et al. 2004; Timm et al. 2006). Analysen zur Mortalität wurden unter Versicherten im Alter zwischen 40 und 69 Jahren für den Zeitraum 1990 bis 2004 durchgeführt (Voges et al. 2004). Für Männer und Frauen weisen die Ergebnisse deutliche Unterscheide nach Berufs- und Versichertenstatus (freiwillig vs. pflichtversichert) auf. Bildungsunterschiede sowie Unterschiede zwischen arbeitslosen und erwerbstätigen Mitgliedern zeigten sich nur bei Männern.

2.2 Wissenschaftliche Erhebungen mit Mortalitäts-follow-up

Im Rahmen wissenschaftlicher Surveys kann das Versterben der Befragten durch Informationen von anderen Haushaltsmitgliedern und über die Einwohnermeldeämter festgestellt werden (Infratest 2002). Dazu muss der Name, das Geburtsdatum und die Anschrift eines Befragten erfasst und gegenüber den Einwohnermeldeämtern ein berechtigtes Interesse nachgewiesen werden. Eine Kontaktaufnahme mit den Angehörigen ist dagegen sensibel, weil eine Güterabwägung zwischen wissenschaftlichem Erkenntnisinteresse und emotionalen Bedürfnissen der Angehörigen getroffen werden muss. Zudem lebt ein großer Teil der älteren Menschen in Ein-Personen-Haushalten, wodurch eventuell keine Angehörigen mehr zu befragen sind.

Anhand des Lebenserwartungssurveys (LES) des Bundesinstituts für Bevölkerungsforschung (BIB) liegen für Westdeutschland Ergebnisse zu sozialen Unterschieden in der Lebenserwartung für den Zeitraum 1984/86 bis 1998 vor (Helmert 2003; Gärtner 2002; Gärtner/Scholz 2005). An der Basisuntersuchung nahmen in Westdeutschland 8.474 Personen im Alter von 31-69 Jahren teil. In Ostdeutschland beteiligten sich 1991 mehr als 1.500 Personen im Alter zwischen 40 und 79 Jahren. Der Vitalstatus wurde anhand von Anfragen bei den zuständigen Meldeämtern im Jahr 1998 ermittelt. In den alten Bundesländern lagen für 86,5 % der Teilnehmerinnen und Teilnehmer Angaben vor, demnach sind zwischen 1984/86 und 1998 17,2 % der Männer und 8,6 % der Frauen verstorben (Helmert 2003).

Die Ergebnisse auf Basis des Lebenserwartungssurveys weisen deutliche Unterschiede nach Bildungsabschluss und Sozialstatus aus (Gärtner 2002; Luy 2006). Der Anteil verstorbener Teilnehmer lag unter Befragten mit einem Hauptschulabschluss fast doppelt so hoch wie unter Abiturienten. So starben im Zeitraum 1984/86 unter den 60- bis 69-jährigen Männern 38,7 % der Hauptschüler, aber nur 26,6 % der Abiturienten. Unter Frauen lagen die entsprechenden Anteile bei 20,8 % und 12,1 %. Die fernere Lebenserwartung im Alter von 45 Jahren lag unter männlichen Hauptschülern bei ca. 27 Jahren, Abiturienten konnten dagegen 32 weitere Lebensjahre erwarten (Luy 2006). Bei Frauen betrugen die entsprechenden Werte 36 und 38 Jahre. Weiterführende Analysen wiesen deutliche Berufsstatus- und Einkommensunterschiede aus. So war die fernere Lebenserwartung von Angestellten gegenüber Arbeitern um etwa 4,2 Jahre bei Männern und Frauen erhöht. Hinsichtlich des Haushaltsnettoeinkommens (<2.000 DM vs. >3.000 DM im Monat) werden Unterschiede von 5,5 und 3,6 Jahren beim Vergleich von Männern und Frauen berichtet. Anhand des Lebenserwartungssurveys wurden auch Analysen zur Entwicklung des Mortalitätsrisikos nach beruflicher Qualifikation durchgeführt (Helmert 2003). Im Vergleich der sich überlappenden Zeiträume 1986 – 1992 und

1986 – 1998 wurde für westdeutsche Männer mit einer niedrigen Berufsausbildung gegenüber der Vergleichsgruppe mit hoher Bildung im Zeitraum ein Anstieg des relativen Mortalitätsrisikos vom 1,8-Fachen auf das 2,4-Fache beobachtet.

Das internationale Projekt ‚Monitoring Trends and Determinants in Cardiovascular Disease' (MONICA) wurde in den 1980er Jahren durch die Weltgesundheitsorganisation (WHO) initiiert (Bothig 1989; Keil 2005). Ziel war eine Erfassung von Entwicklung und Determinanten der Herz-Kreislauf-Mortalität und Morbidität. Für die Studie standen insgesamt 13 Millionen Menschen in 21 Ländern über 10 Jahre lang unter Beobachtung. Im Zuge des deutschen MONICA Teilprojektes wurde in der Region Augsburg (Stadt und Landkreis) eine Kohortenstudie durchgeführt. Die Grundgesamtheit umfasste alle Einwohner der Region, die im Jahr 1984 zwischen 25 und 64 Jahre alt waren (Klein et al. 2001; Schneider 2007). Zum Erhebungsprogramm der Surveys gehörten u. a. kardiovaskuläre Erkrankungen und Risikofaktoren sowie soziodemografische Merkmale der Befragten. Anschließend wurde ein Mortalitäts-follow-up bis zum Jahr 1998 durchgeführt. Es zeigten sich ausgeprägte soziale Unterschiede im Mortalitätsrisiko (Klein et al. 2001; Schneider 2007). Bei Männern nahm das Mortalitätsrisiko mit einem höheren Bildungsabschluss ab. Der Zusammenhang zwischen Bildung und Mortalität wurde in der MONICA Studie insbesondere durch Einkommensunterschiede und Unterschiede im Gesundheitsverhalten im Vergleich der Bildungsgruppen moderiert. Bei Frauen zeigten sich keine Unterschiede nach Bildungsabschluss oder Einkommen. Als wichtigster verhaltensbezogener Prädiktor des Mortalitätsrisikos erwies sich bei Männern und Frauen das Rauchverhalten.

3 Datenbasis und Methode

Das Sozio-oekonomische Panel ist eine Haushaltsbefragung, die seit 1984 jährlich vom Deutschen Institut für Wirtschaftsforschung (DIW) durchgeführt wird (Wagner et al. 2007). Hauptanliegen der Studie ist eine zeitnahe Erfassung des politischen und gesellschaftlichen Wandels in Deutschland. Das Stichprobendesign ist so gewählt, dass sowohl repräsentative Aussagen im Querschnitt getroffen werden können als auch eine längsschnittliche Weiterverfolgung der Studienteilnehmer möglich ist. Stichprobenausfällen, z. B. aufgrund von Teilnahmeverweigerungen, Umzügen ins Ausland oder Todesfällen, wird durch Hochrechnungsfaktoren und die Ziehung von Ergänzungsstichproben begegnet. In der Vergangenheit wurden mehrere Zusatzstichproben gezogen, um bestimmte Bevölkerungsgruppen angemessen zu repräsentieren und aktuellen gesellschaftlichen Ereignissen und Entwicklungen gerecht werden zu können.

Eine sorgfältige Nacherfassung bei Nichterreichbarkeit der Studienteilnehmer ermöglicht, dass auch Todesfälle unter den Teilnehmern erfasst werden (Infratest 2002).

Im folgenden Abschnitt werden die Datenbasis und das methodische Vorgehen bei der Analyse sozialer Unterschiede in der Lebenserwartung beschrieben. Es wird ein Verfahren vorgestellt, das die Überschätzung der Lebenserwartung im SOEP kompensieren soll. Ziel ist es, unverzerrte Lebenserwartungen für sozioökonomische Statusgruppen ausweisen zu können und das Ausmaß der Verzerrung sozialer Differenzen zu beschreiben. Die Datenanalyse erfolgte anhand des Programms Stata in der Version 10 (StataCorp 2007).

3.1 Sozio-oekonomisches Panel

Das SOEP besteht aus mittlerweile 23 abgeschlossenen Befragungswellen (1984-2006), in denen mehr als 450.000 Interviews durchgeführt wurden. Es gibt acht unterschiedliche Teilstichproben (Pischner 2007): A *‚Deutsche Haushalte der Bundesrepublik Deutschland'* (1984), B *‚Ausländische Haushalte der Bundesrepublik Deutschland'* (1984)[13], C *‚Haushalte der DDR'* (1990)[14], D *‚Zuwandererhaushalte'* (1994/95)[15], E *‚Haushalte in Deutschland, Ergänzungsstichprobe'* (1998), F *‚Haushalte in Deutschland, Ergänzungsstichprobe'* (2000), G *‚Hocheinkommenshaushalte in Deutschland, Hocheinkommensstichprobe'*[16] (2002) und H *‚Haushalte in Deutschland, Ergänzungsstichprobe'* (2006).

Die Gewichtung berücksichtigt Ziehungsdesign, Bleibewahrscheinlichkeiten und im letzten Schritt bekannte Merkmale der Grundgesamtheit (Geschlecht, Alter etc.; Pischner 2007).[17] Die bereitgestellten, längsschnittlichen Gewichtungsfaktoren eignen sich allerdings nicht für Mortalitätsanalysen. Teilnehmer, die in ihrem Todesjahr nicht mehr befragt werden konnten (99 % der Verstorbenen) oder nicht an allen Befragungswellen teilgenommen haben, werden darin nicht berücksichtigt. Daher wurde – nach Absprache mit dem DIW – ein einfaches modifiziertes Anpassungsgewicht für Mortalitätsanalysen konstruiert.[18] Die Hochrechnung in Mortalitätsanalysen erfolgt für die Grundgesamtheit der deutschen Bevölkerung im Untersuchungszeitraum. Sie besteht aus zwei Schritten:

13 HH-Vorstand hat die türkische, italienische, spanische, griechische oder jugoslawische Staatsbürgerschaft.
14 HH-Vorstand war im Jahr 1990 Bürger der DDR.
15 Mindestens ein HH-Mitglied ist nach 1984 in die BRD eingewandert.
16 Haushalte mit Haushaltsnettoeinkommen über 7.500 DM bzw. 4.500 EUR in der zweiten Welle.
17 Die Gewichte werden im Zuge der Datenweitergabe für haushalts- und personenbezogene Analysen bereitgestellt.
18 Unser Dank gilt insbesondere Herrn Dr. Peter Krause für die hervorragende Zusammenarbeit.

1. Jedem Teilnehmer wird für alle Beobachtungszeitpunkte das letzte gültige Querschnittsgewicht des Beobachtungszeitraums zugewiesen.
2. Bei Teilnehmern, die im Todesjahr kein gültiges Querschnittsgewicht hatten, wird das letzte gültige Gewicht eingesetzt.[19]

Weil das SOEP derzeit aus acht, zu verschiedenen Zeitpunkten gezogenen Teilstichproben[20] besteht, kann in den Analysen nicht von einer Zufallsauswahl der Teilnehmer ausgegangen werden. Um die statistische Genauigkeit der Ergebnisse angemessen bewerten zu können, müssen die Folgen der Klumpung und Stratifizierung der Teilstichproben berücksichtigt werden. Es wurde dazu ein konservatives Vorgehen gewählt und robuste Standardfehler für nichtzufällige Stichproben ermittelt (Huber 1967; White 1982). Sie werden als konservativ angesehen, weil die ausgewiesenen Standardfehler in der Regel größer als die tatsächlichen statistischen Unsicherheiten der Schätzwerte sind.[21]

Werden Todesfälle im Zuge eines Mortalitäts-follow-up systematisch untererfasst, führt dies zu einer Überschätzung der Lebenserwartung. Das Forschungsinstitut TNS Infratest hat daher im Jahr 2001 im Auftrag des DIW eine Verbleibstudie für 7.902 Personen durchgeführt, die zwischen 1985 – 1998 die weitere Teilnahme am SOEP verweigert haben und so nachträgliche Informationen zu ihrem Vitalstatus gewonnen (Infratest 2002). Etwa 10 % der ehemaligen Teilnehmer waren nicht mehr auffindbar und 9 % bereits verstorben. Die Studie führte für die Wellen bis zum Jahr 2001 zu einer Verbesserung der Grundlagen für Mortalitätsanalysen. Die Ergebnisse von Schnell und Trappmann (2006) deuten allerdings darauf hin, dass die Verzerrung der Überlebensraten weder durch diese und eine frühere Verbleibstudie noch durch die Anwendung einer Anpassungsgewichtung ausgeglichen werden kann.[22] Es zeigte sich, dass die Überlebensraten der Teilnehmer weiterhin deut-

19 Das letzte Beobachtungsjahr lag zwischen 1985-2006 bei etwa 90 % der verstorbenen SOEP-Teilnehmer maximal drei Jahre vor dem Todesjahr, bei 75 % sogar lediglich ein Jahr vor dem Todesjahr.

20 Beim Auswahlverfahren des SOEP werden Haushalte als primäre Erhebungseinheit gezogen, in denen alle Personen befragt werden (Klumpung). Haushaltsmitglieder sind homogener als unabhängig gezogene Personen aus der Grundgesamtheit. Wird dies nicht berücksichtigt, wird die Effizienz der Schätzwerte in Klumpenstichproben überschätzt. Zum Klumpungseffekt trägt auch die nachträgliche Aufnahme von neuen Haushaltsmitgliedern (insbesondere Kinder und neue Partner) bei. Eine Schichtung erfolgt, weil Haushalte innerhalb von ausgewählten Regionen gezogen werden, dies erhöht die Heterogenität der Untersuchungseinheiten sofern sich die Regionen nach relevanten Merkmalen unterscheiden. Eine weitere Schichtungsebene lässt sich an der Zugehörigkeit zu den Teilstichproben festmachen. Vgl. u. a. Pischner (2006) zu Auswahlverfahren im SOEP; zu mehrstufigen Auswahlverfahren vgl. u. a. Schnell et al. (1999); zum Ziehungsdesign vgl. Spieß/Kroh (2007).

21 Eine Alternative zu diesem Vorgehen wäre ein Bootstrapping (Resampling-Verfahren, vgl. Efron 1992) der Standardfehler gewesen, dessen Ergebnisse aber schwer nachzuvollziehen sind.

22 Im Ergebnisteil werden die Unterschiede zwischen SOEP und amtlichen Periodensterbetafeln vor und nach Anwendung der Gewichtung differenziert nach Geschlecht und Altersgruppen dargestellt.

lich höher waren, als auf Basis der amtlichen Sterbetafeln zu erwarten gewesen wäre (Schnell/Trappmann 2006). Ein Grund für die Untererfassung von Todesfällen ist, dass Befragte mit einem schlechten Gesundheitszustand häufiger die weitere Teilnahme an der Studie verweigern (Heller/Schnell 2000). Auch hohes Alter, niedrige Bildung, niedriger beruflicher Status oder geringes Einkommen – weitere Risikofaktoren für vorzeitige Mortalität – erhöhen die Ausfallwahrscheinlichkeit der Teilnehmer (Spieß/Kroh 2008). Diese Probleme machen ein alternatives Vorgehen bei der Analyse sozialer Unterschiede in der Lebenserwartung nötig.

3.2 Sterbetafeln

Wir verwenden in dieser Studie Informationen aus den amtlichen Periodensterbetafeln, um die Überschätzung der Lebenserwartung im SOEP auszugleichen. Sterbetafeln sind demografische Modelle, die eine zusammenfassende Beurteilung der Sterblichkeitsverhältnisse einer Bevölkerung unabhängig von ihrer Größe und Altersstruktur ermöglichen und sich auf einen spezifischen Zeitraum beziehen (Destatis 2006). Es gibt zwei Herangehensweisen bei der Konstruktion von Sterbetafeln.

In der Querschnittsbetrachtung werden alle gestorbenen und lebenden Personen aus dem jeweiligen Zeitraum einbezogen und die so genannten ‚Periodensterbetafeln' berechnet. Sie bilden die Sterblichkeitsverhältnisse einer hypothetischen Bevölkerung ab. Sie gehen implizit davon aus, dass die zwischen 2003 und 2005 geborenen Kinder in 80 Jahren das gleiche Mortalitätsrisiko haben, wie die 80-jährigen zwischen 2003 und 2005. Dadurch wird die tatsächliche Lebenserwartung der heute geborenen Kinder wahrscheinlich unterschätzt. Ihr Vorteil ist, dass sie die periodenbezogenen Sterblichkeitsverhältnisse in einer Population in einem anschaulichen Maß, der mittleren Lebenserwartung, zusammenfassen. Es gibt allgemeine und abgekürzte Periodensterbetafeln: Allgemeine Sterbetafeln werden jeweils im Anschluss an eine Volkszählung erstellt und bis zu der Altersstufe von 100 Jahren veröffentlicht. Abgekürzte Sterbetafeln bilden die Entwicklung zwischen den Volkszählungen ab und werden nur bis zur Altersstufe von 90 Jahren veröffentlicht. Aufgrund der gestiegenen Lebenserwartung werden in Deutschland seit der Sterbetafel 2000/2002 nur noch allgemeine Sterbetafeln bis zur Altersstufe 100 veröffentlicht.[23]

23 Die Interpretation von Periodensterbetafeln für Deutschland ist aufgrund des ‚Healthy-migrant-Effektes' zusätzlich erschwert (Razum 2006). So hat die erste Migrantengeneration zumeist ein geringeres Mortalitätsrisiko als die übrige Bevölkerung ihres Herkunftslandes. Dies kann in Ländern wie Deutschland, mit großen Bevölkerungsanteilen mit Migrationshintergrund, zu einer Überschätzung der mittleren Lebenserwartung führen.

Eine Alternative zur Querschnittsbetrachtung stellt der längsschnittliche Ansatz der ‚Kohortensterbetafeln' dar. Sie weisen die altersspezifische Sterblichkeit einer realen Geburtskohorte aus und können erst nach dem vollständigen Versterben der Kohorte abgeschlossen werden. Die Lebenserwartung in einer Kohortensterbetafel weist die mittlere Lebensdauer der Mitglieder der jeweiligen Kohorte aus. Ein Kohortenvergleich hat bei der Analyse von zeitlichen Entwicklungen gegenüber der querschnittlichen Betrachtung Vorteile, weil er die Unterscheidung von alters-, perioden- und kohortenbezogenen Entwicklungen ermöglicht (Dinkel 1999).

In dieser Studie steht die Darstellung der gegenwärtigen Sterblichkeitsverhältnisse im Vordergrund. Es wird sich daher auf eine periodenbezogene Darstellung beschränkt. Nach den Periodensterbetafeln ist die mittlere Lebenserwartung zwischen 1995 und 2005 bei Männern von 73,6 auf 76,2 um 2,6 Jahre gestiegen (Destatis 2006). Bei Frauen stieg sie von 80,0 auf 81,8 Jahre. Die Heranziehung einer beliebigen Sterbetafel aus dem Untersuchungszeitraum würde dieser Veränderung nicht gerecht werden. Es wurden daher alle Sterbetafeln zwischen 1995 – 1997 und 2003 – 2005 zusammengefasst, indem aus den alters- und geschlechtsspezifischen Mortalitätsrisiken der einzelnen Tafeln eine mittlere Sterbewahrscheinlichkeit für den gesamten Zeitraum berechnet wurde.[24] Anhand der zusammengefassten Sterbetafel für den Zeitraum 1995 bis 2005 ergibt sich für Männer und Frauen eine mittlere Lebenserwartung von 75,3 bzw. 81,3 Jahren (vgl. auch Tabelle 1).

3.3 Statistische Modellierung

Das verwendete Vorgehen bei der Analyse sozialer Unterschiede war zweistufig. In einem ersten Schritt wurden anhand des SOEP soziale Unterschiede in der Sterblichkeit ermittelt. In einem zweiten Schritt wurden diese dann anhand der Periodensterbetafeln auf Unterschiede in der Lebenserwartung hochgerechnet. Allgemein werden nicht-parametrische, semi-parametrische und parametrische ereignisanalytische Analyseverfahren unterschieden, anhand derer soziale Unterschiede im Mortalitätsrisiko beschrieben werden können (Blossfeld/Rohwer 1995; Cleeves et al. 2002).[25]

24 Aufgrund der Veränderung der Altersspanne in den Tafeln (s. o.) gehen in die Mortalitätsrisiken der Altersjahre >90 nur die Informationen aus den Sterbetafeln ab 2000/02 ein. Dies führt zu einer minimalen Unterschätzung der Risiken für diese Gruppe für den Zeitraum 1995-2006.

25 Die verschiedenen Aspekte der Sterblichkeit von Populationen lassen sich dabei anhand von drei Funktionen abbilden. Die Survival- bzw. Überlebensfunktion S(t) beschreibt das Überleben der Population (Überlebensrate) bis zu einem Zeitpunkt t. Die Hazard- bzw. Risikofunktion h(t) beschreibt das Mortalitätsrisiko zum Zeitpunkt t. Die kumulierte Risikofunktion H(t) beschreibt das bis zum Zeitpunkt t akkumulierte Mortalitätsrisko.

Nicht-parametrische Verfahren, wie die Sterbetafelmethode oder die Kaplan-Meier-Methode, treffen keine Annahmen über die Entwicklung von Mortalitätsrisiken über die Zeit. Sie eignen sich für einen ersten exploratorischen Zugang, bei dem die funktionale Form der Hazardfunktion unbekannt ist oder überprüft werden soll. Die Verfahren lassen allerdings nur indirekt Rückschlüsse über den Einfluss von Kovariaten auf das Mortalitätsrisiko zu (durch grafischen oder statistischen Vergleich der resultierenden Funktionen). Mit zunehmender Anzahl von Kovariaten stoßen sie an ihre Grenzen, weil sehr umfangreiche Stichproben benötigt werden, um Überlebensraten für alle Kombinationen der Kovariaten zu ermitteln (Blossfeld/Rohwer 1995).

Das *semi-parametrische* Cox-Regressionsmodell trifft ebenfalls keine Annahmen über die zugrunde liegende Verteilung der Mortalitätsrisiken.[26] Im Unterschied zu nicht-parametrischen Verfahren ermöglicht es die Quantifizierung der Bedeutung verschiedener Einflussfaktoren.[27] Informationen zum Zeitpunkt eines Ereignisses oder zu vorangegangenen Ausprägungen der erklärenden Variablen gehen nicht in das Ergebnis ein. Das Cox-Modell wird sehr häufig verwendet, es eignet sich besonders, wenn Unterschiede im Mortalitätsrisiko statistisch abgesichert werden sollen.

Parametrische Modelle sollten nur verwendet werden, wenn die grundlegende Form der Überlebensfunktion bzw. Risikofunktion bereits bekannt ist. Sie spezifizieren das zeitabhängige Mortalitätsrisiko im Rahmen einer vorgegebenen parametrischen Funktion. Sie eignen sich insbesondere dann, wenn die gleichen Untersuchungsobjekte (Personen) mehrfach beobachtet werden, weil sie diese zusätzlichen Informationen besonders effizient nutzen (Cleeves et al. 2002). Sie nehmen keine Gruppenvergleiche zum Zeitpunkt von Ereignissen vor, sondern schätzen die Überlebensdauer der einzelnen Untersuchungsobjekte (Personen). Neben den Ausprägungen der Kovariaten zum Zeitpunkt von Ereignissen gehen dadurch auch frühere Ausprägungen in die Betrachtung ein. Eine falsche Spezifikation der Risikofunktion kann allerdings zu falschen Schlussfolgerungen über den Einfluss der erklärenden Variablen führen.

Die funktionale Form der Altersabhängigkeit des Mortalitätsrisikos ist für die deutsche Bevölkerung bekannt, daher wird im Folgenden ein parametrisches Exponentialmodell verwendet, um soziale Unterschiede in der Lebenserwartung anhand des SOEP zu analysieren. Diese statistische Modellierung des Mortalitätsrisikos hat sich bereits bei früheren Analysen auf Basis des SOEP bewährt (vgl. u. a. Klein 1993; Unger 2006).

26 Die einzige Annahme des Modells ist, dass die Einflussgrößen im Modell das Mortalitätsrisiko multiplikativ beeinflussen (‚proportional hazard assumption').

27 Das Schätzverfahren maximiert das Produkt aller Gruppenvergleiche zu Zeitpunkten mit Ereignissen (‚Partial Likelihood Funktion'; vgl. u. a. Blossfeld/Rohwer 1995).

Formel 1: Mortalitätsrisiko im Exponentialmodell

$$h(Alter) = \exp(b_0)$$

Im Exponentialmodell wird das Mortalitätsrisiko anhand eines konstanten Faktors beschrieben, es ändert sich damit nicht mit zunehmendem Alter (Formel 1). Um den exponentiellen Anstieg des Mortalitätsrisikos mit steigendem Alter zu berücksichtigen, fügt man einen zusätzlichen Altersterm hinzu. Für das Mortalitätsrisiko ergibt sich dadurch der folgende, zum ebenfalls häufig verwendeten Gompertz-Modell äquivalente Ausdruck:[28]

Formel 2: Mortalitätsrisiko im Exponentialmodell mit Alter als Kovariate

$$h'(Alter) = \exp(b_0 + b_{Alter} \cdot Alter)$$

Die Überlebensfunktion wird durch folgenden Ausdruck beschrieben:

Formel 3: Baseline Überlebensfunktion im Exponentialmodell mit Alter als Kovariate

$$S_0(Alter) = \exp\{-\left(\exp(b_0 + b_{Alter} Alter) Alter\right)\}$$

Soziale Unterschiede in der Sterblichkeit lassen sich als relative Unterschiede im Mortalitätsrisiko (Hazard Ratios) oder als absolute bzw. relative Unterschiede in der Lebenserwartung beschreiben. Beide Maßzahlen werden nicht vom Altersaufbau der zu analysierenden Population beeinflusst. Im nachfolgend verwendeten Exponentialmodell beschreibt der zur Basis e (Eulersche Zahl) potenzierte Effektkoeffizient einer Variablen das altersunabhängige relative Mortalitätsrisiko einer sozioökonomischen Gruppe. Dies kann für das einfache Beispiel einer diskreten Gruppenvariable (z) mit zwei Ausprägungen (z={1,2}) verdeutlicht werden. Das Gruppierungsmerkmal (z) geht anhand von dichotomen Indikatorvariablen ('Dummy-Variablen') in das Modell ein (x_1:z=1; x_2:z=2). Das relative Mortalitätsrisiko der Gruppe x_1 gegenüber x_2 wird dann durch folgenden Ausdruck beschrieben:

Formel 4: Relatives Mortalitätsrisiko auf Basis eines parametrischen Exponentialmodells

$$\frac{h(Alter \mid x_1 = 1)}{h(Alter \mid x_2 = 1)} = \frac{\exp(b_0 + b_{Alter} \cdot Alter + b_{x_1} x_1)}{\exp(b_0 + b_{Alter} \cdot Alter)} = \exp(b_{x_1})$$

28 Das Mortalitätsrisiko steigt dabei exponentiell an, wenn der Koeffizient des Alters (b_{Alter}) positiv ist.

Die mittlere Lebenserwartung repräsentiert die Fläche unter der Überlebensfunktion S(t). Um sie zu berechnen, muss zuvor die Überlebensfunktion bestimmt werden. Beim verwendeten Modell lautet die Formel für die Überlebensfunktionen der beiden Gruppen x_1 und x_2 wie folgt:

Formel 5: Berechnung von Überlebensraten anhand des Exponentialmodells mit Altersterm

$$S_{x_1}(Alter \mid x_1 = 1, x_2 = 0) = \exp\{-(\exp(b_0 + b_{Alter} Alter + b_{x_1} x_1) Alter)\}$$
$$S_{x_2}(Alter \mid x_1 = 0, x_2 = 1) = \exp\{-(\exp(b_0 + b_{Alter} Alter) Alter)\}$$

Anhand der berechneten Überlebensraten wird die mittlere und fernere Lebenserwartung ausgehend von 100.000 fiktiven männlichen bzw. weiblichen Lebendgeborenen (der so genannten ‚Sterbetafelbevölkerung') berechnet. Sie beschreibt die mittlere Lebensdauer einer Population (vgl. u. a. Destatis 2006c):

Formel 6: Berechnung der mittleren und ferneren Lebenserwartung

$$e_t = \frac{e_t l_t}{l_t}$$

mit

$$e_t l_t = \sum_{t \geq x} L_t \; ; L_t = \tfrac{1}{2}(l_t + l_{t+1}); l_t = l_0 \cdot S_x(t); l_0 = 100000$$

e_t : Lebenserwartung im Alter t

t : Alter in Jahren

l_t : Überlebende der Sterbetafelbevölkerung im Alter t

L_t : Von den Überlebenden im Alter t *bis* t+ 1 durchlebte Jahre

$S_x(t)$: Wahrscheinlichkeit für Personen aus Gruppe x bis ins Alter t zu überleben

Um die Verzerrung durch die Überschätzung der Überlebensraten im SOEP auszugleichen, wird die Lebenserwartung anhand einer alternativen Formel berechnet. Anstatt die Überlebensraten der Gruppen anhand des SOEP zu schätzen, werden die tatsächlichen Überlebensraten der Bevölkerung, die in den Periodensterbetafeln des Statistischen Bundesamtes dokumentiert sind, herangezogen. Die relativen Mortalitätsrisiken aus dem ereignisanalytischen Exponentialmodell müssen dazu transformiert werden. Sie sollen nicht den Unterschied des gruppenbezogenen Mortalitätsrisikos im Vergleich zu einer Referenzkategorie ausweisen (Formel 4), sondern den relativen Unterschied zum durchschnittlichen Mortalitätsrisiko im

Datensatz. Die zentrale Annahme dieses Verfahrens ist, dass die relativen Abstände zum Durchschnitt im SOEP den relativen Abständen zum Durchschnitt in der Grundgesamtheit approximativ entsprechen.

Das Vorgehen lässt sich anhand des zuvor verwendeten Beispiels für eine Variable mit zwei Ausprägungen (z={1,2}) darstellen: In einem ersten Schritt werden alle Beobachtungen dupliziert.[29] Dann werden die Werte für die Gruppenvariable (z) bei den duplizierten Fällen auf einen neuen Wert (z=3) gesetzt. Anschließend wird ein neues Modell geschätzt, indem relative Mortalitätsrisiken der Gruppen x_1:z=1 und x_2:z=2 im Verhältnis zur Gruppe x_3:z=3 geschätzt werden:[30]

Formel 7: Transformierte relative Mortalitätsrisiken im Verhältnis zum Durchschnitt des SOEP

$$\frac{h(Alter \mid x_1 = 1, x_2 = 0, x_3 = 0)}{h(Alter \mid x_3 = 1, x_1 = 0, x_2 = 0)} = \frac{\exp(b'_0 + b'_{Alter} \cdot Alter + b'_{x_1} x_1 + b'_{x_2} x_2)}{\exp(b'_0 + b'_{Alter} \cdot Alter)} = \frac{h_0(t) \cdot \exp(b'_{x_1})}{h_0(t)} = \exp(b'_{x_1})$$

und

$$\frac{h(Alter \mid x_1 = 0, x_2 = 1, x_3 = 0)}{h(Alter \mid x_3 = 1, x_1 = 0, x_2 = 0)} = \frac{\exp(b'_0 + b'_{Alter} \cdot Alter + b'_{x_1} x_1 + b'_{x_2} x_2)}{\exp(b'_0 + b'_{Alter} \cdot Alter)} = \frac{h_0(t) \cdot \exp(b'_{x_2})}{h_0(t)} = \exp(b'_{x_2})$$

Die korrigierten altersabhängigen Sterbewahrscheinlichkeiten der beiden Gruppen ergeben sich durch den Austausch der baseline Risikofunktion h_0(Alter):

Formel 8: Korrigiertes Mortalitätsrisiko auf Basis der transformierten relativen Mortalitätsrisiken

$$h'_{x_1}(Alter \mid x_1 = 1, x_2 = 0) = h'_0(Alter) \cdot \exp(b'_{x_1})$$
$$h'_{x_2}(Alter \mid x_1 = 0, x_2 = 1) = h'_0(Alter) \cdot \exp(b'_{x_2})$$

mit

$h'_0(Alter)$: Baseline Mortalitätsrisiko aus den zusammengefassten Periodensterbetafeln

Die korrigierten Überlebensraten lassen sich aufgrund der Relation zwischen Hazard- und Survivalfunktion wie folgt berechnen:

29 Sind – wie im SOEP – mehrere Beobachtungen pro Fall vorhanden, müssen den duplizierten Fällen neue Kennziffern zugewiesen werden.

30 Durch die Verdopplung der Anzahl der Personen und Fälle lassen sich auf Basis der Ergebnisse dieses Modells keine Signifikanztests für die Effekte durchführen.

Formel 9: Überlebensraten auf Basis der kumulierten korrigierten Risikofunktion

$$S'_{x_1}(Alter) = \exp(-H'_{x_1}(Alter \mid x_1 = 1, x_2 = 0))$$
$$S'_{x_2}(Alter) = \exp(-H'_{x_2}(Alter \mid x_1 = 0, x_2 = 1))$$

Sie werden anschließend herangezogen, um die korrigierte mittlere und fernere Lebenserwartung zu berechnen (vgl. Formel 6). Das Vorgehen zur Korrektur der Lebenserwartung bei der Analyse sozialer Unterschiede anhand des SOEP wird nachfolgend zusammengefasst (Abbildung 1):

Abbildung 1 Verfahren bei der Berechnung der Lebenserwartung auf Basis des SOEP

1. **Wahl des Untersuchungszeitraums** und **vollständige Unterteilung (Partition) der Stichprobe** in (disjunkte) Subpopulationen.
2. Berechnung **relativer Mortalitätsrisiken der Subpopulationen im Verhältnis zu einer Referenz-Subpopulation** anhand ereignisanalytischer Modelle und Durchführung von Signifikanztests.
3. Transformation zu **relativen Mortalitätsrisiken im Verhältnis zum Durchschnitt des SOEP.**
4. **Anwendung der transformierten relativen Mortalitätsrisiken** auf Überlebensraten aus den amtlichen Sterbetafeln.
5. **Berechnung der Lebenserwartungen** auf Basis der Überlebensraten (aus 4.) nach der Sterbetafelmethode.

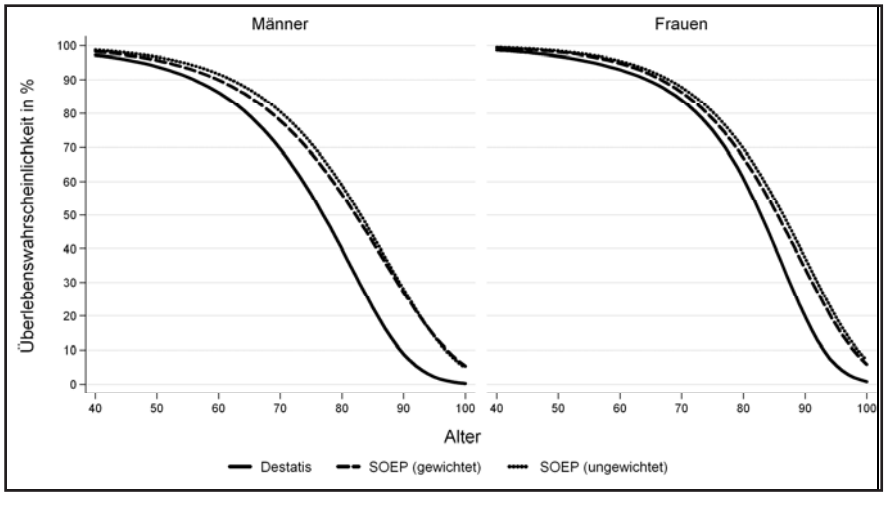
4 Ergebnisse

In die Analysen beziehen wir die Teilnehmer des Sozio-oekonomischen Panels aus dem Zeitraum 1995 bis 2005 ein. Es werden alle Teilstichproben, mit Ausnahme der Stichprobe G (Hocheinkommensbezieher) für die kein regulärer Gewichtungsfaktor vorliegt, analysiert. Die Entscheidung für einen 10-jährigen Untersuchungszeitraum ermöglicht es, regelmäßig aktualisierte Ergebnisse zu berichten und gleichzeitig eine hinreichend große Fallzahl zu erreichen. Zwischen 1995 und 2005 haben ca. 31.800 Personen am SOEP teilgenommen. Etwa 1.900 Teilnehmer (6 %) sind im Untersuchungszeitraum verstorben. Die Zahl der Todesfälle verteilte sich dabei relativ gleichmäßig über alle Jahre und betrug zwischen 150 und 210 Personen.

4.1 Überschätzung der Lebenserwartung im SOEP

Die Überschätzung der Lebenserwartung deutet sich für den Zeitraum 1995 bis 2005 bereits bei einem Vergleich der gewichteten und ungewichteten Überlebensfunktionen mit den Raten aus den amtlichen Sterbetafeln an (Abbildung 2). Die Überlebensraten im SOEP wurden anhand eines exponentiellen ereignisanalytischen Regressionsmodells für den Untersuchungszeitraum generiert (vgl. Formel 3). Die Überlebensraten auf Basis der Teilnehmer des SOEP sind gegenüber den in den Sterbetafeln ausgewiesenen Raten deutlich erhöht. Die Unterschiede zeigen sich auch noch nach der Anwendung einer Ausgleichsgewichtung. Die gewichteten Ergebnisse liegen allerdings näher an den Periodensterbetafeln. Insgesamt ist die Differenz bei Männern deutlich größer als bei Frauen.

Abbildung 2 Vergleich der altersspezifischen Überlebensraten im SOEP mit den amtlichen Periodensterbetafeln 1995-2005 nach Geschlecht



Anhand der Überlebenswahrscheinlichkeiten wurde die mittlere und fernere Lebenserwartung für Männer und Frauen berechnet (Tabelle 1). Die Unterschiede zwischen dem SOEP und den Periodensterbetafeln werden dabei besonders deutlich. Auf Basis der Periodensterbetafeln ergibt sich für den Zeitraum 1995 bis 2005 eine mittlere Lebenserwartung von 75,3 Jahren bei Männern und 81,3 Jahren bei Frauen. Die mittlere Lebenserwartung der Studienteilnehmer des SOEP beträgt dagegen vor bzw. nach Anwendung der Ausgleichsgewichtung 81,8 bzw. 80,8 Jahre

bei Männern und 85,4 bzw. 84,4 Jahre bei Frauen. Auch die gewichteten Ergebnisse überschätzen die Lebenserwartung in der Grundgesamtheit somit um 5,5 bzw. 3,1 Jahre. Durch die Gewichtung verringert sich die Differenz nur um etwa 1,3 % bei Männern und 1,2 % bei Frauen.

Tabelle 1 Vergleich der durchschnittlichen Überlebensraten und Lebenserwartungen auf Basis des SOEP mit den amtlichen Periodensterbetafeln nach Altersgruppen

	Perioden-sterbetafeln		SOEP (ungewichtet)		SOEP (gewichtet)	
Alter	S(t)	e_t	S(t)	e_t	S(t)	e_t
			Männer			
0	100,0	75,3	100,0	81,8	100,0	80,8
1-9	99,9	72,9	100,0	79,3	100,0	78,3
10-19	99,8	66,0	100,0	72,3	99,9	71,4
20-29	99,4	56,2	99,9	62,3	99,8	61,5
30-39	98,6	46,6	99,6	52,5	99,4	51,7
40-49	97,2	37,1	98,8	42,8	98,3	42,1
50-59	94,0	28,1	96,7	33,5	95,8	33,0
60-69	86,4	19,8	91,8	24,9	90,1	24,5
70-79	70,1	12,8	80,9	17,2	78,3	17,1
80-89	41,3	7,3	59,8	10,9	57,2	11,0
90-99	11,1	3,8	29,7	6,1	28,7	6,3
			Frauen			
0	100,0	81,3	100,0	85,4	100,0	84,4
1-9	99,9	78,8	100,0	82,9	100,0	81,9
10-19	99,9	71,9	100,0	75,9	100,0	74,9
20-29	99,7	62,0	100,0	65,9	99,9	65,0
30-39	99,3	52,2	99,8	55,9	99,8	55,0
40-49	98,7	42,4	99,5	46,1	99,4	45,2
50-59	96,9	33,0	98,4	36,5	98,2	35,7
60-69	93,1	24,1	95,5	27,3	94,9	26,6
70-79	84,1	15,8	87,9	18,9	86,5	18,3
80-89	61,5	8,9	70,3	11,9	67,7	11,5
90-99	22,7	4,3	39,0	6,4	36,0	6,2

S(t): Anteil der Überlebenden bis zum Alter t in %.
e_t: (Fernere) Lebenserwartung im Alter t (in Jahren).
SOEP (gewichtet): Geschätzte Lebenserwartung e_t und Überlebensrate S(t) auf Basis des SOEP mit modifizierter Längsschnittgewichtung.
SOEP (ungewichtet): Geschätzte Lebenserwartung e_t und Überlebensrate S(t) auf Basis des SOEP mit ungewichteten Daten.

Datenbasis: SOEP und Periodensterbetafeln 1995-2005.

Die Ergebnisse zeigen, dass die Überlebensraten in der Grundgesamtheit auf Basis des Sozio-oekonomischen Panels – auch nach Anwendung einer Anpassungsgewichtung – deutlich überschätzt werden und mit früheren Analysen übereinstimmen (vgl. Schnell/Trappmann 2006).[31] Dadurch wird auch die Lebenserwartung von Subpopulationen überschätzt und die Ergebnisse zu sozialen Unterschieden verzerrt.

4.2 Soziale Unterschiede in der Lebenserwartung

Frühere Studien zum Ausmaß von sozialen Unterschieden in der Lebenserwartung anhand des SOEP zeigen für verschiedene Indikatoren des sozioökonomischen Status, wie Bildungsabschluss, Berufsstatus oder Einkommen, deutliche Differenzen auf. Für den Zeitraum 1984 bis 1993 wurden für die alten Bundesländer Unterschiede zwischen Abiturienten und Hauptschulabsolventen von 3,3 Jahren bei Männern und 3,9 Jahren bei Frauen ausgewiesen (Klein 1996). Die Differenz in der Lebenserwartung bei Geburt zwischen dem untersten und obersten Einkommensquartil wird für den Zeitraum 1984 bis 1997 mit 6 Jahren bei Männern und 4 Jahren bei Frauen angegeben (Reil-Held 2000). Weitere Ergebnisse, in denen Unterschiede nach relativen Einkommenspositionen berichtet werden, liegen für den Zeitraum 1998 bis 2003 vor (Lampert/Kroll 2006a). Die Ergebnisse weisen zwischen der Armutsrisikogruppe (0-<60 % des mittleren Netto-Äquivalenzeinkommens) und Personen im relativen Wohlstand (>150 %) eine Differenz von 14 Jahren bei Männern und 8 Jahren bei Frauen aus. Die bisher auf Basis des SOEP durchgeführten Studien berücksichtigen den Unterschied zwischen der Mortalität der Studienteilnehmer und der Mortalität der Grundgesamtheit jedoch nicht hinreichend.

In dieser Studie werden soziale Unterschiede auf Basis des Netto-Äquivalenzeinkommens operationalisiert.[32] Das Äquivalenzeinkommen spiegelt die Größe und Zusammensetzung des Haushaltes wider und berücksichtigt damit Einsparungen durch gemeinsames Wirtschaften in Mehrpersonenhaushalten sowie den unterschiedlichen Bedarf von Erwachsenen und Kindern (Hauser 1996; Lampert/Kroll 2006b). Das mittlere Netto-Äquivalenzeinkommen[33] der 18-jährigen und älteren Bevölkerung lag zwischen 1995 und 2005 bei 1.141 EUR. Ausgehend von der re-

31 Weiterführende Analysen der drei Überlebensraten (nicht dargestellt) zeigen, dass das Einsetzen der Mortalität gegenüber den Sterbetafeln anhand der gewichteten Ergebnisse auf Basis des SOEP um 2 % und anhand der ungewichteten Ergebnisse um 5 % überschätzt wird. Die Zunahme des Mortalitätsrisikos im Altersgang wird anhand der ungewichteten Ergebnisse um 1 % überschätzt, während sie anhand der gewichteten Ergebnisse um 5 % unterschätzt wird.

32 Das dargestellte Vorgehen zur Analyse sozialer Unterschiede lässt sich auf alle kategorialen Indikatoren mit hinreichend großen Zellenbesetzungen anwenden.

33 Arithmetisches Mittel der jährlichen Mediane (50. Perzentil) der Einkommensverteilung.

lativen Einkommensposition der Befragten wurden fünf Gruppen[34] unterschieden. Die Zuweisung der Befragten zu den Einkommensgruppen wurde für jedes Jahr aktualisiert. Ein Vorteil von Einkommenspositionen ist, dass Kaufkraftunterschiede zwischen den Jahren keinen Einfluss auf die Einkommensposition haben, weil nur der relative Abstand zum Mittelwert betrachtet wird. Die durchschnittliche Einkommensschwelle, bis zu der Befragte der Armutsrisikogruppe zugeordnet wurden (60 %-Schwelle), lag zwischen 1995 und 2005 bei 685 EUR. Zwischen 1995 und 2005 erzielten durchschnittlich 14 % der Bevölkerung Äquivalenzeinkommen unterhalb der 60 %-Schwelle. Die 150 %-Schwelle betrug durchschnittlich 1.712 EUR und wurde zur Abgrenzung relativer Wohlhabenheit herangezogen. Im Mittel erzielten etwa 19 % der Bevölkerung Einkommen oberhalb der 150 %-Schwelle.

Aufbauend auf den gewichteten Daten des Sozio-oekonomischen Panels wurden die Mortalitätsrisiken der fünf Einkommensgruppen bestimmt (Tabelle 2). In den Analysen wurden sowohl relative Mortalitätsrisiken im Vergleich zur Referenzgruppe mit Einkommenspositionen von mehr als 150 % des gesellschaftlichen Durchschnitts als auch relative Risiken im Vergleich zum durchschnittlichen Risiko im SOEP bestimmt.[35] Das altersstandardisierte Mortalitätsrisiko (HR) von Männern aus der Armutsrisikogruppe ist demnach 2,68-fach und das Risiko von Frauen 2,44-fach im Vergleich zur Referenzgruppe erhöht. Setzt man die relativen Mortalitätsrisiken in Beziehung zum mittleren Risiko der Teilnehmer des SOEP (HR_{MEAN}), ist das Risiko der untersten Einkommensgruppe bei Männern und Frauen um das 1,61-Fache bzw. 1,57-Fache erhöht, während es in der höchsten Einkommensgruppe auf das 0,6-Fache bzw. 0,7-Fache verringert ist. Die vergleichsweise kleine Datenbasis und die Notwendigkeit, aufgrund des komplexen Stichprobendesigns robuste Standardfehler zu berechnen, führen zu großen statistischen Unsicherheiten bei diesen Punktschätzern. Das 95 %-Vertrauensintervall weist im Vergleich von Armutsrisiko- und Referenzgruppe für Männer ein zwischen 1,9- und 3,9-fach und für Frauen ein zwischen 1,7- und 3,6-fach erhöhtes Risiko in der Armutsrisikogruppe aus.

Durch Anwendung der relativen Risiken auf Überlebensraten[36] lässt sich die Lebenserwartung für die Einkommensgruppen berechnen. Nachfolgend wird nur die Lebenserwartung von Männern und Frauen mit Einkommen unter der 60 %-Schwelle und über der 150 %-Schwelle und die Differenz zwischen beiden Gruppen dargestellt (Tabelle 3). Es werden die Werte und Differenzen, die sich auf Basis der amtlichen Periodensterbetafeln 1995-2005, des SOEP 1995-2005 mit Ausgleichsgewichtung und des ungewichteten SOEP 1995-2005 ergeben, verglichen.

34 Unter 60 %, 60 bis unter 80 %, 80 bis unter 100 %, 100 bis unter 150 %, über 150 % des gesellschaftlichen Mittelwertes (Grabka/Krause 2005).
35 Vgl. Formel 4 und Formel 7.
36 Vgl. dazu Abbildung 1 und Formel 5.

Tabelle 2 Relatives Mortalitätsrisiko im Verhältnis zur Referenzgruppe und zum mittleren Mortalitätsrisiko im SOEP nach Einkommensposition und Geschlecht

Einkommensposition	Männer		Frauen	
	HR [95 % KI]	HR_{MEAN}	HR [95 % KI]	HR_{MEAN}
0-<60 %	2,68 [1,88;3,81]	1,61	2,44 [1,67;3,57]	1,57
60-<80 %	1,99 [1,40;2,82]	1,19	1,45 [0,98;2,15]	0,94
80-<100 %	1,69 [1,19;2,39]	1,01	1,43 [0,97;2,11]	0,92
100-<150 %	1,40 [1,00;1,98]	0,84	1,10 [0,75;1,63]	0,72
≥150 %	Ref.	0,60	Ref.	0,65

Einkommensposition: Relative Einkommenspositon auf Basis des bedarfsgewichteten HH-Nettoeinkommens (Neue OECD Skala).
HR [95 % KI]: Hazard Ratio (relatives Mortalitätsrisiko) im Verhältnis zum Risiko von Personen mit einer Einkommenspositon von 150 % und mehr. Konfidenzintervall auf Basis robuster Standardfehler (Huber 1967; White 1982).
HR_{MEAN}: Relatives Mortalitätsrisiko im Verhältnis zum SOEP-Durchschnittsrisiko.
Ref.: Referenzgruppe.

Datenbasis: SOEP und Periodensterbetafeln 1995-2005 (Lampert et al. 2007).

Tabelle 3 Mittlere und fernere Lebenswartung auf Basis verschiedener Überlebensraten nach Einkommensposition und Geschlecht

	Lebenserwartung bei Geburt			Lebenserwartung mit 65 Jahren		
	Amtliche Sterbetafeln	SOEP (gewichtet)	SOEP (ungewichtet)	Amtliche Sterbetafeln	SOEP (gewichtet)	SOEP (ungewichtet)
Männer						
0-<60 %	70,1	74,0	75,3	12,3	15,3	15,6
≥150 %	80,9	85,2	85,8	19,7	23,4	23,5
Differenz	10,8	11,2	10,5	7,4	8,1	7,9
Frauen						
0-<60 %	76,9	78,7	79,7	16,2	17,3	17,9
≥150 %	85,3	87,2	88,0	22,5	24,1	24,7
Differenz	8,4	8,5	8,3	6,3	6,8	6,8

0-<60 %: Männer bzw. Frauen mit Einkommenspositionen von weniger als 60 % des Medianeinkommens.
≥150 %: Männer bzw. Frauen mit Einkommenspositionen von mehr als 150 % des Medianeinkommens.
Amtliche Sterbetafeln: Geschätzte Lebenserwartung auf Basis der Baselinefunktion nach amtlichen Sterbetafeln.
SOEP (gewichtet): Geschätzte Lebenserwartung auf Basis der Baselinefunktion nach SOEP mit modifizierter Längsschnittgewichtung.
SOEP (ungewichtet): Geschätzte Lebenserwartung auf Basis der Baselinefunktion nach SOEP mit ungewichteten Daten.

Datenbasis: SOEP und Periodensterbetafeln 1995-2005.

Für Männer ergeben sich anhand der Periodensterbetafeln Differenzen von 10,8 Jahren in der Lebenserwartung bei Geburt bzw. 7,4 Jahren in der ferneren Lebenserwartung ab 65 Jahren. Legt man den Lebenserwartungen der beiden Gruppen dagegen die Überlebensraten aus dem SOEP zugrunde, verschieben sich die vorgefundenen Unterschiede. Auf Basis der gewichteten Überlebensfunktion wird die Differenz in der Lebenserwartung bei Geburt um 0,4 Jahre überschätzt, anhand der ungewichteten Raten dagegen um 0,3 Jahre unterschätzt. Die Differenz in der ferneren Lebenserwartung wird auf Basis des SOEP anhand gewichteter und ungewichteter Raten sogar um 0,7 bzw. 0,5 Jahre überschätzt. Bei Frauen sind die Differenzen zwischen den Ergebnissen auf Basis des SOEP und den amtlichen Sterbetafeln etwas kleiner. Die Einkommensunterschiede in der Lebenserwartung bei Geburt werden auf Basis der gewichteten und ungewichteten Raten nur um 0,1 bzw. -0,1 Jahre verzerrt. Die Differenzen in der ferneren Lebenserwartung werden allerdings auf Basis beider Raten um etwa 0,5 Jahre überschätzt.

Es hat sich gezeigt, dass anhand des SOEP insbesondere die Einkommensunterschiede in der ferneren Lebenserwartung überschätzt werden. Die Abweichung ist auf Basis der gewichteten Überlebensraten zudem nicht generell geringer als auf Basis ungewichteter Daten. Die Richtung und das Ausmaß der Verzerrung der Differenzen in der Lebenserwartung bei Geburt werden nicht nur vom Umfang der Überschätzung der mittleren Lebenserwartung, sondern auch von der spezifischen Form der Abweichung der Überlebensraten beeinflusst. Sie lassen sich nur schwer prognostizieren. Bei Männern sind diese Abweichungen ebenfalls bedeutend.

5 Fazit

In der vorliegenden Studie wurde ein Überblick über Datenquellen zur Analyse sozialer Unterschiede in der Lebenserwartung gegeben und eine Methode zur Berechnung von sozialen Unterschieden in der Lebenserwartung auf Basis einer Kombination von Daten des Sozio-oekonomischen Panels und der amtlichen Periodensterbetafeln vorgestellt. Die vorgefundenen Ergebnisse weisen auf einen markanten Einkommensgradienten in der Lebenserwartung der Bevölkerung Deutschlands hin: Je höher das Einkommen, desto eher besteht die Aussicht auf ein langes Leben (vgl. Lampert et al. 2007). Bei Männern sind diese Unterschiede noch etwas stärker ausgeprägt als bei Frauen. Die Ergebnisse entsprechen weitgehend denen, die für andere europäische Länder berichtet wurden (Mackenbach et al. 1997).

Bisher bestanden für Deutschland vergleichsweise große Schwierigkeiten, Aussagen über das Ausmaß von sozialen Unterschieden in der Lebenserwartung

zu machen. Die Ergebnisse des Mikrozensus können nicht mit den amtlichen Bevölkerungsregistern verknüpft werden. Zudem sind auch auf den Totenscheinen keine Angaben über sozioökonomische Hintergrundmerkmale enthalten. Die Daten der gesetzlichen Krankenkassen und Rentenversicherung sind ebenfalls nur sehr eingeschränkt für Analysen zur sozial differenziellen Mortalität zu verwenden. Limitationen für das SOEP ergaben sich bisher aufgrund der Überschätzung der Lebenserwartung auf Basis der Daten. Das beschriebene Vorgehen ermöglicht eine Verringerung dieser Verzerrungen bei der Analyse sozialer Unterschiede. Darauf aufbauend können die Ergebnisse auf Basis des SOEP mit Studien aus anderen Ländern verglichen werden, sofern vergleichbare Sozialindikatoren und Altersbereiche vorliegen. Weitere Einschränkungen ergeben sich aus dem geringen Stichprobenumfang und dem komplexen Stichprobendesign des Sozio-oekonomischen Panels. Diese Faktoren führen zu vergleichsweise großen Vertrauensintervallen für relative Unterschiede im Mortalitätsrisiko. Für eine genauere Ermittlung wäre eine deutliche Ausweitung des Stichprobenumfanges erforderlich, diese steht aber nicht in Aussicht. Auf Basis des Mikrozensus könnten soziale Unterschiede mit einer größeren statistischen Genauigkeit ermittelt werden, sofern ein Mortalitäts-follow-up der Teilnehmer durchgeführt würde. Es ist jedoch nicht abzusehen, ob die amtliche Statistik in Zukunft entsprechende Daten zur Verfügung stellen kann.

Vor dem Hintergrund der gesteigerten Aufmerksamkeit, die Unterschieden in der Lebenserwartung in der öffentlichen und politischen Diskussion gewidmet wird, sollten besonders strenge Maßstäbe an ihre Analyse angelegt werden. So sollten insbesondere die Folgen der Überschätzung der Lebenserwartung auf Basis von Stichproben wie dem SOEP berücksichtigt werden. Das beschriebene Vorgehen stellt einen Ansatz dar, um Verzerrungen bei der Berechnung sozialer Unterschiede in der Lebenserwartung anhand des SOEP zu verringern.

Literatur

Badura, B., Schnellschmidt H. und C. Vetter, 2007: Fehlzeitenreport 2006. Heidelberg: Springer.

Blossfeld, H.-P. und G. Rohwer, 1995: Techniques of Event History modeling. New approaches to causal analyses. Hillsdale/NJ: Lawrence Erlbaum Associates.

Bothig, S., 1989: WHO MONICA Project. Objectives and design. International Journal of Epidemiology 18 (supplement 1): 29.

Cleeves, M., W. Gould und R. Gutierrez, 2002: An introduction to survival analyses using Stata. College Station Texas: Stata Press.

Destatis, 2006: Perioden-Sterbetafeln für Deutschland – Allgemeine und abgekürzte Sterbetafeln von 1871/1881 bis 2003/2005. Wiesbaden: Statistisches Bundesamt.

Dinkel, R. H., 1999: Entwicklung und Gesundheitszustand. Eine empirische Kalkulation der Healthy Life Expectancy für die Bundesrepublik auf der Basis von Kohortensterbetafeln. S. 61-84 in: H. Häfner (Hg.): Gesundheit – unser höchstes Gut? Heidelberg: Springer.

Efron, B., 1982: The jackknife, the bootstrap, and other resampling plans. CBMS-NSF Regional conference series in applied mathematics. Philadelphia: Society for Industrial and Applied Mathematics.

Gärtner, K., 2002: Differentielle Sterblichkeit – Ergebnisse des Lebenserwartungssurveys. Zeitschrift für Bevölkerungswissenschaft 27: 185-211.

Gärtner, K. und R. D. Scholz, 2005: Lebenserwartung in Gesundheit. S. 311-331 in: K. Gärtner, E. Grünheid und M. Luy (Hg.): Lebensstile, Lebensphasen, Lebensqualität. Interdisziplinäre Analysen von Gesundheit und Sterblichkeit aus dem Lebenserwartungssurvey des BIB. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gässler, H., S. Geyer und R. Peter, 2005: Sozialstatus und Statusinkonsistenz als Risikofaktoren für ischämische Herzkrankheiten. Eine Studie mit gesetzlich Krankenversicherten. Das Gesundheitswesen 67: Vortrag V21.

Geyer, S. und R. Peter, 1999: Occupational status and all-cause mortality: A study with health insurance data from Nordrhein-Westfalen, Germany. The European Journal of Public Health 9: 114-118.

Geyer, S. und R. Peter, 2000: Income, occupational position, qualification and health inequalities – competing risks? Comparing indicators of social status. Journal of Epidemiology and Community Health 54: 299-305.

Grabka, M. M. und P. Krause, 2005: Einkommen und Armut von Familien und älteren Menschen. Wochenbericht DIW: 155-162.

Greß, S., 2002: Freie Kassenwahl und Preiswettbewerb in der GKV. Effekte und Perspektiven. Vierteljahrshefte zur Wirtschaftsforschung 71: 490-497.

Hauser, R., 1996: Zur Messung individueller Wohlfahrt und ihrer Verteilung. S. 13-38 in: Statistisches Bundesamt (Hg.): Wohlfahrtsmessung – Aufgabe der Statistik im gesellschaftlichen Wandel. Wiesbaden: Statistisches Bundesamt.

Heller, G. und R. Schnell, 2000: The choir invisible. Zur Analyse der gesundheitsbezogenen Panelmortalität im SOEP. S. 115-134 in: U. Helmert, K. Bammann, W. Voges und R. Müller (Hg.): Müssen Arme früher sterben? Soziale Ungleichheit und Gesundheit in Deutschland. Weinheim: Juventa.

Helmert, U., 2000: Der Einfluss von Beruf und Familienstand auf die Frühsterblichkeit von männlichen Krankenversicherten. S. 243-268 in: U. Helmert, K. Bammann, W. Voges und R. Müller (Hg.): Müssen Arme früher sterben? Weinheim: Juventa.

Helmert, U., 2003: Individuelle Risikofaktoren, Gesundheitsverhalten und Mortalitätsentwicklung in Deutschland im Zeitraum 1984 bis 1998. Das Gesundheitswesen 65: 542-547.

Helmert, U., W. Voges, A. Timm und T. Sommer, 2002: Soziale Einflussfaktoren für die Mortalität von männlichen Krankenversicherten in den Jahren 1989 bis 2000. Das Gesundheitswesen 64: 3-10.

Himmelreicher, R. K., H.-M. von Gaudecker und R. D. Scholz, 2006: Nutzungsmöglichkeiten von Daten der gesetzlichen Rentenversicherung über das Forschungsdatenzentrum der Rentenversicherung (FDZ-RV). MPIDR Working Paper WP-2006-018: 1-23.

Hoffmann, H., K. Kaldybajewa und E. Kruse, 2006: Arbeiter und Angestellte im Spiegel der Statistik der gesetzlichen Rentenversicherung. Rückblick und Bestandsaufnahme. Deutsche Rentenversicherung 1/2006: 24-53.

Huber, P. J., 1967: The behavior of maximum likelihood estimates under non-standard conditions. S. 221-233 in: J. Neyman (Hg.): Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability. Berkeley: University of California Press.

Infratest, 2002: Verbesserung der Datengrundlagen für Mortalitäts- und Mobilitätsanalysen: Verbleibstudie bei Panelausfällen im SOEP. München: Infratest Sozialforschung.

Keil, T., 2005: Das weltweite WHO-MONICA-Projekt. Ergebnisse und Ausblick. Das Gesundheitswesen 67 (S1): S38-S45.

Klein, T., 1993: Soziale Determinanten der Lebenserwartung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 45: 712-730.
Klein, T., 1996: Mortalität in Deutschland. Aktuelle Entwicklungen und soziale Unterschiede. S. 366-377 in: W. Zapf, J. Schupp und R. Habich (Hg.): Lebenslagen im Wandel. Sozialberichterstattung im Längsschnitt. Frankfurt a. M.: Campus.
Klein, T. und R. Unger, 2001: Einkommen, Gesundheit und Mortalität in Deutschland, Großbritannien und den USA. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 96-110.
Klein, T., S. Schneider und H. Löwel, 2001: Bildung und Mortalität. Die Bedeutung gesundheitsrelevanter Aspekte des Lebensstils. Zeitschrift für Soziologie 30: 384-400.
Kunst, A. E., M. del Rios, F. Groenhof und J. P. Mackenbach, 1998: Socioeconomic inequalities in stroke mortality among middle-aged men. An international overview. European Union Working Group on Socioeconomic Inequalities in Health. Stroke 29: 2285-2291.
Lampert, T. und L. E. Kroll, 2006a: Einkommensdifferenzen in der Gesundheit und Lebenserwartung. Quer- und Längsschnittliche Befunde anhand des Sozio-ökonomischen Panels (SOEP). Das Gesundheitswesen 68: 219-230.
Lampert, T. und L. E. Kroll, 2006b: Messung des sozioökonomischen Status in sozialepidemiologischen Studien. S. 295-312 in: M. Richter und K. Hurrelmann (Hg.): Gesundheitliche Ungleichheit Grundlagen, Probleme, Konzepte. Wiesbaden: VS-Verlag.
Lampert, T., Kroll, L. E. und A. Dunkelberg, 2007: Soziale Ungleichheit der Lebenserwartung in Deutschland. Aus Politik und Zeitgeschichte 42: 11-18.
Luy, M., 2006: Differentielle Sterblichkeit. Die ungleiche Verteilung der Lebenserwartung in Deutschland. Rostock. Max-Planck-Institut für demografische Forschung. Diskussionspapier 6: 1-26.
Mackenbach, J. P., 2006: Health inequalities. Europe in profile. Rotterdam: Erasmus MC.
Mackenbach, J. P., V. Bos, O. Andersen, M. Cardano, G. Costa, S. Harding, A. Reid, O. Hemstrom, T. Valkonen und A. E. Kunst, 2003: Widening socioeconomic inequalities in mortality in six Western European countries. International Journal of Epidemiology 32: 830-837.
Mackenbach, J. P., A. E. Kunst, A. E. J. M. Cavelaars, F. Groenhof und J. J. M. Geurts, 1997: Socioeconomic inequalities in morbidity and mortality in western Europe. The Lancet 349: 1655-1659.
Martikainen, P., T. Valkonen und T. Martelin, 2001: Change in male and female life expectancy by social class. Decomposition by age and cause of death in Finland 1971-95. Journal of Epidemiology and Community Health 55: 494-499.
Mielck, A., 2000: Soziale Ungleichheit und Gesundheit. Bern/New York: Huber.
Oeppen, J. und J. W. Vaupel, 2002: DEMOGRAPHY. Enhanced. Broken limits to life expectancy. Science 296: 1029-1031.
Pamuk, E. R., 1985: Social class inequality in mortality from 1921 to 1972 in England and Wales. Population Studies 39: 17-31.
Pischner, R., 2007: Die Querschnittsgewichtung und die Hochrechnungsfaktoren des Sozioökonomischen Panels (SOEP) ab Release 2007 (Welle W). Modifikationen und Aktualisierungen. DIW Data Documentation 22: 1-22.
Razum, O., 2006: Migration, Mortalität und der Healthy-Migrant-Effekt. S. 255-270 in: M. Richter und K. Hurrelmann (Hg.): Gesundheitliche Ungleichheit – Grundlagen, Probleme, Perspektiven. Wiesbaden: VS Verlag.
Reil-Held, A., 2000: Einkommen und Sterblichkeit in Deutschland. Leben Reiche länger? DP Sonderforschungsbereich 504 No. 00-14. http://www.sfb504.uni-mannheim.de/publications/dp00-14.pdf (22.4.2009).
Riley, J. C., 2001: Rising life expectancy. A global history. Cambridge/New York: Cambridge University Press.
Schneider, S., 2007: Ursachen schichtspezifischer Mortalität in der Bundesrepublik Deutschland: Tabakkonsum dominiert alle anderen Risikofaktoren. International Journal of Public Health 52: 39-53.

Schnell, R. und M. Trappmann, 2006: Konsequenzen der Panelmortalität im SOEP für Schätzungen der Lebenserwartung. Arbeitspapier des Zentrum für Quantitative Methoden und Surveyforschung 2/2006: 1-16.
Schnell, R., P. B. Hill und E. Esser, 1999: Methoden der Empirischen Sozialforschung. München: Oldenbourg.
Spieß, M. und M. Kroh, 2008: Documentation of sample sizes and panel attrition in the German Socio Economic Panel (SOEP – 1984 until 2006). DIW Data Documentation 39: 1-38.
Spieß, M. und M. Kroh, 2007: Documentation of the dataset design of the Socio-Economic Panel Study (SOEP). DIW: 1-3. http://www.diw.de/documents/dokumentenarchiv/17/60056/designdoku.pdf (22.4.2009).
Stata Corporation, 2007: Stata Statistical Software. Release 10.0. College Station Texas: Stata Corporation.
Timm, A., Helmert, U. und R. Müller, 2006: Berufsstatus und Morbiditätsentwicklung von Krankenversicherten im Zeitraum 1990 bis 2003. Das Gesundheitswesen 68: 517-525.
Unger, R., 2006: Trends in active life expectancy in Germany between 1984 and 2003 – a cohort analysis with different health indicators. Journal of Public Health 14: 155-163.
Voges, W., U. Helmert, A. Timm und R. Müller, 2004: Soziale Einflussfaktoren von Morbidität und Mortalität. Sonderauswertung von Daten der Gmünder Ersatzkasse (GEK) im Auftrag des Robert Koch-Institutes. Bremen: Zentrum für Sozialpolitik.
Wagner, G. G., J. R. Frick und J. Schupp, 2007: The German Socio-Economic Panel Study (SOEP) – Scope, evolution and enhancements. Schmollers Jahrbuch 127: 139-169.
White, H., 1982: Maximum likelihood estimation of misspecified models. Econometrica 50: 1-25.

Korrespondenzadresse: Lars Eric Kroll
Robert Koch-Institut
- Fachgebiet 24 -
Postfach 65 02 61
13302 Berlin
l.kroll@rki.de

Der Einfluss von Häufigkeitsformaten auf die Messung von subjektiven Wahrscheinlichkeiten

The Impact of Frequency Formats on the Measurement of Subjective Probability

Mandy Beuer-Krüssel und Ivar Krumpal

Zusammenfassung
Die Messung subjektiver Wahrscheinlichkeiten ist ein zentrales Anliegen vieler Bevölkerungssurveys zu selbstberichteter Delinquenz (z. B. ALLBUS 1990 und 2000). Ein bekanntes Problem ist hierbei die subjektive Überschätzung von Risiken im Zusammenhang mit seltenen Ereignissen. Fehler in der Risikoeinschätzung (z. B. ‚Nicht-Berücksichtigung von Basisraten' oder ‚Überschätzung') können sowohl auf kognitive Heuristiken der Befragten als auch auf Eigenschaften der Items zurückgeführt werden. Der erste Teil des Beitrags diskutiert und vergleicht Strategien und Formate der Messung von Wahrscheinlichkeiten, insbesondere Häufigkeiten versus Prozente. Hierbei zeigt sich, dass die Abfrage von Basisraten in Form von Häufigkeiten einen biasreduzierenden Effekt auf die Wahrscheinlichkeitseinschätzung seltener Ereignisse ausübt. Im zweiten Teil des Beitrags werden die theoretischen Vorteile von Häufigkeitsskalen in einen zweistufigen Messansatz subjektiver Wahrscheinlichkeiten überführt und durch ein Methodenexperiment empirisch belegt. Am Beispiel von subjektiven Entdeckungswahrscheinlichkeiten beim Schwarzfahren wird ein Kontexteffekt des Häufigkeitsformats demonstriert, der die Risikoüberschätzung seltener Ereignisse verringert. Ein solches Format könnte in der Praxis der empirischen Sozialforschung vermehrt Anwendung finden.

Abstract
In many surveys on deviant behavior the measurement of subjective probability is an important goal (e. g. German General Social Surveys (ALLBUS) 1990 and 2000). A well-known problem is the subjective overestimation of risks in connection with rare events. Errors in risk assessment (e. g. 'base rate neglect' or 'overestimation') can result from cognitive heuristics of the respondents as well as item characteristics. The first part of the article discusses strategies and formats of probability measurement and, in particular, compares frequencies versus percentages. It can be shown that the availability of base rates in the frequency format reduces bias in subsequent probability assessments of rare events. In the second part, the theoretical advantages of frequency scales are transformed into a two-step measurement procedure of subjective probability, and a bias-reducing effect of the frequency format is empirically demonstrated. A methodical experiment on fare dodging shows that subjective probabilities of being caught for dodging the fare (the rare event) are more accurate if base rates are activated via frequencies. This kind of format could be applied more frequently in empirical social research.

1 Einleitung[1]

> „As scientists and as technologists we should discard the idea of a 'true' or 'objective' probability. Instead, we should think of probability judgements as the result of an individual`s feelings of uncertainty, translated into a numerical response by internal decision processes." (Phillips 1970: 254)

Die Relevanz der Messung von subjektiven Wahrscheinlichkeiten ist – trotz bekannten Problemen der Validität – unbestritten: In den Sozial- und Wirtschaftswissenschaften sowie in medizinischen Surveys bilden subjektive Wahrscheinlichkeiten die Basis für Entscheidungsfindungen, Verhaltensprognosen und Risikoeinschätzungen. So werden Befragte[2] im ALLBUS 1990 und 2000 nach ihrer Wahrscheinlichkeit gefragt, bei verschiedenen Delikten, wie Schwarzfahren oder Steuerhinterziehung, entdeckt zu werden. Die Messung dieser subjektiven Risiken erfolgt mit Kategorialskalen, die über sogenannte *Vague Quantifier* verbalisiert sind.[3] Andere Surveys messen subjektive Risiken, so etwa die subjektive Wahrscheinlichkeit Opfer einer Straftat zu werden, ebenso über *Vague Quantifier* (vgl. Wohlfahrtssurveys 1993, 1998; British Crime Surveys 2004, 2005). Studien der Umfragemethodologie und Kognitionspsychologie zeigen, dass die Verwendung von *Vague Quantifiers* häufig zu einer starken inter- und intrapersonellen Variabilität bei der Interpretation der verbalen Wahrscheinlichkeitsausdrücke und folglich zu Mehrdeutigkeiten und Widersprüchlichkeiten bei der Übertragung der *Vague Quantifiers* in numerische Äquivalente führt (vgl. Bradburn/Miles 1979; Krumpal et al. 2008; Wright et al. 1994).[4] Die vorliegende Arbeit untersucht deshalb alternative Messansätze der Erfassung subjektiver Risiken.

Solche Alternativen sind numerische Skalen (Häufigkeiten oder Prozente), die sich durch eine geringere Anfälligkeit für Fehlurteile (Biases) sowie eine konsistentere Verwendung auszeichnen. Allerdings ist auch die Risikoeinschätzung

1 Wir danken Veronika Andorfer, Dorothea Böhr, Heiko Rauhut und Thomas Voss für anregende Gespräche, konstruktive Hinweise sowie ihre Unterstützung bei diesem Forschungsprojekt. Wir danken zudem den beiden anonymen Gutachtern für ihre wertvollen Verbesserungsvorschläge.

2 Im Folgenden wird die männliche Form verwendet. Dies dient ausschließlich der sprachlichen Vereinfachung.

3 Item zu Schwarzfahren im ALLBUS 2000: „Stellen Sie sich vor, Sie würden ein öffentliches Verkehrsmittel benutzen, ohne einen gültigen Fahrausweis zu besitzen. Wie wahrscheinlich wäre es Ihrer Ansicht nach, daß ein Kontrolleur Sie dabei entdecken würde? Benutzen Sie für Ihre Antwort bitte die Liste." Antwortvorgaben als Kategorialskala mit *Vague Quantifiers*: ‚Sehr unwahrscheinlich', ‚Eher unwahrscheinlich', ‚Ungefähr 50 zu 50', ‚Eher wahrscheinlich', ‚Sehr wahrscheinlich'.

4 Je nach Personen-, Item-, und Kontexteigenschaften variieren die Zuweisungen der numerischen zu den verbalen Risikoeinschätzungen, sodass eine Vergleichbarkeit zwischen Personen, Items oder Surveys in vielen Fällen problematisch ist. In einem Experiment von Reuband (2002) ordnen Befragte den *Vague Quantifiers* numerische Wahrscheinlichkeiten zu (Prozente oder Häufigkeiten). Hierbei zeigen sich Inkonsistenzen in den Zuordnungen. Eine Anwendung numerischer Skalen (insbesondere Häufigkeiten) wird deshalb empfohlen.

auf numerischen Skalen, insbesondere auf Prozentskalen, nicht unproblematisch (Hoffrage et al. 2000). So stehen subjektive Einschätzungen der Befragten häufig im Widerspruch zu mathematischen Wahrscheinlichkeitsgesetzen.[5] Subjektive Wahrscheinlichkeiten korrespondieren oftmals nur unzureichend mit tatsächlichen Risiken (Johnson/Bruce 2001). Befragte unterschätzen systematisch die Wahrscheinlichkeit häufiger (bzw. alltäglicher) und überschätzen die Wahrscheinlichkeit seltener Ereignisse (*Overestimation*; vgl. Fischhoff et al. 2000; Pinkerton et al. 2000; Warr 1980; Yamagishi 1997). Zudem ist bekannt, dass Befragte die Wahrscheinlichkeit eines Ereignisses überwiegend danach beurteilen, wie repräsentativ es für ein jeweiliges Vergleichsobjekt ist. Entsprechend ignorieren sie wichtige Basisrateninformationen und sind unempfindlich gegenüber der Stichprobengröße (*Base Rate Neglect*; vgl. Tversky/Kahneman 1974).

Zunächst werden nachfolgend die Vorzüge von Häufigkeitsskalen gegenüber Prozentskalen theoretisch hergeleitet und diskutiert (vgl. Brase 2002; Gigerenzer/Hoffrage 1995; Gigerenzer 1998). Erstere können das wohl größte Problem der Wahrscheinlichkeitsbeurteilung eindämmen: Die subjektive Überschätzung des Risikos seltener Ereignisse. Auch andere Fehler, wie zum Beispiel die *Nicht-Berücksichtigung von Basisraten* oder die Fehlbeurteilung logischer Verknüpfungen, sind über die Häufigkeitsskala weniger stark ausgeprägt als über die Prozentskala. Die theoretischen Vorteile von Häufigkeitsskalen werden in einen zweistufigen Messansatz subjektiver Wahrscheinlichkeiten überführt und durch ein Methodenexperiment empirisch belegt. Am Beispiel von subjektiven Entdeckungswahrscheinlichkeiten beim Schwarzfahren wird ein Kontexteffekt des Häufigkeitsformats demonstriert, der die Risikoüberschätzung seltener Ereignisse verringert. Ein solches Format könnte in der Praxis der empirischen Sozialforschung vermehrt Anwendung finden.

5 Als Beispiel kann hier die Verletzung der Konjunktionsregel angeführt werden. Die Konjunktionsregel besagt, dass die Wahrscheinlichkeit einer logischen UND-Verknüpfung zweier Ereignisse nicht größer sein kann als die einer ihrer Komponenten: $P(A \cap B) \leq P(A)$ und $P(A \cap B) \leq (B)$. Trotz der *Schlüssigkeit* dieses Wahrscheinlichkeitsgesetzes schätzen Befragte Ereigniswahrscheinlichkeiten sowie deren Verknüpfungen falsch ein. Veranschaulicht wird diese Fehleinschätzung durch das sogenannte ‚Linda Problem'. Befragte werden zunächst mit Beschreibungen einer fiktiven Person konfrontiert und anschließend gebeten, die Wahrscheinlichkeit einzuschätzen, mit der ‚Linda' unterschiedliche Berufe und Hobbys ausübt. Es zeigt sich, dass die Verknüpfungen von Aktivitäten von der Mehrzahl der Befragten als wahrscheinlicher eingestuft werden als die einzelnen Aktivitäten selbst. Dies ist dann der Fall, wenn die Konjunktion als *repräsentativer* für die beschriebene Person eingestuft wird als ihre Komponenten (Tversky/Kahneman 1983). Beispielsweise wird ‚Linda' als selbstbewusste, ältere Single-Frau beschrieben, die Philosophie studierte sowie als Studentin an Themen wie Diskriminierung und sozialer Gleichberechtigung großes Interesse zeigte. Befragte wurden gefragt, ob es wahrscheinlicher sei, dass ‚Linda' Bankangestellte [P(B)] ist oder dass sie Bankangestellte [P(B)] UND innerhalb der Frauenbewegung aktiv [P(A)] ist [$P(A \cap B)$]. Da die Konjunktion als wesentlich *repräsentativer* für die Persönlichkeitsbeschreibung von ‚Linda' eingestuft wird, überschätzt die Mehrzahl der Befragten deren Wahrscheinlichkeit. Das empirisch beobachtete Muster war hierbei wie folgt: $P(A) > P(A \cap B) > P(B)$ (vgl. Tversky/Kahneman 1983: 297f.).

2 Fehleinschätzungen von Risiken

2.1 Kognitiver Ansatz

Welche Mechanismen verursachen die zu hohen bzw. zu niedrigen Einschätzungen von Risiken? Hierzu wird zunächst eine terminologische Festlegung vorgenommen: (1) *Allgemeine Wahrscheinlichkeiten* beziehen sich auf Risiken der oder des ‚typischen Anderen'. Befragte beurteilen demnach das Risiko, dass ein Ereignis X für eine Menge von Personen oder eine Einzelperson (nicht sie selbst) auftritt. (2) Der Terminus *subjektive Wahrscheinlichkeit* bezeichnet dagegen das Risiko des Befragten, dass ein Ereignis X für ihn persönlich eintritt. (3) Von beiden unabhängig bezeichnet die *tatsächliche* oder *objektive Wahrscheinlichkeit* das Risiko, mit dem ein Ereignis X in der Realität stattfindet.

Kognitive Ansätze erklären Fehleinschätzungen von Risiken über eine fehlerhafte Wahrnehmung und Informationsverarbeitung der Befragten. Sie nehmen an, dass menschliche Wahrnehmung von Rationalitätsannahmen abweicht, und dass subjektive Wahrscheinlichkeiten systematische Unterschiede zu den normativen Standards der Wahrscheinlichkeitstheorie aufzeigen (Kahneman/Tversky 1973, 1982a). Das Heuristik- und Bias-Konzept von Kahneman und Tversky kann als Grundlage des Kognitiven Ansatzes angesehen werden:

> „In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors." (Kahneman/Tversky 1973: 237)

Die *Repräsentativitätsheuristik* beschreibt die Bewertung des Ausmaßes der Übereinstimmung zwischen einem Sample und einer Population, einem Beispiel und einer Kategorie, d. h. inwieweit die Objekt- oder Ereigniseigenschaften den Kategorieneigenschaften ähneln.[6] Sie beruht nicht ausschließlich auf dem Prinzip der Gleichheit, sondern kann ebenso kausale Annahmen und Wechselwirkungsbeziehungen widerspiegeln. Ihre Verwendung führt oftmals zu Insensibilität gegenüber vorhergehenden Wahrscheinlichkeiten (Basisraten) und Stichprobengrößen sowie

6 Die Gleichheit von Sample und Population wird z. B. über den Vergleich von Verhältnissen ermittelt. Dabei wird das Mengenverhältnis von Objekten in Beziehung gesetzt. Geht man bei Geburten von einer natürlichen Geschlechterverteilung von Mädchen zu Jungen von 50:50 aus, so ist es nicht verwunderlich, dass Befragte eine Geschlechterverteilung von z. B. 3:3 als wahrscheinlicher einschätzen als eine von 1:5. Auch das Prinzip der Zufälligkeit scheint bei letzterem Beispiel verletzt (vgl. Kahneman/Tversky 1972: 432ff.).

zu Unempfindlichkeit hinsichtlich der Vorhersagbarkeit von Ereignissen[7] (Tversky/Kahneman 1974: 1124ff.). Dagegen generiert die *Verfügbarkeitsheuristik* Wahrscheinlichkeitseinschätzungen in Abhängigkeit von der Leichtigkeit, mit der relevante Beispiele erinnert oder vorgestellt werden können.[8] Die Anzahl relevanter Fälle, die schnell und ohne Aufwand kognitiv verfügbar sind, bilden demnach die Grundlage zur Einschätzung von Wahrscheinlichkeiten. Die Leichtigkeit des Erinnerns bzw. der Konstruktion von ähnlichen Szenarien entspricht jedoch häufig nicht der tatsächlichen Auftrittswahrscheinlichkeit der zu beurteilenden Ereignisse (Tversky/Kahneman 1974: 1127f.). Als Folge der Anwendung dieser zwei Heuristiken kann es zu verschiedenen Fehleinschätzungen kommen.[9]

Fehleinschätzungen resultieren u. a. aus der *Nicht-Berücksichtigung von Basisraten*, d. h. Informationen zu vorangegangenen Wahrscheinlichkeiten, Populationsverhältnissen oder zentralen Tendenzen. Mit der *Nicht-Berücksichtigung von Basisraten* (Kahneman/Tversky 1973) einher geht die Überbewertung von spezifischen Informationen über das Zielereignis, über welches Vorhersagen getroffen werden. Jede subjektive Wahrscheinlichkeitseinschätzung basiert auf (1) Einzelfallinformationen sowie (2) Basisraten. Die Gewichtung der beiden Faktoren ist abhängig von der Diagnostizität, die ihnen der Befragte zuschreibt. In vielen Fällen erfolgt eine Einschätzung lediglich basierend auf der *Repräsentativität* der Einzel-

7 Ein Beispiel wird bei Tversky und Kahneman (1974: 1126) beschrieben: Zwei Gruppen von Befragten werden Beschreibungen der Leistung von Lehramtsstudenten in deren praktischer Übungsstunde gegeben. Die eine Gruppe soll die Qualität der beschriebenen Unterrichtsstunde evaluieren, während die andere den Ruf/die Kompetenz der Lehramtsstudenten fünf Jahre nach dieser Übungsstunde vorhersagen soll. Die abgegebenen Urteile unter diesen beiden Bedingungen waren identisch. Die Vorhersage des entfernt liegenden Ereignisses war identisch mit den Evaluationen zur Qualität der Übungsstunde. Die Prognosen richteten sich demnach ausschließlich an der begrenzten Vorhersagekraft dieses Einzelereignisses aus.

8 Ihr empirischer Nachweis kann beispielsweise in Worthäufigkeits-Experimenten erbracht werden. Es werden die Wörter als häufiger in einem Text auftretend eingestuft, die leichter vorzustellen sind. Ist es wahrscheinlicher, dass ein zufällig aus einem englischen Text gezogenes Wort mit einem ‚k' beginnt oder ein ‚k' an dritter Stelle hat? Da Wörter mit dem Anfangsbuchstaben ‚k' einfacher abrufbar/vorstellbar sind, werden diese als häufiger vorkommend eingestuft. Die Realität zeigt jedoch den umgekehrten Fall (Tversky/Kahneman 1973: 211f.). Zentral ist hierbei die subjektiv erfahrene Leichtigkeit der Erinnerung: „Presumably, they monitor their cognitive processes and infer that a given class of events is frequent when relevant exemplars are easy to bring to mind but rare when exemplars are difficult to bring to mind" (Schwarz 1998: 88).

9 Neben der Erklärung des Auftretens von systematischen Fehleinschätzungen über die genannten kognitiven Heuristiken kann als weiterer Erklärungsansatz das Argument Essers (1986) angebracht werden, wonach Antwortverzerrungen stärker im Zusammenhang mit dem Frageinhalt betrachtet werden sollten. Laut Esser treten Antwortverzerrungen vor allem dann auf, wenn die Einstellungsintensität der Befragten schwach ausgeprägt ist, und somit die Relevanz der Frage für die Befragten fehlt bzw. gering ist. So könnten die beobachteten Fehleinschätzungen in den Worthäufigkeits-Experimenten (siehe Fußnote 8) ebenso vor dem Hintergrund einer möglicherweise fehlenden Relevanz der Fragestellung für die Befragten beurteilt werden.

fallinformationen, wohingegen Basisraten kaum Einfluss zeigen. In Experimenten wurden Befragte mit Persönlichkeitsbeschreibungen und Listen mit Studienfächern sowie der Frage konfrontiert, wie groß die Wahrscheinlichkeit sei, dass die fiktive Person ihr Studium im jeweiligen Fach abgeschlossen hat. Es zeigte sich, dass die Abschlussfächer in dem Maße als wahrscheinlich eingestuft wurden, in dem das Persönlichkeitsprofil mit den stereotypen Mitgliedern des jeweiligen Faches korrespondierte, also repräsentativ für dieses war. Dass die Fächer Unterschiede in der reellen Studentenanzahl aufweisen (also verschiedene Basisraten), hatte keinen Einfluss auf die Wahrscheinlichseinschätzungen.[10]

Hier liefert der Kognitive Ansatz jedoch bereits Lösungsstrategien für das Problem der *Nicht-Berücksichtigung von Basisraten*. Dieses könne folglich vermindert werden, indem Basisraten nicht nur als willkürliche Angaben über die Verteilung in der Gesamtpopulation vermittelt werden, sondern ihre Diagnostizität, Relevanz, Spezifizität und Kausalität hervorgehoben wird (Bar-Hillel 1980: 216). Ziel der Messung subjektiver Wahrscheinlichkeiten sollte es demnach sein, die individualisierten Einzelfallinformationen *und die Basisraten* als gleichwertige, relevante Informationen zu kommunizieren, damit insbesondere letztere bei der Wahrscheinlichkeitseinschätzung berücksichtigt werden (Ajzen 1977; Bar-Hillel 1980; Gigerenzer et al. 1988; Ginosar/Trope 1980; Nisbett/Ross 1980). Befragte greifen eher dann auf gegebene Basisraten-Informationen zurück, wenn inkonsistente und nicht kausale Einzelfallinformationen, die nur schwer in Bezug auf existierende stereotype Repräsentationen zu interpretieren sind, zu Mehrdeutigkeiten führen (Ginosar/Trope 1980; Hendrickx et al. 1989; Schwarz et al. 1991). Koehler (1996) kann zeigen, dass von den Befragten direkt erfahrene oder selbst generierte Basisraten eher genutzt werden als statistisch vermittelte, und dass Basisraten in Aufgabenkontexten mit

10 Die fiktive Person ‚Tim' wird als intelligent, ordnungsliebend und systematisch sowie in ihrem Schreibstil mechanisch beschrieben. ‚Tim' ist wenig kreativ, hat wenig Sympathie für andere Menschen und generell wenig Spaß an der Interaktion mit anderen. Die normativen Regeln der Vorhersage (welches ‚Tims' Abschlussfach sei) wurden durch die Anwendung der *Repräsentativitätsheuristik* von der Mehrheit der Befragten verletzt. Über 95 % schätzten die Wahrscheinlichkeit, dass ‚Tim' Informatikabsolvent ist, als größer ein als die, dass er seinen Abschluss in den Geistes- oder Erziehungswissenschaften hat. Das obwohl ihnen klar war, dass es sehr viel mehr Absolventen in den letzteren beiden Fächern gibt (wie es die Basisraten vorgeben). Siehe hierzu die Experimentalbeschreibungen bei Kahneman und Tversky (1973: 238f.). Ob die Basisrate tatsächlich ignoriert oder ihre Diagnostizität von den Befragten lediglich unterschätzt wird, müsste darüber hinaus dadurch geprüft werden, ob sich die Einschätzung bei zusätzlich angegebener Information über Basisraten im Vergleich zum Fehlen dieser Information verändert oder nicht.

einer natürlichen und alltäglichen Problematik stärker verwendet werden als die im Kontext künstlicher oder konstruierter Probleme.[11]

Das Überschätzen der Auftrittswahrscheinlichkeit seltener sowie das Unterschätzen der Wahrscheinlichkeit alltäglicher bzw. häufiger Ereignisse gehört ebenfalls zu den systematischen Fehlern, welche sich aufgrund der Nutzung von Heuristiken ergeben können. Beispielsweise werden bei der Frage nach der Wahrscheinlichkeit verschiedener Todesursachen die seltenen (Unfälle, Selbstmord, Feuer) überschätzt, während die alltäglichen (Diabetes, Schlaganfall) unterschätzt werden (Slovic et al. 2004). *Über-* und *Unterschätzung* sind außerdem stark affektiv bestimmt. Insbesondere Ereignisse, die mit schwerwiegenden Folgen verbunden sind oder in den Medien überbetont werden, erfahren die stärkste Überschätzung (Warr 1980). Begründet wird diese Verzerrung über kognitive Mechanismen des Erinnerns und der Betroffenheit: Die angenommene Wahrscheinlichkeit des Eintretens eines Ereignisses ist abhängig davon, wie *leicht* dieses Ereignis *erinnert* oder vorgestellt werden kann. Je eher ein Ereignis *kognitiv verfügbar* ist und je folgenträchtiger es ist, desto eher wird seine Auftrittswahrscheinlichkeit überhöht beurteilt (Pinkerton et al. 2000).[12]

2.2 Inputorientierter Ansatz

Die Ursachen der Beurteilungsverzerrungen sieht der Inputorientierte Ansatz nicht primär bei den Individuen, sondern bei den Eigenschaften der Items. Die funktionalen Beurteilungsanlagen der Befragten sehen sich mit einem suboptimalen Input konfrontiert, der zu den fehlerhaften Risikoeinschätzungen führt. Dabei spielen vorangehende Fragen, die Itemformulierung und das Skalenformat eine wichtige Rolle (Cosmides/Tooby 1996; Gigerenzer et al. 1991; Koehler 1996).

11 Persönlich erfahrene Informationen (Basisraten) sind anschaulicher und hervorstechender, damit also schneller verfügbar als ‚gelernte' Informationen. Zudem haben Menschen mehr Vertrauen in selbst generierte Basisraten, vor allem wenn diese über Erfahrungen aus erster Hand erworben wurden. Ausführliche Beschreibungen sowie methodische Anleitungen finden sich bei Koehler (1996).

12 In Experimenten mit US-Studenten zum Thema HIV-Infektion zeigte sich, dass diese die Wahrscheinlichkeit einer Ansteckung mit HIV stark überschätzten (im Vergleich zu den tatsächlichen Risiken). Begründet wurden dieser Effekt mit der extensiven Berichterstattung zur HIV-Epidemie in den Medien und den schweren Konsequenzen einer HIV-Infektion. Das Thema HIV wurde durch eine starke Medienpräsenz kognitiv verfügbar gemacht, was in vielen Fällen zu einer Überschätzung der tatsächlichen (geringen) Risiken der allgemeinen Bevölkerung geführt hat (Pinkerton et al. 2000: 16f.).

Prozentskalen

Im Gegensatz zu kategorialen Skalen mit einer geringen Anzahl verbaler Antwortkategorien erlauben Prozentskalen (üblicherweise Skalen von 0 bis 100 %, also 101 wählbaren Alternativen) eine präzisere Kommunikation von subjektiven Wahrscheinlichkeiten (Schnell/Kreuter 2000; Coutts-Heller 2002: 8).[13] Die Kommunikation von Wahrscheinlichkeiten im Prozentformat ist dennoch fehleranfällig, da die von den Befragten angegebenen allgemeinen oder subjektiven Wahrscheinlichkeiten häufig von den tatsächlichen Risiken abweichen. Oftmals werden Ereigniswahrscheinlichkeiten im Prozentformat überschätzt (Black et al. 1995; Dominitz/Manski 1997). Fischhoff et al. (2000) konnten in ihrer Studie über die Erwartungen von Jugendlichen bezüglich wichtiger Lebensereignisse zeigen, dass die Befragten sowohl die Wahrscheinlichkeit positiver als auch negativer Ereignisse[14] überschätzen. *Überschätzung* kann unter anderem mit der übermäßigen Nutzung von 50 %-Antworten erklärt werden. Befragte verwenden diesen Wert weniger in seiner natürlichen numerischen Form, sondern interpretieren ihn vielmehr als Kategorie ‚keine Ahnung'. In einer Studie zu Schwangerschaftsrisiken deuteten die Befragten die Wahrscheinlichkeit von 50 % für einen genetischen Defekt beim Kind als ‚alles ist möglich' (Lippman-Hand/Fraser 1979: 118f.). Die vermehrte ‚Flucht in die Mitte', welche vor allem durch die Unsicherheit der Eltern begründet ist, führte so im prozentualen Wahrscheinlichkeitsformat zu einer Überschätzung der Risiken.[15] Ein weiterer Grund für die überproportionale Verwendung der Mittelposition der Skala ist im sozialen Vergleich zu finden. Befragte orientieren sich bei der Kommunikation ihrer subjektiven Wahrscheinlichkeit häufig an der Referenzpopulation. Dabei interpretieren sie die Skalenmitte als ‚normalen' oder Durchschnittswert für diese Population und ordnen sich entsprechend ein (Schwarz et al. 1985; Wright et al. 1994).

Die Überschätzung von kleinen Ereigniswahrscheinlichkeiten kann ebenso durch die Präferenz der Befragten für runde Zahlen erklärt werden. Befragte neigen dazu ihre Antworten auf Werte wie 0, 10, ..., 50, ... und 100 % zu runden, statt die verfeinerten Wahrscheinlichkeitsangaben auf der Prozentskala zwischen

13 „For example, when subjects say it is 'likely' that someone will get a headache as a result of using a certain medicine and say it is 'likely' that someone else will develop a more severe symptom as a result of using another medicine, they are not referring to the same numeric probability. The 'likely' probability of the less severe symptom will be higher (...)." (Coutts-Heller 2002: 7). Eine weiterführende Diskussion der Probleme von *Vague Quantifiers* findet sich bei Krumpal et al. (2008).

14 Zum Beispiel: Positiv – Abschluss des College bis zum 30. Lebensjahr. Negativ – Inhaftierung oder Tod innerhalb des nächsten Jahres.

15 Eine Möglichkeit zur Reduzierung von 50 %-Antworten ist die Einführung einer zusätzlichen Antwortkategorie ‚keine Ahnung' oder ‚neutral' (Coutts-Heller 2002: 9).

0 und 100 % im vollen Umfang auszunutzen (Dominitz/Manski 1997: 270). Befragte benutzen die Prozentskala häufig wie eine Kategorialskala, die von 0 bis 10 reicht, und geben Antworten oftmals als Vielfache von 10. Da in den Studien aber meist Ereignisse untersucht werden, deren Wahrscheinlichkeiten (weit) unter 10 % liegen, könnte ein solches Befragtenverhalten zu den beobachteten Überschätzungen führen (Slovic/Monahan 1995). Subjektive Wahrscheinlichkeitseinschätzungen spezifischer Ereignisse mit binärem Ausgang (wie die Entdeckungswahrscheinlichkeit eines Deliktes im kommenden Jahr) sind im Prozentformat fehleranfällig. Diese Probleme können auf evolutionäre Prozesse in der Entwicklung der menschlichen kognitiven Fähigkeiten zurückgeführt werden. Menschen besitzen demnach keinen angeborenen kognitiven Mechanismus, der ihnen die Verarbeitung von Einzelereigniswahrscheinlichkeiten im Einklang mit den normativen Standards der Wahrscheinlichkeitstheorie ermöglicht (Brase et al. 1998; Cosmides/Tooby 1996; Gigerenzer 1996b, 1998, 2000).

Häufigkeitsskalen

Laut Gigerenzer (2000) liege der Hauptschwachpunkt der kognitiven und motivationalen Erklärungen für die Fehler bei der Risikoeinschätzung darin begründet, dass diese die Struktur der Aufgabe und deren Beziehung zur Struktur der natürlichen Umwelt des Befragten ausblenden würden. Die Fehleinschätzungen der Befragten seien keine Fehler im statistischen Denken, sondern lediglich bedingt durch eine suboptimale Konstruktion von Informations- und Antwortformaten innerhalb der Studien:

> „The mind acts as if it were a frequentist; it distinguishes between single events and frequencies in the long run – just as probabilists and statisticians do. Despite the fact that researchers in the 'heuristics and biases' program routinely ignore this distinction fundamental to probability theory when they claim to have identified 'errors', it would be foolish to label these judgments 'fallacies'." (Gigerenzer 2000: 253f.)

Die Verwendung von Häufigkeitsformaten würde demnach, verglichen mit Einzelereigniswahrscheinlichkeiten im Prozentformat, zu einer klareren Einschätzung von Risiken führen. Häufigkeiten stellen eine intuitivere Metrik bei der Risikobeurteilung dar als Prozente (Gigerenzer/Hoffrage 1995; Schapira et al. 2001). Der Mensch verfügt über induktive Denkmechanismen, die rationale Prinzipien zum Ausdruck bringen. Diese werden aber nur dann korrekt angewendet, wenn Informationen in Häufigkeiten präsentiert und abgerufen werden. Die Präferenz für natürliche Häufigkeiten als Datenbasis und Skalenformat ist dabei evolutionär bedingt. Eine Datenbank, auf die der Mensch schon in Urzeiten zurückgreifen konnte, waren

seine eigenen, zählbaren Beobachtungen. Die beobachteten und aufsummierten Ereignishäufigkeiten dienten der Verbesserung von Entscheidungsprozessen, wie dem Finden erfolgversprechender Jagdgebiete. Beim *natürlichen Sampling* werden keine Basisraten benötigt um konditionale Wahrscheinlichkeiten zu ermitteln (vgl. Brase 2002; Brase et al. 1998; Gigerenzer 1996b; Gigerenzer/Hoffrage 1995).[16] In konstruierten Experimenten braucht es jedoch Basisraten um die fehlenden Ausgangsinformationen zu liefern. Die Kommunikation von Risiken über Prozente, Chancen oder Odds-Ratios enthält genau diese Informationen nicht. Im Gegensatz zu normalisierten (relativen) Häufigkeiten und Wahrscheinlichkeiten beinhalten natürliche Häufigkeiten Informationen über Basisraten, weshalb sie als Informationsbasis für die Wahrscheinlichkeitsbeurteilung adäquater sind; sie stimmen mit den natürlichen Enkodierungsmechanismen[17] überein (Cosmides/Tooby 1996; Gigerenzer 2000; Gigerenzer/Hoffrage 1999; Hoffrage et al. 2000, 2002; Jones et al. 1995). Natürliche Häufigkeiten sind (verglichen mit Prozenten) weniger abstrakt, einfacher zu verstehen sowie zu visualisieren und benötigen weniger kognitiven Aufwand bei der Einschätzung von Risiken (Brase 2002; Gigerenzer 1996b, 2000; Griffin/Buehler 1999). Zudem erhöhen sie die Anwendung der *Bayes Regeln*.[18]

Die Informationspräsentation in Häufigkeiten statt Prozenten steigert die Beurteilungsleistung der Befragten. Studien zur Qualität von Antwortskalen im Hinblick auf Wahrscheinlichkeitseinschätzungen zeigen, dass das Ersetzen von Prozent- durch Häufigkeitsskalen Fehleinschätzungen deutlich verringert und so zu Leistungssteigerungen bei der Genauigkeit der Wahrscheinlichkeitseinschätzungen führt (Brase et al. 1998; Evans et al. 2000; Gigerenzer et al. 1991; Hoffrage et al. 2000). Während Basisraten bei Prozentinformationen und bei Abfrage auf einer Prozentskala kaum von den Befragten genutzt werden, kann die *Nicht-Berücksichtigung von Basisraten* durch die Präsentation von natürlichen Häufigkeiten und die Nutzung von Häufigkeitsskalen als Antwortmöglichkeit reduziert werden (Gigerenzer 1998,

16 *Natürliches Sampling* bezeichnet die aufeinanderfolgende und fortlaufende Erfassung von Informationen über das Aktualisieren von Ereignishäufigkeiten, also das Zählen von erlebten oder überlieferten Objekten und Ereignissen (Gigerenzer/Hoffrage 1995: 686).

17 Fehler, welche sich bei der Einschätzung von Wahrscheinlichkeiten ergeben, seien demnach vorwiegend auf das Informationsformat zurückführbar (vgl. Gigerenzer/Hoffrage 1995: 685ff.). Die von Menschen erworbenen mathematischen Algorithmen seien lediglich für bestimmte Repräsentationen angelegt – die Eingangsinformationen müssten demnach im richtigen Format vorliegen, damit die Algorithmen optimal arbeiten können (wie auch ein Taschenrechner **nicht** mit binären Zahlen arbeiten kann).

18 Die Berechnung von Ereigniswahrscheinlichkeiten über Bayesianische Algorithmen fällt im Häufigkeitsformat leichter, da weniger und einfachere Rechenschritte notwendig sind. Detaillierte Informationen und Beispielstudien zu den *Bayes Regeln* und ihrem Zusammenhang zum Informationsformat geben Gigerenzer (1996a, 1996b, 1998, 2000), Gigerenzer/Hoffrage (1995, 1999), Hoffrage et al. (2002) sowie Mellers/McGraw (1999).

2000; Gigerenzer/Hoffrage 1999; Price 1998). Das Überschätzen seltener und Unterschätzen häufiger Ereignisse kann durch die Präsentation von Informationen im Häufigkeitsformat nicht vollständig unterbunden werden. Jedoch können Häufigkeiten das Ausmaß dieser Fehleinschätzungen verringern und zu realistischeren Werten führen (Teigen 1974: 62).

3 Möglichkeiten der Biasreduktion

Häufigkeitsskalen können vor allem bei der Ermittlung von Basisraten, also beispielsweise der *allgemeinen* Auftrittswahrscheinlichkeit eines Ereignisses für eine Population, viele Vorteile bieten. Basisraten stehen jedoch nicht immer im Mittelpunkt des Interesses. Vielmehr sollen individuelle Entscheidungsprozesse und Handlungen erklärt werden, weshalb die Abfrage *subjektiver* Wahrscheinlichkeiten notwendig ist (z. B. Viktimisierungs- und Entdeckungswahrscheinlichkeiten). Da deren Messung mit Häufigkeitsskalen nur schwer realisierbar ist (Coutts-Heller 2002: 12), kann lediglich auf Prozentskalen zurückgegriffen werden. Hierbei wäre es von großem Vorteil den Beurteilungsprozess dergestalt zu beeinflussen, dass die Nutzung der Prozentskala mehr im Einklang mit der Nutzung der Häufigkeitsskala steht. Studien zu Kontexteffekten bei der Einstellungsmessung zeigen, dass formal gleiche Fragen in Abhängigkeit des Kontextes zu unterschiedlichen Antworten führen (Schwarz/Sudman 1992). Über die Variation des Kontextes wird also die Beantwortung von Folgefragen beeinflusst.

3.1 Antwortprozess und Kontexteffekte

In Surveys kann der Prozess der Beantwortung von Fragen in vier Phasen unterteilt werden: (1) Informationsaufnahme und Verstehen des Frageinhalts, (2) Abrufen relevanter Informationen aus dem Gedächtnis, (3) Benutzung der erinnerten Informationen zur Beantwortung der Frage sowie (4) Selektion und Wiedergabe einer Antwort (Tourangeau et al. 2000: 7). Informationsverarbeitung und Handeln sind Prozesse, die häufig unbewusst erfolgen. Gewissermaßen wird automatisch ein Konstrukt gewählt. Welches Konstrukt verfügbar ist, ist abhängig von aktiven Erinnerungshinweisen – unabhängig davon, ob sich der Befragte dieser Operation bewusst ist. Verfügbarkeit beschreibt dabei die Leichtigkeit, mit der eine Episode, emotionale Reaktion, vorangegangene Beurteilung, Wissensstruktur oder ein kognitives Konstrukt ins Bewusstsein gerufen wird (Feldman 1992: 51ff.). Die Erinnerung relevanter Informationen wird bestimmt von der Frageformulierung, den Frageinstruktionen, der Verfügbarkeit früherer Überlegungen und Urteile sowie dem Inhalt vorangegangener Items (Tourangeau 1992: 36).

Kontexteffekte sind Antworteffekte, die durch eine oder mehrere vorangegangene Fragen (und Antworten) oder durch die *Skalen* vorhergehender Fragen hervorgerufen werden (Billiet et al. 1992: 131). Diese vorangegangenen Fragen haben zwei Funktionen: (1) Informationen zu aktivieren, sodass diese zu einem späteren Zeitpunkt leichter (automatisch und unbewusst) abgerufen werden können, und (2) eine Informationsbasis zu liefern, über die sich die Fragebedeutung erschließt. Über einen *Priming Effekt* sind Informationen, die entweder direkt in der vorangegangenen Frage enthalten sind oder vom Befragten bei deren Beantwortung aktiviert wurden, zugänglicher und strahlen auf die Folgefrage aus (Strack 1992: 25). Eine Beurteilung, die einmal abgegeben wurde, dient entsprechend als *Anker*, dem nachfolgende Beurteilungen angeglichen oder kontrastiert werden (Knowles et al. 1992: 223).

3.2 Gesteuerte Kontexteffekte bei der Messung subjektiver Wahrscheinlichkeiten

Die Generierung subjektiver Wahrscheinlichkeiten hinsichtlich des Eintretens seltener Ereignisse ist häufig geprägt von *Überschätzung*, d. h. Befragte schätzen diese Wahrscheinlichkeiten im Vergleich zu objektiven Standards zu hoch ein (Pinkerton et al. 2000; Teigen 1974; Warr 1980).[19] Eine Erklärung für diesen Bias ist, dass Befragte oftmals Basisraten, die die allgemeine Wahrscheinlichkeit repräsentieren, nicht in ihre Beurteilungen einschließen[20], obwohl sie einen adäquaten Anker darstellen. Durch gezielte Manipulation des Befragungskontextes kann die Verfügbarkeit solch ‚erwünschter' Informationen jedoch erhöht und so die Beurteilung des Befragten beeinflusst werden.[21]

Kreuter (2002) stellt fest, dass die Wahrnehmung von Wahrscheinlichkeiten und Risiken ein mehrdimensionales Konstrukt ist. Bei den Überlegungen zu ihren eigenen Risiken nannten Befragte „zunächst die allgemeinen Risiken, besannen sich auf ihre üblichen Alltagsroutinen und leiteten daraus ihre Antwort auf das eigene Risiko ab" (Kreuter 2002: 228). Da jedoch nicht davon ausgegangen werden kann, dass alle Befragten diese kognitive Kalkulation ausführen, somit unterschiedliche

19 Aus den theoretischen Überlegungen von Esser (1986) könnte zudem argumentiert werden, dass das Ausmaß der Fehleinschätzungen vom Frageinhalt moderiert wird. Demnach wären bei Befragten mit einer stark ausgeprägten Einstellungsintensität geringere Fehleinschätzungen zu erwarten.

20 Viel wichtiger noch: Basisraten werden in vielen Studien als Referenzanker weder abgefragt noch vorgegeben.

21 Beispielstudien zu Kontexteffekten bei der Wahrscheinlichkeitsmessung finden sich u. a. bei Coutts-Heller (2002), Hoorens/Buunk (1993), Rottenstreich/Tversky (1997) sowie Windschitl (2002).

Interpretationsgrundlagen bestehen würden und folglich die Ergebnisse schwer vergleichbar wären, ist es sinnvoll, die Abfrage allgemeiner Wahrscheinlichkeiten (Basisraten) generell *vor* der subjektiver Wahrscheinlichkeiten vorzunehmen. Aufgrund der erhöhten kognitiven Verfügbarkeit würde dies die Nutzung der Basisraten bei der individuellen Wahrscheinlichkeitseinschätzung erhöhen und damit Fehleinschätzungen verringern. Dabei ist es besonders hilfreich, die Basisraten nicht lediglich vorzugeben, sondern von den Befragten zu erfragen. So werden nicht (scheinbar) willkürlich Werte festgelegt, sondern die subjektiven Annahmen und Erfahrungen der Befragten berücksichtigt, was die Glaubhaftigkeit, Relevanz und Eindeutigkeit der Basisraten unterstreicht und ihre Weiterverwendung in Folgefragen zusätzlich verstärken kann (vgl. Koehler 1996).

3.3 Allgemeine Hypothesen

In einem ersten Schritt führt die Abfrage allgemeiner Wahrscheinlichkeiten (Ankerfrage) zur Aktivierung von Informationen über die Verteilung bestimmter Eigenschaften oder Ereignisse in der Referenzpopulation. In einem zweiten Schritt nutzen die Befragten diese selbst generierten Basisrateninformationen zur Beurteilung ihrer subjektiven Wahrscheinlichkeit (Zielfrage). Somit liegt ein Kontexteffekt der vorangegangenen Ankerfrage auf die nachfolgende Zielfrage vor. Zur Abfrage von Basisraten sollten zudem Häufigkeitsskalen benutzt werden. Denn ihre Verwendung führt zu einer weiteren Reduzierung vieler systematischer Einschätzungsfehler, wie *Überschätzung*.

Durch eine experimentelle Variation des Skalenformats (Häufigkeitsskala versus Prozentskala) in der Ankerfrage kann somit überprüft werden: (1) Ob über die Häufigkeitsskala die allgemeine Wahrscheinlichkeit tatsächlich adäquater[22] gemessen werden kann als mit Hilfe der Prozentskala. (2) Ob sich ein Kontexteffekt des Skalenformats der Ankerfrage auf die Beurteilung der subjektiven Wahrscheinlichkeit insofern ergibt, als die Zielfrage realistischer eingeschätzt wird, wenn die allgemeine Wahrscheinlichkeit zuvor über eine Häufigkeits- statt Prozentskala abgefragt wurde.

Die Nutzung von Häufigkeitsskalen bei der Wahrscheinlichkeitsbeurteilung anstelle von Prozentskalen erhöht die Eichung und Kohärenz des Urteils bezogen auf objektive Werte (vgl. Coutts-Heller 2002: 10). Außerdem stellt die Verschiebung –

22 D. h. weniger überschätzt bei seltenen Ereignissen (bzw. weniger unterschätzt bei häufigen Ereignissen) und somit näher an objektiven Wahrscheinlichkeiten. Die Häufigkeitsskala ruft bei der Abfrage allgemeiner Wahrscheinlichkeiten einen regelgeleiteten Denkmodus hervor, während die Prozentskala oftmals lediglich Heuristiken (wie die *Repräsentativitätsheuristik*) aktiviert.

weg von der Beurteilung von Einzelereigniswahrscheinlichkeiten (Prozenten) hin zu Einschätzungen der Ereigniswahrscheinlichkeiten ganzer Klassen von Menschen oder Objekten (Häufigkeiten) – einen Perspektivenwechsel im Wahrscheinlichkeitsdenken dar: Während Häufigkeitsskalen einen distributionalen Modus aktivieren, lösen Prozentskalen einen singulären Modus aus. Die Vorhersagen im distributionalen Modus basieren auf Wissen über Basisraten in Referenzpopulationen. Im singulären Modus hingegen beruhen sie auf Einstellungen zum und Erfahrungen über die Eigenschaften des spezifischen, zu beurteilenden Objekts/Ereignisses und sind damit anfälliger für systematische Fehleinschätzungen. Wahrscheinlichkeitsbeurteilungen im distributionalen Modus (Häufigkeitsformat) führen zu akkurateren Einschätzungen:[23]

> *Hypothese 1*
> *Die Ankerfrage im Häufigkeitsformat liefert realistischere Einschätzungen der allgemeinen Wahrscheinlichkeit als die im Prozentformat. Erstere führt zu geringeren Wahrscheinlichkeitseinschätzungen seltener Ereignisse.*

Zudem wird vermutet, dass die Messung der allgemeinen Wahrscheinlichkeiten im Häufigkeitsformat (verglichen mit dem Prozentformat) einen biasverringernden Effekt auf die darauffolgende Abfrage der subjektiven Wahrscheinlichkeiten ausübt (vgl. Billiet et al. 1992: 131). Die abhängige Variable ‚subjektive Wahrscheinlichkeit' wird hierbei in beiden Fällen über eine Prozentskala gemessen. Es wird vermutet, dass der durch die vorgeschaltete Häufigkeitsskala aktivierte distributionale Denkprozess und die damit einhergehende stärkere Berücksichtigung von Basisraten, sich auf die nachfolgende Zielfrage übertragen und dort ebenso zu adäquateren Wahrscheinlichkeitseinschätzungen führen. Bei der Ankerfrage mit Prozentskala werden dagegen weniger günstige kontextuelle Auswirkungen auf die Zielfrage erwartet. Auch Reeves und Lockhart (1993) konnten in ihrer Studie zeigen, dass die Abfrage von Wahrscheinlichkeiten über das Häufigkeitsformat vor der Abfrage von Einzelereigniswahrscheinlichkeiten (diese schließen subjektive Wahrscheinlichkeiten ein) zu deutlichen Leistungsverbesserungen führt. Entsprechend kann eine Hypothese zum vermuteten Kontexteffekt abgeleitet werden:

23 Beispielstudien zur Unterscheidung des distributionalen und singulären Modus finden sich bei Bruine de Bruin et al. (2000), Kahneman/Tversky (1982b), Klar et al. (1996), Koehler (2001) sowie Reeves/Lockhart (1993): „Frequency problems evoke a distributional approach of probability and are mentally represented in such a way that the relevance of extensional rules is more compelling. Case-specific problems evoke a singular approach and are modeled in ways that support the use of nonextensional heuristics such as representativeness. […] Presumably, solving frequency problems first not only cued extensional rules but also provided a model for how those rules could be applied. […] Evidently, simply having subjects solve problems that naturally evoke extensional rules will lead them to apply extensional rules to problems that would otherwise evoke nonextensional strategies." (Reeves/Lockhart 1993: 212)

Hypothese 2
Befragte, welche die allgemeine Wahrscheinlichkeit zunächst auf einer Häufigkeitsskala beurteilen, schätzen ihre eigene Wahrscheinlichkeit in einer darauffolgenden Frage subjektiv geringer ein als Befragte, die zuvor die Ankerfrage im Prozentformat beantwortet haben.

4 Analyse einer experimentellen Schwarzfahrerstudie

4.1 Design

Unsere Analysen basieren auf einem Methodenexperiment, welches an der Universität Leipzig durchgeführt wurde. Im Zentrum der Studie steht neben allgemeinen Fragen zur Nutzung öffentlicher Nahverkehrsmittel die subjektive Entdeckungswahrscheinlichkeit beim Schwarzfahren. Dabei wurden Fragebögen an insgesamt 405 Studenten innerhalb zweier Vorlesungen zu zwei verschiedenen Messzeitpunkten verteilt: 242 Befragte nahmen an der ersten Welle im Juni sowie 163 an der zweiten im Juli 2007 teil. Im Abstand von zwei Wochen wurden in den Vorlesungen zwei unterschiedliche Fragebogenversionen im Wechsel an die Gruppen ausgegeben, vor Ort von den Studenten ausgefüllt und wieder abgegeben. Innerhalb dieser Fragebogenversionen wurde das Format zur Messung der allgemeinen Entdeckungswahrscheinlichkeit experimentell variiert (Häufigkeits- versus Prozentskala). Durch die Erhebung zu zwei Zeitpunkten mit zwei Gruppen wird es möglich, beide Befragtengruppen mit beiden Fragebogenversionen zu konfrontieren (faktorielles ‚between-within-Design'[24]). Das Wirken beider Anker auf die subjektive Entdeckungswahrscheinlichkeit kann also sowohl zwischen als auch innerhalb der Gruppen verglichen werden.

Die Fragebögen[25] enthalten neben dieser experimentellen Variation identische Fragen zu demografischen Aspekten (Alter, Geschlecht, Wohndauer in Leipzig, Hauptfach, Hochschulsemester), zu Themen der spezifischen Nutzung von öffentlichen Nahverkehrsmitteln (Transportmittel, Jahreszeit, Uhrzeit, Fahrscheinart, durchschnittliche Fahrtdauer), zur Beurteilung der Ticketpreise und des Schwarzfahrens sowie zur Einschätzung des aktiven Schwarzfahrens von Freunden/Bekannten und des Befragten selbst. Der Einschätzung der subjektiven Entdeckungswahrscheinlichkeit beim Schwarzfahren geht eine Ankerfrage voraus, die die allgemeine Entdeckungs-

24 Detaillierte Informationen zum between- und within-Design sowie deren Anwendungen finden sich bei Birnbaum (1999), Kahneman/Tversky (1996), Price (1998) sowie Shadish et al. (2002).
25 Der vollständige Fragebogen in seinen zwei Versionen, der Datensatz und alle Analyseroutinen sind auf Anfrage erhältlich.

wahrscheinlichkeit thematisiert und deren Format sich zwischen den Versionen unterscheidet. Mittels dieser wird der Kontext vor der Zielfrage variiert, indem Basisraten über zwei unterschiedliche Skalen (Häufigkeits- und Prozentskala) abgefragt werden. Die Zielfrage befindet sich abgetrennt auf der nächsten Seite des Fragebogens, sodass die Ankerfrage bei deren Beantwortung nicht direkt sichtbar ist. Der Einfluss der vorangegangenen Frage wirkt also vorwiegend über die Aktivierung und kognitive Bereitstellung von Informationen, statt unmittelbar visuell zugänglich zu sein; das Rückblättern der Befragten kann dennoch nicht ausgeschlossen werden.

Abbildung 1 Die zwei Ankerfragen und die Zielfrage des Fragebogens

Ankerfrage im Häufigkeitsformat (experimenteller Split 1)

Angenommen, 100 beliebige Personen würden tagsüber (zwischen 7:00 und 17:00 Uhr) ohne einen gültigen Fahrschein mit öffentlichen Nahverkehrsmitteln (z. B. Straßenbahn oder Bus) in Leipzig jeweils eine Strecke von ca. 25 Minuten Dauer zurücklegen. Wie viele von diesen 100 Personen würden Ihrer Ansicht nach dabei von einem Kontrolleur entdeckt werden?

Geben Sie hierzu ihre Einschätzung auf einer Skala von 0 = „es wird keiner entdeckt" bis 100 = „es werden alle entdeckt" an. Sie können bei Ihrer Antwort jeden beliebigen Wert dazwischen angeben.

_____ *Personen von 100 Personen*

Ankerfrage im Prozentformat (experimenteller Split 2)

Angenommen, eine beliebige Person (nicht Sie selbst) würde tagsüber (zwischen 7:00 und 17:00 Uhr) ohne einen gültigen Fahrschein mit öffentlichen Nahverkehrsmitteln (z. B. Straßenbahn oder Bus) in Leipzig eine Strecke von ca. 25 Minuten Dauer zurücklegen. Für wie wahrscheinlich halten Sie es, dass diese Person dabei von einem Kontrolleur entdeckt wird?

Geben Sie hierzu ihre Einschätzung auf einer Skala von 0 % = „sie wird auf keinen Fall entdeckt" bis 100 % = „sie wird auf jeden Fall entdeckt" an. Sie können bei Ihrer Antwort jeden beliebigen Wert dazwischen angeben.

_____ *Prozent*

Zielfrage (beide Splits)

Angenommen, Sie würden tagsüber (zwischen 7:00 und 17:00 Uhr) ohne einen gültigen Fahrschein mit öffentlichen Nahverkehrsmitteln (z. B. Straßenbahn oder Bus) in Leipzig eine Strecke von ca. 25 Minuten Dauer zurücklegen. Für wie wahrscheinlich halten Sie es, dass Sie dabei von einem Kontrolleur entdeckt werden?

Geben Sie hierzu ihre Einschätzung auf einer Skala von 0 % = „ich werde auf keinen Fall entdeckt" bis 100 % = „ich werde auf jeden Fall entdeckt" an! Sie können bei Ihrer Antwort jeden beliebigen Wert dazwischen angeben.

_____ *Prozent*

In Anwendung der allgemeinen Hypothesen auf die Entdeckungswahrscheinlichkeiten beim Schwarzfahren resultieren die folgenden spezifischen Vorhersagen: (1) Die Ankerfrage im Häufigkeitsformat liefert realistischere Einschätzungen der allgemeinen Entdeckungswahrscheinlichkeit als die im Prozentformat. Erstere führt zu geringeren Wahrscheinlichkeitseinschätzungen. (2) Befragte, welche die allgemeine Entdeckungswahrscheinlichkeit zunächst auf einer Häufigkeitsskala beurteilen, schätzen in der darauffolgenden Frage ihre subjektive Entdeckungswahrscheinlichkeit geringer ein als Befragte, die zuvor die Ankerfrage im Prozentformat beantwortet haben.[26]

4.2 Empirische Befunde

In der Fragebogenversion mit Häufigkeitsformat werden generell signifikant niedrigere Werte angegeben als in der mit Prozentformat, womit beide Vorhersagen zunächst bestätigt werden können.[27] Tabelle 1 liefert eine Übersicht der Ergebnisse des Mittelwertvergleichs:

Tabelle 1 Differenzen der Mittelwerte der allgemeinen und subjektiven Entdeckungswahrscheinlichkeit in Abhängigkeit des Skalenformats der Ankerfrage

	Allgemeine Entdeckungswahrscheinlichkeit	**Subjektive Entdeckungswahrscheinlichkeit**
Häufigkeitsformat (H)	23,10 (217)	33,57 (217)
Prozentformat (P)	35,09 (180)	39,87 (180)
Differenz (H-P)	-11,99	-6,30
p-Wert (H<P)	0,00	0,02

Anmerkung: Dargestellt sind die Mittelwerte (Fallzahlen in Klammern) der allgemeinen und subjektiven Entdeckungswahrscheinlichkeit, unterteilt nach Prozent- versus Häufigkeitsformat der Ankerfrage. Die empirischen Signifikanzniveaus (p-Werte) der Mittelwertdifferenzen wurden auf Grundlage von einseitigen T-Tests (Hypothese: H<P) berechnet.

26 Im Rahmen unseres Experiments lautet das seltene Ereignis ‚beim Schwarzfahren entdeckt werden'. Unseren beiden Hypothesen liegt die Annahme zugrunde, dass die Einschätzung der allgemeinen und subjektiven Entdeckungswahrscheinlichkeit dann realistischer ist, wenn geringere Werte angegeben werden. Diese ‚weniger ist besser – Annahme' kann nicht explizit getestet werden, da die tatsächliche Entdeckungswahrscheinlichkeit nicht bekannt ist. Wir nehmen jedoch an, dass die tatsächliche Entdeckungswahrscheinlichkeit hinreichend klein ist und in beiden Formaten subjektiv überschätzt wird.

27 Informationen über die empirischen Häufigkeitsverteilungen der allgemeinen und subjektiven Entdeckungswahrscheinlichkeiten sind im Anhang zu finden.

Zunächst ist auffällig, dass sich die Angaben zwischen den allgemeinen und subjektiven Entdeckungswahrscheinlichkeiten unterscheiden, wobei letzteren im Durchschnitt höhere Werte zugeordnet werden.[28] Die Einschätzung der allgemeinen Entdeckungswahrscheinlichkeit im Häufigkeitsformat fällt verglichen mit der Abfrage im Prozentformat signifikant geringer aus (Differenz -11,99). Zudem verringert die Nutzung der Häufigkeitsskala als Anker, im Vergleich zur Nutzung der Prozentskala als Anker, die darauffolgende subjektive Entdeckungswahrscheinlichkeit beim Schwarzfahren (Differenz -6,30).

Die Befunde zu den allgemeinen Entdeckungswahrscheinlichkeiten können unter Kontrolle der Variablen (1) Befragtengruppe und (2) Welle bestätigt werden: Befragte, die die Ankerfrage im Prozentformat beantwortet haben, zeigen signifikant höhere Werte als die Vergleichsgruppe mit Häufigkeitsformat (Inter-Gruppenvergleich, Differenzen -16,55 und -5,97). Der Effekt des Skalenformats kann demnach für beide Gruppen unabhängig nachgewiesen werden (vgl. Tabelle 2).

Tabelle 2 Differenzen der Mittelwerte der allgemeinen Entdeckungswahrscheinlichkeit in Abhängigkeit des Skalenformats, der Welle und Befragtengruppe

		Gruppe 1 (G1)	Gruppe 2 (G2)	Differenz (H-P)	p-Wert (H<P)
Welle 1	Prozentformat (P)	39,02 (100)			
	Häufigkeitsformat (H)		22,47 (138)	-16,55	0,00
Welle 2	Prozentformat (P)		30,17 (80)		
	Häufigkeitsformat (H)	24,20 (79)		-5,97	0,05
Differenz (H-P)		-14,82	-7,70		
p-Wert (H<P)		0,00	0,00		

Anmerkung: Dargestellt sind die Mittelwerte (Fallzahlen in Klammern) der allgemeinen Entdeckungswahrscheinlichkeit, unterteilt nach Prozent- versus Häufigkeitsformat der Ankerfrage. Die empirischen Signifikanzniveaus (p-Werte) der Mittelwertdifferenzen wurden auf Grundlage von einseitigen T-Tests (Hypothese: H<P) berechnet.

Auch innerhalb der Gruppen ergeben sich in Abhängigkeit des Skalenformats deutliche Unterschiede (Intra-Gruppenvergleich, Differenzen -14,82 und -7,70). Gruppe 1 hat in der ersten Welle die Ankerfrage im Prozent- und in der zweiten im Häu-

28 Vgl. hierzu Studien zur Überschätzung eigener Risiken und zum *Optimismus/Pessimismus Bias* von Dolinski et al. (1987), Franic/Pathak (2000), Hoorens/Buunk (1993), Klar et al. (1996), Pinkerton et al. (2000).

figkeitsformat bearbeitet. Bei Gruppe 2 ergibt sich entsprechend die umgekehrte Reihenfolge. Innerhalb beider Gruppen zeigt sich jeweils, dass Befragte, die die allgemeine Entdeckungswahrscheinlichkeit auf der Häufigkeitsskala beantwortet haben, geringere Werte für diese Variable angeben als solche, die die Prozentskala genutzt haben. Die Datenanalyse zeigt über alle Gruppenvergleiche hinweg starke Effekte der Häufigkeitsskala auf die allgemeine Entdeckungswahrscheinlichkeit in die erwartete Richtung. Somit kann Hypothese 1 bestätigt werden.

Bei den subjektiven Entdeckungswahrscheinlichkeiten zwischen den Gruppen fallen die erwarteten Kontexteffekte schwächer aus (Inter-Gruppenvergleich): In der ersten Erhebungswelle zeigt sich der erwartete Effekt des Skalenformats der Ankerfrage auf die darauffolgende subjektive Entdeckungswahrscheinlichkeit. Befragte, die die Kontextfrage mit Häufigkeitsskala bearbeitet haben, zeigen bei der Zielfrage signifikant niedrigere Werte als solche, die zuvor mit Hilfe der Prozentskala geantwortet haben (Differenz -8,85). Dieser Kontexteffekt des Skalenformats der Ankerfrage auf die Zielfrage zeigt in der zweiten Welle ebenso in die erwartete Richtung, fällt jedoch recht schwach aus (Differenz -2,81) und ist zudem nicht signifikant (auf den konventionellen 1 %-, 5 %- bzw. 10 %-Niveaus). Tabelle 3 visualisiert diese Befunde.

Tabelle 3 Differenzen der Mittelwerte der subjektiven Entdeckungswahrscheinlichkeit in Abhängigkeit des Skalenformats, der Welle und Befragtengruppe

		Gruppe 1 (G1)	Gruppe 2 (G2)	Differenz (H-P)	p-Wert (H<P)
Welle 1	Prozentanker (P)	41,81 (100)			
	Häufigkeitsanker (H)		32,96 (138)	-8,85	0,01
Welle 2	Prozentanker (P)		37,44 (80)		
	Häufigkeitsanker (H)	34,63 (79)		-2,81	0,28
Differenz (H-P)		-7,18	-4,48		
p-Wert (H<P)		0,05	0,14		

Anmerkung: Dargestellt sind die Mittelwerte (Fallzahlen in Klammern) der subjektiven Entdeckungswahrscheinlichkeit, unterteilt nach Prozent- versus Häufigkeitsformat der Ankerfrage. Die empirischen Signifikanzniveaus (p-Werte) der Mittelwertdifferenzen wurden auf Grundlage von einseitigen T-Tests (Hypothese: H<P) berechnet.

Auffällig ist die Differenz (37,44 – 41,81 = -4,37) bei der subjektiven Entdeckungswahrscheinlichkeit mit jeweils vorgeschaltetem Prozentanker zwischen den Grup-

pen.[29] Befragte der Gruppe 2, denen der Prozentanker in Welle 2 vorgelegt wurde, zeigen eine geringere subjektive Entdeckungswahrscheinlichkeit (nicht signifikant: einseitiger T-Test; p=0,16) verglichen mit Befragten der Gruppe 1, denen der Prozentanker bereits in Welle 1 vorgelegt wurde. Somit kann ein Kontexteffekt von Welle 1 auf Welle 2 nicht ausgeschlossen werden: Es besteht die Vermutung, dass in Gruppe 2 durch die Benutzung des Häufigkeitsankers und der sich daraus ergebenden geringen subjektiven Entdeckungswahrscheinlichkeit in Welle 1, diese auch in Welle 2 – trotz Prozentanker – nach unten korrigiert wurde. Solch ein Reihenfolgeeffekt (Häufigkeiten in Welle 1 und Prozente in Welle 2) wäre eine mögliche Erklärung des nur schwachen Kontexteffektes in Gruppe 2.

Innerhalb der Gruppen zeigen alle Vergleiche der subjektiven Wahrscheinlichkeiten ebenso in die erwartete Richtung (Intra-Gruppenvergleich), wobei auch hier die angenommenen Effekte (verglichen mit Hypothese 1) schwächer ausfallen. Befragte der Gruppe 1, welche die Ankerfrage im Häufigkeits- statt Prozentformat beantwortet haben, zeigen eine geringere subjektive Entdeckungswahrscheinlichkeit (Differenz -7,18). Auch bei Befragten der Gruppe 2 zeigt sich der Effekt in die erwartete Richtung, allerdings schwächer (Differenz -4,48) und nicht signifikant (p=0,14).

Erneut drängt sich die Vermutung auf, dass ein Reihenfolgeeffekt (Gruppe 1: Prozente in Welle 1 und Häufigkeiten in Welle 2; Gruppe 2: Häufigkeiten in Welle 1 und Prozente in Welle 2) den Kontexteffekt der Anker- auf die Zielfrage moderiert. Tabelle 2 und Tabelle 3 zeigen, dass die Schätzungen in der zweiten Welle sowohl durch das Format der ersten wie der zweiten Welle beeinflusst werden. Befragte der Gruppe 1, die in der ersten Welle im Prozentformat antworteten, berichten in der zweiten Welle höhere Werte im Häufigkeitsformat als Befragte der Gruppe 2, die in der ersten Welle unmittelbar im Häufigkeitsformat antworteten.[30] Umgekehrt resultiert in Gruppe 2 das Häufigkeitsformat in Welle 1 in niedrigeren Werten im Prozentformat in Welle 2.[31] Dies deutet auf einen Ankereffekt über Welle und Format hin. Dieser Befund ist konsistent mit Ergebnissen bisheriger Studien, die zeigen, dass Manipulationen zum Zeitpunkt *t1* die kognitive Repräsentation verändern, auf die zum Zeitpunkt *t2* zurückgegriffen wird (vgl. Carlston 1980; Sherman

29 Bezüglich der Häufigkeitsanker sind die Angaben zur subjektiven Entdeckungswahrscheinlichkeit für die Gruppen nahezu identisch (Differenz: 34,63 – 32,96 = 1,67; einseitiger T-Test: p=0,34).

30 In Tabelle 2 beträgt die Differenz 1,73 (einseitiger T-Test; p=0,28) und in Tabelle 3 beträgt die Differenz 1,67 (einseitiger T-Test; p=0,34).

31 In Tabelle 2 beträgt die Differenz -8,85 (einseitiger T-Test; p=0,01) und in Tabelle 3 beträgt die Differenz -4,37 (einseitiger T-Test; p=0,16).

1980).[32] Allerdings zeigen drei von vier Tests keine signifikanten Differenzen. Die von uns aufgestellten Vermutungen zum Reihenfolgeeffekt sollten daher in einer weiteren Untersuchung (mit höherer Fallzahl) repliziert werden.

Zusammenfassend kann auch Hypothese 2 der Tendenz nach bestätigt werden, wobei weitere Untersuchungen eines möglichen Interaktionseffektes von (1) Skalenformat (Häufigkeitsanker versus Prozentanker) und (2) Präsentationsabfolge der Formate (Abfolge: Häufigkeitsanker-Prozentanker versus Prozentanker-Häufigkeitsanker) auf die Einschätzung der subjektiven Entdeckungswahrscheinlichkeit lohnenswert scheinen.[33]

5 Diskussion

Im theoretischen Teil der Arbeit wurde die Problematik der Messung von subjektiven Wahrscheinlichkeiten besprochen. Auffällig sind die sich aus ihr ergebenden zahlreichen Fehleinschätzungen, wie *Überschätzung* und *die Nicht-Berücksichtigung von Basisraten*. Die Gründe hierfür sind zum einen in den kognitiven Prozessen der

32 Als eine mögliche Erklärung für diesen Effekt diskutiert Sherman (1980: 218) die erhöhte Verfügbarkeit von stereotypen Antwortsequenzen (sog. kognitiven ‚Skripten'), die bei der erstmaligen Beantwortung der Frage gebildet, anschließend im Gedächtnis gespeichert und zu späteren Zeitpunkten (bei ähnlichen Fragestellungen) leichter aktiviert werden können. Ähnlich argumentiert Carlston (1980: 324): „Cognitive processes occurring after stimulus observation can alter the information subjects have available for making later impression judgements, and consequently, can alter the impression judgments that are made."

33 Neben dem Vergleich der konditionalen Mittelwerte wurden zudem 2 multiple OLS-Regressionen mit Interaktionstermen geschätzt: Im ersten Modell ist die abhängige Variable die allgemeine Entdeckungswahrscheinlichkeit (AEW), im zweiten Modell ist die abhängige Variable die subjektive Entdeckungswahrscheinlichkeit (SEW). In beiden Modellen stehen jeweils auf der rechten Seite der Gleichung die beiden Experimentalvariablen Skalenformat (SF: 1 = Häufigkeiten, 0 = Prozente) und Präsentationsabfolge (PA: 1 = Prozente zuerst, 0 = Häufigkeiten zuerst) sowie der entsprechende Interaktionsterm Skalenformat (SF) * Präsentationsabfolge (PA). Im ersten Modell zeigen die Regressionskoeffizienten der beiden Experimentalvariablen in die theoretisch erwartete Richtung: AEW = 30,17 − 7,70*SF + 8,85*PA − 7,12*SF*PA. Die Koeffizienten für das Skalenformat (p=0,01) und die Präsentationsabfolge (p=0,01) sind jeweils signifikant, der Interaktionsterm Skalenformat * Präsentationsabfolge (p=0,12) ist dagegen nicht signifikant. Im zweiten Modell zeigen die Regressionskoeffizienten der beiden Experimentalvariablen ebenfalls in die theoretisch erwartete Richtung: SEW = 37,44 – 4,48*SF + 4,37*PA – 2,70*SF*PA. Im Vergleich zum Prozentformat, resultiert die Fragebogenversion mit der Ankerfrage im Häufigkeitsformat in niedrigere subjektive Entdeckungswahrscheinlichkeiten. Im Vergleich zur Reihenfolge ‚Häufigkeiten zuerst', resultiert die Reihenfolge ‚Prozente zuerst' in höhere subjektive Entdeckungswahrscheinlichkeiten. Im Vergleich zum ersten Modell deuten die Koeffizienten im zweiten Modell auf schwächere Effekte hin. Zudem sind die Koeffizienten nicht signifikant (p=0,28 für das Skalenformat; p=0,32 für die Präsentationsabfolge; p=0,66 für den Interaktionsterm Skalenformat * Präsentationsabfolge). In einer separaten Analyse wurden zudem robuste Standardfehler geschätzt (Huber-White-Sandwich-Varianzschätzer). Die hieraus resultierenden p-Werte bleiben nahezu unverändert.

Befragten zu finden, die ihre Einschätzungen häufig mittels einfacher Heuristiken, wie der *Repräsentativitätsheuristik* und der *Verfügbarkeitsheuristik*, vornehmen. Zum anderen führt aber auch die Anwendung inadäquater Präsentations- und Abfrageformate zu den Fehleinschätzungen. Wie die Durchsicht früherer theoretischer und empirischer Forschungsarbeiten gezeigt hat, stellen die in zahlreichen Surveys verwendeten, mit *Vague Quantifiers* verbalisierten Kategorialskalen eine nur ungenaue und häufig inkonsistente Form der Risikoeinschätzung und -kommunikation dar (vgl. Bradburn/Miles 1979; Krumpal et al. 2008; Wright et al. 1994). Ausgangspunkt des vorliegenden Artikels war es deshalb, adäquatere Alternativen zur Erfassung von Wahrscheinlichkeiten zu untersuchen und experimentell zu vergleichen. Es wurden zwei numerische Formate (Häufigkeitsskala versus Prozentskala) vergleichend untersucht. Hierbei zeigte sich ein Einfluss des Formats auf die Wahrscheinlichkeitseinschätzungen der Befragten. Im Vergleich zum Prozentformat führt das Häufigkeitsformat zu niedrigeren Wahrscheinlichkeitseinschätzungen hinsichtlich des Eintretens seltener Ereignisse.

Im Rahmen einer Schwarzfahrerstudie wurde ein experimentelles Design angewendet, welches in einem zweistufigen Ansatz die Abfrage allgemeiner Entdeckungswahrscheinlichkeiten vor der Abfrage subjektiver Entdeckungswahrscheinlichkeiten erfordert und das Wirken von Kontexteffekten annimmt. Die Ergebnisse dieses Methodenexperiments zeigen, dass die Ankerfrage im Häufigkeitsformat zu niedrigeren allgemeinen Wahrscheinlichkeitseinschätzungen führt. Überdies ergab sich ein Kontexteffekt der Ankerfrage mit Häufigkeitsskala auf die Zielfrage, wobei die subjektive Entdeckungswahrscheinlichkeit hier kleiner eingeschätzt wurde als bei der Ankerfrage mit Prozentskala.

Die vorgestellten Ergebnisse können für die empirische Sozialforschung und die Umfragemethodologie von Nutzen sein: Erstens können sie in der Phase der Fragebogenentwicklung bzw. der Konstruktion von Messinstrumenten zur Erfassung und Kommunikation allgemeiner sowie subjektiver Entdeckungswahrscheinlichkeiten Anregungen liefern. Zweitens wurde eine Alternative zu den häufig verwendeten *Vague Quantifiers* bei der Erfassung von Entdeckungswahrscheinlichkeiten aufgezeigt. Drittens wurde argumentiert, dass bei der Informationspräsentation sowie Abfrage von Entdeckungswahrscheinlichkeiten in numerischen Formaten die Nutzung von natürlichen Häufigkeiten intuitiver und verständlicher sei und die subjektive Überschätzung von kleinen Wahrscheinlichkeiten abnehme, wenn Befragte Risiken in Häufigkeiten anstatt Prozenten schätzen. Dies könnte zunächst als Nachteil gedeutet werden, wenn man davon ausgeht, dass viele Umfragen durchgeführt werden, um aus den Antworten der Stichprobe auf Wahrnehmungen in der Population zu schließen. So wäre denkbar, dass Häufigkeitsformate die

Wahrnehmungen und Antworten in der Stichprobe derart beeinflussen, dass diese von den spontanen Wahrnehmungen der Personen in der Population abweichen und dadurch als handlungsrelevante Variablen an Erklärungskraft verlieren würden. Bisherige Forschungsarbeiten deuten allerdings darauf hin, dass das Häufigkeitsformat dem alltäglichen Denken näher ist:

> „Natural frequencies facilitate inferences because they carry implicit information about base rates (...). They also correspond to the way in which humans have experienced statistical information over most of their history." (Hoffrage et al. 2000: 2261)

Literatur

Ajzen, I., 1977: Intuitive theories of events and the effects of base-rate information on prediction. Journal of Personality and Social Psychology 35: 303-314.

ALLBUS 1990/2000: Fragebögen und Codebücher. http://www.gesis.org/dienstleistungen/daten/umfragedaten/allbus/suche-in-fragetexten/?L= (02.12.2008).

Bar-Hillel, M., 1980: The base-rate fallacy in probability judgments. Acta Psychologica 44: 211-233.

Billiet, J. B., L. Waterplas und Geert Loosveldt, 1992: Context effects as substantive data in social surveys. S. 131-147 in: N. Schwarz und S. Sudman (Hg.): Context effects in social and psychological research. New York: Springer-Verlag.

Birnbaum, M. H., 1999: How to show that 9 > 221. Collect judgments in a between-subjects design. Psychological Methods 4: 243-249.

Black, W. C., R. F. Nease Jr. und A. N. A. Tosteson, 1995: Perceptions of breast cancer risk and screening effectiveness in women younger than 50 years of age. Journal of the National Cancer Institute 87: 720-731.

Bradburn, N. M. und C. Miles., 1979: Vague quantifiers. Public Opinion Quarterly 43: 92-101.

Brase, G. L., 2002: Which statistical formats facilitate what decisions? The perception and influence of different statistical information formats. Journal of Behavioral Decision Making 15: 381-401.

Brase, G. L., L. Cosmides und J. Tooby, 1998: Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty. Journal of Experimental Psychology General 127: 3-21.

Bruine de Bruin, W., B. Fischhoff, S. G. Millstein und B. L. Halpern-Felsher, 2000: Verbal and numerical expressions of probability: "It`s a fifty-fifty chance". Organizational Behavior and Human Decision Processes 81: 115-131.

Carlston, D. E., 1980: The recall and use of traits and events in social inference processes. Journal of Experimental Social Psychology 16: 303-328.

Cosmides, L. und J. Tooby, 1996: Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 58: 1-73.

Coutts-Heller, E., 2002: Context effects in the measurement of subjective probabilities in surveys. Universität Konstanz: Diplomarbeit.

Dolinski, D., W. Gromski und E. Zawisza, 1987: Unrealistic pessimism. Journal of Social Psychology 127: 511-516.

Dominitz, J. und C. F. Manski, 1997: Perceptions of economic insecurity. Evidence from the survey of economic expectations. Public Opinion Quarterly 61: 261-287.

Esser, H., 1986: Können Befragte lügen? Zum Konzept des „wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 38: 314-336.
Evans, J. S., S. J. Handley, N. Perham, D. E. Over und V. A. Thompson, 2000: Frequency versus probability formats in statistical word problems. Cognition 77: 197-213.
Feldman, J. M., 1992: Constructive processes as a source of context effects in survey research. Explorations in self-generated validity. S. 49-61 in: N. Schwarz und S. Sudman (Hg.): Context effects in social and psychological research. New York: Springer-Verlag.
Fischhoff, B., A. M. Parker, W. B. De Bruin, J. Downs, C. Palmgren, R. Dawes und C. F. Manski, 2000: Teen expectations for significant life events. Public Opinion Quarterly 64: 189-205.
Franic, D. M. und D. S. Pathak, 2000: Communicating the frequency of adverse drug reactions to female patients. Drug Information Journal 34: 251-272.
Gigerenzer, G., 1996a: On narrow norms and vague heuristics. A reply to Kahneman und Tversky (1996). Psychological Review 103: 592-596.
Gigerenzer, G., 1996b: The psychology of good judgment. Frequency formats and simple algorithms. Medical Decision Making 16: 273-280.
Gigerenzer, G., 1998: Ecological intelligence. An adaption for frequencies. S. 9-29 in: D. D. Cummins und C. Allen (Hg.): The evolution of mind. New York: Oxford University Press.
Gigerenzer, G., 2000: Adaptive thinking. Rationality in the real world. New York: Oxford University Press.
Gigerenzer, G., U. Hoffrage und H. Kleinbölting, 1991: Probabilistic mental models. A Brunswikian theory of confidence. Psychological Review 98: 506-528.
Gigerenzer, G. und U. Hoffrage, 1995: How to improve Bayesian reasoning without instruction. Frequency formats. Psychological Review 102: 684-704.
Gigerenzer, G. und U. Hoffrage, 1999: Overcoming difficulties in Bayesian reasoning. A reply to Lewis and Keren (1999) and Mellers and McGraw (1999). Psychological Review 106: 425-430.
Gigerenzer, G., W. Hell und H. Blank, 1988: Presentation and content. The use of base rates as a continuous variable. Journal of Experimental Psychology: Human Perception and Performance 14: 513-525.
Ginosar, Z. und Y. Trope, 1980: The effects of base rates and individuating information on judgment about another person. Journal of Experimental Social Psychology 16: 228-242.
Griffin, D. und R. Buehler, 1999: Frequency, probability, and prediction. Easy solutions to cognitive illusions? Cognitive Psychology 38: 48-78.
Hendrickx, L., C. Vlek und H. Oppewal, 1989: Relative importance of scenario information and frequency information in the judgment of risk. Acta Psychologica 72: 41-63.
Hoffrage, U., G. Gigerenzer, S. Krauss und L. Martignon, 2002: Representation facilitates reasoning. What natural frequencies are and what they are not. Cognition 84: 343-352.
Hoffrage, U., S. Lindsey, R. Hertwig und G. Gigerenzer, 2000: Communicating statistical information. Science 290: 2261-2262.
Hoorens, V. und B. P. Buunk, 1993: Social comparison of health risks. Locus of control, the person-positivity bias, and unrealistic optimism. Journal of Applied Social Psychology 23: 291-302.
Johnson, J. E. V. und A. C. Bruce, 2001: Calibration of subjective probability judgments in a naturalistic setting. Organizational Behavior and Human Decision Processes 85: 265-290.
Jones, S. K., K. T. Jones und D. Frisch, 1995: Biases of probability assessment. A comparison of frequency and single-case judgments. Organizational Behavior and Human Decision Processes 61: 109-122.
Kahneman, D. und A. Tversky, 1972: Subjective probability. A judgment of representativeness. Cognitive Psychology 3: 430-454.
Kahneman, D. und A. Tversky, 1973: On the psychology of prediction. Psychological Review 80: 237-251.
Kahneman, D. und A. Tversky, 1982a: On the study of statistical intuitions. Cognition 11: 123-141.

Kahneman, D. und A. Tversky, 1982b: Variants of uncertainty. Cognition 11: 143-157.
Kahneman, D. und A. Tversky, 1996: Theoretical notes on the reality of cognitive illusions. Psychological Review 103: 582-591.
Klar, Y., A. Medding und D. Sarel, 1996: Nonunique Invulnerability. Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. Organizational Behavior and Human Decision Processes 67: 229-245.
Knowles, E. S., M. C. Coker, D. A. Cook, S. R. Diercks, M. E. Irwin, E. J. Lundeen, J. W. Neville und M. E. Sibicky, 1992: Order effects within personality measures. S. 221-236 in: N. Schwarz und S. Sudman (Hg.): Context effects in social and psychological research. New York: Springer-Verlag.
Koehler, J. J., 1996: The base rate fallacy reconsidered. Descriptive, normative, and methodological challenges. Behavioral and Brain Sciences 19: 1-53.
Koehler, J. J., 2001: When are people persuaded by DNA match statistics? Law and Human Behavior 25: 493-513.
Kreuter, F., 2002: Kriminalitätsfurcht. Messung und methodische Probleme. Opladen: Leske+Budrich.
Krumpal, I., H. Rauhut, D. Böhr und E. Naumann, 2008: Wie wahrscheinlich ist „wahrscheinlich"? Zur subjektiven Einschätzung und Kommunikation von Viktimisierungswahrscheinlichkeiten. Methoden, Daten und Analysen: Zeitschrift für empirische Sozialforschung 2: 3-27. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.2_Heft_1/2008_MDA1_Krumpal_et_al.pdf (15.5.2009).
Lippman-Hand, A. und F. C. Fraser, 1979: Genetic counselling. Provision and Reception of information. American Journal of Medical Genetics 3: 113-127.
Mellers, B. A. und A. Peter McGraw, 1999: How to improve Bayesian reasoning. Comment on Gigerenzer and Hoffrage (1995). Psychological Review 106: 417-424.
Nisbett, R. und L. Ross, 1980: Human inference. Strategies and shortcomings of social judgment. Englewood Cliffs: Prentice-Hall.
Phillips, L. D., 1970: The ´true probability` problem. Acta Psychologica 34: 254-264.
Pinkerton, S. D., L. I. Wagner-Raphael, C. A. Craun und P. R. Abramson, 2000: A quantitative study of the accuracy of college students` HIV risk estimates. Journal of Applied Biobehavioral Research 5: 1-25.
Price, P. C., 1998: Effects of a relative-frequency elicitation question on likelihood judgment accuracy. The case of external correspondence. Organizational Behavior and Human Decision Processes 76: 277-297.
Reeves, T. und R. S. Lockhart, 1993: Distributional versus singular approaches to probability and errors in probabilistic reasoning. Journal of Experimental Psychology 122: 207-226.
Reuband, K.-H., 2002: Subjektive Wahrscheinlichkeiten und Antwortmuster. Der Einfluss von Personenbezug und Skalierungsart. ZA-Information 50: 46-58.
Rottenstreich, Y. und A. Tversky, 1997: Unpacking, repacking, and anchoring. Advances in support theory. Psychological Review 104: 406-415.
Schapira, M. M., A. B. Nattinger und C. A. McHorney, 2001: Frequency or probability? A qualitative study of risk communication formats used in health care. Medical Decision Making 21: 459-467.
Schnell, R. und F. Kreuter, 2000: Das DEFECT-Projekt. Sampling-Errors und Nonsampling-Errors in komplexen Bevölkerungsstichproben. ZUMA-Nachrichten 47, 89-101. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_47.pdf (15.5.2009).
Schwarz, N., 1998: Accessible content and accessibility experiences. The interplay of declarative and experiential information in judgment. Personality and Social Psychology Review 2: 87-99.
Schwarz, N. und S. Sudman, 1992: Context effects in social and psychological research. New York: Springer-Verlag.

Schwarz, N., F. Strack, D. J. Hilton und G. Naderer, 1991: Base-rates, representativeness, and the logic of conversation. The contextual relevance of "irrelevant" information. Social Cognition 9: 67-84.
Schwarz, N., H.-J. Hippler, B. Deutsch und F. Strack, 1985: Response scales. Effects of category range on reported behavior and comparative judgments. Public Opinion Quarterly 49: 388-395.
Shadish, W. R., T. D. Cook und D. T. Campbell, 2002: Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.
Sherman, S. J., 1980: On the self-erasing nature of errors of prediction. Journal of Personality and Social Psychology 39: 211-221.
Slovic, P., M. L. Finucane, E. Peters und D. G. MacGregor, 2004: Risk as analysis and risk as feelings. Some thoughts about affect, reason, risk, and rationality. Risk Analysis 24: 311-322.
Slovic, P. und J. Monahan, 1995: Probability, danger, and coercion. A study of risk perception and decision making in Mental Health Law. Law and Human Behavior 19: 49-65.
Strack, F., 1992: „Order effects" in survey research. Activation and information functions of preceding questions. S. 23-34 in: N. Schwarz und S. Sudman (Hg.): Context effects in social and psychological research. New York: Springer-Verlag.
Teigen, K. H., 1974: Overestimation of subjective probabilities. Scandinavian Journal of Psychology 15: 56-62.
Tourangeau, R., 1992: Context effects on responses to attitude questions. Attitudes as memory structures. S. 35-47 in: N. Schwarz und S. Sudman (Hg.): Context effects in social and psychological research. New York: Springer-Verlag.
Tourangeau, R., L. J. Rips und K. Rasinski, 2000: The psychology of survey response. Cambridge: Cambridge University Press.
Tversky, A. und D. Kahneman, 1973: Availability. A heuristic for judging frequency and probability. Cognitive Psychology 5: 207-232.
Tversky, A. und D. Kahneman, 1974: Judgment under uncertainty. Heuristics and biases. Science 185: 1124-1131.
Tversky, A. und D. Kahneman, 1983: Extensional versus intuitive reasoning. The conjunction fallacy in probability judgment. Psychological Review 90: 293-315.
Warr, M., 1980: The accuracy of public beliefs about crime. Social Forces 59: 456-470.
Windschitl, P. D., 2002: Judging the accuracy of a likelihood judgment. The case of smoking risk. Journal of Behavioral Decision Making 15: 19-35.
Wright, D. B., G. D. Gaskell und C. A. O`Muircheartaigh, 1994: How much is "Quite a bit"? Mapping between numerical values and vague quantifiers. Applied Cognitive Psychology 8: 479-496.
Yamagishi, K., 1997: Upward versus downward anchoring in frequency judgments of social facts. Japanese Psychological Research 39: 124-129.

Korrespondenzadresse: Ivar Krumpal
Universität Leipzig
Institut für Soziologie
Beethovenstraße 15
04107 Leipzig
krumpal@sozio.uni-leipzig.de

Anhang Empirische Häufigkeitsverteilungen der allgemeinen und subjektiven Entdeckungswahrscheinlichkeiten (absolute Häufigkeiten)

Wahrscheinlichkeiten (Wertebereich)	Allgemeine Entdeckungs-wahrscheinlichkeit (Häufigkeitsformat)	Subjektive Entdeckungs-wahrscheinlichkeit (Häufigkeitsformat)	Allgemeine Entdeckungs-wahrscheinlichkeit (Prozentformat)	Subjektive Entdeckungs-wahrscheinlichkeit (Prozentformat)
0	**1**	**1**	**0**	**2**
1-4	26	17	6	5
5	29	27	12	15
6-9	8	4	3	2
10	**29**	**25**	**24**	**22**
11-14	2	0	1	1
15	13	8	6	6
16-19	1	0	0	0
20	**30**	**24**	**15**	**15**
21-24	0	0	1	1
25	6	13	9	10
26-29	1	0	1	1
30	**19**	**16**	**19**	**12**
31-34	2	1	3	2
35	6	2	2	2
36-39	0	0	2	1
40	**7**	**6**	**12**	**9**
41-44	0	0	0	0
45	2	0	1	0
46-49	1	1	0	0
50	**11**	**23**	**24**	**20**
51-54	0	0	0	0
55	0	0	1	0
56-59	1	1	0	1
60	**6**	**4**	**9**	**9**
61-64	0	0	0	0
65	2	0	4	0
66-69	0	0	0	0
70	**6**	**13**	**10**	**11**
71-74	0	0	0	0
75	2	4	5	4
76-79	0	0	0	0
80	**4**	**12**	**7**	**10**
81-84	0	0	0	0
85	0	1	1	3
86-89	0	0	0	0
90	**1**	**6**	**2**	**5**
91-94	0	0	0	0
95	1	2	0	3
96-99	0	0	0	1
100	**0**	**6**	**0**	**7**
N	217	217	180	180
25 % Quantil	5	10	12	10
50 % Quantil	17	25	30	30
75 % Quantil	30	50	50	60

Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey

Complexity, Learning Effects and Plausibility of Vignettes in the Factorial Survey Design

Katrin Auspurg, Thomas Hinz und Stefan Liebig

Zusammenfassung

Der Faktorielle Survey gilt als eine Erhebungsmethode, bei der sich die Vorteile der Umfrageforschung mit denen experimenteller Designs verbinden. Statt einzelner Items bewerten die Befragten hypothetische Objekt- oder Situationsbeschreibungen. Indem in diesen ‚Vignetten' einzelne Merkmalsausprägungen experimentell variiert werden, lässt sich ihr Einfluss auf die abgefragten Urteile oder Entscheidungen exakt bestimmen und damit das Gewicht von Faktoren isolieren, die in der Realität oftmals konfundiert sind. Bislang liegen allerdings nur sehr wenige Methodenstudien zur Validität der erzielten Messungen vor. Der Beitrag gibt zunächst einen knappen Überblick zum Einsatz des Faktoriellen Surveys in der sozialwissenschaftlichen Forschung und benennt anschließend bislang ungeklärte methodische Probleme. Die mit einer eigenen experimentellen Datenerhebung durchgeführten Analysen beziehen sich auf die Stabilität des Urteilsverhaltens der Befragten in Abhängigkeit von der Anzahl der in den Vignetten abgebildeten Dimensionen, möglichen Lerneffekten sowie von ‚unplausiblen' oder ‚unlogischen' Fällen (Vignettentexte für Situationen, die in der Realität sehr selten oder gar nicht vorkommen und die Befragten daher irritieren könnten). Getestet werden verschiedene Hypothesen zur Komplexität der Erhebungssituation und der Kohärenz der Urteile. Nach

Abstract

The factorial survey is a method of data collection that combines the advantages of survey research and the advantages of experimental designs. Respondents react to hypothetical descriptions of objects or situations (vignettes) instead of answering single-item questions. By varying each dimension of the vignettes in an experimental design, the dimensions' impact on respondents' judgments or decisions can be estimated accurately. Thus, the method is able to identify the effect of single factors which are often confounded in reality. So far, only few methodological studies address questions of measurement validity when a factorial survey design is used. The article provides a brief overview of the use of the factorial design in the social sciences and points our still unresolved methodological questions. Using experimental data specifically designed for this purpose our analyses consider the stability of respondents' judgments with respect to the number of dimensions presented in the vignettes, possible learning effects and ‚implausible' or ‚illogical' cases (vignettes describing objects or situations which are rare or even impossible). We test several hypotheses regarding the complexity of vignettes and the consistency of judgments. According to our results, a high complexity of vignettes and implausible cases cause respondents to consider

unseren Ergebnissen führen eine hohe Komplexität der Vignetten und unplausible Fälle zu einem weniger Vignettendimensionen einbeziehenden Urteilsverhalten, damit geringeren Einflussstärken einzelner Vignettenmerkmale bei gleich bleibender Konsistenz. Abschließend diskutieren wir die praktischen Konsequenzen dieser Befunde.

fewer dimensions in their judgments; we find smaller influences of vignette variables while the consistency of the judgments remains the same. Finally, we discuss the practical consequences of these results.

1 Einleitung[1]

Der Faktorielle Survey ist eine in Umfragen einsetzbare experimentelle Methode, bei der den Befragten hypothetische Objekt- oder Situationsbeschreibungen (*Vignetten*) vorgelegt werden.[2] Die Vignetten unterscheiden sich nach Merkmalen (*Dimensionen*), die in ihren Ausprägungen (*levels*) variieren. Solche hypothetischen Fälle und Szenarien, die Befragte beurteilen oder bewerten, werden heute in verschiedenen akademischen und nicht-akademischen Forschungszusammenhängen vermehrt eingesetzt, neben den Sozialwissenschaften etwa auch in den Gesundheitswissenschaften, der Rechtswissenschaft, der Psychologie und der Marktforschung. Thematisch zeigen die Studien in der Soziologie eine beachtliche Breite. In der Norm- und Werteforschung beschäftigen sie sich mit der Messung von Status und Prestige von Individuen und Haushalten (Rossi 1979; Rossi et al. 1974; Meudell 1982; Nock 1982), den Vorstellungen über ein gerechtes Erwerbseinkommen (Alves/Rossi 1978; Hermkens/Boerman 1989; Jann 2003; Jasso 1994; Jasso/Webster 1997, 1999; Shepelak/Alwin 1986), der Bewertung von Armutsdimensionen (Will 1993), den Kriterien zur Festlegung wohlfahrtsstaatlicher Unterstützungszahlungen (Liebig/Mau 2002), gerechten Steuersätzen (Liebig/Mau 2005) und Entlassungsverfahren (Struck et al. 2008). Ebenso liegen Arbeiten vor zur Bewertung von sexuellem Missbrauch/sexueller Belästigung (Garrett 1982; Rossi/Anderson 1982; O´Toole et al. 1999), zu der Bestrafung und dem Umgang mit Straftätern (Berk/Rossi 1977;

1 Der Beitrag entstand im Rahmen des von der DFG geförderten Forschungsprojekts ‚Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen'. Projektleiter sind Thomas Hinz (Universität Konstanz) und Stefan Liebig (Universität Bielefeld). Die Autoren danken Peter Steiner sowie einem anonymen Gutachter für wertvolle Hinweise und Anmerkungen. Für die Unterstützung bei der Organisation der Feldphase bedanken wir uns bei Judith Tonner.

2 Ursprünglich wurde der faktorielle Survey in den Sozialwissenschaften 1951 von Peter H. Rossi in seiner Dissertation entwickelt und zur Einschätzung des sozialen Status von Haushalten verwendet (Alves/Rossi 1978; Rossi 1979; Rossi/Nock 1982). Rossis zentrales Anliegen war es, ein Messverfahren zu entwickeln, das es ermöglicht herauszufinden, welche Objekteigenschaften in welchem Ausmaß für soziale Einstellungen relevant sind (Rossi/Anderson 1982: 15ff.; Rossi/Nock 1982: 9ff.).

Hembroff 1987; Miller/Rossi/Simpson 1986), unterschiedlichen Kriterien der Einbürgerung (Jasso 1988), zur Vergabe medizinischer Hilfen (Hechter et al. 1999), zur Qualität von Kinderbetreuungsmaßnahmen (Shlay et al. 2005) und zum sozialen Kontext von Normgeltung (Beck/Opp 2001; Diefenbach/Opp 2007; Horne 2003; Jasso/Opp 1997). Ferner existieren Arbeiten, die der Frage möglicher Diskriminierungen nachgehen (Jann 2003; John/Bates 1990), Effekte sozialer Einbettung analysieren (Buskens/Weesie 2000) oder familiensoziologische Theorien untersuchen (Auspurg/Abraham 2007). Angesichts des großen und vielfältigen Interesses für diese Erhebungsmethode verwundert es, dass methodische und modelltheoretische Fragen sehr selten diskutiert werden (Ausnahmen: Dülmer 2001, 2007; Dülmer/Klein 2003; Steiner/Atzmüller 2006). Wenn sie thematisiert werden, so besteht das Anliegen meistens darin, die Vorteile dieser Befragungsmethode gegenüber itembasierten Abfragen oder den traditionellen experimentellen Vorgehensweisen zu unterstreichen (Hechter/Kim/Baer 2005; Jasso 1988). Die im Verfahren angelegten methodischen Probleme waren dagegen kaum Gegenstand einer expliziten Untersuchung. Dies gilt insbesondere für Probleme, die sich aus der Anlage und Durchführung eines Faktoriellen Surveys und dem Einsatz von Vignetten in Umfragen ergeben.

Wir verfolgen daher das Ziel, drei miteinander verbundene und als besonders relevant geltende methodische Probleme zu diskutieren und anhand von empirischen Tests zu untersuchen. Dies sind *erstens* die Effekte der Komplexität der den Befragten geschilderten Situation und *zweitens* die hiermit in Verbindung stehenden Lerneffekte bei wiederholter Präsentation von Vignetten. Da das Risiko von unplausiblen Fällen mit der Komplexität steigt und diese zudem als ursächlich für Lerneffekte in Form vereinfachter Entscheidungsheuristiken gelten, analysieren wir *drittens* die Auswirkungen unplausibler Vignetten auf das Urteilsverhalten. Diese Aspekte wurden nach unserem Wissen für Vignettenstudien allesamt noch nicht gezielt untersucht. Eine Beschäftigung mit methodischen Effekten scheint für eine Verbesserung der Datenqualität jedoch dringend angeraten, auch um möglichen Fehlschlüssen gezielt vorzubeugen (die andernfalls beim Vergleich verschieden komplexer Vignettenstudien oder kognitiv unterschiedlich belastbarer Probandengruppen zu befürchten sind) – sei es durch ihren Einbezug bei der Vignettenkonstruktion, Datenauswertung und/oder Ergebnisinterpretation.

Die Gliederung ist wie folgt: Zunächst werden die Verfahrensweise des Faktoriellen Surveys sowie der Stand der Methodendiskussion knapp vorgestellt (Abschnitt 2). Dann werden ausgehend vom Forschungsstand Hypothesen zu den genannten Problemstellungen abgeleitet (Abschnitt 3) und auf der Grundlage einer experimentellen Online-Erhebung getestet (Abschnitte 4 und 5). Schließlich werden die Ergebnisse diskutiert und weiterer Analysebedarf aufgezeigt (Abschnitt 6).

2 Faktorieller Survey: Aufbau, Motivation und Probleme

Der Faktorielle Survey zielt darauf ab, die *relativen* Gewichte einzelner Objekt- oder Situationsmerkmale für Einstellungen, Bewertungen oder Entscheidungen zu bestimmen (für detaillierte Einführungen Beck/Opp 2001; Jasso 2006; Rossi/Anderson 1982). Dazu sind zunächst die in den Vignetten enthaltenen Merkmalsdimensionen und ihre Ausprägungen nach theoretischen Vorüberlegungen auszuwählen. In den Befragungssituationen werden diese Ausprägungen dann experimentell variiert, um zu prüfen, ob die gezielt erzeugte Variation der Objekt- und Situationsmerkmale eine entsprechende Variation der Urteile der Befragten nach sich zieht. In den Auswertungen lassen sich damit die exakten Beziehungen zwischen den Merkmalen und den Urteilen der Befragten ermitteln.

In der Durchführung Faktorieller Surveys werden die Befragten in der Regel also mit mehreren, zufällig oder systematisch ausgewählten Vignetten konfrontiert.[3] Die Befragungsmethode hat gegenüber itembasierten Survey-Studien vier wesentliche Vorteile. *Erstens* erlaubt sie eine Konstruktion von Objekten und Situationen, bei denen eine Mehrzahl solcher Merkmale zusammentreten, die in der Realität oft stark miteinander korrelieren und deswegen keine getrennte Einschätzung ihrer Bedeutung erlauben. Im experimentellen Design des Faktoriellen Surveys lassen sich diese Faktoren isolieren, im technischen Sinn zueinander orthogonal setzen. Die so erzeugte Unkorreliertheit der Merkmale ermöglicht eine separate Bestimmung ihres jeweiligen Einflusses auf Urteil und Entscheidung. *Zweitens* können entsprechende Forschungshypothesen im Unterschied zur klassischen Laborforschung auf der Grundlage größerer (Zufalls-)Stichproben in Bevölkerungsumfragen überprüft werden. *Drittens* eröffnen sich interessante Analysemöglichkeiten, wenn den Befragten mehrere Vignetten vorgelegt und deshalb pro Befragten mehrere Urteile erzielt werden. Dadurch entsteht eine hierarchische Mehrebenenstruktur, die genutzt werden kann, um zwischen ‚between-' und ‚within-subject'-Faktoren zu unterscheiden. Es ist möglich, die Kovariation des Einflusses von Vignetten- und

3 Dies stellt auch das Standardvorgehen in der vornehmlich in der Marktforschung verwendeten und dem Faktoriellen Survey ähnlichen Conjoint-Analyse dar (Carroll/Green 1995). Hier werden den Probanden meist simulierte oder echte Produktbeschreibungen vorgelegt und anschließend die relativen Nutzenwerte je Produktmerkmal ermittelt. Die Produkte weisen wie die Vignetten ein mehrfaktorielles Merkmalsbündel auf (Klein 2002; Orme 2006). Geht es um die Ermittlung von Entscheidungen, werden dagegen zum Teil nur wenige, in manchen Fällen nur eine einzige Vignette präsentiert. Es gibt durchaus Argumente, bei randomisierter Verteilung der Vignetten auf die Befragten nur eine einzige Vignette zu präsentieren: Die Effekte sozialer Erwünschtheit sowie die in diesem Aufsatz thematisierten Lerneffekte werden vermindert. In solchen Studien muss auch kein Mehrebenendesign bemüht werden (Jann 2003).

Befragtenmerkmalen auf die Urteile zu ermitteln. *Viertens* kann mit Faktoriellen Surveys einem gewichtigen Vorwurf an die konventionelle Einstellungsmessung begegnet werden, die Analyse lediglich einzelner Item-Werte würde der komplexen Struktur von Einstellungen nicht gerecht (Jasso/Opp 1997: 949; Liebig/Mau 2002: 114-116). Im Faktoriellen Survey sind komplexe Beurteilungs- und Entscheidungsprobleme simulierbar, indem eine Vielzahl von Merkmalen gekreuzt wird. Dies gilt insbesondere für solche Objekte und Situationen, bei denen verschiedene Objekt- oder Situationsmerkmale in unterschiedlichem Grad urteilsrelevant werden und bei denen der soziale Kontext einer Entscheidungssituation eine wichtige Rolle spielt. So wird beispielsweise die Höhe eines als gerecht empfundenen Erwerbseinkommens für einen Erwerbstätigen oder das gerechte Strafmaß für einen Verurteilten an das Vorliegen verschiedener Bedingungen gekoppelt sein. Genau diese Bedingungen können im Rahmen des Faktoriellen Surveys berücksichtigt und ‚alltagsnah' simuliert werden. Durch eine solche ‚Verbundmessung' könne – so die Argumentation einiger Autoren (Hechter/Kim/Baer 2005; Jasso 1988; Dülmer/Klein 2003) – eine validere Messung von Einstellungen erzielt werden als durch itembasierte Verfahren. Denn die Einstellungen zu den einzelnen Dimensionen werden nicht sequenziell, sondern in der Situationsbeschreibung gemeinsam erfragt. Darüber hinaus verhindere die wiederholte Bewertung einer größeren Anzahl von Objekten und Situationen, dass Befragte ein ‚falsches' oder ‚künstliches' Bild ihrer Einstellungen zeichnen (Hechter et al. 1999). Tatsächlich haben Vergleiche von item- und vignettenbasierten Messungen gezeigt, dass über Faktorielle Surveys erfasste Einstellungen weniger durch soziale Erwünschtheit verzerrt werden (Jann 2003; Liebig/Mau 2002; Smith 1986). Vor diesem Hintergrund resümieren Dülmer/Klein (2003), dass über die Vignettenanalyse eine vergleichsweise exakte Einstellungsmessung möglich sei (siehe auch Hechter/Kim/Baer 2005: 103; Jasso 1988).

Von Kritikern des Faktoriellen Surveys werden aber auch eine ganze Reihe von Nachteilen bzw. Unzulänglichkeiten genannt. Grundsätzliche Einwände beziehen sich zunächst auf den vergleichsweise hohen zeitlichen Befragungsaufwand und die daraus resultierenden Opportunitätskosten bezüglich der Erhebung alternativer Items (Sniderman/Grob 1996). Die Bewertung von zehn und mehr Vignetten ist zeitlich aufwändiger als eine entsprechende itembasierte Abfrage der Dimensionen (Dülmer/Klein 2003; Liebig/Mau 2002). Als problematischer wird jedoch angesehen, dass bei Vignettenstudien vergleichsweise starke Antworteffekte plausibel sind, die sich aus der Auswahl der Beispiele (z. B. Kontrasteffekte), deren Reihenfolge (*carry-over*-Effekte) oder aus der Komplexität der präsentierten Beispiele ergeben können. Mit Faktoriellen Surveys erhobene Einstellungen wären daher höchst instabil und letztlich Artefakte. Letzteres sei insbesondere dann zu erwarten, wenn

die Befragten aufgrund der hohen Komplexität der Bewertungsaufgabe überfordert seien. Sie würden mitunter solche Dimensionen in ihr Antwortverhalten einfließen lassen, denen sie ‚eigentlich' gar keine Bedeutung zumessen. Kritisch angemerkt wird diesbezüglich zudem die Gefahr einer zu starken oder ausschließlichen Konzentration der Befragten auf ein in sich stimmiges Antwortverhalten (Faia 1980; Seyde 2005). Ferner können mögliche Kontexteffekte durch Namen, Begriffe oder Bezeichnungen entstehen und Störeffekte hervorrufen, die aus den individuellen Erfahrungen der Befragten stammen (welche den unterschiedlichsten Alltagssituationen inhärent sind) und kaum kontrollierbar sind.[4] Diese Einwände konnten bislang aufgrund fehlender Methodenstudien weder bestätigt noch entkräftet werden.

Der vorliegende Beitrag bezieht sich auf derartige Forschungslücken und entstand in der ersten Phase eines breiter ansetzenden, von der Deutschen Forschungsgemeinschaft (DFG) finanzierten Projekts der Universitäten Konstanz und Bielefeld.[5] Die hier präsentierten Analysen konzentrieren sich auf folgende drei Aspekte: (1) Zunächst geht es um die Bestimmung einer noch handhabbaren Komplexität der geschilderten Situationen (Beck/Opp 2001: 287; Rossi/Anderson 1982: 59). Diese wird anhand der Menge an variablen Dimensionen untersucht. Mögliche kognitive Über- bzw. Unterforderungen sind allerdings nicht unabhängig von Lerneffekten durch die wiederholte Bearbeitung von Vignetten zu beurteilen, weshalb wir als weiteren Aspekt (2) die Konsistenz des Urteilsverhaltens im Bearbeitungsverlauf analysieren. Schließlich adressieren wir (3) die Auswirkung von unplausiblen Fällen, die – wie noch ausführlicher begründet – ebenfalls in Wechselwirkung mit diesen beiden anderen Aspekten zu sehen ist.

4 Fraglich ist schließlich auch die prognostische Validität des Verfahrens (Rooks et al. 2000), da die Befragten nur hypothetische und nicht aktuelle Entscheidungen treffen (dazu Hechter/Kim/Baer 2005; für Versuche einer externen Validierung Eifler 2007; Groß/Börensen 2009; Nisic/Auspurg 2009).

5 Im Rahmen dieses DFG-Forschungsprojekts werden vielfältige experimentelle Variationen zur Komplexität der Erhebungssituation (Anzahl der Merkmalsdimensionen und Vignetten sowie Relevanz von möglichen Reihenfolgeeffekten) und zur Bedeutung von Darstellungsformen (Bandbreite der Ausprägungen bzw. ‚range'-Effekte, Einflüsse verschiedener Beurteilungsskalen und Präsentationsformen) untersucht. Außerdem gilt die Aufmerksamkeit der zeitlichen Stabilität der Messungen. Umfangreiche Experimentalreihen werden mit einem Studierenden-Sample bearbeitet, es geht darauf aufbauend im Projekt aber ebenso um die Tauglichkeit der Befragungsmethode in *allgemeinen* Bevölkerungsumfragen. Um die Belastbarkeit der Befragten und den Zeitaufwand alters- und bildungsübergreifend einschätzen zu können, werden unterschiedlich komplexe Designs an einer bevölkerungsrepräsentativen Stichprobe in zwei Surveysituationen (‚face-to-face' und schriftlich) getestet. Für nähere Informationen: http://www.uni-konstanz.de/hinz/?cont=faktorieller_survey&lang=de.

3 Forschungsstand und Hypothesen

Im Folgenden berichten wir den Forschungsstand zu den drei benannten methodischen Problemen und leiten daraus Hypothesen zu den Effekten auf das Antwortverhalten ab. Aufgrund der unzureichenden Forschungslage zu Faktoriellen Surveys ziehen wir mitunter Literatur zu verwandten Verfahren der Marktforschung und der Umwelt- und Gesundheitsökonomie heran (Conjoint- und Choice-Experimente).

3.1 Komplexität der Vignetten: Anzahl der Dimensionen

Wie bereits erwähnt, ist der Faktorielle Survey insbesondere für Fragestellungen geeignet, bei denen komplexe Bewertungen vorzunehmen sind. Der Wunsch, über viele Dimensionen eine möglichst detaillierte und ‚alltagsnahe' Beschreibung zu erhalten, kollidiert allerdings mit der eingeschränkten Verarbeitungskapazität der Befragten. Die Entscheidung für eine bestimmte Anzahl von Dimensionen ist somit von weit reichender Bedeutung (Rossi/Anderson 1982). Dies gilt, weil die Anzahl der Dimensionen über die Länge der Situationsbeschreibungen und damit die Komplexität der Bewertungsaufgabe entscheidet. Eine Vielzahl von Dimensionen erzeugt für die Befragten eine möglicherweise nicht mehr oder nur schwer handhabbare Komplexität. Die Folge wäre, dass die entsprechenden Urteile – falls es nicht zum vorzeitigen Abbruch kommt – im ungünstigsten Fall nur noch Artefakte darstellen. Jasso (2006) schlägt vor, nur solche Dimensionen auszuwählen, von denen eine Relevanz für die Bewertung bekannt ist. Dies kann durch theoretische Überlegungen, vorherige Untersuchungen oder aufgrund von Alltagsbeobachtungen geschehen. In Anknüpfung an kognitionspsychologische Arbeiten argumentiert sie zudem, dass Personen nur wenige Dimensionen zur Meinungsbildung heranziehen. Rossi und Anderson (1982) empfehlen, sich auf sechs Dimensionen zu beschränken. In den bislang durchgeführten Faktoriellen Surveys reicht die Anzahl der verwendeten Dimensionen unseres Wissens von drei (Berk/Rossi 1977) bis 21 (Shlay et al. 2005). In der Mehrzahl der Studien werden fünf bis sieben Dimensionen verwendet. Man stützt sich dabei allerdings nur auf eine ‚Daumenregel' aus den Informations- und Kognitionswissenschaften, wonach Menschen sieben plus/minus zwei Informationen am besten verarbeiten können (Zimbardo 1988: 275). Es zeigt sich also, dass die bisherige Forschungspraxis durch sehr unterschiedliche Vorgehensweisen bestimmt ist. Die in der Literatur zu findenden Empfehlungen gehen über allgemeine Ratschläge nicht wirklich hinaus, etwa wenn Beck und Opp

(2001: 287) raten, die Ausprägungen aus Hypothesen zu generieren und nur solche zu verwenden, bei deren Variation man einen tatsächlichen Einfluss vermutet.[6]

Die zunächst nahe liegende, grundsätzliche Annahme lautet, dass die kognitive Anforderung für die Befragten mit der Anzahl der Dimensionen steigt, bis hin zu einer eventuell nicht mehr handhabbaren Komplexität (Rossi/Anderson 1982; für Choice- und Conjoint-Analysen Melles 2001; DeShazo/Fermo 2002). Weitaus weniger klar ist, wie sich die dann zu erwartende Tendenz zur Vereinfachung äußert. Neben einem kompletten Befragungsabbruch und Item-Nonresponses kommt ebenso ein inkonsistenteres Antwortverhalten in Frage. Alternativ sind Heuristiken in Form eines vollständigen Ausblendens inhaltlich weniger relevanter (oder vergleichsweise unauffällig operationalisierter, da z. B. mit weniger Ausprägungen vorgegebener) Dimensionen erwartbar (Wason/Polonsky/Hyman 2002; für Befunde bei Choice- und Conjoint-Analysen Swait/Adamowicz 2001; Melles 2001; DeShazo/Fermo 2002). Vertreten wird bei Choice- und Conjoint-Analysen zudem auch die Gegenhypothese eines *konsistenteren* Antwortverhaltens bei mehr Dimensionen (Sauer 2009). Die dahinter stehende Annahme ist, dass den wenig-dimensionalen Vignetten urteilsrelevante Informationen fehlen, die daher von den Befragten selbst ‚konstruiert' werden müssen.[7] Gegenüber der expliziten Vorgabe durch den Forscher bedeutet die ‚Unterkomplexität' eine geringere inhaltliche Kontrolle über das Vignettenexperiment, was zumindest befragtenübergreifend eine geringere Präzision der Schätzungen erwarten lässt (DeShazo/Fermo 2002; Caussade et al. 2005: 632; Johnson 2006: 46f.). Ähnlich wird vermutet, dass unkontrollierte ‚Framing'-Effekte wahrscheinlicher werden (dazu z. B. Melles 2001: 186). Und schließlich gilt auch ein Informationsmangel als kognitiv belastend, weil es beispielsweise bei wenigen Merkmalsvorgaben schwieriger ist, Unterschiede in den Fallbeispielen zu erkennen und damit zwischen ihnen zu differenzieren (für dieses Argument bei Choice-Experimenten Hensher 2006). Als ein erster Beleg für einen solchen ‚information-underload' können die Befunde einer Wiederholungsbefragung gewertet werden, bei der Studierende zu drei Messzeitpunkten mit den jeweils selben Vignetten befragt wurden: Die Stabilität der Urteile erwies sich bei acht Dimensionen höher als bei fünf Dimensionen (Liebig/Meyermann/Schulze 2006).

Für alle Effekte ist jedenfalls unklar, ab welcher Dimensionszahl mit ihnen zu rechnen ist. Für die vorliegende Untersuchung wird daher mit fünf versus zwölf Di-

6 Neben der Anzahl der Dimensionen ist auch die Zahl der Ausprägungen pro Dimension relevant, weil damit die Größe des ‚Vignettenuniversums' festgelegt wird. Als Vignettenuniversum wird die Gesamtheit aller möglichen Varianten der Situations- bzw. Objektbeschreibungen bezeichnet.

7 In Vignettenstudien zur Einkommensgerechtigkeit könnte ein solches Informationsdefizit z. B. in der Berufserfahrung der Einkommensbezieher bestehen.

mensionen bewusst ein starker Kontrast gewählt. Die – gemessen an den vorliegenden Studien mit überwiegend fünf bis neun Dimensionen – überdurchschnittliche maximale Dimensionszahl von zwölf lässt ein Durchschlagen des ‚Überforderungseffektes' erwarten. Es ergeben sich zwei Unterhypothesen:

H_{1a}: *Bei zwölf Dimensionen sind Befragungsabbrüche häufiger als bei fünf Dimensionen.*

H_{1b}: *Das Urteilsverhalten ist bei zwölf Dimensionen inkonsistenter als bei fünf Dimensionen.*

Alternativ ist von einer vereinfachten Urteilsstrategie in Form einer Ausblendung einzelner Merkmale auszugehen (zu dieser ‚dimensional-reductions'-Strategie bei Choice-Anaylsen: Swait/Adamowicz 2001: 137):

H_{1c}: *Bei zwölf Dimensionen sind einzelne Vignettenvariablen weniger urteilsrelevant, zeigen also geringere Einflüsse auf die Urteile als bei fünf Dimensionen.*

3.2 Lern- und Ermüdungseffekte

In fast allen Vignettenstudien sollen die einzelnen Befragten mehrere Vignetten beurteilen. Gängig sind zehn bis 20 Vignetten, in einer Studie waren es ganze 95 Vignetten pro einzelnem Befragten (Beck/Opp 2001; Rossi et al. 1974). Die mehrfache Präsentation von Vignetten ermöglicht es, selbst bei geringen Befragtenzahlen noch ausreichend viele Urteilszahlen zur Hypothesentestung zu sammeln (Auspurg/Abraham/Hinz 2009). Zudem erlaubt sie, befragtenspezifische Urteils- und Entscheidungsregeln (sog. ‚within-subject'-Effekte) aufzudecken. Mit der wiederholten Bewertungsaufgabe sind allerdings Lerneffekte zu erwarten, die mit anderen Kennzeichen der Erhebungssituation in Wechselwirkung stehen. Sehr deutlich ist dies bei der Anzahl der Dimensionen. Bei einer höheren Dimensionszahl benötigen Lernprozesse länger, gleichzeitig könnten Ermüdungserscheinungen früher einsetzen. Lern- und Ermüdungseffekte sind wechselseitige Aspekte von Komplexität. Beim Lernen geht es um ein zunehmend konsistentes Antwortverhalten sowie um das Vermögen, mehr Dimensionen gleichzeitig in ein Urteil zu integrieren.[8] Ermü-

8 Eine im Befragungsverlauf zunehmende Beachtung von Dimensionen wird zudem damit begründet, dass die Probanden die in der Realität korrelierten Merkmale zu Beginn als redundant ansehen. Erst wenn sie nach einer ganzen Reihe von präsentierten Vignetten erkennen, dass sie im experimentellen Design unabhängig voneinander variieren, schenken sie ihnen mehr Aufmerksamkeit bzw. lassen sie separat in ihr Urteil einfließen (für Conjoint-Analysen Melles 2001: 118).

dungs- und Langeweile-Effekte schlagen sich umgekehrt in einer sinkenden Konsistenz und in einer Beachtung weniger Merkmale oder anderen vereinfachten Entscheidungsregeln nieder (für Choice-Analysen: Carson et al. 1994: 335f.).[9] Die Rolle und das Ausmaß von Lern- und Ermüdungseffekten sind für Vignettenstudien bislang unerforscht. Ebenso ist es eine noch völlig ungeklärte Frage, ab welcher Vignettenzahl mit einem Umkippen von Lern- in Ermüdungseffekte zu rechnen ist.

Als ein erster Orientierungspunkt können Erfahrungen aus den verwandten Choice-Experimenten herangezogen werden. Demnach nimmt die Urteilskonsistenz bis etwa zum zehnten Urteil zu, um danach wieder abzusinken (z. B. Bradley/Daly 1994: 180; Caussade et al. 2005: 631f.). Da selbst bei Vignettenstudien mit 50 oder mehr Vignetten bislang keine nennenswerten Probleme im Hinblick auf die Urteilsgüte berichtet werden (Jasso 2006), scheint bei der vorliegenden Fallzahl von maximal zehn Vignetten pro Befragten (dazu unten Abschnitt 4) eine Dominanz der Lerneffekte plausibel. Es ergeben sich die folgenden Annahmen:

H_{2a}: *Mit der Position der Vignetten steigt die Konsistenz des Antwortverhaltens und/oder die Anzahl berücksichtigter Dimensionen.*

H_{2b}: *Diese Lerneffekte treten stärker bei zwölf als bei fünf Dimensionen auf.*

3.3 Behandlung unlogischer Fälle

Bevor die tatsächlich zu bewertenden Vignetten zusammengestellt werden (also eine Auswahl aus dem Universum aller möglichen Kombinationen von Merkmalsausprägungen getroffen wird; dazu Beck/Opp 2001; Steiner/Atzmüller 2006; Dülmer 2007), ist es bisher gängige Praxis, ‚unlogische und unplausible Fälle' zu eliminieren. Es werden also solche Vignetten ausgeschlossen, die offensichtlich ungewöhnliche oder unsinnige Merkmalskombinationen enthalten. Ein Beispiel dafür wären erwerbstätige Personen ohne Schul- oder Berufsausbildung in einem Beruf, bei dem eine Ausbildung unabdingbar ist (etwa Richter, Hochschullehrer). Der Ausschluss solcher Fälle wird vor allem mit den zu erwartenden Folgen für das Antwortverhalten begründet. Offensichtlich unsinnige Fälle würden die Ernsthaftigkeit der Bewertungsaufgabe in Frage stellen und zu einem Anstieg der Item-Non-Response-Quote, oder gar zum völligen Befragungsabbruch (Faia 1980; Jasso 2006) führen.

9 Grafisch ist also ein umgekehrt u-förmiger Zusammenhang zwischen der Bearbeitungsabfolge der Vignetten und der Konsistenz bzw. Anzahl berücksichtiger Dimensionen zu erwarten.

Dieses Argument ist durchaus plausibel, doch sind die Kriterien, was als unlogisch oder unsinnig zu gelten hat, sehr vage. In vielen Faktoriellen Surveys geht es darum, möglichst unabhängig von den gängigen Normen, bestehenden Gesetzen und empirischen Beobachtungen Bewertungen vornehmen zu lassen, um so auch die kontrafaktischen Meinungen und Überzeugungen der Befragten zu erheben. Die Norm eines ‚logischen Falles' wird durch empirische Regelmäßigkeiten und damit zusammenhängenden Erwartungshaltungen geprägt. Faktorielle Surveys bieten jedoch die seltene Möglichkeit, die Probanden bewusst mit abweichenden Fällen zu konfrontierten – und gerade in der Reaktion auf solche ‚abweichende' Fälle kann ein Erkenntnisziel liegen. In dieser Hinsicht sind Eingriffe in die Merkmalskombinationen problematisch, engen sie doch die Variation der Situations- und Objektbeschreibungen *a priori* auf ein empirisch vorfindbares Maß ein (Beck/Opp 2001).

Solides methodisches Wissen besteht bislang ausschließlich im Hinblick auf die *statistischen* Folgen. Durch den gezielten Ausschluss einzelner Fälle wird die Orthogonalität der Dimensionen im Vignettenuniversum eingeschränkt, Multikollinearität wird also erzwungen (zu deren Konsequenzen für Schätzverfahren: Greene 2003: 56-59; Wooldridge 2003: 96-100). Die Relevanz des Ausschlusses von Fällen für die Balanciertheit und Unkorreliertheit von Vignettensamples ist inzwischen gut einschätzbar (Kuhfeld/Randall/Garratt 1994: 551; Dülmer 2007: 391f.; Steiner/Atzmüller 2006) und es liegen Algorithmen vor, welche die Einbußen an Effizienz gezielt minimieren (dazu Kuhfeld 2005). Aufgrund des andernfalls drohenden Effizienzverlustes lautet daher die eindeutige Empfehlung, diese Algorithmen auch einzusetzen.

Die Auswirkungen der unplausiblen oder unlogischen Fälle auf das *Antwortverhalten* sind dagegen weitaus strittiger, was vor allem durch fehlende einschlägige Untersuchungen bedingt ist.[10] Trifft die oben angesprochene Vermutung zu, dass durch unplausible Vignetten der grundsätzliche Glaube an den Wert der Befragung und damit den Nutzen eigener Mitarbeit beeinträchtigt wird, sind Befragungsabbrüche und invalide Antworten zu erwarten (Response-Sets oder flüchtige und inkonsistente Urteile). Es ergeben sich daher zunächst die folgenden Hypothesen:

H_{3a}: *Werden den Befragten unplausible Fälle vorgelegt, sind Befragungsabbrüche häufiger, als wenn dies nicht der Fall ist.*

H_{3b}: *Werden die Befragten mit unplausiblen Fällen konfrontiert, ist die Konsistenz ihres Antwortverhaltens geringer, als wenn dies nicht der Fall ist.*

10 Die zwischen dem Autorenteam Rossi/Alves (1980) und Faia (1980) ausgetragene Diskussion über die ‚Sinnigkeit' bzw. den Nutzen unplausibler Vignetten ist daher nach wie vor nicht mit empirischen Argumenten zu entscheiden.

Faia (1980) erwartet zudem, dass die für die Unplausibilität ursächlichen Dimensionen in den Vordergrund geraten – die Befragten würden die Aufgabe in einen reinen ‚Intelligenztest' zur Entlarvung von ‚Anomalien' uminterpretieren. Gerade dies würde die Gültigkeit der Urteile beinträchtigen und verdient daher eine Überprüfung:

H_{3c}: *Nach einer Konfrontation mit unplausiblen Fällen beziehen die Befragten primär die für die Unplausibilität verantwortlichen Dimensionen in ihre Urteile ein, gewinnen diese somit relativ zu allen anderen Dimensionen an Bedeutung.*

Als alternative Begründung hierfür lässt sich ein Lerneffekt anführen: Die Befragten bemerken erst bei einer empirisch seltenen Kombination, dass die Merkmale unabhängig voneinander variieren und damit nicht redundant sind. Ähnlich könnte sich so eine sinkende Bereitschaft zu differenzierten Urteilen manifestieren: Dimensionen verlieren durch ein Umschwenken auf ein vereinfachtes, weniger Merkmale einbeziehendes und daher kognitiv weniger belastendes Antwortverhalten an Relevanz.

Die Diskussion dieser drei Problemstellungen verdeutlicht, dass komplexe Wechselwirkungen zwischen den methodischen Aspekten von Faktoriellen Surveys zu erwarten sind. Wir können hier schon aus Platzgründen nur die besonders nahe liegenden Zusammenhänge analysieren, im genannten DFG-Projekt wird derzeit ein weitaus größeres Spektrum methodischer Effekte untersucht.

4 Methodik und Datengrundlage

Die drei methodischen Probleme lassen sich nicht analytisch lösen, sondern erfordern eine empirische Herangehensweise. Ideal dazu ist ein Methodenexperiment, bei dem die Bedeutung von Designelementen für das Antwortverhalten durch ihre gezielte Variation beobachtbar wird. Wichtig ist, dass die methodischen Splits zufällig auf die Befragten verteilt werden und sie zudem nicht mit einzelnen Vignetten(decks) korreliert sind – wie bei jedem Experiment erlaubt erst diese Randomisierung, unbekannte Drittvariablen der Befragten zu neutralisieren und ungewünschte Konfundierungen mit den inhaltlichen Dimensionen der Vignetten zu vermeiden.[11] In den vorliegenden Experimenten wird die Komplexität der Vignetten über die Zahl der Dimensionen variiert: Etwa die Hälfte der Befragten bekommt durchgehend

11 Bei Choice-Experimenten werden derartige Studien unter dem Namen ‚Design of Design' geführt (z. B. Hensher 2004, 2006; Caussade et al. 2005). Im Prinzip handelt es sich um eine mehrfaktorielle Erweiterung des ‚split-ballot'-Designs: Es werden gleich *mehrere* Designelemente unabhängig voneinander variiert (dazu Sniderman/Grob 1996).

Vignetten mit fünf, die andere mit zwölf Dimensionen vorgelegt (es handelt sich also um ein reines ‚between-subject'-Design).[12] Zunächst wurden jedem Teilnehmer sieben Vignetten zugeteilt; aufgrund der geringen Abbruchquote wurde diese Zahl in einer zweiten (kleineren) Befragungswelle auf zehn erhöht.

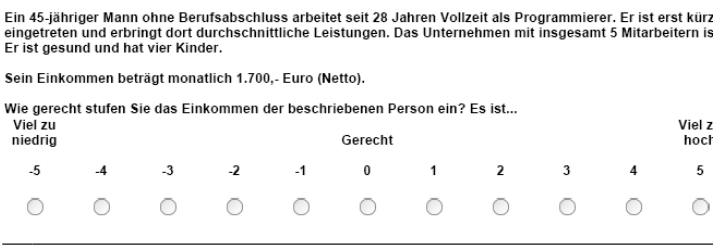
Als inhaltliche Fragestellung dient der besonders gut erforschte ‚Klassiker' von Vignettenstudien – die Erhebung von Einkommensgerechtigkeit (z. B. Alves/Rossi 1978; Jasso/Webster 1997, 1999; Jann 2003; Hermkens/Boerman 1989, Shepelak/Alwin 1986). Den Befragten werden jeweils fiktive Personen vorgestellt, die sich in einer Reihe von einkommensrelevanten Merkmalen unterscheiden, wie dem Geschlecht, Alter, Bildungsstand oder Beruf. Zusätzlich enthält jede Vignette das monatliche Netto-Einkommen der beschriebenen Person. Dieses soll dann auf einer elf-stufigen Ratingskala danach beurteilt werden, ob und in welchem Ausmaß es (un-)gerecht erscheint. Abbildung 1 zeigt eine Beispielvignette mit zwölf Dimensionen. Die Ausprägungen der Dimensionen sind darunter im Überblick aufgeführt.[13] Bei der Auswahl der Merkmale wurde darauf geachtet, dass ihre Relevanz für das Urteilsverhalten bereits belegt ist. Damit sollte sichergestellt werden, dass eine mögliche Nicht-Beachtung methodisch und nicht inhaltlich zu deuten ist (ähnlich für Choice-Experimente Hensher 2006: 16).

Um die zufällige Variation der experimentellen Splits und Vignetten mit verhältnismäßig wenig Aufwand umsetzen zu können, fiel die Wahl auf eine Online-Befragung. Ein weiterer Grund für diesen Befragungsmodus ist die gute Erfassbarkeit von Metadaten (z. B. Beantwortungszeiten), welche zusätzlichen Aufschluss über die Bearbeitungsstrategien versprechen. Bei Experimenten kommt es nicht auf eine repräsentative und zufällige Stichprobe der Probanden an, sondern es sind zumindest bei kleinen Stichproben homogene Experimentalgruppen vorteilhaft (da diese ein geringeres Risiko ungleich verteilter Drittvariablen bergen; z. B. Diekmann 2007: 337ff.). Ihre relativ große Homogenität und ihre gute Erreichbarkeit sprachen für die Wahl von Studierenden verschiedener Universitäten, die über E-Mail-Verteiler der Fachschaften kontaktiert und mit einem Link zur Befragung um ihre Teilnahme gebeten wurden.

12 Dies ist nicht ganz korrekt, denn als zweiter experimenteller Faktor wurde eine der Dimensionen, das Geschlecht der Vignettenpersonen, nur bei einem Teil der Befragten zwischen den Vignetten variiert (‚within'-Variation). Den anderen Befragten wurden stets nur Vignetten eines Geschlechts vorgelegt (‚between'-Variation), sie bewerteten also durchgehend jeweils nur Beschreibungen mit männlichen oder weiblichen Protagonisten, womit sich für sie die Anzahl variabler Dimensionen auf vier bzw. elf Merkmale reduziert. Der Hintergrund dieses Splits ist der, dass sich damit Effekte sozialer Erwünschtheit bzw. eines bewussten vs. unbewussten Urteilsverhaltens untersuchen lassen. Da dieser Faktor aber vollständig unabhängig variiert wurde, kann er an dieser Stelle und den nachfolgenden Analysen ausgeblendet werden – er verdient eine eigenständige Betrachtung.

13 Diese Aufstellung aller Dimensionen dient hier nur als Information für den Leser; den Befragten wurde diese Übersicht nicht vorgelegt.

Abbildung 1 Beispielvignette mit zwölf Dimensionen



Vignettendimensionen und Ausprägungen:

1) *Alter:* 25, 35, 45, 55 Jahre
2) *Geschlecht:* Mann, Frau
3) *Berufsabschluss:* ohne Berufsabschluss, mit abgeschlossener Berufsausbildung, mit Hochschulabschluss
4) *Beruf:* 10 Ausprägungen von Hilfsarbeiter/in bis Anwalt (Auswahl nach Dezentilen der Magnitude-Prestige-Skala)
5) *Einkommen:* 10 Ausprägungen von 250,- bis 15.000,- Euro Netto
6) *Berufserfahrung:* keine, 25%, 50%, 100% der potenziellen Erwerbszeit
7) *Betriebszugehörigkeit:* erst kürzlich eingetreten, schon seit langem im Unternehmen beschäftigt
8) *Leistung:* unterdurchschnittlich, durchschnittlich, überdurchschnittlich
9) *Betriebsgröße:* 5, 20, 200, 2.000 Mitarbeiter
10) *Wirtschaftliche Lage des Unternehmens:* vom Konkurs bedroht, ausgeglichene Bilanz, hohe Gewinne
11) *Gesundheitszustand:* gesund, 30% schwerbehindert
12) *Kinder:* 6 Ausprägungen von keine bis 5 Kinder.

Bei den Vignetten handelt es sich um eine fraktionalisierte Auswahl aus dem kompletten Universum für zwölf Dimensionen, wobei auf eine Orthogonalisierung aller Haupteffekte geachtet wurde (sog. ‚resolution III-Design', Kuhfeld/Randall/Garratt 1994: 546). Mit dieser Anforderung sind bei der vorliegenden Spezifikation von Dimensionen und Ausprägungen etwa 100 Vignetten für eine effiziente Stichprobe hinreichend.[14] Durch den Ausschluss logisch unmöglicher Kombinationen (wie Personen ohne Berufserfahrung, die schon lange im Betrieb arbeiten), reduzierte sich das Sample weiter zu insgesamt 93 unterschiedlichen Vignetten (empirisch seltene, aber gleichwohl mögliche Fälle wurden dagegen bewusst beibehalten – mehr dazu unten). Bei dem Split mit fünf Dimensionen wurde exakt dieselbe Vignettenstichprobe eingesetzt (es wurden einfach die überzähligen Dimensionen gelöscht). Zwar ließen sich für diese ‚sparsameren' Vignetten weitaus effizientere Designs bilden, gerade diese statistischen Effizienzwerte sollten aber konstant gehalten werden, um eine reine Abschätzung der *methodischen* Effekte zu ermöglichen. Nur unter Kontrolle der statistischen Effizienz lassen sich Unterschiede in den Signifikanzen von

14 Es wird eine D-Effizienz von 98,2 erreicht, wobei Werte über 90 als zufrieden stellend gelten (Kuhfeld 2005). Allerdings reduziert sich die Effizienz mit dem Ausschluss unlogischer Fälle wieder.

Regressionskoeffizienten tatsächlich auf das Antwortverhalten zurückführen.[15] Zugleich wird mit der Verwendung identischer Vignettensamples für die Splits mit fünf und zwölf Dimensionen einer Vermischung von inhaltlichen und Designeffekten vorgebeugt. Es lassen sich durch dieses Vorgehen auftretende Unterschiede im Antwortverhalten eindeutiger auf die differente Anzahl an Dimensionen zurückführen statt auf unterschiedliche inhaltliche Kombinationen der Vignettendimensionen.

Alle Teilnehmer wurden zufällig einem der beiden methodischen Splits sowie einem Subset an Vignetten zugewiesen. Pro Befragten wurde eine eigene Zufallsziehung von Vignetten (Ziehung ohne Zurücklegen) vorgenommen. Mit dieser randomisierten Setbildung sollte eine möglichst hohe Ausschöpfung der Stichprobe von 93 Vignetten gewährleistet werden. Zudem wurde eine befragtenspezifische, zufällige Reihenfolge der Vignetten gewählt, um Kontrast- und Reihenfolge-Effekte auszuschließen: ‚Extreme' Vignetten verteilen sich dann zufällig auf die Bearbeitungspositionen, womit über alle Befragten hinweg zu beobachtende Einflüsse der Reihenfolge eindeutiger als Lern- bzw. Ermüdungseffekte zu deuten sind. Die befragtenspezifische Zufallsauswahl von Vignetten hat zudem den Vorteil, dass sich automatisch weitere methodische Variationen zwischen den Befragten ergeben, etwa im Auftreten und der Häufigkeit von unplausiblen Fällen.[16]

Die Befragung fand im Zeitraum Dezember 2007 bis März 2008 statt. Die Vignetten wurden in einen Rahmenfragebogen integriert, in dem neben soziodemographischen Merkmalen politische und soziale Einstellungen über ‚klassische'

15 Schließlich ist für die Präzision der Schätzungen die statistische Effizienz der Vignettenstichprobe ähnlich wichtig wie die ‚kognitive Effizienz' der von den Befragten abgegebenen Urteile (für entsprechende Argumente in Bezug auf Choice- und Conjoint-Analysen Melles 2001: 109; Louviere 2001b).

16 Bei der Alternative einer bewussten bzw. fraktionalisierten Setbildung wären zwar Konfundierungen besser kontrollierbar, aber angesichts der geringen Setgröße von sieben bzw. zehn Vignetten auch nicht vermeidbar – gerade für die komplexere Variante mit zwölf Dimensionen wären der Preis unweigerlich starke Kontexteffekte der einzelnen Sets (selbst Haupteffekte sind innerhalb der einzelnen Sets untereinander korreliert). Aus diesen Gründen ist der Einsatz einer möglichst hohen Anzahl an unterschiedlichen Sets vorzuziehen, zumal angesichts des homogenen Samples und der hohen Befragtenzahl die Gefahr der Konfundierung von Vignetten- mit Befragtenmerkmalen gering erscheint (siehe für eine ausführliche Diskussion der Vor- und Nachteile unterschiedlicher Setbildungen Steiner/Atzmüller 2006). Hinzu kommt, dass fraktionalisierte Setbildungen einem der Analyseziele zuwiderlaufen: Sie arbeiten mit einer möglichst gleichmäßigen Verteilung von Extremfällen, was impliziert, dass auch unplausible Fälle sehr regelmäßig auf die Sets bzw. Befragten verteilt werden und daher die ‚between'-Varianz zu gering ausfallen dürfte, um ihren Einfluss verlässlich zu prüfen. Insgesamt lassen diese Abwägungen somit bei den vorliegenden Analysezielen eine randomisierte Setbildung als vorteilhaft erscheinen. Das mit ihr verbundene Risiko einer unbalancierten Verteilung von Vignetten auf die Splits mit fünf vs. zwölf Dimensionen (bzw. sieben vs. zehn Vignetten) ist angesichts der hohen Set- und Befragtenzahlen gering. Problematisch wären für die angestrebten Analysen insbesondere Unterschiede in den Korrelationsstrukturen. Diese stimmen jedoch in der Tat sehr gut zwischen den einzelnen Splits überein, wie die im Anhang aufgeführten Korrelationsmatrizen belegen (Tabellen A1 und A2).

Itemabfragen erhoben wurden. Den Befragungslink haben 558 Personen aufgerufen, für die Vignetten liegen 3.480 Urteile von insgesamt 460 Probanden vor.[17] Tabelle 1 zeigt die für die einzelnen experimentellen Varianten realisierten Fallzahlen.

Tabelle 1 Realisierte Fallzahlen für Vignettenurteile und Befragte[a]

	5 Dimensionen		12 Dimensionen		Summe	
	Vignetten	Befragte	Vignetten	Befragte	Vignetten	Befragte
Sieben Vignetten pro Befragten	1.213	176	1.109	162	2.322	338
Zehn Vignetten pro Befragten	574	59	584	63	1.158	122
Summe	1.787	235	1.693	225	3.480	460

[a] *Nur Befragte, die mindestens eine Vignette beantwortet haben.*

Bei der Datenauswertung ist die Mehrebenenstruktur zu beachten. Werden Befragten mehrere Vignetten vorgelegt, entsteht ein hierarchischer Datensatz (für eine anschauliche Darstellung Beck/Opp 2001). Auf der untersten Ebene stehen die Vignettenurteile, eine zweite Analyseebene bilden die Merkmale der Befragten. Da wir nur auf die Analyse von Vignettendimensionen (der ersten Ebene) abstellen und zudem ein homogenes Befragtensample verwenden, berücksichtigen wir die Datenstruktur lediglich durch die Schätzung von robusten Standardfehlern (Wooldridge 2003: 258ff., Wooldridge 2002; zur Modellwahl speziell bei Vignettenstudien: Jasso 2006; Auspurg/Abraham/Hinz 2009; Hox/Kreft/Hermkens 1991). Befragtenspezifische Schwankungen der Urteile und ihre mögliche Erklärung interessieren hier nicht. Die für die einzelnen Hypothesen eingesetzten Analysestrategien und Operationalisierungen werden im folgenden Abschnitt erläutert.

17 Aufgrund der verwendeten Samplingprozedur lassen sich keine Rücklaufquoten berichten. An dieser Stelle ist nochmals zu betonen, dass wir lediglich einen experimentellen Hypothesentest, nicht aber deskriptive Aussagen zu Gerechtigkeitseinstellungen anstreben. Dafür scheint der Verzicht auf eine Zufallsstichprobe unproblematisch. Mehrfachteilnahmen wurden so gut wie möglich ausgeschlossen.

5 Ergebnisse

5.1 Deskriptive Befunde

Bevor die Hypothesen mit multivariaten Analysen geprüft werden (Abschnitt 5.2), gibt ein Blick auf die deskriptiven Verteilungen und Rücklaufquoten erste Aufschlüsse über das Antwortverhalten. Insgesamt haben 124 der 558 Teilnehmer (22,2 %) die Umfrage nicht beendet. Die Abbrüche konzentrieren sich zu einem sehr großen Teil auf die Begrüßungsseite oder den Rahmenfragebogen vor den Vignetten; direkt im Vignettenteil haben lediglich 23 Befragte (4,1 % der Gesamt-Teilnehmerschaft) abgebrochen, im anschließenden Befragungsteil sind es weitere 19 Personen (3,4 %). Eine Differenzierung der Abbrüche nach experimentellen Splits erscheint angesichts dieser geringen Fallzahlen kaum sinnvoll. Festhalten lässt sich jedenfalls, dass selbst die umfangreiche Bewertungsaufgabe bei zwölf Dimensionen (der immerhin ca. die Hälfte der Befragten ausgesetzt war) und das Auftreten ungewöhnlicher Fälle (wie Anwälten ohne Hochschulabschluss) nicht zu auffallend hohen Abbruchquoten führen. Dies gilt ähnlich für Antwortverweigerungen: Lediglich 68 Vignetten, damit 1,9 % blieben unbeantwortet.[18]

Die vorangegangenen Ausführungen haben jedoch bereits gezeigt, dass sich eine mangelnde Kooperationsbereitschaft oder Überforderung ebenso in einem veränderten Antwortverhalten bei fortgesetzter Befragung niederschlagen kann – speziell dessen Verkennung wäre für die Ergebnisinterpretationen kritisch.[19] Einen ersten Hinweis auf mögliche Response-Sets liefern die Verteilungen der Vignettenurteile, wie sie in Tabelle 2 für die unterschiedlichen experimentellen Splits aufgeschlüsselt sind. Über alle Befragten hinweg (mittlere Spalte) als auch pro Befragten berechnet (letzte Spalte), wird eine etwas geringere Streuung (Standardabweichung) der Vignettenurteile, damit stärkere Konstanz des Antwortverhaltens bei zwölf gegenüber fünf Dimensionen offensichtlich. Allerdings verfehlt dieser Unterschied das Signifikanzniveau von fünf Prozent.[20]

18 Diese Quote an Missings entspricht etwa der von ‚herkömmlichen' Itemabfragen in der gleichen Erhebung. Die Befragten wurden direkt im Anschluss an die Vignetten gebeten, die Bedeutung der Vignettendimensionen für eine gerechte Entlohnung jeweils einzeln auf siebenstufigen Itemskalen einzustufen (von sollte ‚überhaupt keine Bedeutung' bis sollte ‚sehr große Bedeutung' spielen). Die Missings bewegen sich bei diesen Items zwischen 0,9 und 2,2 %, im Mittel sind es 1,2 % (vorherige Befragungsabbrüche nicht mitgezählt).

19 „If tasks are too long or too difficult or lack sufficient realism and credibility, data quality will suffer in the sense of not containing the information sought. Unfortunately, respondents generally answer the questions asked and seldom go out of their way to point out problems with tasks posed" (Carson et al. 1994: 355).

20 T-Test für die Mittelwertdifferenz der befragtenspezifischen Urteilsvarianz zwischen fünf und zwölf Dimensionen: $t = 1{,}48$; $p = 0{,}140$ bei zweiseitigem Test und Adaption für die verletzte Annahme der Varianzungleichheit (vorherige Prüfung mit Levene's Test).

Tabelle 2 Deskriptive Übersicht über die Vignettenurteile[a]

Experimentelle Variante	Anzahl	Mittelwert	S.D.	Mittlerer Mittelwert pro Befragten	Mittlere S.D. pro Befragten
5 Dimensionen, 7 Vignetten	1.213	5,21	3,10	5,20	2,94
5 Dimensionen, 10 Vignetten	574	5,44	3,21	5,45	3,12
12 Dimensionen, 7 Vignetten	1.109	5,51	2,96	5,51	2,87
12 Dimensionen, 10 Vignetten	584	5,36	2,98	5,35	2,86

[a] *Skala von 1 ‚ungerechterweise zu niedrig' bis 11 ‚ungerechterweise zu hoch'. Der Wert 6 kennzeichnet eine als gerecht empfundene Entlohnung.*

5.2 Multivariate Analysen

Erwartete Folgen einer zu hohen Komplexität sind ein inkonsistenteres Antwortverhalten (H_{1b}), statistisch eine geringere erklärte Varianz bzw. höhere Fehlervarianz, und ein Ausblenden einzelner Dimensionen (H_{1c}), was sich statistisch in geringeren Einflussstärken bzw. weniger signifikanten Effekten äußert. Zur Prüfung dieser beiden Annahmen dienen die in Tabelle 3 aufgeführten OLS-Regressionen, die wegen der hierarchischen Datenstruktur jeweils mit robusten Standardfehlern geschätzt sind. Um die methodischen Effekte besser von möglichen Drittvariableneffekten trennen zu können, werden die Regressionen für die ‚Zwölfer-Vignetten' ohne (Modell 2) und mit (Modell 3) Kontrolle der zusätzlichen Dimensionen präsentiert.[21]

Zunächst zur inhaltlichen ‚Lesart' der Ergebnisse: Bei der vorliegenden Kodierung der abhängigen Variablen bedeuten positive (negative) Koeffizientenwerte, dass das Einkommen als ungerechterweise zu hoch (niedrig) empfunden wird. Negative Effekte lassen sich somit als eine Erhöhung des als angemessen empfundenen Nettoeinkommens deuten. Nach allen drei Modellschätzungen wird beispielsweise Personen mit einem Berufsabschluss ein höheres Einkommen zugestanden als solchen ohne Abschluss. Für unser methodisches Forschungsinteresse ist aber interessanter, ob sich Unterschiede zwischen den Koeffizientenwerten der drei Modelle zeigen.

21 Prinzipiell sind die Vignettendimensionen bei fraktionalisierten Auswahlen unkorreliert. Sie geben also ihren reinen ‚Nettoeffekt' selbst dann wieder, wenn nicht auf Drittvariablen kontrolliert wird. Gerade hierin liegt ja eine wesentliche Stärke dieses Verfahrens. Einschränkung erfährt dies allerdings mit dem gezielten Ausschluss von Kombinationen, der unweigerlich zu Korrelationen führt. Dies betrifft auch das vorliegende Sample, von dem die logisch völlig unmöglichen Fälle ausgeschlossen wurden (wie z. B. Personen ohne Berufserfahrung, die schon lange in einem Betrieb arbeiten, vgl. Abschnitt 4). Eine Übersicht über die Korrelationen zwischen den einzelnen Dimensionen findet sich in Tabelle A3 im Anhang.

Tabelle 3 OLS-Regressionen der Vignettenurteile[a] (robuste Standardfehler in Klammern; sign. Unterschiede der Koeffizienten zwischen Modell 1 und 2 hervorgehoben)[b]

	Modell 1 5 Dimensionen	Modell 2 12 Dimensionen	Modell 3 12 Dimensionen
Weibliche Vignettenperson	-0,057 (0,122)	-0,136 (0,115)	-0,105 (0,113)
Alter [Jahre]	-0,021*** (0,005)	-0,029*** (0,005)	-0,020*** (0,005)
Abschluss (Ref.: kein Abschluss)			
– Berufsabschluss	-0,654*** (0,133)	-0,472*** (0,131)	-0,429*** (0,129)
– Hochschulabschluss	-1,126*** (0,129)	-0,623*** (0,126)	-0,830*** (0,130)
Berufprestige [10 MPS-Score]	-0,157*** (0,011)	-0,097*** (0,012)	-0,106*** (0,012)
Nettoeinkommen [100,- Euro]	0,060*** (0,002)	0,055*** (0,002)	0,058*** (0,002)
Berufserfahrung [Prozent der potenziellen Erwerbszeit]			0,066 (0,048)
Schon seit langem im Betrieb beschäftigt (Ref.: erst seit kurzem)			-0,645 (0,131)***
Leistung (Ref.: unterdurchschnittlich)			
– durchschnittlich			-0,813*** (0,129)
– überdurchschnittlich			-0,788*** (0,138)
Anzahl Mitarbeiter [100]			0,028*** (0,006)
Betriebssituation (Ref.: vom Konkurs bedroht)			
– ausgeglichene Bilanz			-0,037 (0,130)
– hohe Gewinne			-0,292** (0,122)
Zu 30% schwerbehindert (Ref.: gesund)			0,049 (0,114)
Anzahl Kinder			-0,152*** (0,029)
Konstante	6,465*** (0,280)	6,274*** (0,236)	6,820*** (0,272)
Beobachtungen:			
– Vignetten	1.787	1.693	1.693
– Befragte	235	225	225
R^2	0,47	0,45	0,49

[a] Skala von 1 ‚ungerechterweise zu niedrig' bis 11 ‚ungerechterweise zu hoch'. Der Wert 6 kennzeichnet eine als gerecht empfundene Entlohnung.
[b] Prüfung mittels Interaktionstermen zwischen den Vignettendimensionen und der Dimensionszahl in einem gepoolten Modell, Signifikanzniveau von fünf Prozent.
**** p<0,01, ** p<0,05, * p<0,1 bei zweiseitigem Test; Schätzungen mit robusten Standardfehlern.*

Dies ist im Hinblick auf die Vorzeichen nicht der Fall, jedoch zeigen die Vignettenmerkmale bei den komplexeren zwölfdimensionalen Varianten oftmals einen betragsmäßig schwächeren Einfluss. Ein Chow-Test bestätigt signifikante Differenzen zwischen den Modellen 1 und 2 ($F = 4{,}04$ bei $df = 7$ und 459; $p = 0{,}000$).[22] Einzeln geprüft erweisen sich die Einflussstärken des Hochschulabschlusses und des Prestiges als signifikant verschieden. Da gerade die Einflüsse dieser beiden Variablen bei Kontrolle für die weiteren Dimensionen stabil bleiben (die Koeffizienten unterschieden sich nur marginal zwischen Modell 2 und 3), ist dieser Unterschied nicht durch Drittvariableneffekte bedingt, sondern er deutet vielmehr darauf hin, dass mit höherer Komplexität tatsächlich Dimensionen tendenziell ausgeblendet werden.[23] Die Anteile erklärter Varianz (R^2-Werte), welche als Maß für die Konsistenz des Antwortverhaltens herangezogen werden können, unterscheiden sich dagegen nicht substantiell zwischen den Modellen.[24]

Insgesamt wird unsere erste Hypothese (H_{1a}, H_{1b}, H_{1c}) damit nur in dem Teilaspekt H_{1c} bestätigt.[25] Die Anzeichen für eine kognitive Überforderung sind – trotz der hohen Dimensionszahl – gering. Solange nicht Wege gefunden werden, die Komplexität von Vignetten zu kontrollieren, sollten absolute Effektstärken dennoch vorsichtig interpretiert werden (für sich genommen und beim Vergleich von Studien). An dieser Stelle ist auf eine weitere, bei anderen Autoren zu findende, problematische Interpretation zu verweisen: Oftmals werden ‚hohe' R^2-Werte als Beleg dafür gewertet, dass es gelungen sei, alle für die Befragten relevanten Merkmale in die Vignetten aufzunehmen (es bleibt kaum mehr etwas unerklärt, somit

22 Technisch besteht dieser Test darin, ein gepooltes Modell zu schätzen, in das zusätzlich eine Dummyvariable für die zu prüfende Designvariante (hier die Anzahl der Dimensionen) sowie Interaktionsterme aller Vignettendimensionen mit dieser Designvariante aufgenommen werden. Geprüft wird dann, ob die Aufnahme dieser Variablen *insgesamt* zu einer signifikanten Modellverbesserung führt; im vorliegenden Falle einer OLS-Regression, ob es zu einem signifikanten Anstieg der erklärten Varianz kommt (für Details: Wooldridge 2003: 238f.).

23 Der Vergleich zwischen fünf- und zwölfdimensionalen Vignetten ist statistisch nicht trivial. Mit einer höheren Variablenzahl steigt automatisch die Wahrscheinlichkeit von Korrelationen der Variablen untereinander oder von Konfundierungen mit Wechselwirkungen. Aufgrund der hohen Anzahl an möglichen Wechselwirkungen (bei der Variante mit zwölf zum Teil kategorialen Dimensionen liegen allein mehr als 70 mögliche Interaktionen erster Ordnung vor) sind diese nicht alle modellierbar (mitunter wird dies bereits durch die Stichprobenbildung verhindert). Damit ist nicht gänzlich auszuschließen, dass die Effekte im Falle der höher-dimensionalen Vignetten leicht verzerrt geschätzt werden (‚omitted-variable-bias'; wir danken P. Steiner für diesen wertvollen Hinweis). Es sollten daher künftig nochmals Replikationen mit anderen Vignettenstichproben durchgeführt werden. Darauf wird in der Schlussbetrachtung (Abschnitt 6) zurückgekommen.

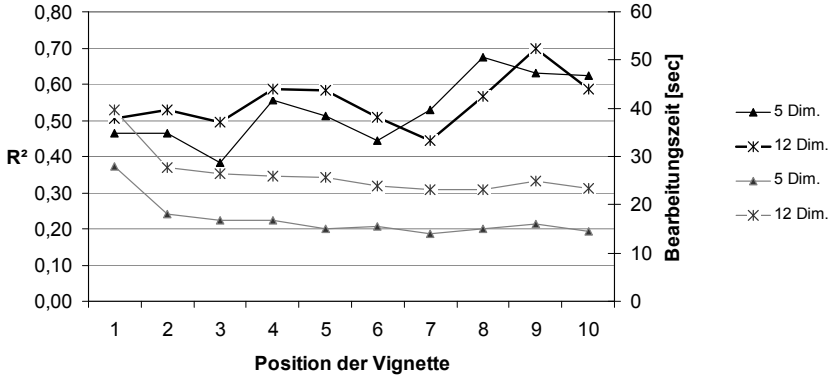
24 Der Vergleich von R^2-Werten zwischen Modellen ist nicht unproblematisch (Wooldridge 2003). Im vorliegenden Fall scheinen die Voraussetzungen jedoch erfüllt: Die Fallzahlen sind vergleichbar und ebenso bestehen nur minimale Unterschiede in der Varianz der abhängigen Variablen.

25 Wobei sich die These H_{1a} (häufigere Befragungsabbrüche bei höher-dimensionalen Vignetten) aufgrund der geringen Abbruchquoten nicht *statistisch* prüfen lässt.

seien alle urteilsrelevanten Informationen berücksichtigt; z. B. Beck/Opp 2001: 302). Wie unsere Ergebnisse zeigen, kann dies ein Trugschluss sein, denn die hinzukommenden Merkmale in Modell 3 erweisen sich fast ausnahmslos als signifikant, ohne dass es zu einem bedeutenden Anstieg der Varianzaufklärung käme. Eine hohe Modellanpassung ist somit zwar ein Maß für ein in sich konsistentes Urteilsverhalten, damit aber noch nicht unbedingt ein Indikator dafür, dass *alle* inhaltlich relevanten Dimensionen berücksichtigt sind.[26]

Zu beachten ist ferner, dass unseren Befragten mit maximal zehn Vignetten vergleichsweise wenige Urteile abverlangt wurden. Möglicherweise fallen kognitive Überforderungen und Ermüdungen erst bei weitaus höheren Vignettenzahlen ins Gewicht, oder schwächen sich umgekehrt mit zunehmender Übung ab. Damit sind die Hypothesen 2a und 2b angesprochen, die eine mit der Beantwortungssequenz zunehmende Konsistenz des Antwortverhaltens postulieren, speziell bei den komplexeren ‚Zwölfer'-Vignetten. Zur Überprüfung stellen wir Regressionsschätzungen getrennt für die einzelnen Bearbeitungspositionen der Vignetten an.

Abbildung 2 R^2-Werte (dicker gedruckte, obere Linien) und Bearbeitungszeiten pro Vignette (schwächere, untere Linien) in Abhängigkeit von der Position der Vignette und Anzahl ihrer Dimensionen



26 Womit auch Aussagen wie die folgende eine Relativierung finden: „The factorial survey method makes it possible to assess the number and identity of the characteristics a person uses in reaching a judgement." (Jasso 2006: 342)

In Abbildung 2 sind die resultierenden R^2-Werte für die beiden Designvarianten (fünf- vs. zwölfdimensional) gegen die Positionen der Vignetten abgetragen (dunklere, obere Linien). Da diese ebenfalls Aufschluss über Lern- bzw. Ermüdungseffekte geben, sind zugleich die mittleren Bearbeitungszeiten pro Vignette[27] (untere bzw. hellere Linien) dargestellt.

Was die Varianzaufklärung bzw. R^2-Werte betrifft, ist im Bearbeitungsverlauf ein leichter Anstieg zu erkennen. Die durchschnittliche Bearbeitungszeit pro Vignette sinkt dagegen insbesondere nach der ersten Vignette sprunghaft und mit abnehmender Rate weiter bis zur siebten Vignette. Zusammen genommen deutet dies auf einen Lerneffekt hin: Die Befragten können die Vignetten in zunehmend kürzerer Zeit beantworten, ohne dass es zu Einbußen ihrer Antwortkonsistenz käme. Entgegen unserer Erwartungen (H_{2b}) gilt dies nicht verstärkt für die komplexeren Vignetten: Die Linien für die fünf- und zwölfdimensionalen Vignetten verlaufen jeweils parallel zueinander, was bedeutet, dass die Lerneffekte für beide Versionen etwa gleich stark ausfallen. Für die vermutete Wechselwirkung zwischen Komplexitäts- und Lerneffekten findet sich also kein Beleg. Um die Interpretation als einen Lerneffekt abzusichern, ist zusätzlich noch zu prüfen, ob die steigende (oder zumindest gleich bleibende) Konsistenz nicht einer verstärkten Ausblendung von Dimensionen, also einer vereinfachten Entscheidungsheuristik, geschuldet ist. Um dies auszuschließen, wurden separate Regressionen mit dem ersten, zweiten und letzten Drittel der Vignetten berechnet. Die hier aus Platzgründen nicht dargestellten Modellschätzungen unterscheiden sich nicht signifikant voneinander,[28] d. h. die Anzahl einflussreicher Dimensionen, ihre Effektstärken und allgemein das Antwortmuster bleiben in der Bearbeitungssequenz stabil. Trotz der hohen Komplexität von zwölf Dimensionen führen also bereits die ersten Vignettenurteile zu sehr reliablen Urteilen – was bedeutet, dass sie nicht als ‚Übungsfälle' betrachtet

27 Die verwendete Online-Programmierung erlaubt es, die Bearbeitungszeit pro Vignette auf die Sekunde genau zu messen; exakter handelt es sich um die Zeit, die zwischen dem Abschicken der jeweiligen Vignettenseite und der Beendigung der vorherigen Seite verstrichen ist. Für derartige Zeitmessungen ist die bei Online-Befragungen geringe Kontrolle über das Setting nachteilig: Pausen der Befragten werden unweigerlich mit zur Bearbeitungszeit gerechnet. Aus diesem Grunde wurde jeweils das obere Fünf-Prozent-Perzentil der Antwortzeiten aus den Berechnungen ausgeschlossen (zur grundsätzlichen Empfehlung einer Bereinigung um ‚outliers' bei Befragungszeiten: Urban/Mayerl 2007; Mayerl/Selke/Urban 2005).

28 Entsprechende Chow-Tests fallen nicht signifikant aus. Einzeln betrachtet nehmen die Dimensionen mit den Vignettenpositionen in ihren Effektstärken tendenziell zu (wiederum Vergleich des ersten mit den beiden anderen Dritteln an Vignetten), kommt es also zu einer immer stärkeren Beachtung der Dimensionen, was den Lerneffekt eher noch untermauert. Allerdings wird die Signifikanzschwelle von fünf Prozent keinesfalls erreicht. Ebenfalls finden sich keine signifikanten Modellunterschiede, wenn die Berechnungen getrennt für die beiden Splits mit fünf und zwölf Dimensionen wiederholt werden.

werden müssen, die aus den Ergebnisanalysen ausgeschlossen werden sollten (zu einer entsprechenden Empfehlung bei Choice-Analysen: Caussade et al. 2005: 632, Anm. 6). An dieser Stelle ist aber darauf hinzuweisen, dass sich diese Aussagen nicht über die hier vorliegende, geringe Vignettenzahl und das sehr alters- und bildungshomogene Befragtensample hinaus verallgemeinern lassen. Ob bei höheren Vignettenanzahlen und/oder anderen Befragtengruppen nicht doch Ermüdungserscheinungen durchschlagen, bleibt künftigen Untersuchungen vorbehalten.

Die Wirkung unplausibler Fälle kann ähnlich untersucht werden. Dafür ist zunächst eine Festlegung erforderlich, was überhaupt als ‚unplausibel' zu gelten hat (vgl. Abschnitt 3.3) – aufgrund der hier bestehenden fließenden Übergänge und sicherlich auch subjektiv differierenden Einschätzungen keine triviale Aufgabe. Um möglichst objektiv vorzugehen, ziehen wir die Rückmeldungen aus ca. 60 mündlichen Pretestinterviews heran, die mit demselben Vignettensample im Herbst 2007 im Rahmen eines Forschungs-Projektseminars an der Universität Konstanz durchgeführt wurden. Für diese Interviews wurden bewusst sehr heterogene Personen der Allgemeinbevölkerung ausgewählt. Besonders häufig monierten die Befragten die Unsinnigkeit von Kombinationen der beiden Dimensionen ‚Beruf' und ‚Ausbildungsabschluss': Speziell ‚Anwälte ohne Ausbildung und Hochschulabschluss' sorgten oftmals geradezu für Verärgerung.[29] Wir werten entsprechend alle Vignetten als unplausibel, bei denen die geschilderten Personen nicht über einen Ausbildungs- oder Hochschulabschluss verfügen, der für ihren Beruf in Deutschland eigentlich Voraussetzung oder zumindest sehr üblich wäre.[30] Dies trifft auf insgesamt 22,5 % (N=800) unserer Vignetten zu.

Wir ziehen zudem eine alternative Operationalisierung auf Basis der Dimensionen ‚Einkommen' und ‚Beruf' heran. Speziell für Vignetten mit einem *berufsspezifisch* ungewöhnlichen Einkommen finden sich ebenfalls mehrere Pretestkommentare, welche ihre Ernsthaftigkeit in Zweifel ziehen (z. B. sorgten Vollzeit arbeitende, leitende Manager mit einem monatlichen Gehalt von nur 250 Euro netto für starke Irritationen, oder Friseure mit 15.000 Euro netto). Überdies besteht für diese beiden Dimensionen wiederum die Möglichkeit einer weitgehend objektiven Definiti-

29 Es fielen Äußerungen wie „man fühle sich auf die Palme gebracht"; „wer denkt sich so einen Unsinn aus". Befragt wurden Personen unterschiedlichen Alters und Bildungsgrades, die Rekrutierung und Durchführung der Interviews geschah durch die Seminarteilnehmer – ihnen sei an dieser Stelle unser Dank ausgesprochen.

30 Konkret sind dies: Anwälte ohne Hochschulabschluss oder nur mit Berufsabschluss; Verwaltungsfachkräfte, Elektroingenieure, Sozialarbeiter, Lokführer und leitende Manager ohne Ausbildung. Bei den übrigen (allesamt auch nicht gesetzlich geschützten) Berufsangaben (Friseure, Pförtner, Programmierer und ungelernte Arbeiter) scheinen dagegen Tätigkeiten ohne Ausbildungsabschluss plausibler.

on. Unplausible Vignetten lassen sich durch einen Abgleich der Vignetteneinkommen mit den realen Nettoeinkommen der jeweiligen Berufsgruppen identifizieren. Dazu bestimmen wir zunächst anhand der Stichprobe des Sozio-ökonomischen Panels (SOEP) von 2007 die mittleren tatsächlichen Nettoeinkommen der zehn in den Vignetten verwendeten Berufsgruppen und berechnen dann für jede Vignette die absolute Differenz zwischen ihrem ‚virtuellen' Vignetteneinkommen und dem tatsächlichen mittleren Berufseinkommen nach dem SOEP. Atypisch sind dann Fälle mit einer betragsmäßig besonders großen Differenz nach oben oder unten; konkret werten wir alle Vignetten mit einer absoluten Abweichung von mindestens 3.000 Euro als unplausibel (das sind 23,9 % aller Vignetten).[31]

Zu wählen ist noch ein geeignetes Analyseverfahren. Der zunächst nahe liegende Abgleich von Regressionsschätzungen auf Basis von plausiblen versus unplausiblen Fällen wird durch die unterschiedliche Varianz der unabhängigen Variablen in diesen beiden Gruppen beeinträchtigt: Bei den unplausiblen Fällen liegt *per se* eine andere Korrelation und Varianz der sie definierenden Dimensionen vor. Zudem würden die deutlichen Diskrepanzen in den Fallzahlen den Vergleich erschweren (die unplausiblen Fälle machen jeweils nur eine Minderheit aus). Theoretisch zu erwarten ist aber ohnehin ein Effekt, der sich nicht nur auf die unplausiblen Vignetten selbst bezieht, sondern ebenso auf die nachfolgenden: Wir gehen schließlich davon aus, dass die Konfrontation mit unrealistischen Fällen den *grundsätzlichen* Glauben an die Ernsthaftigkeit der Befragung und damit die *generelle* Kooperationsbereitschaft schmälert – d. h. genau genommen rechnen wir *ab* dem Auftreten unplausibler Vignetten mit einem weniger konsistenten (H_{3b}) oder stärker vereinfachten (d. h. weniger, bzw. allein die unplausiblen Dimensionen einbeziehenden) Urteilsverhalten (H_{3c}).

Dies legt es nahe, die Antwortmuster *bis* und *ab* dem Auftreten eines ersten unplausiblen Falls miteinander zu vergleichen. Wann das erste Mal eine unplausible Vignette erscheint, und wie viele unplausible Vignetten es pro Befragten sind, variiert aufgrund der zufälligen Deckzusammenstellung zwischen den Befragten. Insbesondere die zufällige Reihenfolgeposition pro Befragten erlaubt es, die Wirkung der

31 Neben dieser Definition über das Quartil haben wir Kontrollrechnungen mit dem Quintil und Dezentil durchgeführt – also mit einer noch stärkeren Eingrenzung unplausibler Fälle. Diese bestätigen unsere Befunde vollends und werden daher nicht separat dargestellt. Zur Berechnung der Netto-Berufseinkommen wurde die generierte Einkommensvariable des SOEP 2007 verwendet (für nähere Informationen zum SOEP: Wagner/Frick/Schupp 2007).

unplausiblen Vignetten von Lerneffekten zu trennen.[32] Tabelle 4 beinhaltet Regressionsschätzungen einerseits für die Definition der Unplausibilität über die Ausbildung (Modelle 1 und 2), andererseits für ihre Operationalisierung über das Einkommen (Modelle 3 und 4). Es werden jeweils Schätzungen für die Urteile vor (Modelle 1 und 3) sowie ab dem Auftreten eines ersten ‚unplausiblen' Falles (Modelle 2 und 4) gegenübergestellt. Ein Chow-Test weist die Modellunterschiede bei beiden Operationalisierungen als signifikant aus ($F = 2{,}06$; $p = 0{,}046$ bei $df = 7$ und 459; bzw. $F = 71{,}37$; $p = 0{,}000$ bei $df = 7$ und 459). Bei der Ausbildungs-Operationalisierung geht dies neben einem Niveaueffekt (die Konstante ist in Modell 2 geringer als in Modell 1) auf die Ausbildungsdimension zurück. Der Einfluss des Berufsabschlusses unterscheidet sich zu einem Zehn-, der Hochschulabschluss zu einem Fünf-Prozent-Niveau signifikant (Prüfung durch Schätzung eines gemeinsamen Modells mit entsprechenden Interaktionstermen).[33] Dagegen erweist sich bei der Operationalisierung über das Einkommen (Modelle 3 und 4) allein der Einfluss dieser Dimension als signifikant verschieden. Eine Analogie besteht auch in der Richtung der Unterschiede: Die Effekte fallen jeweils für die Modelle ab dem Auftreten eines unplausiblen Falles (also in Modell 2 statt 1 und 4 statt 3) schwächer aus. Dies entspricht nicht

32 Hierzu wird in den Modellen zudem für die Bearbeitungsposition kontrolliert. Bei beiden Operationalisierungen treten unplausible Vignetten zu etwa 60 % auf einer der ersten drei Rangpositionen das erste Mal auf, und es haben jeweils ca. ein Drittel der Befragten eine bis maximal zwei unplausible Vignetten erhalten. Maximal sind es fünf unplausible Vignetten pro einzelnem Befragten. Das Bestehen eines Lerneffektes wird überdies durch die entsprechenden Analysen im vorherigen Abschnitt entkräftet, schließlich haben sich die dort geprüften Unterschiede der Koeffizienten nach Bearbeitungsposition der Vignetten nicht als signifikant erwiesen. Um Fehlschlüsse zu vermeiden, sind dennoch ein paar weitere Überlegungen zur Vergleichbarkeit der beiden Gruppen erforderlich. Das Problem der unterschiedlichen Varianz der Dimensionen ist noch nicht völlig ausgeräumt – die Antworten ab dem Auftreten unplausibler Fälle beziehen sich unweigerlich auf stärker variierende und geringer korrelierte Vignettendimensionen, schließlich sind bei diesen Vignetten (im Gegensatz zur Vergleichsgruppe ohne unplausible Fälle) *alle* Merkmalskombinationen zulässig und tritt somit beispielsweise das berufsspezifische Einkommen in einer höheren Bandbreite auf. Diese stärkere Varianz und Unkorreliertheit bewirken aber *per se* eine höhere Schätzpräzision, damit ‚Power' von Signifikanztests. Finden sich größere Einflussstärken unplausibler Dimensionen, stellen diese daher noch nicht unbedingt ein Beleg für die von Faia (1980) vermutete Fokussierung der Befragten auf diese Merkmale dar, sie können ebenso allein mathematisch-statistisch bedingt sein. Und selbst wenn sich eine höhere kognitive Aufmerksamkeit der Befragten feststellen ließe, wäre diese dann noch nicht notwendigerweise der ‚Irrealität' der Fälle geschuldet, sie könnte ebenso Anzeichen eines ‚number-of-levels'- oder ‚range'-Effektes sein. Hierunter werden höhere kognitionspsychologische Aufmerksamkeiten der Befragten für stärker (und in größeren Spannweiten) variierende Merkmale gefasst, unabhängig von deren Inhalten. Diese Effekte, die zumindest für die verwandten Conjoint-Verfahren bereits gut belegt sind (z. B. Ohler et al. 2000; Louviere 2001a; Wittink/Krishnamurthi/Nutter 1982; Wittink/Krishnamurthi/Reibstein 1989; Perrey 1996), bieten also alternative Interpretationen für die Wirkung unplausibler Fälle. Auf diese Vermischung weisen ähnlich bereits Creyer/Ross 1988 hin, mitunter durch empirische Befunde gestützt (dazu auch Klein 2002: 15).

33 Es wurden Interaktionsterme zwischen allen Vignettendimensionen und einer Dummy-Variable gebildet, welche die Vignette einordnet (vor vs. ab unplausiblem Fall).

unserer Annahme, lässt sich aber gleichwohl als ein schlüssiger Befund deuten: Die Befragten beziehen ab der Konfrontation mit einer unplausiblen Kombination die ursächlichen Dimensionen weniger in ihr Urteil ein – man könnte auch sagen, sie nehmen diese weniger ‚ernst'.[34] Sollte sich dieser überraschende Befund in künftigen Untersuchungen (die aufgrund des zugegebenermaßen ‚ad hoc'-Charakters unserer Interpretation angeraten scheinen) replizieren, kann die Empfehlung nur lauten, sparsam mit solchen Fällen umzugehen; oder zumindest wie hier mit Zufallsreihenfolgen der Vignetten zu arbeiten, um Konfundierungen der Auswirkungen unplausibler Fälle (die eben erst ab ihrem Erscheinen auf den hinteren Bearbeitungspositionen auftreten können) mit inhaltlichen Effekten zu vermeiden.

Unsere Erwartung hinsichtlich der Antwortkonsistenz werden ebenfalls kaum erfüllt – mit dem Auftreten unplausibler Vignetten kommt es lediglich bei der Definition über das Einkommen zu einer leichten Abnahme der Konsistenz, gemessen am R^2. Bei der Operationalisierung über die Ausbildung ist dagegen sogar ein leichter Anstieg dieses Wertes zu verzeichnen. Die Konsistenz der Urteile wird hier durch die Abstraktion von den unrealistischen Dimensionen also sogar erhöht, was nochmals auf die mit Vorsicht zu behandelnde Interpretation des Wertes als einen Indikator für ‚erschöpfende Urteilsregeln' verweist (zu entsprechenden Interpretationen Beck/Opp 2001: 302; Jasso 2006: 416).

Bevor wir zu einem Fazit kommen, wollen wir die Einflüsse auf die Antwortkonsistenz und die Bearbeitungszeit noch in zwei abschließenden, multivariaten Modellen zusammenfassen. Als Maß für die Konsistenz verwenden wir die unerklärt gebliebene Varianz, genauer gesagt die quadrierten Residuen und schätzen OLS-Regressionen mit den Designvariablen als unabhängigen Variablen. Ein negativer Effekt bedeutet dann eine geringere Fehlervarianz bzw. höhere Konsistenz des Antwortverhaltens. Wie Tabelle 5 zeigt, finden sich lediglich zwei derartige (und nur zum Zehn-Prozent-Niveau) signifikante Effekte (Modell 1): Ab dem Auftreten von unplausiblen Fällen verringert sich die Fehlervarianz bzw. erhöht sich die Antwortkonsistenz (was wie oben gezeigt einem weniger Dimensionen einbeziehenden, damit vereinfachten Antwortverhalten geschuldet ist), und ebenso ist die Antwortkonsistenz umso höher, je mehr Zeit sich die Befragten für das Beantworten der einzelnen Vignetten nehmen.[35]

34 Der Annahme von Faia (1980) ist dies ebenso diametral entgegengesetzt wie einem ‚number-of-levels-' oder ‚range'-Effekt.

35 Um von grundsätzlichen Unterschieden in der individuellen ‚Basisgeschwindigkeit' (der vom Frageinhalt unabhängigen Grundgeschwindigkeit der Befragten) zu abstrahieren, wurden die Bearbeitungszeiten der Vignetten bei den hier vorliegenden Analysen jeweils pro Befragten mit seiner Antwortzeit bei ‚herkömmlichen' Itembatterien gewichtet (sog. ‚Latenzzeiten'). Die Geschwindigkeiten wurden dabei vorab um ‚Ausreißer' (oberes Fünf-Prozent-Perzentil) bereinigt. Zur Empfehlung eines ähnlichen Vorgehens siehe Mayerl/Selke/Urban 2005; Urban/Mayerl 2007.

Tabelle 4 OLS-Regressionen der Vignettenurteile[a] in Abhängigkeit des Auftretens unplausibler Fälle (robuste Standardfehler in Klammern; sign. Unterschiede in den Koeffizienten zwischen Modell 1 und 2 bzw. 3 und 4 hervorgehoben)[b]

	Definition über Ausbildung		Definition über Einkommen	
	Modell 1 Vor unpl. Vignetten	Modell 2 Ab unpl. Vignetten	Modell 3 Vor unpl. Vignetten	Modell 4 Ab unpl. Vignetten
Weibliche Vignettenperson	-0,244*	-0,023	-0,178	-0,072
	(0,130)	(0,104)	(0,117)	(0,100)
Alter [Jahre]	-0,030***	-0,025 ***	-0,030 ***	-0,022 ***
	(0,006)	(0,004)	(0,005)	(0,004)
Abschluss (Ref.: kein Abschluss)				
– Berufsabschluss	-0,972***	-0,571 ***	-0,671 ***	-0,670 ***
	(0,202)	(0,105)	(0,145)	(0,107)
– Hochschulabschluss	-1,386***	-0,784 ***	-1,101 ***	-0,962 ***
	(0,182)	(0,117)	(0,134)	(0,109)
Berufsprestige [10 MPS-Score]	-0,135 ***	-0,114 ***	-0,126 ***	-0,149 ***
	(0,015)	(0,010)	(0,013)	(0,010)
Nettoeinkommen [100,- Euro]	0,058 ***	0,057 ***	0,161 ***	0,055 ***
	(0,002)	(0,002)	(0,005)	(0,001)
Position der Vignette[c]	0,004	0,063 ***	-0,029	0,071 ***
	(0,032)	(0,021)	(0,029)	(0,021)
Konstante	7,152 ***	5,813 ***	5,021 ***	6,124 ***
	(0,353)	(0,241)	(0,277)	(0,252)
Beobachtungen:				
– Vignetten	1.301	2.179	1.197	2.283
– Befragte	355	400	344	409
R^2	0,44	0,48	0,56	0,52

[a] Skala von 1 ‚ungerechterweise zu niedrig' bis 11 ‚ungerechterweise zu hoch'. Der Wert 6 kennzeichnet eine als gerecht empfundene Entlohnung.

[b] Prüfung mittels Interaktionstermen zwischen den Vignettendimensionen und der Dimensionszahl in einem gepoolten Modell, Signifikanzniveau von fünf Prozent.

[c] Hier als lineare Variable ausgewiesen, da ausschließlich Kontrollfunktion. Bei einer alternativen Modellierung als Dummy-Variablen bleiben die Ergebnisse stabil.

**** p<0,01, ** p<0,05, * p<0,1 bei zweiseitigem Test; Schätzungen mit robusten Standardfehlern.*

Tabelle 5 OLS-Regressionen der Quadrierten Residuen[a] und Bearbeitungszeiten[b] pro Vignette (Robuste Standardfehler in Klammern)

	Modell 1 Quadrierte Residuen[a]	Modell 2 Bearbeitungszeit pro Vignette[b]
Position Vignette	-0,106	-0,064***
	(0,197)	(0,004)
Position Vignette, quadriert	0,006	0,005***
	(0,018)	(0,000)
Zwölf Dimensionen (Ref.: fünf)	-0,123	0,119***
	(0,300)	(0,008)
Ab unplausiblem Fall[c]	-0,564*	-0,004
	(0,306)	(0,008)
Bearbeitungszeit pro Vignette[a]	-1,461*	
	(0,865)	
Konstante	6,102***	0,371***
	(0,554)	(0,010)
Beobachtungen:		
– Vignetten	3.095	3.095
– Befragte	416	416
R^2	0,00	0,26

[a] *Residuen einer OLS-Regression des Vignettenurteils auf die ersten fünf Vignettendimensionen.*
[b] *Bearbeitungszeit in Sekunden, pro Befragten mit der Bearbeitungszeit einer entsprechenden Itembatterie gewichtet. Bei beiden Bearbeitungszeiten wurde das obere Fünf-Prozent-Perzentil ausgeschlossen. Hieraus resultieren die geringeren Fallzahlen in diesen Modellen.*
[c] *Hier definiert über den Ausbildungsabschluss.*
*** p<0,01, ** p<0,05, * p<0,1 bei zweiseitigem Test; Schätzungen mit robusten Standardfehlern.*

Die in Modell 2 betrachtete Bearbeitungszeit sinkt dagegen mit der Position der Vignetten, d. h. den Befragten gelingen zunehmend zeiteffiziente Urteile. Wie an dem positiven Effekt der quadrierten Bearbeitungszeit ersichtlich ist, handelt es sich um einen Effekt mit abnehmender Rate (grafisch einen ‚u-förmigen' Effekt).[36] Zudem bestätigt sich nochmals die ‚zeitraubendere' Bearbeitung von Vignetten mit einer höheren Dimensionszahl. Unplausible Fälle führen zumindest bei der hier ver-

36 Der Wendepunkt wäre nach der vorliegenden Modellschätzung bei der zehnten Vignette erreicht. Da für darüber hinausgehende Vignetten keine Beobachtungen vorliegen, ist dieser Befund aber nochmals mit umfangreicheren Vignettendecks pro Befragten zu validieren.

wendeten Operationalisierung über die Ausbildung nicht zu einem *zeitlich* flüchtigeren Antwortverhalten.[37]

6 Zusammenfassung und Schlussfolgerungen

Der Faktorielle Survey hat sich in der soziologischen Norm- und Einstellungsforschung inzwischen als Erhebungsmethode sehr gut durchgesetzt. Aber auch in anderen soziologischen Forschungszusammenhängen (wie z. B. der Diskriminierungsforschung) wird das Verfahren in den letzten Jahren vermehrt als eine ideale und innovative Methodik entdeckt. In Diskrepanz zu dieser guten Etablierung steht die geringe Erforschung des Verfahrens selbst. Es fehlen anwendungsbezogene Kriterien für die Konzeption Faktorieller Surveys, was ihre Durchführung erschwert (Beck/Opp 2001: 283f.). Ferner bestehen substantielle Zweifel an ihrer (internen) Validität fort. Invalide Urteile könnten etwa aus einer kognitiven Überforderung und/oder der Anwendung vereinfachter Entscheidungsstrategien (Heuristiken) resultieren. Ohne gezielte Methodenstudien ist kaum zu entscheiden, inwieweit die mit Vignetten gewonnenen Ergebnisse belastbar oder als methodische Artefakte zu interpretieren sind.

Der vorliegende Beitrag zielt daher auf eine erste Untersuchung der Stabilität und Konsistenz des Antwortverhaltens in Abhängigkeit von Designmerkmalen, konkret der Komplexität, Reihenfolge und Plausibilität von Vignetten. Mittels einer experimentellen Onlinebefragung von 460 Studierenden wurden systematische Analysen zum Einfluss der Anzahl der Dimensionen, Lerneffekten und der Wirkung von unplausiblen Fällen vorgenommen. Als ein erster übergreifender Befund lässt sich festhalten, dass diese drei methodischen Aspekte für die Antwortmuster und damit die Interpretationen der inhaltlichen Ergebnisse durchaus relevant sind.

So zeigen unsere Analysen, dass die Komplexität von Vignetten, jedenfalls gemessen an ihrer Dimensionszahl, die Urteile signifikant beeinflusst. Die Effekte einzelner Merkmale schwächen sich mit der Anzahl der Dimensionen ab – was zunächst bedeutet, dass die Interpretationen der absoluten Effektstärken nicht überstrapaziert werden sollten. Der Einfluss der einzelnen Merkmale scheint mitunter eine Funktion ihrer ‚Einzelständigkeit' zu sein, was zumindest beim Vergleich von

37 Bei einer Operationalisierung über das Einkommen findet sich allerdings eine zum Fünf-Prozent-Niveau signifikante Verringerung der Beantwortungszeit (um durchschnittlich 0,016 Sekunden). Diese differenten Befunde je nach Operationalisierung unplausibler Fälle fordern zu weiteren Untersuchungen heraus.

unterschiedlichen Studien zu berücksichtigen ist. Anders gesagt: aussagekräftige Vergleiche von Effekten in verschiedenen Erhebungen bedürfen ähnlich komplexer Vignetten. Die bei höherer Komplexität zu beobachtenden Heuristiken führen zumindest dann zu Artefakten, wenn sie keine Entsprechung mehr zu realen Urteilen aufweisen (ähnlich Swait/Adamowicz 2001: 147). Nicht-signifikante Einflüsse sind möglicherweise nochmals mit weniger-dimensionalen Fallbeispielen zu validieren. Die gute Botschaft lautet aber, dass die Auswirkungen auf das Antwortverhalten insgesamt gering sind und die Befragten selbst die hier vorgelegte, hohe Komplexität von zwölf Dimensionen insgesamt noch gut zu bewerkstelligen scheinen. Aufgrund der zu vermutenden Wechselwirkungen mit anderen Designvariablen (wie der Anzahl an Ausprägungen) und den Merkmalen der Befragten (kognitive Leistungsfähigkeit) sind allerdings vertiefende Untersuchungen angebracht.

Methodenstudien zu den verwandten Conjoint- und Choice-Analysen sprechen für ein komplexes Verhältnis von Lern- und Ermüdungseffekten. Um dieses vollständig abzubilden, ist die hier verwendete Fallzahl von maximal zehn Vignetten pro Befragten zu gering. Ein interessanter Befund ist aber schon einmal, dass bis zu unserer letzten, zehnten Vignette Lerneffekte dominieren, welche sich primär in einer zunehmenden Antwortgeschwindigkeit bei gleich bleibender Konsistenz (R^2) äußern. Die mit den ersten Vignetten gewonnenen Urteile sind in unserer Stichprobe reliabel, sie werden inhaltlich also durch die nachfolgenden bestätigt. Dies spricht für die grundsätzliche Verwertbarkeit dieser ‚ungeübten' ersten Urteile und damit die Validität von Studien, die mit einem reinen ‚between-subject'-Design arbeiten (nur eine Vignette pro Befragten). Ab welcher Anzahl an Vignetten die quantitativen Zugewinne an Urteilen mit merklichen Einbußen ihrer Datenqualität bezahlt werden, ist dagegen erst mit umfangreicheren Vignettendecks zu klären. Zudem sollte geprüft werden, ob diese Befunde auch einer anderen Komplexität von Vignetten und einem heterogeneren Befragtensample Stand halten.[38]

Von den einen als Stärke des Verfahrens gelobt, sehen Kritiker gerade durch empirisch seltene, daher besonders ‚virtuelle' Vignetten artifizielle Urteile herbeigeführt. Eine Skepsis, die nach unseren Analysen durchaus angebracht ist. Unplausible Merkmalskombinationen scheinen zwar zu keinen drastischen Befragungsabbrüchen oder Antwortverweigerungen zu führen (die Quote an Abbrüchen und Non-Responses ist insgesamt sehr gering), aber sie provozieren eine geringere

38 Anzunehmen ist, dass die Komplexität für die Befragten auch mit dem inhaltlichen Thema variiert, genauer gesagt mit ihrer Vertrautheit mit dem zu beurteilenden Gegenstand. Dies wurde zumindest für Choice-Experimente vereinzelt bereits untersucht (in der gesundheitsökonomischen Panel-Studie von Bryan et al. 2000 ist allerdings *kein* Effekt der Erfahrung auf die Reliabilität der Antworten festzustellen).

Berücksichtigung (bis möglicherweise vollständige Ausblendung) der für die Unplausibilität ursächlichen Dimensionen. Dieser Befund ist bei inhaltlichen Ergebnisinterpretationen zu beachten: Fehlende oder geringe Signifikanzen können statt einer genuinen Irrelevanz für das Urteilsverhalten ebenso anzeigen, dass die Dimensionen allein *in Folge ihrer Irrealität* weniger ernst genommen werden. Sollte sich dieses Ergebnis in weiteren Untersuchungen bestätigen, kann die praktische Empfehlung nur lauten, auf unplausible Fälle zu verzichten, oder zumindest sparsam mit ihnen umzugehen. Dank computerbasierter Verfahren lassen sich die durch ihren Ausschluss hervorgerufenen Einbußen an Effizienz der Vignettenstichproben (Balanciertheit und Unkorreliertheit der Dimensionen) auf ein vertretbares Maß reduzieren.

Unsere Analysen haben zudem gezeigt, dass es zur Feststellung methodischer Effekte multipler Kriterien bedarf. Vereinfachte Entscheidungsregeln tragen tendenziell zu einer hohen Messgüte der Ergebnisse bei (bewertet an der Varianzaufklärung), die Abweichung von den tatsächlichen Einstellungen und Urteilsregeln der Befragten kann gleichwohl groß sein.[39] Die zentrale Schlussfolgerung ist hier, dass die R^2-Werte allein als ein Indikator für die Konsistenz der Urteile zu gebrauchen sind, nicht aber als ein Maß dafür, inwieweit es gelungen ist, alle urteilsrelevanten Dimensionen ausfindig zu machen. Die Befragten scheinen sich bei einer drohenden Überforderung eher auf ein weiterhin konsistentes, aber gerade darum weniger detailliertes Urteilsverhalten zu konzentrieren. Faktorielle Surveys sind demnach primär ein geeignetes Verfahren für die Feststellung der Signifikanz *einzelner* Merkmale (etwa für entsprechende Hypothesentests), und weniger für die Aufdeckung inhaltlich *erschöpfender* Urteilsregeln. Anders ausgedrückt lassen sich mit ihnen Aussagen über den Einfluss der berücksichtigten Dimensionen treffen, nicht aber über die zusätzliche (Ir-)Relevanz weiterer Merkmale. Dies verweist nochmals auf die hohe Bedeutung einer sorgfältigen Auswahl der Dimensionen.

Aufgrund der Vielzahl weiterer methodischer Problemlagen und der zu erwartenden Wechselwirkungen mit anderen Designmerkmalen (wie z. B. der Anzahl an Ausprägungen und deren Variation und Bandbreite) ist mit den hier vorgelegten Untersuchungen erst ein Anfang gemacht. Empfehlenswert erscheinen zunächst Replikationen mit anderen Vignettenstichproben, etwa mit einem fraktionalisierten Design mit zusätzlicher Konfundierung aller Interaktionen erster Ordnung und/oder mit fraktionalisierten statt randomisierten Setbildungen. Dies erscheint ange-

39 Bei Berücksichtigung lediglich eines (oder weniger) Merkmale ist eine hohe Antwortkonsistenz schließlich keine kognitive ‚Herausforderung' – als Indikator für eine valide Messung ist sie gerade darum nicht hinreichend.

bracht, weil durch die Auswahl und Zusammenstellung von Vignetten unweigerlich die im Vignettenuniversum gegebene, vollständige Orthogonalität aller Dimensionen und ihrer Interaktionen verloren geht. In diesem Zusammenhang ist nicht *gänzlich* auszuschließen, dass sich die dadurch hervorgerufene Verringerung der statistischen Effizienz unterschiedlich auf die hier gegenübergestellten Gruppen mit weniger oder mehr Vignettendimensionen bzw. auf die Gruppen von Vignetten vor und ab dem Auftreten unplausibler Fälle verteilt, was ihre Vergleichbarkeit etwas beeinträchtigen könnte. Zudem sind weitere, hier aus Platzgründen nicht angesprochene methodische Aspekte von Interesse, wie beispielsweise die Wahl möglichst geeigneter Präsentationsformen (Fließtext oder tabellarische Darstellung) und der Einsatz von unterschiedlichen Antwortskalen. Darüber hinaus wären andere statistische Auswertungsverfahren zu erproben. Der gängigen Praxis folgend wurden die Vignettenurteile als metrisch behandelt, genau genommen weisen sie lediglich ordinales Skalenniveau auf. Die Wahl von OLS-Schätzungen wird zwar allgemein durch ihre hohe Robustheit und bessere Interpretierbarkeit gerechtfertigt (Winship/Mare 1984); speziell für Vignettenstudien wurde bislang aber noch zu wenig ausgelotet, welche Analysegewinne sich mit adäquateren – besser mit der Datenstruktur korrespondierenden – (Mehr-)Ebenenverfahren erzielen lassen.[40]

Mit den Daten des DFG-geförderten Projekts ‚Der Faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen' können eine Vielzahl der benannten Aspekte untersucht werden. Nach den hier präsentierten, ersten Befunden verdienen sie in methodischer wie inhaltlicher Hinsicht stärkere Beachtung.

Literatur

Alves, W. M. und P. H. Rossi, 1978: Who should get what? Fairness judgments of the distribution of earnings. American Journal of Sociology 84: 541-564.

Auspurg, K., M. Abraham und Th. Hinz, 2009: Wenig Fälle, viele Informationen: Die Methodik des faktoriellen Surveys als Paarbefragung. S. 179-210 in: P. Kriwy und C. Groß (Hg.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS Verlag für Sozialwissenschaften.

Auspurg, K. und M. Abraham, 2007: Die Umzugsentscheidung von Paaren als Verhandlungsproblem. Eine quasiexperimentelle Überprüfung des Bargaining-Modells. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 271-293.

40 Aufgrund der mehrfachen Bewertungsaufgabe sind zumindest bei den hier verwendeten, geschlossenen Antwortskalen Zensierungen der Urteile zu befürchten (wurden bereits extreme Urteile abgegeben, lässt sich das Antwortverhalten nicht mehr hinreichend abstufen). Dies kann ebenfalls zu verzerrten Ergebnissen führen und legt den Einsatz von einschlägigen Regressionsverfahren (z. B. Tobit-Modellen) nahe. Auch hier sind Wechselwirkungen mit dem Auftreten von unplausiblen Fällen zu erwarten, da diese vermehrt zu extremen Antworten motivieren dürften.

Beck, M. und K.-D. Opp, 2001: Der faktorielle Survey und die Messung von Normen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 283-306.
Berk, R. A. und P. H. Rossi, 1977: Prison reform and state elites. Cambridge: Ballinger.
Buskens, V. und J. Weesie, 2000: An experiment on the effects of embeddedness in trust situations. Buying a used car. Rationality and Society 12: 227-253.
Bradley, M. und A. Daly, 1994: Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21: 167-184.
Bryan, S., L. Gold, R. Sheldon und M. Buxton, 2000: Preference measurement using conjoint methods. An empirical investigation of reliability. Health Economics 9: 385-395.
Carroll, D. J. und P. E. Green, 1995: Psychometric methods in marketing research. Part I, Conjoint analysis. Journal of Marketing Research 32: 358-391.
Carson, R., J. J. Louviere, D. A. Anderson, P. Arabie, D. Bunch, D. A. Hensher, R. M. Johnsons, W. F. Kuhfeld, D. Steinberg, J. Swait und H. Timmerman, 1994: Experimental analysis of choice. Marketing Letters 5: 351-368.
Caussade, S., J. de D. Ortúzar, L. I. Rizzi und D. A. Hensher, 2005: Assessing the influence of design dimensions on stated choice experiment estimates. Transportation research part B: Methodological 39: 621-640.
Creyer, E. und W. T. Ross, 1988: The effect of range-frequency manipulations on conjoint importance weight stability. Advances in Consumer Research 15: 505-509.
DeShazo, J. R. und G. Fermo, 2002: Designing choice sets for stated preference methods. The effects of complexity on choice consistency. Journal of Environmental Economics and Management 44: 123-143.
Diefenbach, H. und K.-D. Opp, 2007: When and why do people think there should be a divorce? An application of the factorial survey. Rationality and Society 19: 485-517.
Diekmann, A., 2007. Empirische Sozialforschung. Reinbek bei Hamburg: Rowohlt.
Dülmer, H., 2001: Bildung und der Einfluss von Argumenten auf das moralische Urteil. Eine empirische Analyse zur moralischen Entwicklungstheorie Kohlbergs. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 1-27.
Dülmer, H. und M. Klein, 2003: Die Messung gesellschaftlicher Wertorientierungen via Conjoint- und Vignettenanalyse: Ein Ansatz zur adäquaten Operationalisierung von Ingelharts materialistischen und postmaterialistischen Wertorientierungen. Unveröffentlichter Abschlussbericht an die Fritz-Thyssen Stiftung.
Dülmer, H., 2007: Experimental plans in factorial surveys. Random or quota design? Sociological Methods & Research 35: 382-409.
Eifler, S., 2007: Evaluating the validity of self-reported deviant behavior using vignette analyses. Quality & Quantity 41: 303-318.
Faia, M., 1980: The vagaries of the vignette world. A comment on Alves and Rossi. American Journal of Sociology 85: 951-954.
Garrett, K., 1982: Child abuse: problems of definition. S. 177-204 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Greene, W. H., 2003: Econometric Analysis. Prentice Hall: New York.
Groß, J. und C. Börensen, 2009: Wie valide sind Verhaltensmessungen mittels Vignetten? Ein methodischer Vergleich von faktoriellem Survey und Verhaltensbeobachtung. S. 149-178 in: P. Kriwy und C. Groß (Hg.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS Verlag für Sozialwissenschaften.
Hechter, M., J. Ranger-Moore, G. Jasso und C. Horne, 1999: Do values matter? An analysis of advance directives for medical treatment. European Sociological Review 15: 405-430.
Hechter, M., H. Kim und J. Baer, 2005: Prediction versus explanation in the measurement of values. European Sociological Review 21: 91-108.
Hembroff, L. A., 1987: The seriousness of acts and social contexts. A test of Black's theory of the behavior of law. American Journal of Sociology 93: 322-347.

Hermkens, P. L. J. und F. A. Boerman, 1989: Consensus with respect to the fairness of incomes. Differences between social groups. Social Justice Research 3: 201-215.
Hensher, D. A., 2004: How do respondents handle stated choice experiments? Information processing strategies under varying information load. Working paper 04-14. University of Sydney: Institute of Transport Studies.
Hensher, D. A., 2006: Revealing differences in willingness to pay due to the dimensionality of stated choice designs. An initial assessment. Environmental & Resource Economics 34: 7-44.
Horne, C., 2003: The internal enforcement of norms. European Sociological Review 19: 335-343.
Hox, J. J., I. Kreft und P. Hermkens, 1991: The analysis of factorial surveys. Sociological Methods & Research 19: 493-510.
Jann, B., 2003: Lohngerechtigkeit und Geschlechterdiskriminierung. Experimentelle Evidenz. Unveröffentlichtes Manuskript an der Eidgenössischen Technischen Hochschule Zürich.
Jasso, G., 1988: Whom Shall We Welcome? Elite judgments of the criteria for the selection of immigrants. American Sociological Review 53: 919-932.
Jasso, G., 1994: Assessing individual and group differences in the sense of justice. Framework and application to gender differences in the justice of earnings. Social Science Research 23: 368-406.
Jasso, G. und M. Webster, 1997: Double standards in just earnings for male and female workers. Social Psychology Quarterly 60: 66-78.
Jasso, G. und K.-D. Opp, 1997: Probing the character of norms. A factorial survey analysis of the norms of political action. American Sociological Review 62: 947-964.
Jasso, G. und M. Webster, 1999: Assessing the gender gap in just earnings and its underlying mechanisms. Social Psychology Quarterly 62: 367-380.
Jasso, G., 2006: Factorial survey methods for studying beliefs and judgments. Sociological Methods and Research 34: 334-423.
John S., C. und N. A. Bates, 1990: Racial composition and neighborhood evaluation. Social Science Research 19: 47-61.
Johnson, R. F., 2006: Comment on "Revealing differences in willingness to pay due to the dimensionality of stated choice designs. An initial assessment". Environmental & Resource Economics 34: 45-50.
Klein, M., 2002: Die Conjoint-Analyse. Eine Einführung in das Verfahren mit einem Ausblick auf mögliche sozialwissenschaftliche Anwendungen. ZA-Information 50: 7-45. http://www.gesis.org/forschung-lehre/gesis-publikationen/zeitschriften/archiv/za-information/ (16.3.2009).
Kuhfeld, W. F., T. D. Randall und M. Garratt, 1994: Efficient experimental design with marketing research applications. Journal of Marketing Research 31: 545-557.
Kuhfeld, W. F., 2005: Marketing research methods in SAS. Experimental design, choice, conjoint and graphical techniques. Cary: SAS Institute.
Liebig, S. und S. Mau, 2002: Einstellungen zur sozialen Mindestsicherung. Ein Vorschlag zur differenzierten Erfassung normativer Urteile. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 109-134.
Liebig, S. und S. Mau, 2005: Wann ist ein Steuersystem gerecht? Zeitschrift für Soziologie 34: 468-491.
Liebig, S., A. Meyermann und A. Schulze, 2006: Temporal stability of justice evaluations. Paper presented at the 11th conference of the international society for justice research. Berlin: Humboldt Universität.
Louviere, J. J., 2001a: What if consumer experiments impact variances as well as means? Response variablility as a behavioral phenomen. Journal of Consumer Research 28: 506-511.
Louviere, J. J., 2001b: Choice experiments. An overview of concepts and issues. S. 13-36 in: Bennett, J. und R. Blamey (Hg.): The choice modelling approach to environmental valuation. Cheltenham/Northhampton: Edward Elgar.
Mayerl, J., P. Selke und D. Urban, 2005: Analyzing cognitive processes in CATI-Surveys with response latencies. An empirical evaluation of the consequences of using different

baseline speed measures. Schriftenreihe des Instituts für Sozialwissenschaften der Universität Stuttgart: SISS 2/2005.
Melles, Th., 2001: Framing-Effekte in der Conjoint-Analyse. Ein Beispiel für Probleme der Merkmalsdefinition. Aachen: Shaker.
Meudell, M. B., 1982: Household and social standing. Dynamic and static dimensions. S. 69-94 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Miller, J. L., P. H. Rossi und J. E. Simpson, 1986: Perceptions of justice. Race and gender differences in judgments of appropriate prison sentences. Law & Society Review 20: 313-334.
Nisic, N. und K. Auspurg, 2009: Faktorieller Survey und klassische Bevölkerungsumfrage im Vergleich – Validität, Grenzen und Möglichkeiten beider Ansätze. S. 211-246 in: P. Kriwy und C. Groß (Hg.): Klein aber fein! Quantitative empirische Sozialforschung mit kleinen Fallzahlen. Wiesbaden: VS Verlag für Sozialwissenschaften.
Nock, S. L., 1982: Family social standing. Consensus on characteristics. S. 95-118 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Ohler, T., A. Le, J. Louviere und J. Swait, 2000: Attribute range effects in binary response tasks. Marketing Letters 11: 249-260.
Orme, B., 2006: Getting started with conjoint analysis. Strategies for product design and pricing research. Madison/Wisconsin: Research Publishers LLC.
O`Toole, R., S. W. Webster, A. W. O`Toole und B. Lucal, 1999: Teachers´ recognition and reporting of child abuse. A factorial survey. Child Abuse & Neglect 23: 1083-1101.
Perrey, J., 1996: Erhebungsdesign-Effekte bei der Conjoint-Analyse. Marketing – Zeitschrift für Forschung und Praxis 18: 105-116.
Rooks, G., W. Raub, R. Selten, und F. Tazelaar, 2000: How inter-firm co-operation depends on social embeddedness: A vignette study. Acta Sociologica 43: 123-137.
Rossi, P. H., W. A. Sampson, C. E. Bose, G. Jasso und J. Passel, 1974: Measuring household social standing. Social Science Research 3: 169-190.
Rossi, P. H., 1979: Vignette analysis. Uncovering the normative structure of complex judgments. S. 176-186 in: R. K. Merton, J. S. Coleman und P. H. Rossi (Hg.): Qualitative and Quantitative Social Research. Papers in honour of Paul F. Lazarsfeld. New York: Free Press.
Rossi, P. H. und W. M. Alves, 1980: Rejoinder to Faia. The American Journal of Sociology 85: 954-955.
Rossi, P. H. und A. B. Anderson, 1982: The factorial survey approach. An introduction. S. 15-67 in: P. H. Rossi und S. L. Nock (Hg.): Measuring social judgments. The factorial survey approach. Beverly Hills u.a.: Sage.
Rossi, P. H. und S. L. Nock, 1982: Measuring social judgements. The factorial survey approach. Beverly Hills u. a.: Sage.
Sauer, C., 2009: Methodische Probleme von Conjoint- und Vignettenanalysen – Literaturreview. Arbeitsbericht Nummer 1 des Projekts „Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen". Bielefeld/Konstanz: Universität Bielefeld/Universität Konstanz.
Seyde, C., 2005: Beiträge und Sanktionen in Kollektivgutsituationen. Ein faktorieller Survey. Arbeitsbericht 42 des Instituts für Soziologie. Leipzig: Universität Leipzig.
Shepelak, N. J. und D. F. Alwin, 1986: Beliefs about Inequality and Perceptions of Distributive Justice. American Sociological Review 51: 30-46.
Shlay, A. B., H. Tran, M. Weinraub und M. Harmon, 2005: Teasing apart the child care conundrum. A factorial survey analysis of perceptions of child care quality, fair market price and willingness to pay by low-income, African American parents. Early Childhood Research Quarterly 20: 393-416.
Smith, T. W., 1986: A study of non-response and negative values on the factorial vignettes on welfare. GSS Methodological Report 44. Chicago: NORC.
Sniderman, P. M. und D. B. Grob, 1996: Innovations in experimental design in attitude surveys. Annual Review of Sociology 22: 377-399.

Steiner, P. M. und C. Atzmüller, 2006: Experimentelle Vignettendesigns in faktoriellen Surveys. Kölner Zeitschrift für Soziologie und Sozialpsychologie 58: 117-146.
Struck, O., A. Krause und C. Pfeifer, 2008: Entlassungen: Gerechtigkeitsempfinden und Folgewirkungen. Theoretische Konzepte und empirische Ergebnisse. Kölner Zeitschrift für Soziologie und Sozialpsychologie 60: 102-122.
Swait, J. und W. Adamowicz, 2001: The influence of task complexity on consumer choice. A latent class model of decision strategy switching. Journal of Consumer Research 28: 135-148.
Urban, D. und J. Mayerl, 2007: Antwortlatenzzeiten in der survey-basierten Verhaltensforschung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 692-713.
Wagner, G. G., J. R. Frick und J. Schupp, 2007: The German Socio-Economic Panel Study (SOEP) – Evolution, scope and enhancements. Schmollers Jahrbuch (Journal of Applied Social Science Studies) 127 (1): 139-169.
Wason, K. D., M. J. Polonsky und M. R. Hyman, 2002: Designing vignette studies in marketing. Australasian Marketing Journal 10: 41-58.
Will, J. A., 1993: The dimensions of poverty. Public perceptions of the deserving poor. Social Science Research 22: 312-332.
Winship, C. und R. D. Mare, 1984: Regression models with ordinal variables. American Sociological Review 49: 512-525.
Wittink, D. R., L. Krishnamurthi und J. B. Nutter, 1982: Comparing derived importance weights across attributes. Journal of Consumer Research 8: 471-474.
Wittink, D. R., L. Krishnamurthi und D. J. Reibstein, 1989: The effect of differences in the number of attribute levels on conjoint results. Marketing Letters 1: 113-123.
Wooldridge, J. M., 2002: Econometric analysis of cross section and panel data. Cambridge/Mass.: MIT Press.
Wooldridge, J. M., 2003: Introductory econometrics. A modern approach. Mahson, Ohio: Thomson.
Zimbardo, P. G., 1988: Psychologie. Berlin u. a.: Springer.

Korrespondenzadresse: Katrin Auspurg
Universität Konstanz
Fach D40
78457 Konstanz
katrin.auspurg@uni-konstanz.de

Statistischer Anhang

Tabelle A1 Übersicht über die Korrelationen der Dimensionen in den Splits mit 5 und 12 Dimensionen[a]

	Split mit 5 Dimensionen				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	-0,056	1,000			
Abschluss	-0,019	0,016	1,000		
Berufsprestige	0,017	0,096	-0,023	1,000	
Einkommen	-0,027	0,008	0,042	0,148	1,000
	Split mit 12 Dimensionen				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	0,034	1,000			
Abschluss	-0,007	-0,022	1,000		
Berufsprestige	0,032	0,049	-0,031	1,000	
Einkommen	0,030	-0,031	0,073	0,172	1,000

[a] *Korrelationskoeffizient nach Pearson bzw. bei der ordinalen Dimension ‚Abschluss' Rang-Korrelationskoeffizient nach Spearman.*

Tabelle A2 Übersicht über die Korrelationen der Dimensionen in den Splits mit 7 und 10 Vignetten pro Befragten[a]

	Split mit 7 Vignetten				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	-0,033	1,000			
Abschluss	0,003	0,002	1,000		
Berufsprestige	0,039	0,070	-0,031	1,000	
Einkommen	-0,015	-0,016	0,055	0,182	1,000
	Split mit 10 Vignetten				
	Geschlecht	**Alter**	**Abschluss**	**Berufsprestige**	**Einkommen**
Geschlecht	1,000				
Alter	0,032	1,000			
Abschluss	-0,050	-0,011	1,000		
Berufsprestige	-0,006	0,080	-0,016	1,000	
Einkommen	0,038	-0,004	0,059	0,115	1,000

[a] *Korrelationskoeffizient nach Pearson bzw. bei der ordinalen Dimension ‚Abschluss' Rang-Korrelationskoeffizient nach Spearman.*

Tabelle A3 Übersicht über die Korrelationen[a] aller zwölf Vignettendimensionen (nur Split mit 12 Dimensionen)

	Geschlecht	Alter	Abschluss	Berufs-prestige	Einkommen	Berufs-erfahrung	Betriebs-zugehörigkeit	Leistung	Betriebs-größe	Wirtsch. Lage	Gesundheit	Kinder
Geschlecht	1,000											
Alter	0,034	1,000										
Abschluss	-0,007	-0,022	1,000									
Berufsprestige	0,032	0,049	-0,031	1,000								
Einkommen	0,030	-0,031	0,073	0,172	1,000							
Berufserfahrg.	0,018	0,187	-0,107	0,109	0,045	1,000						
Betriebszug.	0,024	0,279	-0,101	0,093	0,117	0,416	1,000					
Leistung	-0,015	0,051	-0,025	-0,015	0,001	0,027	0,013	1,000				
Betriebsgröße	-0,036	0,072	0,156	-0,092	-0,153	-0,012	0,212	-0,083	1,000			
Wirtsch. Lage	0,049	0,059	-0,031	0,091	-0,032	0,042	0,059	0,012	0,075	1,000		
Gesundheit	0,027	0,015	0,013	0,086	-0,165	-0,013	-0,041	-0,072	-0,014	-0,023	1,000	
Kinder	0,018	0,101	0,023	-0,051	0,001	0,088	-0,064	-0,005	0,063	0,023	-0,087	1,000

[a] *Bei metrischen und binären Variablen Korrelationskoeffizient nach Pearson; bei ordinalen Variablen (Abschluss, Leistung und wirtschaftliche Lage) Rang-Korrelationskoeffizient nach Spearman.*

Die Erhebung biometrischer Daten im Survey of Health, Ageing and Retirement in Europe

Befunde und Perspektiven

The Collection of Biomarkers in the Survey of Health, Ageing and Retirement in Europe

Findings and Perspectives

Karsten Hank, Hendrik Jürges und Barbara Schaan

Zusammenfassung

Im Rahmen des 2004 erstmals durchgeführten *Survey of Health, Ageing and Retirement in Europe* (SHARE) wurden vielfältige Informationen über den psychischen und physischen Gesundheitszustand der Befragten erhoben. Trotz ihres unbestrittenen Wertes haben sich subjektive bzw. selbstberichtete Gesundheitsindikatoren jedoch insbesondere für international vergleichende Analysen als nicht unproblematisch erwiesen. Die Erfassung biometrischer Daten leistet einen wichtigen Beitrag, um diesem Problem zu begegnen. Im vorliegenden Beitrag sollen zunächst am Beispiel der isometrischen Greifkraft Analysen mit bereits heute in SHARE erfassten biometrischen Daten präsentiert werden. Anschließend wird die Einbeziehung weiterer biometrischer Daten (insbesondere über Blutproben) in das Erhebungsprogramm des längsschnittlich angelegten SHARE diskutiert. Hier werden neben in diesem Zusammenhang relevanten soziologischen Fragestellungen (z. B. Bedeutung von Biomarkern für die Untersuchung des Zusammenhangs von sozio-ökonomischem Status und Gesundheit) auch Erfahrungen aus mit SHARE vergleichbaren Studien, insbesondere aus dem angelsächsischen Raum, betrachtet.

Abstract

A large variety of information of respondents' physical and mental health has been collected within the context of the *Survey of Health, Ageing and Retirement in Europe* (SHARE) from its first wave in 2004 on. Despite their undisputable value, self-reported and subjective health indicators turned out not to be unproblematic in international comparative analyses. The collection of biometric data contributes to remedying such problems. This research paper presents analyses with measures of isometric grip-strength – one of the biometric measures already available in SHARE to date. Further, the authors discuss the inclusion of other biometric measures (especially via blood samples) into the investigational program of the longitudinally designed SHARE. Relevant sociological problems and questions (e. g. the importance of biomarkers for analyses of the correlation between socio-economic status and health) as well as experiences with biometric data in studies comparable to SHARE (especially from Anglo-Saxon countries) are described.

1 Einleitung: What Biology Do Sociologists Need to Know?

In der vergangenen Dekade hat die Debatte um die Bedeutung der Biologie – und hier insbesondere genetischer Einflussfaktoren – für sozialwissenschaftliche Fragestellungen wachsende Aufmerksamkeit erfahren (z. B. Freese et al. 2003; Udry 1995; vgl. auch neuere Sonderhefte führender Zeitschriften wie American Journal of Sociology [Bearman 2008], Social Forces [Guo 2006] oder Sociological Methods and Research [Guo 2008]). Im Zentrum dieser Debatte steht nicht allein der Zusammenhang zwischen Biomarkern und Mortalität bzw. Morbidität im höheren Lebensalter (z. B. Vaupel 1998; Weinstein et al. 2003), sondern es wird auch diskutiert, inwieweit sich etwa genetische Anlagen auf menschliches Verhalten auswirken (z. B. Diewald 2008; Kohler et al. 1999).

Mit diesen neuen inhaltlichen Fragestellungen steht die sozialwissenschaftliche Umfragforschung vor der Herausforderung, ihr traditionelles Fragenprogramm durch die Einbeziehung biometrischer Daten[1] zu ergänzen (vgl. National Research Council 2008, für einen aktuellen Überblick). Der wissenschaftliche Nutzen der Aufnahme solcher Informationen in sozialwissenschaftliche Umfragen ist vielfältig und kann im Rahmen dieses Beitrags nur exemplarisch belegt werden. Der wohl wesentlichste Vorteil, und zwar unabhängig von spezifischen Fragestellungen, besteht in der Möglichkeit zur Verknüpfung von biologischen Merkmalen mit sozio-demographischen und sozio-ökonomischen Charakteristika der Befragten aus repräsentativen (d. h. nicht-klinischen) Bevölkerungsstichproben zum Zwecke interdisziplinärer Forschung (z. B. Finch/Vaupel 2001; Lillard/Wagner 2006).

Grundsätzlich lassen sich verschiedene Arten von Biomarkern unterscheiden, die in Bevölkerungsumfragen erhoben werden können: „direct measures of physical or physiological characteristics (e. g., hip circumference, blood pressure), functional testing (e. g., cognitive function, balance, grip strength), or collection of specimens that require laboratory processing in order to generate analyzable data" (Lindau/McDade 2008: 252; vgl. auch Lillard/Wagner 2006). Die Erhebung solcher biometrischer Daten ist in unterschiedlichem Maße bereits in wichtigen sozialwissenschaftlichen Umfragen implementiert worden. Im Mittelpunkt des vorliegenden Beitrags stehen exemplarische Befunde und Perspektiven der Erhebung von Biomarkern im S*urvey of Health, Ageing and Retirement in Europe* (SHARE).

1 Wir schließen uns hier der Definition von Lillard/Wagner (2006: 1, Hervorhebungen im Original) an, indem wir „interchangeably use the terms 'biomarkers' and 'biometric data' to refer to data that measure physical characteristics of an *anonymous* respondent who gave an *informed consent*."

2 Die Erhebung biometrischer Daten im SHARE und vergleichbaren sozialwissenschaftlichen Umfragen heute

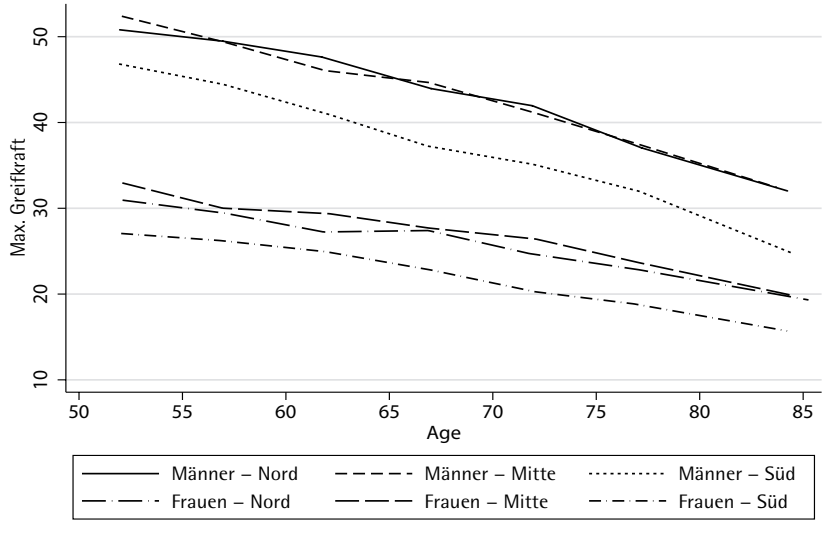
Im Rahmen des 2004 erstmals durchgeführten *Survey of Health, Ageing and Retirement in Europe* (SHARE; http://www.share-project.org) – einer repräsentativen Stichprobe (n = 31.000) der Bevölkerung im Alter 50+ in 13 kontinentaleuropäischen Ländern und Israel – werden vielfältige Informationen über den psychischen und physischen Gesundheitszustand der Befragten erhoben (siehe Börsch-Supan et al. 2005: Kapitel 3). Neben Selbstberichten über Körpergröße und Gewicht, diagnostizierte chronische Krankheiten, seelische Probleme und körperliche Beeinträchtigungen sowie verschiedenen kognitiven Tests, enthält SHARE auch Messungen der Gehgeschwindigkeit und der Greifkraft (vgl. Hank et al. 2009; Jürges 2005). In der zweiten SHARE-Welle 2006 – 2007 wurde darüber hinaus die Lungenkapazität sowie die Zeit, die die Befragten zum fünfmaligen Aufstehen von einem Stuhl benötigen, gemessen (Börsch-Supan et al. 2008).

SHARE ist damit Teil eines dichter werdenden Netzes nationaler und internationaler Dateninfrastrukturen, die neben sozialwissenschaftlichen Daten auch Biomarker erheben (vgl. National Research Council 2008). Beispielhaft seien hier das *Sozio-oekonomische Panel* (SOEP) sowie die *English Longitudinal Study of Ageing* (ELSA) genannt. Für Deutschland werden im Rahmen des SOEP neben allgemeinen Gesundheitsinformationen seit 2002 auch das Gewicht und die Körpergröße (zur Berechnung des Body Mass Index; Kroh 2005) sowie seit 2006 die Handgreifkraft (Hank et al. 2009) gemessen. Im ELSA-Projekt wurden 2004 durch ausgebildete Krankenschwestern erstmals auch invasive Maße biologischer Merkmale erhoben, bei denen es etwa durch den Einsatz von Nadeln zu minimalen Verletzungen von Haut oder Weichteilen kommt (vgl. Banks/Breeze et al. 2006: Kapitel 5; Marmot/Steptoe 2008). Die den Befragten abgenommene Blutprobe diente der Laboranalyse z. B. von Cholesterin, C-reaktivem Protein, Fibrinogen und Hämoglobin. Dieses erfolgreiche europäische Beispiel soll als Vorbild für die zukünftig geplante Erhebung weiterer biometrischer Daten im SHARE dienen (siehe unten; vgl. auch die Empfehlungen zur Erhebung von Biomarkern in der geplanten UK Longitudinal Household Study [Kumari et al. 2006]).

3 Ein Beispiel: Die isometrische Greifkraft als nicht-invasiver Biomarker

Die Messung der Handgreifkraft hat sich als ein in persönlichen Befragungen einfach zu erhebendes, nicht-invasives und verlässliches ‚objektives' Gesundheitsmaß erwiesen. Die Greifkraft der Hände ist zudem mit der Stärke anderer Muskelgruppen hoch korreliert und zur Identifikation von Genvarianten geeignet, die für die körperliche Funktionsfähigkeit im mittleren und hohen Lebensalter relevant sind (z. B. Carmelli/Reed 2000). Eine Vielzahl von Studien belegt, dass die Greifkraft in Folge einer fortschreitenden Abnahme der Muskelkraft und –masse (*Sarkopenie*) mit zunehmendem Alter generell schwächer wird (z. B. Frederiksen et al. 2006) – mit entsprechenden Konsequenzen für die körperliche Leistungsfähigkeit der Betroffenen, und damit auch für deren Lebensqualität und Unabhängigkeit im Alter. Darüber hinaus haben Längsschnittstudien gezeigt, dass Muskelschwäche im mittleren Lebensalter – gemessen über die isometrische Greifkraft – ein sehr guter Prädiktor für zukünftige Behinderungen, etwa bei den so genannten ‚activities of daily living' (z. B. Rantanen et al. 1999), oder für Mortalitätsrisiken (z. B. Metter et al. 2002) im höheren Alter ist.

Abbildung 1 Mittlere maximale Greifkraft nach Alter, Geschlecht und Ländergruppe



Quelle: Börsch-Supan et al. (2005: 87)

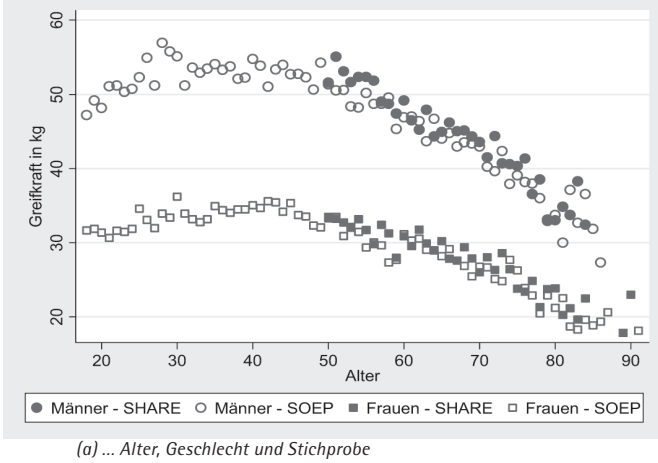
Die Auswertung von Querschnittsdaten der ersten Welle des SHARE zeigt, dass die deutschen SHARE-Befragten im Durchschnitt eine den anderen nord- und mitteleuropäischen Ländern vergleichbare Greifkraft aufweisen, damit aber über den jeweiligen Werten der Mittelmeerländer liegen (vgl. Abbildung 1). Dieser Befund bleibt auch nach Kontrolle konfundierender individueller Merkmale stabil. Die Variable Geschlecht erweist sich, unabhängig von weiteren individuellen Eigenschaften, als das mit Abstand am deutlichsten diskriminierende Merkmal hinsichtlich der Stärke der Handgreifkraft. Männer im Alter zwischen 70 und 80 Jahren können etwa mit Frauen im Alter von 50 oder jünger verglichen werden, und männliche Befragte, die 165 cm groß sind, erweisen sich im Durchschnitt immer noch als etwas kräftiger, als um 10 cm größere Frauen. Innerhalb der beiden Geschlechter findet sich jedoch ein klar linearer Zusammenhang zwischen Greifkraft und Alter bzw. Körpergröße (vgl. Abbildung 2, die auf Ergebnissen aus der deutschen SHARE-Teilstichprobe und dem SOEP basiert). Zudem wird ein deutlicher Zusammenhang zwischen der isometrischen Greifkraft und dem sozio-ökonomischen Status sowie dem allgemeinen Gesundheitszustand der Befragten berichtet, der im Mittelpunkt zukünftiger Längsschnittanalysen stehen wird (vgl. Hank et al. 2009).

4 Zukünftige Forschungsfelder für ‚verknüpfte' biometrische und sozialwissenschaftliche Daten

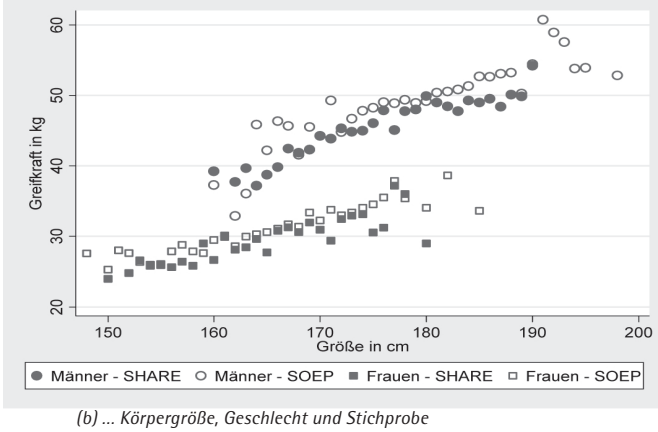
Den noch sehr begrenzten Auswertungsmöglichkeiten auf Basis der aktuell verfügbaren Biomarker des SHARE stehen eine Vielzahl von Forschungsfragen gegenüber, die erst mit der Verfügbarkeit längsschnittlicher Informationen über weitere biometrische Merkmale der Befragten untersucht werden können. Fünf vielversprechende inhaltlich wie methodisch hochrelevante Forschungsfelder seien im Folgenden beispielhaft genannt:

1. Selbst wenn sozio-demographische Variablen und selbstberichtete Angaben zum Gesundheitszustand in statistischen Modellen der Überlebenswahrscheinlichkeit berücksichtigt werden, erhöhen Biomarker signifikant die Präzision der Schätzungen und gewähren wichtige zusätzliche Einblicke in die Verläufe der *Langlebigkeit* und Determinanten der *Mortalität* im höheren Lebensalter (z. B. Turra et al. 2005; Vaupel et al. 1998). Neben DNA-Informationen haben sich hier u. a. solche Biomarker als relevant erwiesen, die im Zusammenhang mit dem Vorliegen eines metabolischen Syndroms – als einem entscheidenden Faktor für koronare Herzkrankheiten – stehen.

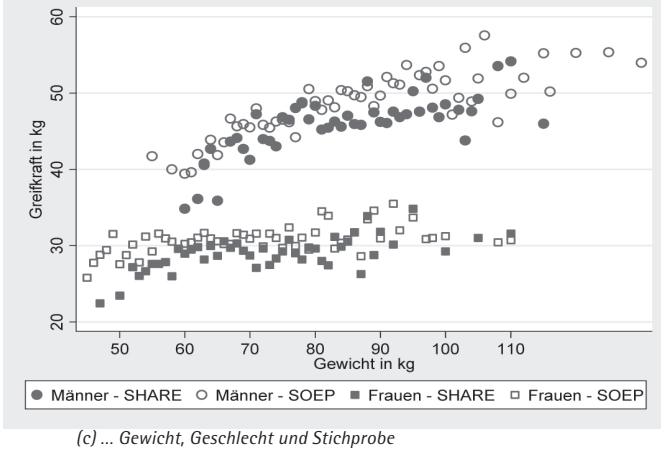
Abbildung 2 Mittlere maximale Greifkraft nach ...



(a) ... Alter, Geschlecht und Stichprobe



(b) ... Körpergröße, Geschlecht und Stichprobe



(c) ... Gewicht, Geschlecht und Stichprobe

Quelle: Hank et al. (2009).

2. Neuere Studien berücksichtigen Biomarker auch bei der Untersuchung des Zusammenhangs von *sozio-ökonomischem Status* (SES) und *Gesundheit* (z. B. Banks/Marmot et al. 2006; Dowd et al. 2006). Besondere Aufmerksamkeit hat hier u. a. die Frage erfahren, inwieweit nachhaltige Effekte von chronischem psychosozialem Stress auf das Nervensystem – gemessen z. B. über den Adrenalinspiegel – eine vermittelnde Rolle in der Beziehung zwischen SES und Gesundheit einnehmen. International vergleichende Untersuchungen weisen darauf hin, dass in diesem Zusammenhang auch unterschiedliche soziale und kulturelle Umweltprägungen zu berücksichtigen sind, die zu unterschiedlichen physiologischen Reaktionen auf Stressoren oder einen niedrigen sozialen Status führen können (z. B. Martikainen et al. 2001).
3. Die biologischen Mechanismen, die für den häufig beobachteten Zusammenhang zwischen einem hohen Maß an *sozialer Integration* und guter Gesundheit verantwortlich sind, werden bislang nur sehr unvollständig verstanden. Biomarker können hier neue Erkenntnisse liefern (z. B. Loucks et al. 2006; Uchino 2006). So konnte z. B. gezeigt werden, dass ältere Männer mit intensiveren sozialen Kontakten eine geringere Konzentration des Capsel-reaktiven Proteins aufweisen, das wesentlich mitverantwortlich für koronare Herzerkrankungen ist (Seeman et al. 2004).
4. Biomarker sind nicht nur im Hinblick auf physische Gesundheit relevant, sondern haben sich auch hinsichtlich des *seelischen Wohlbefindens* als bedeutsam erwiesen (z. B. Moffitt et al. 2006; Seplaki et al. 2004). Vor allem bei älteren Menschen sinkt z. B. die Serumskonzentration von Dehydroepiandrosteronsulfat (DHEAS), was neben funktionalen Einschränkungen auch mit einem erhöhten Depressionsrisiko und einer schlechteren Bewertung des eigenen Gesundheitszustandes sowie der eigenen Lebenszufriedenheit einhergeht (vgl. Berr et al. 1996).
5. Schließlich ist auf die Bedeutung von Biomarkern als ‚objektivem' Gesundheitsmaß für die *Validierung* selbstberichteter Gesundheitsangaben in Bevölkerungsumfragen hinzuweisen (z. B. Beckett et al. 2000; Goldman et al. 2003). Trotz ihres unbestrittenen Wertes haben sich selbstberichtete Indikatoren der Gesundheit – z. B. Selbsteinschätzungen auf einer Skala von ‚sehr gut' bis ‚sehr schlecht' – insbesondere dann als problematisch erwiesen, wenn verschiedene Subpopulationen innerhalb eines Landes oder mehrere Länder miteinander verglichen werden sollen und das Antwortverhalten nicht einheitlich ist (vgl. Jürges 2007).

5 Die Kosten der Erhebung biometrischer Daten in sozialwissenschaftlichen Umfragen

Dem hier skizzierten enormen Potential der Verknüpfung sozialwissenschaftlicher Umfragedaten mit biometrischen Informationen stehen jedoch auch verschiedene Arten von Kosten gegenüber (vgl. Weinstein/Willis 2001: 265ff.):

Erstens müssen im Vorfeld der Datenerhebung die relevanten rechtlichen und ethischen Fragen geklärt und das Surveydesign sowie Datennutzungsbestimmungen entsprechend angepasst werden (z. B. Botkin 2001; Durfy 2001). Von elementarer Bedeutung sind in diesem Zusammenhang eine umfassende Einverständniserklärung der Studienteilnehmer, absolute Vertraulichkeit und die Verpflichtung bzw. das Angebot, die Studienteilnehmer über die Untersuchungsergebnisse – und sich hierin möglicherweise widerspiegelnde Gesundheitsgefährdungen – zu informieren (vgl. etwa das entsprechende Formular des ELSA-Projektes unter http://www.ifs.org.uk/elsa/docs_w2/consent_booklet_ nurse_office.pdf).

Zweitens ist der im Zusammenhang mit der Erhebung, Lagerung und Laborauswertung von Biodaten einher gehende logistische und finanzielle Aufwand zu berücksichtigen. Im einfachsten Szenario werden nur solche Informationen erhoben, die von einem medizinisch ungeschulten Interviewer ohne Gefährdung der Befragten zuverlässig erhoben werden können (wie bereits jetzt in den ersten beiden SHARE-Wellen). Umfassendere Möglichkeiten und deutlich höhere Kosten ergeben sich hingegen aus dem Einsatz von Krankenschwestern (z. B. in ELSA) oder bei der Datenerhebung im Krankenhaus (wie im taiwanesischen SEBAS-Projekt; vgl. Weinstein/Willis 2001). Hinsichtlich des Transportes und der Lagerung etwa von Blutproben entstehen dann kaum zusätzliche Kosten, wenn die Probe per Post an ein Labor geschickt werden kann und nach der unmittelbaren Auswertung vernichtet wird. Bestimmte Analysen müssen jedoch innerhalb weniger Stunden nach Entnahme einer Blut- oder Speichelprobe durchgeführt werden; darüber hinaus kann es wünschenswert sein, Teilproben für zukünftige Analysen zu lagern (vgl. auch hierzu die Beispiele ELSA und SEBAS).

Schließlich ist, *drittens*, auf die mit der Messung (invasiver) biometrischer Daten einhergehende Belastung der Befragten hinzuweisen, die sich insbesondere bei Längsschnittstudien wie SHARE als problematisch erweisen könnte, wenn sie sich nämlich negativ auf die Panelmortaliät – d. h. das Ausscheiden von Studienteilnehmern aus der Befragung – auswirkt. Die bislang überwiegend positiven Erfahrungen vergleichbarer Untersuchungen (siehe National Research Council 2008: Teil I) geben diesbezüglich jedoch berechtigten Anlass zu einer optimistischen Einschätzung der möglichen Konsequenzen einer Ausweitung der Erhebung von Biomarkern im Rahmen des SHARE.

6 Perspektiven der Erhebung biometrischer Daten im SHARE

Die Messung weitergehender – d. h. innerer – biometrischer Merkmale von Befragten des SHARE-Projektes durch invasive Maße wird für die im Jahr 2010 – 2011 geplante vierte Erhebungswelle anvisiert. In nationalen Pilotstudien sollen Erfahrungen hinsichtlich der optimalen Balance zwischen finanziellem und logistischem Aufwand einerseits und wissenschaftlichem Ertrag andererseits gesammelt werden, die dann in den Aufbau einer europaweiten Dateninfrastruktur zur Erforschung biosozialer Aspekte des Alter(n)s einfließen sollen.

Im Rahmen der deutschen Teilstichprobe ist – zusätzlich zu den bereits im Standardprogramm von SHARE enthaltenen funktionellen Biomarkern – die Erhebung folgender biometrischer Informationen geplant:

1. Körpergröße und -gewicht (Messung zusätzlich zum Selbstbericht der Befragten),
2. Verhältnis von Hüft- zu Taillenumfang,
3. Blutdruck (am Anfang, in der Mitte und am Ende des Interviews) sowie
4. getrocknete Blutstropfen für die Laboranalyse von HbA1c, Cholesterin und C-reaktivem Protein.

Die Auswahl dieser Biomarker folgt vor allem der Maßgabe, dass sie gute Indikatoren für Krankheiten sein sollen, die sowohl im Alter stark prävalent sind als auch potentiell sozio-ökonomische Ursachen haben können. Dies sind Adipositas, Diabetes, Herzkreislauferkrankungen und Stress. *Adipositas* wird durch die Messung von Gewicht und Körpergröße sowie Hüft- zu Taillenumfang gemessen (z. B. Spencer et al. 2004). *Diabetes* wird durch HbA1c in getrockneten Blutstropfen gemessen. Im getrockneten Blut lassen sich ebenso Werte für Cholesterin (*Herzkreislauferkrankungen*) und C-reaktivem Protein (Herzkreislauferkrankungen, akute Entzündungen, *Stress*) nachweisen (vgl. hierzu ausführlich McDade et al. 2007).

Diese Beispiele zeigen, dass bereits auf Basis von relativ kostengünstigen und nur minimal invasiven Verfahren, die von Interviewern ohne professionelle medizinische Schulung im Rahmen einer sozialwissenschaftlichen Studie wie SHARE eingesetzt werden können, die Erhebung biometrischer Daten über wichtige Erkrankungen des frühen Alters möglich ist. Damit steht das Tor für eine zukünftig bessere Verknüpfung medizinisch-biologischer und sozialwissenschaftlicher Forschung weit offen.

Literatur

Banks, J., E. Breeze, C. Lessof und J. Nazroo (Hg.), 2006: Retirement, health and relationships of the older population in England: The 2004 English longitudinal study of ageing. London: IFS.

Banks, J., M. Marmot, Z. Oldfield und J. P. Smith, 2006: Disease and disadvantage in the United States and England. Journal of the American Medical Association 295: 2037-2045.

Bearman, P., 2008: Introduction – Exploring genetics and social structure. American Journal of Sociology 114 Supplement: v-x.

Beckett, M., M. Weinstein, N. Goldman und Y.-H. Lin, 2000: Do health interview surveys yield reliable data on chronic illness among older respondents? American Journal of Epidemiology 151: 315-323.

Berr, C., S. Lafont, B. Debuire, J. F. Dartigues und E. E. Baulieu, 1996: Relationships of dehydroepiandrosterone sulfate in the elderly with functional, psychological and mental status, and short-term mortality. A French community based study. Proceedings of the National Academy of Sciences of the United States of America 93: 13410-13415.

Börsch-Supan, A., A. Brugiavini, H. Jürges, J. Mackenbach, J. Siegrist und G. Weber (Hg.), 2005: Health, ageing and retirement in Europe. First results from the survey of health, ageing and retirement in Europe. Mannheim: MEA.

Börsch-Supan, A., A. Brugiavini, H. Jürges, A. Kapteyn, J. Mackenbach, J. Siegrist und G. Weber (Hg.), 2008: First results from the survey of health, ageing and retirement in Europe (2004-2007). Starting the longitudinal dimension. Mannheim: MEA.

Botkin, J. R., 2001: Informed consent for the collection of biological samples in household surveys. S. 276-302 in: C. E. Finch et al. (Hg.): Cells and Surveys. Should biological measures be included in Social Research? Washington D. C.: National Academy Press.

Carmelli, D. und T. Reed, 2000: Stability and change in genetic environmental influences on hand-grip strength in older male twins. Journal of Applied Physiology 89: 1879-1883.

Diewald, M., 2008: Zwillings- und Adoptivkinder-Stichproben für soziologische Analysen? Zum Ertrag verhaltensgenetischer Ansätze für sozialwissenschaftliche Fragestellungen und Erklärungen. DIW Research Notes 27: 1-46.

Dowd, J. B. und N. Goldman, 2006: Do biomarkers of stress mediate the relation between socioeconomic status and health? Journal of Epidemiology and Community Health 60: 633-639.

Durfy, S. J., 2001: Ethical and social issues in incorporating genetic research into survey studies. S. 303-328 in: C. E. Finch et al. (Hg.): Cells and surveys. Should biological measures be included in Social Research? Washington, D. C.: National Academy Press.

Finch, C. E. und J. W. Vaupel, 2001: Collecting biological indicators in household surveys. S. 1-8 in: C. E. Finch et al. (Hg.): Cells and Surveys. Should biological measures be included in Social Research? Washington, D. C.: National Academy Press.

Frederiksen, H., J. Hjelmborg, J. Mortensen, M. McGue, J. W. Vaupel und K. Christensen, 2006: Age trajectories of grip strength. Cross-sectional and longitudinal data among 8,342 Danes aged 46 to 102. Annals of Epidemiology 16: 554-562.

Freese, J., J.-C. A. Li und L. D. Wade, 2003: The potential relevances of biology to social inquiry. Annual Review of Sociology 29: 233-256.

Goldman, N., I.-F. Lin, M. Weinstein und Y.-H. Lin, 2003: Evaluating the quality of self-reports of hypertension and diabetes. Journal of Clinical Epidemiology 56: 148-154.

Guo, G., 2006: The linking of Sociology and Biology. Social Forces 85: 145-149.

Guo, G., 2008: Introduction to the special issue on society and genetics. Sociological Methods & Research 37: 159-163.

Hank, K., H. Jürges, J. Schupp und G. G. Wagner, 2009: Isometrische Greifkraft und sozialgerontologische Forschung: Ergebnisse und Analysepotentiale des SHARE und SOEP. Zeitschrift für Gerontologie und Geriatrie, in Druck.

Jürges, H., 2005: Handkraft und Gehgeschwindigkeit als Beispiele neuer gesundheitsbezogener Messinstrumente in der Survey-Forschung. Erfahrungen aus SHARE. S. 43-49 in: J. Schupp (Hg.): Befragungsgestützte Messung von Gesundheit. Bestandsaufnahme und Ausblick (DIW Event Documentation 2). Berlin: DIW.

Jürges, H., 2007: True health vs. response styles: Exploring cross-country differences in self-reported health. Health Economics 16: 163-178.

Kohler, H.-P., J. L. Rodgers und K. Christensen, 1999: Is fertility behavior in our genes? Findings from a Danish twin study. Population and Development Review 25: 253-288.

Kroh, M., 2005: Intervieweffekte bei der Erhebung des Körpergewichts in Bevölkerungsumfragen. Das Gesundheitswesen 67: 646-655.

Kumari, M., M. Wadsworth, M. Blake, J. Bynner und G. G. Wagner, 2006: Biomarkers in the proposed UK longitudinal household study. Economic & Social Research Council.

Lillard, D. und G. G. Wagner, 2006: The Value Added of biomarkers in household panel studies. DIW Data Documentation 14, 1-12.

Lindau, S. T. und T. W. McDade, 2008: Minimally invasive and innovative methods for biomeasure collection in population-based research. S. 251-277 in: National Research Council (Hg.): Biosocial Surveys. Washington, D. C.: National Academy Press.

Loucks, E. B., L.-F. Berkman, T. L. Gruenewald und T. E. Seeman, 2006: Relation of social integration to inflammatory marker concentrations in men and women 70 to 79 years. American Journal of Cardiology 97: 1010-1016.

Marmot, M. und A. Steptoe, 2008: Whitehall II and ELSA. Integrating epidemiological and psychobiological approaches to the assessment of biological indicators. S. 42-59 in: National Research Council (Hg.): Biosocial Surveys. Washington, D. C.: National Academy Press.

Martikainen, P., M. Ishizaki, M. Marmot, H. Nakagawa und S. Kagamirnori, 2001: Socioeconomic differences in behavioural and biological risk factors. A comparison of a Japanese and an English cohort of employed men. International Journal of Epidemiology 20: 833-838.

McDade, T.W., S. Williams und J. J. Snodgrass, 2007: What a drop can do. Dried blood spots as a minimally invasive method for integrating biomarkers into population-based research. Demography 44: 899-925.

Metter, E. J., L.A. Talbot, M. Schrager und R. Conwit, 2002: Skeletal muscle strength as a predictor of all-cause mortality in healthy men. Journals of Gerontology – Biological Sciences 57A: 359-365.

Moffitt, T. E., A. Caspi und M. Rutter, 2006: Measured gene-environment interactions in psychopathology. Perspectives on Psychological Science 1: 5-27.

National Research Council (Hg.), 2008: Biosocial surveys. Washington, D. C.: National Academy Press.

Rantanen, T., J. M. Guralnik, D. Foley, K. Masaki, S. Leveille, J. D. Curb und L. White, 1999: Midlife hand grip strength as a predictor of old age disability. Journal of the American Medical Association 281: 558-560.

Seeman, T., D. Glei, N. Goldman, M. Weinstein, B. Singer und Y.-H. Lin, 2004: Social relationships and allostatic load in taiwanese elderly and near elderly. Social Science & Medicine 59: 2245-2257.

Seplaki, C. L., N. Goldman, M. Weinstein und Y.-H. Lin, 2004: How are biomarkers related to physical and mental well-being? Journals of Gerontology – Biological Sciences 61A: B201-B217.

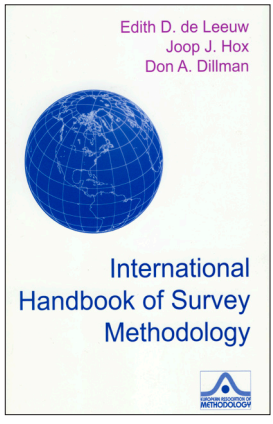
Spencer, E. A., A. W. Roddam und T. J. Key, 2004: Accuracy of self-reported waist and hip measurements in 4492 EPIC–Oxford participants. Public Health Nutrition 7: 723-727.

Turra, C. M., N. Goldman, C. L. Seplaki, D. A. Glei, Y.-H. Lin und M. Weinstein, 2005: Determinants of mortality at older ages. The role of biological markers of chronic disease. Population and Development Review 31: 675-698.

Uchino, B. N., 2006: Social support and health. A review of physiological processes potentially underlying links to disease outcomes. Journal of Behavioral Medicine 29: 377-387.
Udry, J. R., 1995: Sociology and Biology. What biology do sociologists need to know? Social Forces 73: 1267-1278.
Vaupel, J. W., et al., 1998: Biodemographic trajectories of longevity. Science 280: 855-860.
Weinstein, M., N. Goldman, A. Hedley, L. Yu-Husan und T. Seeman, 2003: Social linkages to biological markers of health among the elderly. Journal of Biosocial Science 35: 433-453.
Weinstein, M. und R. J. Willis, 2001: Stretching social surveys to include bioindicators. Possibilities for the health and retirement study, experience from the Taiwan study of the elderly. S. 250-275 in: C. E. Finch et al. (Hg.): Cells and surveys. Should biological measures be included in social research? Washington, D. C.: National Academy Press.

Korrespondenzadresse: PD Dr. Karsten Hank
Prof. Dr. Hendrik Jürges
Barbara Schaan
Mannheim Research Institute for the Economics of Aging (MEA)
Universität Mannheim
L 13, 17
68131 Mannheim
hank@mea.uni-mannheim.de

Rezensionen



E.D.DE LEEUW, J.J HOX und D.A. DILLMAN (EDS.), 2008: International Handbook of Survey Methodology. New York/London: Erlbaum/Taylor & Francis. ISBN-10: 0805857532, ISBN-13: 978-0805857535; 560 Seiten, 91,99 EUR.

Alle, die sich mit Umfragemethodologie beschäftigen oder konkret vor der Entwicklung einer (kleineren oder größeren) Umfrage stehen, haben sich wahrscheinlich schon mit Fragen dieser Art auseinandergesetzt: „How many people need to be surveyed in order to be able to describe fairly accurately the entire group? How should the people be selected? What questions should be asked and how should they be posed to respondents?" Oder haben sich gefragt: „What data collection method should one consider using, and are some of those methods of collecting data better than others?" (S. 1)

Die Herausgeber des im Jahr 2008 erschienenen International Handbook of Survey Methodology stellen sich der nicht einfachen Herausforderung, diese Fragen zu beantworten und damit alle bedeutsamen methodischen und statistischen Aspekte des Designs und der Auswertung von Umfragen abzudecken. Dementsprechend finden sich zahlreiche prominente Methodiker und Statistiker in diesem Buch, die die ihnen vertrauten Forschungsgebiete kapitelweise abdecken. Der dadurch umfänglich geratene Sammelband besteht aus den fünf thematischen Abschnitten Foundations, Design, Implementation, Data Analysis und Special Issues, die wiederum jeweils vier bis sechs Beiträge einschließen (insgesamt 26 Kapitel).

Im ersten Abschnitt (Foundations) werden zunächst in einem einleitenden Kapitel die Eckpfeiler der Umfrageforschung (Coverage, Sampling, Response, Measurement) von den Herausgebern abgesteckt. Der inhaltliche Aufbau des Sammelbandes erschließt sich demnach metaphorisch: Die Planung und Durchführung einer Umfrage wird mit dem Bau eines Hauses verglichen; nur wenn die Eckpfeiler beim Hausbau solide gesetzt werden, wird das Resultat nicht in sich zusammenstürzen (S. 3). In diesem Abschnitt zusätzlich enthalten ist Kapitel 2 von Norbert Schwarz, Bärbel Knäuper, Daphna Oyserman und Christine Stich über die kognitionspsychologischen Aspekte des Frage-Antwortprozesses; umfassend wenn auch sehr knapp wird auf (meist experimentelle) Forschungsbefunde zu Kontexteffekten, Antwortalternativen, Frageformulierungen, Frageanordnungen und Erinnerungseffekte Bezug genommen. Die praktische Handlungsanweisung am Ende des Kapitels fällt leider etwas zu kurz aus. Obwohl dem Nonresponse-Problem gleich zu Beginn des Sammelbandes ein eigenes Kapitel 3, verfasst von Peter Lynn, gewidmet ist (und das zu Recht, stellen Ausfälle durch Nonresponse einen fundamentalen und kritischen Aspekt in Umfragen dar), lässt dieses jedoch manche Wünsche offen. Verweise auf die zu diesem Themengebiet zahlreich vorliegende und einschlägige Literatur fehlen fast vollständig und auch die aktuelle Diskussion, ob Nonresponse nun unbedingt zu einem Nonresponse-Bias führt, bleibt unberücksichtigt. Dem internationalen Anspruch des Bandes am nächsten kommt Kapitel 4 von Janet A. Harkness, in dem relevante Aspekte der international und national vergleichenden Umfrageforschung diskutiert werden (Fragebogenübersetzung, Standardisierung und Vergleichbarkeit von Umfragedaten). Praktische Handlungsanweisungen (‚good practice' für Fragebogen-

übersetzungen, S. 68-70) runden das Kapitel gelungen ab. Erfreulich ist das abschließende Kapitel 5, verfasst von Eleanor Singer, über ethische Ansprüche der Umfrageforschung, ein Thema, das (in Lehrbüchern der Umfragemethodologie) zu selten angesprochen wird.

Im zweiten Abschnitt (Design) werden wesentliche Punkte vorgestellt, die es speziell im Vorfeld einer Umfrage zu beachten gilt. Zunächst wird von Sharon L. Lohr in Kapitel 6 – in klassischer Manier – auf Aspekte der Zufallsstichprobenziehung, der Abdeckung der Grundgesamtheit (Coverage) und auch – leider nur recht knapp – auf Möglichkeiten der Ziehung seltener Populationsanteile eingegangen. Entscheidungshilfen zur Wahl der adäquaten Methode zur Datensammlung werden in Kapitel 7 von Edith D. de Leeuw aufgezeigt, ein wichtiger Punkt gerade auch wenn es um die Frage der Erreichbarkeit und Verweigerung von Befragten geht. Vieles aus Kapitel 8, die Autoren sind Floyd J. Fowler Jr. und Carol Cosenza, erinnert an Kapitel 2 (The Psychology of Asking Questions); Redundanzen lösen sich aber schnell durch zahlreiche konkrete Beispiele auf. Passend schließt sich Kapitel 9 von Don A. Dillman an, indem die Implementation der gebildeten Fragen in den jeweiligen Fragebogen und nach entsprechender Methode der Datenerhebung im Vordergrund steht. Schließlich wird in Kapitel 10, verfasst von Pamela Campanelli, auf die Wichtigkeit von Pretests hingewiesen und gleichzeitig konventionelle und neuere Pretest-Techniken vorgestellt. Erwähnenswert ist, dass die meisten Kapitel dieses Abschnittes mit einer Art Zusammenfassung abschließen (‚good practice' oder Abwägung der Vor- und Nachteile), die dem eiligen Leser schnell komprimierte Informationen liefern.

Als sehr gelungen kann der dritte größere Abschnitt (Implementation) bezeichnet werden. Hier werden verschiedene Möglichkeiten aufgezeigt, eine Umfrage durchzuführen. Enthalten sind jeweils Kapitel über persönliche Befragungen mittels Interviewer (Kapitel 11, Autor ist Geert Loosveldt), telefonische Befragungen (plus ein knapper Absatz zu Handys) (Kapitel 12, verfasst von Charlotte Steeh), schriftliche Befragungen (plus ein knapper Absatz zu Schulbefragungen) (Kapitel 13, Autoren sind Edith D. de Leeuw und Joop J. Hox), web-basierte Befragungen (Kapitel 14, verfasst von Katja Lozar Manfreda und Vasja Vehovar) und Interactive Voice Response (IVR) (Kapitel 15, verfasst von Darby Miller Steiger und Beverly Conroy). Von besonderer Aktualität ist das von Edith D. de Leeuw, Don A. Dillman und Joop J. Hox verfasste abschließende Kapitel 16. Hier werden Chancen und Risiken diskutiert, multiple Methoden der Datensammlung in Umfragen einzusetzen. Dies ist von besonderer Relevanz, gerade wenn es um die Frage der Stichprobenabdeckung (Coverage) und Nonresponse geht.

Der vierte Abschnitt (Data Analysis) lässt die Leser erwarten, dass nun verschiedene multivariate Analyseverfahren (Mehrebenenanalyse, Analyse von kategorialen Daten, Zeitreihenanalysen) zur Anwendung kommen. Stattdessen – und nicht zum Nachteil des Sammelbandes – setzen die Herausgeber auf statistische Belange, die im Zusammenhang mit der Umfrageforschung im Vordergrund stehen. Behandelt werden Gewichtungsverfahren (Kapitel 17, Autoren sind Paul P. Biemer und Sharon L. Christ), Analysemöglichkeiten komplexer Umfragedaten (Kapitel 18, verfasst von Laura M. Stapleton), der Umgang mit fehlenden Werten (Item-Nonresponse) (Kapitel 19, Autoren sind Susanne Rässler, Donald B. Rubin und Nathaniel Schenker) und der Umgang mit Messfehlern (in Bezug auf Reliabilität) (Kapitel 20, verfasst von Joop J. Hox). Dieser Abschnitt setzt – entgegen den vorherigen Abschnitten – mindestens ein statistisches Grundlagenwissen und eine gewisse Vertrautheit mit formalen Darstellungen voraus.

Der fünfte und letzte Abschnitt (Special Issues) präsentiert sich als interessante Sammlung verschiedener Gebiete, die zusätzlich als wichtige Eckpfeiler der Umfrageforschung angesehen werden können: Dokumentation von Umfragen (Kapitel 21, Autoren sind Peter Mohler,

Beth-Ellen Pennell und Frost Hubbard), Qualitätssicherung und -kontrolle (Kapitel 22, verfasst von Lars E. Lyberg und Paul P. Biemer), Interviewertraining (Kapitel 23, Autoren sind Judith T. Lessler, Joe Eyerman und Kevin Wang), Datensammlung bei heiklen Themen (Kapitel 24, verfasst von Gerty Lensvelt-Mulders) sowie Panelstudien (inklusive Access Panels) (Kapitel 25, Autoren sind Dirk Sikkel und Adriaan Hoogendoorn). Der Sammelband schließt mit dem von Jelke Bethlehem verfassten Kapitel 26 über die ergänzende Zuspielung von Registerdaten (beispielsweise Zensusdaten oder amtliche Statistiken) zu erhobenen Datensätzen. Ergänzt werden könnte dieser Abschnitt durch ein Kapitel, das sich explizit mit Experimenten oder experimentellen Designs in den Sozialwissenschaften beschäftigt. Gerade vor dem Hintergrund, dass in Umfragen und Pretests immer häufiger experimentelle Versuchsanordnungen eingebaut werden und wurden (beispielsweise zur Wirkung von Incentives) und damit zum wissenschaftlichen Erkenntnisgewinn gerade auch in der Umfrageforschung beitragen.

Die einzelnen Abschnitte umfassen zwar die wichtigsten Bereiche der Umfrageforschung, könnten allerdings erweitert oder anders aufgebaut werden. Vor dem Hintergrund der zunehmenden Bedeutung länggsschnittlicher Datenerhebungen in der Umfragepraxis wäre es angemessen, dem Bereich der Panelstudien (und den damit verbundenen Vorteilen und Problembereichen) einen eigenständigen Abschnitt zu gewähren. Die einzelnen Kapitel sind verständlich zu lesen, in sich schlüssig und präsentieren den Stand der Forschung auf eine adäquate und ansprechende Art und Weise. Am Ende jedes Kapitels werden die zentralen Begriffe des Textes in einem Glossar nochmals kurz erläutert. Zudem hilft ein Index der wichtigsten Schlüsselbegriffe am Ende des Bandes, sich schnell zurechtzufinden. Sinnvoll gesetzt wurden auch Querverweise zwischen den einzelnen Kapiteln. Trotz der informativ gestalteten Kapitel bleibt den interessierten Lesern oder Anwendern nur die Möglichkeit, sich vertieftes Wissen durch die zusätzliche Lektüre spezieller Schwerpunktliteratur zu erschließen. Hierfür wird auf einer eigens dafür eingerichteten Website zusätzliches Material mit Hinweisen auf weiterführende Literatur bereitgestellt: http://www.xs4all.nl/~edithl/surveyhandbook. Sinnvoller wäre es meines Erachtens gewesen, die weiterführenden Literaturhinweise gleich am Ende des jeweiligen Kapitels zu integrieren; zudem wäre es wünschenswert, die verwendete Literaturliste gleich im Anschluss an das entsprechende Kapitel anzufügen. Damit könnte den Lesern unnötiges Nachschlagen im gesamten Literaturverzeichnis erspart werden.

Zum Schluss soll noch auf den Titel des Sammelbandes eingegangen werden. Eine explizit internationale Orientierung – wie sie den Lesern durch den Buchtitel suggeriert wird – weisen leider nur die Kapitel 4 (Comparative Survey Research), Kapitel 12 (Telephone Surveys), Kapitel 21 (Survey Documentation) und Kapitel 22 (Quality Assurance and Control) auf. Gerade Probleme mit Nonresponse oder Fragen zur Stichprobenziehung könnten im Hinblick auf internationale Umfragemethodologie ausführlicher behandelt werden.

Trotz dieser vereinzelten formalen und inhaltlichen Kritikpunkte ist es den Herausgebern gelungen, einen informativen und reichhaltigen Sammelband zur Umfragemethodologie zusammenzustellen. Das ‚International Handbook of Survey Methodology' erweist sich als nützliche Grundlage für Durchführung von methodischen Seminarveranstaltungen für fortgeschrittene Studierende oder auch als hilfreiches Nachschlagewerk für Sozialforscher, die Umfragen – auf nationaler oder internationaler Ebene – planen, erheben oder auswerten.

Sigrid Haunberger, Bern

* * * * *

Frank Faulbaum, Peter Prüfer und Margrit Rexroth, 2009: Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. GWV Fachverlage: Wiesbaden. ISBN 978-3-531-15824-2, 264 Seiten, 19,90 EUR.

Es ist bereits einige Zeit her, dass die Meinung aufgeschrieben wurde, das Entwickeln von Fragen für sozialwissenschaftliche Erhebungen sei eine Kunstlehre (Payne 1951). Damit war wohl gemeint, dass es einer gewissen individuellen Veranlagung sowie langjähriger Erfahrungen bedarf, um ordentliche Fragebogenfragen bzw. einen ganzen Fragebogen zu entwickeln. Auch dürfte damit wohl gemeint gewesen sein, dass es nicht einfach ist, solche Regeln in Worte zu fassen. Nun ist es das erklärte Anliegen des vorliegenden Buches, mit der Hilfe eines Fragebogenbewertungssystems „Frage für Frage eines Fragebogens in Hinblick auf mögliche Gefährdungen der Fragequalität zu überprüfen und damit eine sukzessive Mängelbeseitigung herbeizuführen" (S. 9). Offenbar wird hier der Versuch unternommen, im Interesse der Umfrageforschung die ursprüngliche Kunstlehre systematisch zu formalisieren.

Das Buch besteht aus einem theoretischen (drei Abschnitte) und einem praktischen (vier Abschnitte) Teil. Zunächst wird im ersten Abschnitt bei der Darstellung darüber, was denn überhaupt eine Frage sei, relativ weit ausgeholt. So werden die verschiedenen Fragetypen behandelt, diverse Varianten für Antwortvorgaben (Skalen) vorgestellt und die Befragungsmodi besprochen. Der Abschnitt hat eher den Charakter einer Auffrischung als eines Artikels innerhalb eines Lehrbuchs. In dem Bemühen, möglichst vollständig beispielsweise alle Administrationsverfahren von Fragebögen (CASI, SCAQ, DBM, ACASI, T-ACASI und CAPAR) in aller zu Gebote stehenden Kürze (d. h. auf knapp vier Seiten) vorstellen zu wollen, bleibt der Text entsprechend allgemein. Vermutlich wird es für einen Anfänger auf dem Gebiet schwer sein, ihn voll zu verstehen. Immerhin finden sich zahlreiche nützliche Verweise auf weiterführende Quellen.

Es schließt sich ein Abschnitt an, der – ebenso randvoll mit den verschiedensten Informationen wie der vorangegangene – die theoretischen Grundlagen für die Diskussion der Qualität von Fragebogenfragen legen soll. Dazu bildet das auf Tourangeau (1984) zurück gehende kognitionspsychologische Modell für die Darstellung der Antwortfindung bei sozialwissenschaftlichen Befragungen die Grundlage. Es unterstellt die Existenz eines wahren Wertes sowie von verschiedenen Einflüssen, die die Messung dieses wahren Wertes behindern bzw. modifizieren. Allerdings wird eine solche Idee bis zum heutigen Tag nicht von allen Autoren geteilt. Folgt man beispielsweise Esser (1986), so gibt es einen solchen wahren Wert gar nicht. Stattdessen richten die Zielpersonen ihre Antworten nach der sozialen Erwünschtheit aus. Im Zusammenhang mit der Entwicklung von Fragebogenfragen stellt Essers Standpunkt eher eine pessimistische Perspektive dar. Nur dürfte diese an dieser Stelle in der Tat wenig hilfreich sein. Eventuell hätte man aber auf diese Kontroverse kurz verweisen sollen.

Immerhin gelingt es an dieser Stelle, den Leser zu sensibilisieren, an welchen Stellen der Fragebogenkonstruktion bzw. aus welchen zahlreichen Quellen die Qualität eines Fragebogens negativ beeinflusst werden kann. Angereichert ist dieser Abschnitt bereits mit verschiedenen wertvollen praktischen Hinweisen für den Fragebogenentwickler.

Den Verfahren zur Evaluation von Fragebogenfragen ist der dritte Abschnitt des theoretischen Teils des Buches gewidmet. Die

verschiedenen kognitiven Pretestverfahren werden darin ebenso kurz besprochen wie die statistischen Verfahren zur Bewertung der Fragebogenqualität, etwa das Modell der konfirmatorischen Faktorenanalyse. Auch hier muss vermutet werden, dass die knappe Darstellung nicht ausreicht, um diese Verfahren lehrbuchgerecht einem Neuling zu vermitteln. Zur Wissensauffrischung ist die Darstellung allerdings gut geeignet. Zudem wird wiederum auf weiterführende Quellen verwiesen.

Der gesamte theoretische Teil des Buches ist zusammenfassend als sehr hilfreich zu bewerten. Er ist unbedingt erforderlich zum Verständnis des folgenden zweiten Hauptteils des Buches, in dem es um die praktische Seite der Bewertung von Fragen geht.

Das Fragebogenbewertungssystem (FBS) bildet den innovativen Mittelpunkt des zweiten Teils des Buches. Das FBS dient der Qualitätsprüfung einzelner Fragebogenfragen, es ist ein „Instrument zur systematischen Mängelminimierung" (S. 111). Bei einem eiligen Lesen könnte nun eventuell der Eindruck entstehen, als würde sich mit Hilfe dieses Systems die Fragebogenüberprüfung automatisieren lassen und somit deutlich einfacher und schneller gehen. Das FBS ist jedoch nicht vergleichbar beispielsweise mit einer systematischen Anleitung zur Pflanzenbestimmung, die mit etwas Übung gehandhabt dem Suchenden einen relativ schnellen Erfolg beschert. Dies wäre eine verfehlte Hoffnung. So verweisen die Autoren völlig zu Recht darauf, dass beispielsweise auch weiterhin kognitive Pretests erforderlich sein werden, um schließlich die Qualität eines Fragebogenentwurfs empirisch zu ermitteln.

Das FBS selbst funktioniert so: Man analysiert jeden einzelnen zu prüfenden Indikator eines Fragebogens nach einer umfangreichen Checkliste mit insgesamt 36 Kriterien. Diese Kriterien bewerten den Indikator wiederum anhand von 12 Dimensionen, beispielsweise A: „Probleme mit Worten/Texten", B: „Unzutreffende Annahmen über Befragte" usw. Eine Frage aus dem Komplex H „Kontext der Fragen/Fragensukzession" lautet: „Es besteht die Gefahr, dass die Frage aufgrund vorangegangener Fragen nicht in der intendierten Weise interpretiert wird" (S. 120). Dem Problemgegenstand angemessen handelt es sich bei solchen Fragen an die Fragen ganz offensichtlich nicht um die längst bekannten trivialen Faustregeln, wie sie bereits genügend oft vorgelegt wurden, sondern um eine systematische Fehlersuche mithilfe gezielt operationalisierter Kriterien.

Die Autoren verweisen darauf, dass „Die Probleme/Fehler ... in der FBS-Checkliste auf Grund ihrer knappen Formulierung für den Anwender beim ersten Durchlesen vielleicht nicht immer unmittelbar verständlich" (S. 124) sind. Dies dürfte in der Tat so sein. Somit fällt dann auch die Fehlersuche bzw. die Antwortfindung für den Nutzer des FBS sicherlich nicht immer leicht. Zur Unterstützung verweisen die Autoren jedoch auf zahlreiche zu jeder Frage angefügte Beispiele, die im folgenden Abschnitt mit den entsprechenden Kommentaren präsentiert werden.

Gerade diese Beispiele sollten beim Leser besonderes Interesse wecken. Hier liegt dann auch die besondere Stärke des Buches. Die Autoren demonstrieren anhand bekannter Erhebungsreihen (ALLBUS, ISSP, ESS usw.) viele erstaunliche Qualitätsmängel von teilweise altbekannten Fragebogenfragen.

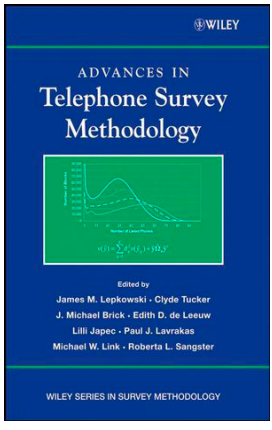
Insbesondere in diesem Abschnitt kommt den Autoren ihre immense Erfahrung beim Umgang mit sozialwissenschaftlichen Fragebögen aller Art zugute. Es gelingt ihnen, diesen Erfahrungsschatz zu komprimieren und so anderen Fragebogenentwicklern zugänglich zu machen. Umfassend, systematisch und mit einem gewissen Anspruch auf Vollständigkeit werden Fehlerquellen offen gelegt. Sie zu vermeiden wird damit einfacher.

Zusammenfassend bleibt erstens festzustellen, dass die Zukunft zeigen wird, ob und in welchem Maße sich das nicht ganz einfach zu handhabende FBS bewähren wird. Zweitens erfüllt nicht jeder Abschnitt gleichermaßen die herkömmlichen Erwartun-

gen an ein Lehrbuch. Besonders lesenswert ist das Buch drittens aber vor allem wegen der Erfahrungssammlung. Hier wird an den Fragebogenentwickler anhand gut nachvollziehbarer und eindrucksvoller Beispiele ein solides Wissen vermittelt. Das Buch stellt viertens einen wertvollen Beitrag dar, wenn es gilt, aus der Kunst, einen Fragebogen für eine sozialwissenschaftliche Umfrage zu entwerfen, schrittweise eine strukturierbare, nachvollziehbare und damit erlernbare Tätigkeit werden zu lassen.

Michael Häder, Dresden

* * * * *



J.M. Lepkowski, C. Tucker, J.M. Brick, E.D.de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link und R.L. Sangster (Eds.), 2007: Advances in Telephone Survey Methodology. Wiley Series in Survey Methodology. ISBN: 978-0-471-74531-0, 683 Seiten, 70,80 EUR.

Die Monographie ist Joseph Waksberg und Warren Mitofsky gewidmet, die auf dem Gebiet der Telefonumfragen mit der Mitofsky-Waksberg-Methode einen Meilenstein in der Entwicklung von Telefonstichproben in den USA gesetzt haben. Das Buch basiert auf den Invited Papers der zweiten Internationalen Konferenz vom 11.-15. Januar 2006 in Florida, die Methoden bei Telefonumfragen behandelte und das Ziel hatte, Arbeiten zu fördern, die sich mit der Messung und Reduzierung von Fehlern befassen, die im Zusammenhang mit Telefonerhebungen auftreten. Zudem sollten der State of the Art dokumentiert und neue Ideen für zukünftige Forschungen in Gang gebracht werden. 75 Forscher und Praktiker auf dem Gebiet der Telefonumfragen haben an dem Sammelband mitgewirkt. Dieser kann aufgrund seines Umfangs im Folgenden nur überblicksartig vorgestellt werden. Es sei aber schon an dieser Stelle vermerkt, dass die Lektüre für in der Umfrageforschung arbeitende Empiriker überaus lohnenswert ist.

Teil 1 befasst sich mit der Veränderung der Stichprobenlandschaft, die sich ausgehend von Quotenstichproben in Richtung Zufallsstichproben entwickelte. Dabei gab es bis Ende der 60er Jahre praktisch nur schriftliche oder persönlich mündliche Umfragen. Mit zunehmender Ausstattung der Haushalte mit Telefonen wuchs der Anteil der Telefonumfragen, so dass bis 1980 diese in der Praxis ein möglicher Standard wurden. Insbesondere die Möglichkeiten, die das computerunterstützte Telefoninterview mit sich brachte, eröffneten neue Chancen, ergaben aber auch neue Fragestellungen. Zentrale CATI-Labors wurden gebaut und geeignete CATI-Fragebogen erstellt. Neue Probleme ergaben sich jedoch aus der zunehmenden Weigerung von Personen, an Befragungen teilzunehmen. Außerdem änderte sich die Technologie auf diesem Gebiet sehr schnell, so dass viele Personen über mehr als nur einen Telefonanschluss verfügten. Auch die Identifikation von Geschäftsanschlüssen wurde immer schwieriger. Dazu kommen in neuerer Zeit die Haushalte, die über keinen Festnetzanschluss mehr verfügen und nur noch über Mobiltelefon erreichbar sind. All diese Probleme führten zu einer Ineffizienz des Random-Digit-Dialing-Verfahrens. Die technologische Explosion auf diesem Gebiet führte auch zu Überlegungen, Telefon- und WEB-basierte Umfragen zu kombinieren. Mobilfunk- und Festnetzstichproben sind eine andere Möglichkeit, zukünftig eine bessere Überdeckung der Bevölkerung zu erreichen. Der Aspekt der Gewichtung der Stichproben nimmt dabei an Wichtigkeit zu.

Teil 2 enthält sieben Kapitel und beginnt mit einem Überblick der Methoden, die bisher bei

der Auswahl und Schätzung im Zusammenhang mit telefonischen Umfragen verwendet wurden. Die wachsende Zahl der Haushalte, die in den USA und anderen Ländern nur noch über Mobilfunk erreicht werden, wird in mehreren Beiträgen untersucht. Auch dual- oder multiple-frame-Ansätzen sind einige Beiträge gewidmet. Die Möglichkeit, seltene Populationen über Telefon durch Screening zu finden und zu befragen, wird in einem eigenen Kapitel untersucht. In einem weiteren Kapitel behandeln Lee und Valliant die Gewichtung mittels Propensity Scores. Diese dienen nicht nur der Korrektur des Nonresponse, sondern helfen auch, Überdeckungsprobleme zu mindern.

Der Teil 3 widmet sich dem Problemkreis Datensammlung und umfasst sechs Beiträge. Dabei handelt es sich allerdings nicht um eine systematische Darstellung des Datensammlungsprozesses, sondern um eine Reihe inhaltlich unverbundener, speziellen Problemen gewidmeter Aufsätze. Es geht – mehr oder weniger verallgemeinerungswürdig – um Fragen wie z. B. Interviewerfehler, mündliche Übersetzungen am Telefon, visuelle Elemente bei der Fragebogengestaltung sowie Modeeffekte jeweils bei Telefonbefragungen. Allgemeiner ist insbesondere der Aufsatz von Lilli Japec (Statistics Sweden) gehalten. Sie beschäftigt sich mit dem Thema „Interviewer Error and Interviewer Burden" (Kapitel 9). Darin entwickelt sie ein Modell des im Interview ablaufenden Frage-Antwort-Prozesses, das deutlich über Ansätze wie etwa von Tourangeau (1984) hinausgeht. In Japecs Modell werden auch diejenigen kognitiven Prozesse berücksichtigt, die beim Interviewenden ablaufen. Damit wird eine Systematik für die Analyse von Interviewer-Fehlern geschaffen. Weitere Überlegungen richten sich auf eine Übersicht zu den Interviewer-Lasten, die auf den Daten zweier Befragungen der Interviewer von Statistics Sweden basieren. Diese sind im Anhang dokumentiert, so dass die Analysen sehr gut nachvollziehbar sind.

Sehr nützlich ist der Beitrag von Brad Edwards und weiteren Kollegen von Westat (Kapitel 13). Sie widmen sich der Gestaltung der Fragebögen bzw. dem Aufbau der Bildschirme bei CATI-Befragungen. Nach einer Übersicht über vorhandene Literatur zu diesem Thema werden Beispiele für den Bildschirmaufbau und Designvarianten diskutiert. Der Aufsatz endet mit einigen Richtlinien zum CATI User Interface Design.

Im Teil 4 „Operations" werden sowohl technische Probleme bei Telefonumfragen als auch Interviewereigenschaften behandelt. Bei den technischen Fragestellungen handelt es sich z. B. um Erfahrungen beim Aufbau und der Einrichtung eines neuen Call Centers (Kapitel 15) und um die Ausgestaltung von CATI Management Systemen (Kapitel 16) für die logistische Steuerung bei Telefonumfragen. Hier werden praktische Tipps für die Realisierung von telefonischen Befragungen gegeben, die insbesondere an Sozialwissenschaftler gerichtet und für diese äußerst empfehlenswert sind – von den notwendigen Einrichtungsgegenständen für ein Call Center bis hin zur optimalen Anrufzeit.

Die nächsten drei Kapitel besprechen unterschiedliche Aspekte des Interviewereinsatzes. Sehr informativ ist z. B. Kapitel 17, das einen profunden Überblick über Interviewertraining sowie Qualitätskontrolle und Effizienzmessung bei Interviewern bietet. Insgesamt hinterlässt der Teil 4 einen schlüssigeren, zusammenhängenderen Eindruck als Teil 3 mit seinen weitgehend isolierten Kapiteln. Positiv hervorzuheben sind auch die zahlreichen praktischen Tipps, die bei einer Einarbeitung in die Problematik Telefonumfragen von unmittelbarem Wert sein werden.

In Teil 5 wird schließlich das Thema „Nonresponse" aufgegriffen, ein Thema, das für die Umfrageforschung angesichts der weltweit sinkenden Teilnahmebereitschaft bei Befragungen überlebenswichtig sein dürfte. So widmet sich auch der erste Aufsatz in diesem Teil (Kapitel 21) den Gründen für die Nichtteilnahme bei Umfragen: „Privacy, Confidentiality, and Response Burden as Factors in Telephone Survey Nonresponse." Eleanor

Singer und Stanley Presser geben hier eine Übersicht über Einstellungen zu Umfragen – und dies nicht nur aus US-Perspektive, sondern auch unter Berücksichtigung europäischer Trends. Sie kommen zu folgendem Ergebnis: „We suspect that the impact of response rates of all three factors – privacy, confidentiality, and perceived burden – will increase in the future because of increasing impatience with the intrusion of telephones into private space and the increasing salience of privacy and confidentiality issues" (S. 470). Gerade weil die Öffentlichkeit keine Trennung „between marketing calls and legitimate survey requests" (S. 470) vornimmt, scheint den Rezensenten eine Image-Kampagne der mit seriöser Markt- und Sozialforschung befassten Institute unumgänglich zu sein. Wenn die öffentliche Meinung nicht positiver auf Umfrageforschung eingestellt ist, werden in Zukunft kaum mehr akzeptable Responseraten zu verzeichnen sein. Dagegen ist auch die Vergabe von Incentives kein probates Mittel (Kapitel 22). Insgesamt ist dieser Teil 5 äußerst informativ und deckt den Bereich des Nonresponse gut ab.

Zusammenfassend gilt es zu vermerken, dass der vorliegende Sammelband einen sehr guten Überblick über die theoretischen und praktischen Probleme der Telefonumfragen bietet. Nützlich sind insbesondere auch die konkreten Tipps, wie etwa zur Einrichtung eines CATI-Labors. Da das Buch allerdings inzwischen schon zwei Jahre alt ist, muss zu einigen speziellen Themen, wie z. B. zur Methodik von Mobilfunkbefragungen, ergänzende Literatur herangezogen werden.

Sabine Häder und Siegried Gabler, Mannheim

* * * * *

Michael Häder und Sabine Häder (Hrsg.), 2009: Telefonbefragungen über das Mobilfunknetz: Konzept, Design und Umsetzung einer Strategie zur Datenerhebung. VS-Verlag, Wiesbaden. ISBN 978-3-531-15790-0, 303 Seiten, 34,90 EUR.

Der Untertitel deutet an, dass das Buch eine Studie beschreibt, in der Mobilfunkbefragungen getestet wurden. Diese Studie, namens CELLA, kombinierte Mobilfunk- (CEL = cell phone) und Festnetzbefragungen (LA = land line). In ihrem Vorwort legen die Autoren das Ziel der Publikation dar. Sie möchten die Leser an ihren Erfahrungen mit einer relativ neuen Erhebungstechnik teilhaben lassen. Das Buch richtet sich somit an Leser, die diese Erfahrungen kritisch diskutieren sollen. Der Inhalt ist also weniger ein ausgereiftes Konzept einer Umfragetechnik, das der Praxis als Kopiervorlage übergeben werden kann, als vielmehr der erste Schritt in einem neuen Forschungsfeld in Deutschland. Ich verstehe dies als Auftrag an die Forschungsgemeinschaft, die Ergebnisse zu replizieren, die neu aufgeworfenen Fragen mit anderen Studiendesigns versuchen zu beantworten und die neue Umfragetechnik zur Marktreife zu entwickeln.

Das Buch gliedert sich in fünf Teile. Der erste Teil schildert die Entwicklung der Umfrage mittels Telefon in Deutschland, sowie die Vor- und Nachteile telefonischer Befragungen. Dies mündet in eine Empfehlung, zukünftig Mixed-Mode-Studien mit Mobilfunk- und Festnetzbefragungen durchzuführen, um möglichst alle Personen mit Telefon erreichen zu können, also auch die, die nur (noch)

per Handy telefonieren (Mobile-only), und diejenigen, die nur Festnetztelefonie nutzen. Es werden die Themenfelder benannt, die für eine Mixed-Mode-Studie relevant sind und die in der CELLA-Studie systematisch untersucht wurden: Stichprobenziehung, Gewichtung, Mode-Effekte und Teilnahmebereitschaft.

Im zweiten Teil wird die Zahl der Mobile-only-Personen in verschiedenen europäischen Ländern genannt. Dazu wird aber lediglich eine Studie zitiert. Der relativ große Anteil von 11 Prozent führt zur Empfehlung, zukünftig Mixed-Mode-Studien durchzuführen. Mir sind andere Zahlen bekannt, die eine Mixed-Mode-Studie nicht notwendig erscheinen lassen. Aber unabhängig davon, wann der Mixed-Mode-Ansatz notwendig sein wird, verdient die Erforschung des Ansatzes Lob. In diesem Teil wird weiterhin die Stichprobenziehung detailliert beschrieben und der Dual-Frame-Ansatz begründet. Bei diesem Ansatz wird kein Screening durchgeführt, um nach speziellen Teilnehmern zu suchen, sondern es werden Umfrageteilnehmer aus zwei Auswahlgrundlagen gezogen und kombiniert ausgewertet. Die beiden Auswahlgrundlagen sind das nach dem Gabler-Häder-Verfahren konstruierte Universum der Festnetznummern und das ähnlich konstruierte Universum der Mobilfunknummern. Letzteres wird ausreichend beschrieben, um erkennen zu können, dass die Auswahlgrundlage für den Studienansatz geeignet ist. In Ermangelung anderer, insbesondere regionaler Kriterien wird das Universum der Mobilfunknummern nach Provider geschichtet. Dieses Merkmal wird aus der Vorwahlnummer generiert, wohl wissend, dass die Trennschärfe des Merkmals wegen der Rufnummernportierung mangelhaft sein könnte. Die Autoren zitieren aber die Bundesnetzagentur, die konstatiert, dass nur ein Prozent der Mobilfunkteilnehmer zwischen 2003 und 2006 von dieser Möglichkeit Gebrauch gemacht hätten. Dann wäre dieser Fehler vernachlässigbar. Allerdings, so wird im dritten Teil des Buches geschrieben, behaupten über zehn Prozent der CELLA-Stichprobe, sie hätten ihre Rufnummer von einem Anbieter zum nächsten portiert. Diese Diskrepanz wird an keiner Stelle des Buches diskutiert. Sie deutet darauf hin, dass entweder nach 2006 die Rufnummerportierung ein bedeutsames Phänomen darstellt, das die Schichtung stark beeinträchtigt, oder die Stichprobe einen erheblichen Bias aufweist. Für letzteres spricht, dass die Auswahlchance lediglich über die Zahl der Rufnummern operationalisiert wurde. Es wird zwar argumentiert, dass ein erheblicher Anteil in beiden Substichproben ihr Handy immer angeschaltet lassen, aber dennoch ist die Differenz in diesem Merkmal zwischen beiden Substichproben bedeutsam. Es scheint, dass zu viele Handynutzer ausgewählt wurden, die ihr Handy immer angeschaltet haben. Das könnte mit der Rufnummerportierung zusammenhängen.

Für die Kombination der beiden Substichproben wird ein Gewichtungsverfahren mittels Nivellierung der Inklusionswahrscheinlichkeiten vorgeschlagen. Wie oben angeführt, scheint das Verfahren noch verbesserungswürdig zu sein. Zudem wurden Annahmen über die Zahl der Rufnummern getroffen, die zweifelhaft sind. Bei Angabe eines ISDN-Anschlusses werden beispielsweise 2,5 Rufnummern geschätzt. Es ist nach meinen eigenen Erfahrungen tatsächlich schwierig, die valide Anzahl von Rufnummern per Befragung zu ermitteln, aber dennoch sollte man zumindest versuchen, eine geeignete Befragungsroutine zu entwickeln. Auch das könnte Aufgabe zukünftiger Forschung sein.

Der zweite Teil des Buches schließt mit der Diskussion der Stichprobenqualität. Dazu werden einerseits Verteilungen bestimmter soziodemografischer Merkmale mit denen des Mikrozensus verglichen und andererseits der Ausschöpfungsbericht betrachtet. Beim Vergleich mit dem Mikrozensus werden die Mobile-only-Personen näher untersucht. Es zeigt sich, dass sich dieser Personenkreis von der restlichen Bevölkerung in soziodemografischen Merkmalen erheblich unterscheidet: sie sind vornehmlich männlich, zwischen 20 und 29 Jahre alt, überwiegend ledig und le-

ben hauptsächlich in Einpersonenhaushalten. Dieses Ergebnis der CELLA-Studie verdeutlicht die Notwendigkeit, Mobile-only-Personen in Befragungen zu berücksichtigen, wenn deren Anteil an der Gesamtbevölkerung tatsächlich über zehn Prozent beträgt.

Der weitere Vergleich von Randverteilungen soziodemografischer Merkmale der CELLA-Studie mit dem Mikrozensus macht deutlich, dass sich zwar die beiden Substichproben teilweise erheblich von der Mikrozensus-Stichprobe unterscheiden, die kombinierte Gesamtstichprobe der CELLA-Studie aber meist kaum noch vom Mikrozensus differiert. Die Autoren schließen daraus, dass sich mögliche Fehler der Substichproben gegenseitig aufheben und daher eine Mixed-Mode-Stichprobe eine Verbesserung darstellt. Hinsichtlich des Noncoverage-Fehlers ist diese Aussage sicher zutreffend, aber sonst greift diese Schlussfolgerung zu kurz. Überspitzt formuliert könnte man diese Aussage so zu einer Empfehlung zusammenfassen: Kombiniere zwei schlechte Stichproben und du erhältst eine gute. Man muss aber Folgendes beachten: während Differenzen zwischen Stichproben und Mikrozensus oder amtlichen Daten auf ernste Probleme in der Abbildungsgüte hinweisen – und das ist in den Substichproben der Fall –, ist das Fehlen von solchen Differenzen keine Gewähr für die Abbildungsgüte vor allem in den interessierenden Merkmalen. Tatsächlich wird ja beispielsweise der Anteil derjenigen, die ihre Rufnummer von einem Anbieter zum nächsten mitgenommen haben, erheblich überschätzt. Ein Hinweis, warum dies in der CELLA-Studie zu beobachten ist, findet sich in der nachfolgenden Darstellung des Ausschöpfungsberichts. Bei 21 Prozent der kontaktierten Handynummern meldete sich die Mailbox und es konnte kein Kontakt hergestellt werden. Die Autoren schlagen vor, dies wegen widersprüchlicher Befunde bei mehrfachen Kontaktversuchen als stichprobenneutraler Ausfall zu werten. Angemessen wäre hier aber sicher in der Mehrzahl der Fälle die Annahme eines stichprobenrelevanten Ausfalls, denn es ist zu vermuten, dass diejenigen Personen, die ihr Handy nicht immer eingeschaltet haben, deutlich schlechter erreicht wurden. Der Anteil der Handy-affinen Personen ist damit deutlich überschätzt, denn über 50 Prozent der Befragten konnten mit dem ersten Kontaktversuch interviewt werden.

Im dritten Teil wird deutlich, dass die Verweigerungsrate bei Mobilfunkbefragungen kein besonderes Problem darstellt. Die über das Handy kontaktierten Personen sind sogar eher bereit zu kooperieren als die über das Festnetz kontaktierten. Dies wird erklärt durch den Neuigkeitsgrad dieser Umfragetechnik. Die Teilnahmebereitschaft wird zusätzlich erhöht, wenn die Interviewanfrage per SMS angekündigt wird.

Bei der Frage nach der Teilnahmebereitschaft zu Mobilfunkbefragungen in Abhängigkeit verschiedener Situationen und Lokalitäten erklären die Befragten über Mobilfunk in allen Situationen eine höhere Bereitschaft als die anderen Befragten. Das ist natürlich insofern nicht überraschend, weil für die über Festnetz kontaktierten Personen eine Mobilfunkbefragung eine ungewöhnliche Erhebungstechnik darstellt. Interessant ist der Befund, dass fast 40 Prozent der über das Handy kontaktierten Personen nicht zu Hause, sondern anderswo interviewt wurden. Die Lokalität, in dem das Interview stattfindet, scheint für die Teilnahmebereitschaft nicht sonderlich nachträglich zu sein. Das gilt auch für Situationen, in denen dritte Personen anwesend sind.

Da das Handy fast ausnahmslos von den befragten Personen alleine genutzt wird, sind die Einheiten der Mobilfunkstichprobe Personen. Die Einheiten der Festnetzstichprobe dagegen sind Haushalte. Bei diesen Haushalten erfolgt beim Erstkontakt eine Personenauswahl. Hierzu hat Siegfried Gabler (S. 93) die Geburtstagsmethode durch einen Zufallschritt ergänzt, indem ein Datum per Zufall bestimmt und dann die Person des Haushalts ausgewählt wird, die entweder als letztes vor diesem Datum oder als erstes nach diesem Datum Geburtstag hat. Dieses Verfahren hat sich in der CELLA-Studie bewährt und simuliert m. E. hervorragend eine Zufallsauswahl.

Im dritten Teil werden auch die Ergebnisse der Fragen zum Mobilfunktelefonverhalten

und die allgemeinen Erfahrungen mit den Handy-Interviews in der CELLA-Studie beschrieben. Die Autoren berichten von durchaus positiven Erfahrungen. Die Personen waren nicht in besonderem Maße verärgert und es gab keine großen technischen Probleme. Die CELLA-Studie belegt also, dass Mobilfunkbefragungen möglich sind.

Interessant ist der Zugang der Autoren zur Erfassung der Anwesenheit Dritter und die kurze Befragung von Nonrespondenten. Die Interviewer sollten möglichst alle Nebengeräusche erfassen. Tatsächlich war die Zahl der Verweigerungen und Terminvereinbarungen größer, wenn Nebengeräusche registriert wurden. Da im Fragebogen auch explizit nach der Anwesenheit Dritter gefragt wurde, konnte auch festgestellt werden, dass Kleinkinder und Partner des Befragten das Interview beeinflussen. Insgesamt gilt, der Einfluss Dritter fällt im Telefoninterview geringer aus als in persönlich-mündlichen Befragungen, wie ein Vergleich mit anderen Studien ergab. Der Fragebogen für die Nonrespondenten ergab die üblichen Verweigerungsgründe und konnte meist nicht, wie erhofft, als Incentive wirken und zu einer Teilnahme führen. Hier regen die Autoren weitere Forschung an. Weiterhin werden im dritten Teil auch die Pretests und Vorstudien besprochen.

Im vierten Teil werden die Mode-Effekte besprochen. Bemerkenswert ist, dass die Autoren eine Vielzahl von Messparametern, Instrumenten und experimentellen Designs für die Analyse von möglichen Effekten entwickelt haben. Erfreulicherweise – für die Praxis der Umfrageforschung – treten kaum Mode-Effekte auf. In Mobilfunkbefragungen gibt es allenfalls eine geringere Tendenz zur sozialen Erwünschtheit des Antwortverhaltens und eine bessere Erinnerungsleistung. Nachteilig für diese Methode ist die etwas größere Anzahl von Abbrüchen während des Interviews. Einen deutlichen Effekt gibt es bei Einstellungsfragen. Allerdings wurden in diesem Zusammenhang nur Fragen bezüglich des Mobilfunks gestellt. Es überrascht nicht, dass die über das Handy kontaktierten Personen sich positiver zu Themen des Mobilfunks äußern. Etwas vorschnell werten die Autoren dies als Mode-Effekt und interpretieren dieses Verhalten als durch das Handy-Interview verursacht. Sie beachten dabei nicht, dass sich – wie oben angeführt – auch die Stichproben in mehreren Merkmalen unterscheiden. Ich nehme eher an, die positivere Einstellung ist darauf zurückzuführen, dass in der Mobilfunkstichprobe mehr Handy-affine Personen interviewt wurden als in der Festnetzstichprobe. Hierzu müssen nachfolgende Studien durchgeführt werden mit Einstellungsfragen zu anderen Themen und eventuell mit echten Experimenten, das heißt mit Zufallsaufteilung auf die Substichproben. Dazu kann man die in der CELLA-Studie entwickelten Instrumente und Parameter verwenden.

Der fünfte Teil schließlich enthält die verwendeten Quellen und den Fragebogen. Für Forscher auf diesem Gebiet ist die sehr umfangreiche Literaturliste wertvoll. Vergleicht man das Literaturverzeichnis mit den entsprechenden Zitaten im Text, fallen einige Unstimmigkeiten auf. Beispielsweise sind Autoren vertauscht oder mein Name ist im Text falsch geschrieben. Letzteres ist absolut verzeihlich, zieht man die vielen möglichen Schreibweisen in Betracht. Weniger verzeihlich ist, dass im Text Literatur erwähnt wird, die im Verzeichnis fehlt. Hier hätte man etwas sorgfältiger schreiben und redigieren sollen. Dann wären sicher auch die Fehler aufgefallen wie beispielsweise in Tabelle 13.10, in der eine Variable als besonders bedeutsam beschrieben wird, obwohl ihr beta-Gewicht Null ist.

Das sind aber nur Kleinigkeiten, die den Lesegenuss nur minimal schmälern. Alles in allem ist dieses Buch von hohem Wert für alle Forscher, die auf diesem Gebiet arbeiten. Den Autoren gebührt Dank dafür, dieses Forschungsgebiet für Wissenschaftler in Deutschland erschlossen zu haben. Ihre Studie beantwortet bereits einige Fragen, wirft aber – und das ist in einem Forschungsprozess die Regel – noch mehr Fragen auf. Die Forscher können sich nun darauf freuen, hier weiter zu arbeiten. Ihnen sei dieses Buch sehr empfohlen.

Gerd Meier, Lüneburg

Birgit Pfau-Effinger, Slađana Sakač Magdalenić und Christof Wolf (Hrsg.), 2009: International vergleichende Sozialforschung. VS-Verlag, Wiesbaden. ISBN 978-3-531-16524-0, 235 Seiten, 39,90 EUR.

Der von Birgit Pfau-Effinger, Slađana Sakač Magdalenić und Christof Wolf herausgegebene Sammelband ‚International vergleichende Sozialforschung' enthält Beiträge einer Tagung, die 2007 von der DGS Methodensektion und der Arbeitsgemeinschaft sozialwissenschaftlicher Institute durchgeführt wurde. Die ersten drei Beiträge befassen sich stärker als die restlichen Beiträge aus inhaltlicher Perspektive mit den Konsequenzen von Globalisierungsprozessen und den institutionellen Konstellationen in einzelnen Ländern, die diese Wirkungen puffern und damit länderspezifisch verändern. Die fünf weiteren Beiträge des Bandes befassen sich mit den Möglichkeiten und Grenzen ganz bestimmter Messinstrumente in international vergleichenden Studien. Der Aspekt der Globalisierung spielt hier kaum eine Rolle. Dementsprechend ist die Diskussion von methodischen Problemen und Aspekten in diesen Beiträgen deutlich ausgeprägter.

Hans-Peter Blossfeld und seine Kollegen befassen sich in ihrem Beitrag mit dem Einfluss von Globalisierungsprozessen auf Lebensverläufe. Dabei fassen sie die Ergebnisse aus zwei größeren Forschungsprojekten zu dieser Thematik zusammen. Ohne methodisch in die Tiefe zu gehen, gibt der Beitrag einen sehr informativen Überblick über die Befunde dieser Studien. Allerdings ist eine ähnliche Zusammenfassung auch bereits an anderer Stelle erschienen. Die Autoren zeigen, dass Globalisierung keineswegs einheitliche Wirkungen zeigt, in Abhängigkeit von Geschlecht und Lebensverlauf ergeben sich je unterschiedliche Risiken, zudem sorgen die unterschiedlichen institutionellen Rahmenbedingungen der Länder (Wohlfahrtsregime) zu im internationalen Vergleich unterschiedlichen Konsequenzen. Leider nur locker mit den Befunden verbunden, ist das im ersten Teil des Beitrages kurz vorgestellte multidimensionale Messkonzept von Globalisierung. Jürgen Beyer befasst sich in seinem Beitrag kritisch mit der Frage, inwieweit die institutionellen Rahmenbedingungen verschiedener Länder zu einer Verfestigung bestehender marktwirtschaftlicher Strukturen beitragen. Grundthese ist, dass die Unternehmen die jeweiligen Standortvorteile gezielt für ihre Produktion nutzen. Er untersucht hierfür organisatorische Veränderungen deutscher Unternehmen und legt zudem eine Analyse von Makrodaten von 25 Ländern vor, in der die sektorale Spezialisierung als Indikator für eine Verfestigung von Strukturen herangezogen wird. Der international vergleichende Teil bezieht sich auf eine relativ lange Zeitspanne, die Paarvergleiche zwischen liberalen und koordinierten Ökonomien allerdings primär auf Deutschland und die USA. Die Ergebnisse beider Analysen sprechen gegen die Verfestigungsthese. Raj Kollmorgen geht in seinem Beitrag der Frage nach, inwieweit die Esping-Andersen Typologie der Wohlfahrtsregime auf postsozialistische Übergangsgesellschaften in Mittel- und Osteuropa zu übertragen ist. Er untersucht dazu zum einen Wohlfahrtsausgaben im Vergleich, zum anderen Merkmale der Wohlfahrtssysteme von Übergangsgesellschaften. Auf der Basis sehr kleiner Fallzahlen bildet er neue Subtypen (zwei Länder bilden bereits einen Typ) und kommt zu dem Schluss, dass zumindest zum jetzigen Zeitpunkt die Globalisierung keine Uniformierung von Wohlfahrtsregimen nach sich zieht, sondern in den Ländern neue Kombinationen eigener Traditionsbe-

stände und verschiedener institutioneller Elemente auftreten.

Die nun folgenden Beiträge konzentrieren sich stärker auf die Diskussion methodischer Probleme. Henning Lohmann sucht in seinem Beitrag nach Konzepten und Messinstrumenten für Defamilisierungsprozesse, um Familienpolitik international vergleichen zu können. Bei der Durchsicht einer Vielzahl von Studien stellt er fest, dass weder auf der konzeptuellen noch auf der methodischen Ebene Ansätze vorliegen, die den methodischen Kriterien einer Vergleichbarkeit genügen. Dennoch zeigen die Studien auf inhaltlicher Ebene ähnliche Befunde trotz unterschiedlicher Konzepte. Cornelia Weins diskutiert in ihrem Beitrag die Probleme der vergleichenden Analyse von Vorurteilen, dabei bezieht sie sich primär auf die ISSP-Studie von 2003 und einer von Semyonov et al. 2006 in der ASR publizierten Analyse dieser Daten. Sie zeigt, wie die Nicht-Berücksichtigung der sehr hohen Zahl fehlender Werte und eine mangelnde Prüfung der Messinvarianz zu verzerrten Ergebnissen führt.

Volker Müller-Benedict beschäftigt sich mit der Frage, ob vorhandene indikatorengesteuerte entwicklungspolitische Maßnahmen zur Veränderung der Bildungspolitik von Entwicklungsländern in der Lage sind, die intendierten Wirkungen zu erzielen. Er bezieht sich dabei auf die Initiative Education For All (EFA) bzw. die EFA-Fast-Track Initiate der UNESCO, die Indikatoren der Schulwirksamkeit entwickelt haben, um eine Primarschulbildung für alle Kinder der Erde bis zum Jahr 2015 zu realisieren. Zunächst werden unterschiedliche Definitionen von Schulqualität und Modelle der Schulwirksamkeit, Ergebnisse der Schulwirksamkeitsforschung sowie die Indikatoren der EFA-FTI erläutert. Dabei wird deutlich, dass davon ausgegangen wird, dass über die Stärkung der Eigenverantwortlichkeit der Schulen und der Lehrer eine Verbesserung der Bildungsqualität erreicht werden soll. Als Indikatoren sind nur Einschulungs-, Erfolgs- und Abbrecherqouten vorgesehen, die über die Einführung der Maßstäbe des ‚guten Unterrrichts' und der Konkurrenz unter den Schulen verbessert werden sollen. Dieses Modell der UNESCO wird dann mit der Schulwirklichkeit in Honduras konfrontiert. Dabei bezieht sich Müller-Benedict nicht auf eine selbst durchgeführte Studie, sondern auf eine von Claudia Richter durchgeführte qualitative Studie aus dem Jahr 2007. Auf der Grundlage von Dokumentenanalysen und von qualitativen Experteninterviews kommt die Studie zu dem Ergebnis, dass die Probleme der Grundversorgung in Honduras (problematischer Ernährungszustand der Kinder, fehlende Lern- und Lehrmaterialien, fehlende sanitäre Einrichtungen sowie die häufige Abwesenheit von Lehrern und Schülern) durch die EFA-FTI-Maßnahmen nicht entscheidend verbessert werden. Vielmehr zeigt sich, dass die Indikatoren und die am Bildungsprozess beteiligten authochtonen Experten selbst wesentliche Probleme des Bildungssystems in Honduras ausblenden. Leider wird in diesem Beitrag weder begründet, warum eine qualitative Studie durchgeführt wird, noch erläutert, ob die Fragestellung nur durch einen kulturimmanten Ansatz beantwortet werden kann.

Die beiden letzten Aufsätze befassen sich in sehr spezialisierten Beiträgen mit der Entwicklung von Operationalisierungen sozialstruktureller Merkmale in international vergleichenden Surveystudien. Pollack und seine Kollegen beziehen sich dabei auf die Entwicklung der neuen sozioökonomischen Klassifikation für Europa (ESeC – European Socio-economic Classification), dabei wird ein dieser Klassifikation zugrunde liegender Aspekt, nämlich die Problematik der Erfassung und Abgrenzung der sogenannten Supervisoren-Verantwortlichkeit genauer analysiert. Die Autoren zeigen anhand einer Pilotstudie, dass abweichende Itemformulierungen die Anteilschätzungen der als Supervisor identifizierten Befragten verzerren. Abschließend beschäftigen sich auch Uwe Warner und Jürgen Hoffmeyer-Zlotnik mit sehr speziellen Problemen der Operationali-

sierung, in diesem Fall mit der Schwierigkeit der Abgrenzung des Konzeptes des privaten Haushaltes in international vergleichenden Studien. Anhand von Daten des ESS und des European Community Household Panel werden die Folgen unterschiedlicher Definitionen dargelegt. Abschließend wird ein Vorschlag für die Operationalisierung des Konzeptes vorgestellt.

Im Rahmen des vorgelegten Sammelbandes werden interessante Themen der aktuellen international vergleichenden Forschung behandelt. Allerdings ergibt sich ein sehr breites Spektrum von Themenfeldern, das nur sehr locker über den Begriff des internationalen Vergleichs zusammengehalten wird. Auch das im Untertitel aufgegriffene Thema der Globalisierung findet sich nicht wirklich in allen Beiträgen wieder. Insbesondere die ersten drei Beiträge widmen sich dieser Thematik und dürften daher auch für Leser, die sich für die inhaltliche Diskussion dieser Thematik interessieren, interessante Befunde liefern. Die weiteren Beiträge sind hingegen wohl eher für methodisch interessierte und spezialisierte Leser interessant, die sich mit den jeweiligen Themengebieten selbst in der Forschung befassen. Insgesamt krankt der Sammelband daran, dass allein die sehr allgemein gehaltene Forschungsperspektive des internationalen Vergleichs noch keine zumindest im Groben kohärenten Inhalte liefert. Das je nach Interesse selektive Lesen der Beiträge ist für den Leser sicherlich von Gewinn, da alle Beiträge für sich, interessante Befunde und Diskussionsanstöße liefern.

Susanne Rippl, Chemnitz und
Christian Seipel, Hildesheim

Ankündigungen

6. Nutzerkonferenz ‚Forschung mit dem Mikrozensus'

German Microdata Lab, GESIS &
Statistisches Bundesamt
15. - 16. Oktober 2009

Konferenzort: Hotel Wartburg
Quadrat F 4, 4-11, 68159 Mannheim

Analysen zur Sozialstruktur und zum sozialen Wandel

Die 6. Nutzerkonferenz widmet sich der Untersuchung der Sozialstruktur und des sozialen Wandels in Deutschland. Auf der Basis von Mikrozensusdaten gewonnene Forschungsergebnisse werden vorgestellt und diskutiert. Darüber hinaus ist die Konferenz ein Forum für den Erfahrungsaustausch der Datennutzer/innen untereinander sowie mit den Vertreter/innen der amtlichen Statistik. Sie wendet sich an Wissenschaftler/innen, die mit den Scientific Use Files des Mikrozensus arbeiten oder zukünftig mit diesen Daten arbeiten wollen.

Eine Anmeldung zu der Konferenz ist ab sofort unter folgender Adresse möglich: workshop-mannheim@gesis.org
Der Konferenzbeitrag beträgt €120 (Studierende € 90).
Weitere Informationen finden Sie unter: www.gesis.org/gml/veranstaltungen
Bei Fragen wenden Sie sich bitte an die Organisatorinnen bei GESIS: andrea.lengerer@gesis.org und julia.schroedter@gesis.org

Programm

Donnerstag, 15. Oktober 2009

10:00 – 10:45 Begrüßung und Einführung

Sozialer Wandel in Deutschland: Analysen mit dem Mikrozensus
Christof Wolf (GESIS, Mannheim)

Weiterentwicklung des Mikrozensus
Hermann Seewald (Destatis, Bonn)

10:45 – 12:15 Erwerbstätigkeit und Familie

Die Erwerbstätigkeit junger Mütter - Entwicklungen und Muster in Ost- und Westdeutschland
Barbara Hanel und Regina Riphahn (Universität Erlangen-Nürnberg)

Väter in Elternzeit. Eine Analyse mit den Mikrozensen 1999-2007
Esther Geisler und Michaela Kreyenfeld (MPIDR, Rostock)

Aufteilung der Erwerbsarbeit bei Paaren mit Kindern in Ost- und Westdeutschland
Jeanette Bohr (GESIS, Mannheim)

12:15 – 13:30 Mittagspause

13:30 – 15:30 Armut

Immer ärmer? Zum Wandel der Wohlstandsposition von Haushalten mit Kindern von 1962 bis 2004
Peter Hartmann (Universität Düsseldorf)

Wandel und Ursachen familialer Armut in Deutschland
Mara Boehle (GESIS, Mannheim)

Working Poor: Erwerbstätigkeit in Haushalten mit Transferleistungen
Helmut Rudolph (IAB, Nürnberg)

Analysen zur Einkommensarmut mit dem Mikrozensus
Sabine Köhne-Finster (Destatis, Bonn)

15:30 – 16:00 Kaffeepause

16:00 – 18:30 Arbeitsmarkt und Einkommen

Einkommensrenditen beruflicher Weiterbildung: Kausal- oder Selektionseffekt? Empirische Analysen mit dem Mikrozensus-Panel 1996–1999
Felix Wolter und Jürgen Schiener (Universität Mainz)

Einkommensungleichheit von EU- und Nicht-EU-BürgerInnen in Deutschland. Ergebnisse auf Basis der Mikrozensen 1973 bis 2004
Peter Kriwy (Universität Kiel)

Frauen als Stille Reserve im Ingenieurbereich – eine ökonometrische Analyse
Eva Schlenker (Universität Hohenheim)

Hochschulabschluss gleich fester Arbeitsplatz? Neue und alte Risiken für AkademikerInnen
Katja Rackow (WZB, Berlin)

Berufliche Selbständigkeit in Deutschland: Gibt es Unterschiede zwischen den Regionen?
Dieter Bögenhold (Universität Bozen) und Uwe Fachinger (Universität Vechta)

19:00 Empfang und Buffet im Hotel Wartburg

Freitag, 16. Oktober 2009

09:00 – 11:00 Migration und Integration

Effekte der sozialen und räumlichen Einbettung auf das Unternehmertum von Migranten
Reinhard Schunck und Michael Windzio (Universität Bremen)

Neuzuwanderer auf dem deutschen Arbeitsmarkt. Analysen zur Arbeitsmarktintegration erwachsener ausländischer Zuwanderer seit Mitte der 1990er Jahre
Dietmar Hobler (Universität Göttingen)

Integrationsmessung mit dem Mikrozensus
Wolfgang Seifert (IT.NRW, Düsseldorf)

Migration und Gesundheit: Ermittlung der Rauchprävalenz bei Aussiedlern aus der ehemaligen Sowjetunion in Abhängigkeit von der Aufenthaltsdauer. Eine Untersuchung des Mikrozensus 2005
Katharina Reiss, Jacob Spallek, Doris Bardehle und Oliver Razum (Universität Bielefeld)

11:00 – 11:15 Kaffeepause

11:15 – 12:45 Wandel der Lebensformen und Partnerwahl

Zum langfristigen Wandel der Sozialstruktur partnerschaftlicher Lebensformen
Andrea Lengerer (GESIS, Mannheim)

Familiärer Wandel in Zeiten sich ändernder Erwerbsstrukturen: Ein Vergleich Deutschlands mit den USA 1973-2004
Hans Bertram, Christian Ledig und Wiebke Rösler (HU Berlin)

Regionale Ungleichheit auf dem Partnermarkt? Makrostrukturelle Rahmenbedingungen der Partnerwahl in regionaler Perspektive
Johannes Stauder (Universität Heidelberg)

12:45 – 13:45 Mittagspause

13:45 – 14:30 Datenerhebung und Datenqualität

Ergebnisse der Interviewerbefragung im Mikrozensus
Andreas Lingnau (Destatis, Bonn)

Datenqualität beruflicher Ausbildungsabschlüsse im Mikrozensus & Regionalkennung im Scientific Use File: Gemeindegrößenklasse versus BBR-Gemeindetypisierung
Robert Herter-Eschweiler (Destatis, Bonn)

14:30 – 15:00 Abschlussdiskussion

Call for Papers

XVII ISA World Congress of Sociology

ISA Research Committee on Logic and Methodology RC33

Gothenburg, Sweden
July 11 - 17, 2010

Logic and Methodology RC33

Anyone interested in presenting a paper should contact a session organizer before *December 15, 2009*. Any individual may participate on up to two sessions. Once your presentation is approved by the session chair, you must then submit an abstract of your paper on-line. Abstracts are only accepted by the system from those who are already registered for the Congress. The deadline for submission of approved abstracts is *May 1, 2010*.

Programme Coordinators: Jörg Blasius, RC33 President, University of Bonn, Germany (jblasius@uni-bonn.de); Katja Lozar Manfreda, RC33 Secretary, University of Ljubljana, Slovenia (katja.lozar@fdv.uni-lj.si).

Proposed Sessions

The proposed session invites papers which address theoretical and conceptual problems of research to deal with fundamental social change and which outline practical strategies how to deal with such phenomena. For detailed information about the sessions see: http://www.isa-sociology.org/congress2010/rc/rc33.htm.

Session 1: *Methods of social network analysis.* Organizers: Anuška Ferligoj, University of Ljubljana, Slovenia (anuska.ferligoj@fdv.uni-lj.si and Peter J. Carrington, University of Waterloo, Canada (pjc@uwaterloo.ca).

Session 2: *Analysis of large social networks.* Organizers: Anuška Ferligoj, University of Ljubljana, Slovenia (anuska.ferligoj@fdv.uni-lj.si) and Vladimir Batagelj, University of Ljubljana, Slovenia (vladimir.batagelj@fmf.uni-lj.si).

Session 3: *Assessing equivalence of constructs in a cross-cultural or over time perspective.* Organizers: Eldad Davidov, University of Cologne, Germany (Davidov@wiso.uni-koeln.de) and Peter Schmidt, University of Gießen, Germany (peter.schmidt@sowi.uni-giessen.de).

Session 4: *Data analysis strategies for cross-cultural research.* Organizers: Michael Braun, GESIS – Leibniz-Institute for

the Social Sciences, Germany (michael.braun@gesis.org) and Timothy Johnson, University of Illinois, Chicago, U.S. (timj@uic.edu).

Session 5: *Why do polls go wrong... sometimes?* Organizers: Claire Durand, Université de Montréal, Canada (Claire.Durand@umontreal.ca), John Goyder, Waterloo University, Canada and Martial Foucault, Université de Montréal, Canada (martial.foucault@umontreal.ca).

Session 6: *Experimental techniques in sociological research.* Organizer: Stefanie Eifler, University of Bielefeld, Germany (stefanie.eifler@uni-bielefeld.de).

Session 7: *Measurement error in panel surveys.* Organizers: Annette Jäckle, Institute for Social and Economic Research, University of Essex, UK (aejack@essex.ac.uk), Emanuela Sala, Institute for Social and Economic Research, University of Essex, UK (esala@essex.ac.uk) and S.C. Noah Uhrig, Institute for Social and Economic Research, University of Essex, UK (scnuhrig@essex.ac.uk).

Session 8: *Issues in the teaching of research methods in the social sciences.* Organizers: Barbara Kawulich, University of West Georgia, USA (bkawulic@westga.edu), Claire Wagner, University of Pretoria, South Africa and Mark Garner, University of Aberdeen, UK.

Session 9: *New technologies and data collection in social sciences.* Organizers: Katja Lozar Manfreda, Faculty of Social Sciences, University of Ljubljana, Slovenia (katja.lozar@fdv.uni-lj.si) and Vasja Vehovar, Faculty of Social Sciences, University of Ljubljana, Slovenia (vasja.vehovar@fdv.uni-lj.si).

Session 10: *Methodological issues in survey research.* Organizer: Ken Reed, Deakin University, Melbourne, Australia (kreed@deakin.edu.au).

Session 11: *Analysis of social change with survey data.* Organizer: Christof Wolf, GESIS – Leibniz-Institute for the Social Sciences, Germany (Christof.wolf@gesis.org).

Session 12: *Data quality in surveys among the elderly.* Organizer: Marek Fuchs, University of Kassel, Social Science Department, Germany (marek.fuchs@uni-kassel.de).

Session 13: *Transfer of socio-economic variables and of response scales in international comparable survey research.* Organizers: Juergen H.P. Hoffmeyer-Zlotnik, GESIS – Leibniz-Institute for the Social Sciences, Germany (juergen.hoffmeyer-zlotnik@gesis.org) and Dagmar Krebs, Institute for Sociology, University of Gießen, Germany (dagmar.krebs@sowi.uni-giessen.de).

Session 14: *Assessing the quality of data.* Organizer: Jörg Blasius, Department of Political Science and Sociology, University of Bonn, Germany (jblasius@uni-bonn.de).

Session 15: *Business Meeting.* Organizer: Jörg Blasius, Department of Political Science and Sociology, University of Bonn, Germany (jblasius@uni-bonn.de).

Session 16: *Reflective modeling.* Joint session of RC33 Logic and Methodology in Sociology and RC51 Sociocybernetics [host committee].

Session 17: *Methodological and conceptual issues in risk.* Joint sessions of RC33 Logic and Methodology in Sociology and TG04 Sociology of Risk and Uncertainty [host committee].

Session 18: *How to do research in a changing world?* Additional session on the Congress theme. Organizer: Jens O. Zinn, School of Social and Political Sciences, University of Melbourne, Australia (jzinn@ormond.unimelb.edu.au).

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 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) – Müller (2002: 75) – (vgl. Müller 2002: 75) – (Müller 2002; Mayer/Müller/Schulze 2003).

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