

methoden daten analysen

ZEITSCHRIFT FÜR EMPIRISCHE SOZIALFORSCHUNG

mda

2008, Jahrgang 2, Heft 1



*Ivar Krumpal, Heiko Rauhut,
Dorothea Böhr und
Elias Naumann* Wie wahrscheinlich ist ‚wahrscheinlich‘?
Zur subjektiven Einschätzung und
Kommunikation von Viktimisierungswahrscheinlichkeiten

*Isabelle Stadelmann-Steffen
und Marc Bühlmann* Space and Time in Comparative Political
Research. Pooled Time-series Cross-section
Analysis and Multilevel Designs Compared

Andreas Pöge Persönliche Codes „reloaded“

Herausgegeben von *Christof Wolf
Marek Fuchs
Bärbel Knäuper
Petra Stein*

Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung

Die Zeitschrift wird herausgegeben von der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS). Die GESIS ist eine Einrichtung der Leibniz-Gemeinschaft.

Herausgeber: Christof **Wolf** (Mannheim, geschäftsführend), Marek **Fuchs** (Kassel), Bärbel **Knäuper** (Montreal), Petra **Stein** (Duisburg-Essen)

Wissenschaftlicher

Beirat: Hans-Jürgen **Andreß** (Köln), Andreas **Diekmann** (Zürich), Sabine **Häder** (Mannheim), Udo **Kelle** (Marburg), Dagmar **Krebs** (Gießen), Frauke **Kreuter** (College Park, Maryland), Edith **de Leeuw** (Utrecht), Norbert **Schwarz** (Ann Arbor)

Redaktion: Paul Lüttinger
GESIS
Postfach 12 21 55
68072 Mannheim
Tel.: 0621 – 1246-268
E-Mail: mda@gesis.org
Internet: www.gesis.org/Publikationen/Zeitschriften/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 der GESIS:

http://www.gesis.org/Publikationen/Zeitschriften/MDA/MDA_subscribe.asp

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

ISSN 1864-6956

2. Jahrgang 2008 © GESIS, Mannheim, Juni 2008

Inhalt

FORSCHUNGSBERICHTE

- 3 Wie wahrscheinlich ist ‚wahrscheinlich‘?
Zur subjektiven Einschätzung und Kommunikation
von Viktimisierungswahrscheinlichkeiten
*Ivar Krumpal, Heiko Rauhut, Dorothea Böhr und
Elias Naumann*
- 29 Space and Time in Comparative Political Research.
Pooled Time-series Cross-section Analysis and
Multilevel Designs Compared
Isabelle Stadelmann-Steffen und Marc Bühlmann
-

PRAXISBERICHTE

- 59 Persönliche Codes „reloaded“
Andreas Pöge
-

REZENSIONEN

- 71 Die Integration qualitativer und quantitativer Methoden in
der empirischen Sozialforschung. Theoretische Grundlagen
und methodologische Konzepte. Udo Kelle, 2007
Hartmut Esser
- 75 Designing and Constructing Instruments for Social Research
and Evaluation. David Colton/Robert W. Covert, 2007
Marek Fuchs
- 80 Methoden der Politikwissenschaft. Neuere qualitative und
quantitative Analyseverfahren. Joachim Behnke/Thomas
Gschwend/Delia Schindler/Kai-Uwe Schnapp (Hg.), 2006
Bernd Weiß
- 82 Quantitative Forschung in der Sozialstrukturanalyse.
Anwendungsbeispiele aus methodischer Perspektive.
Nicole Burzan, 2008
Peter Hartmann
- 84 Transatlantische Bereicherungen. Zur Erfindung der
empirischen Sozialforschung. Christian Fleck, 2007
Anthony Oberschall

A N K Ü N D I G U N G E N

88 Call for Papers: European Labour Force Survey (EU-LFS) and
European Union Statistics on Income and Living Conditions
(EU-SILC): European User Conference

90 Hinweise für unsere Autorinnen und Autoren

Wie wahrscheinlich ist ‚wahrscheinlich‘?

How likely is 'likely'?

*Zur subjektiven
Einschätzung und
Kommunikation von
Viktimisierungswahr-
scheinlichkeiten*

*Subjective perception and
communication of
victimization probabilities*

*Ivar Krumpal, Heiko Rauhut,
Dorothea Böhr und Elias Naumann*

Zusammenfassung

Was bedeutet ‚wahrscheinlich‘ wenn Be-fragte das Risiko einschätzen, Opfer einer Straftat zu werden? Solche Risiken können einerseits als Gewinne interpretiert und dargestellt werden, indem man hofft, von Straftaten verschont zu bleiben. Andererseits können solche Risiken als Verluste interpretiert werden, indem Respondenten ihre Wahrscheinlichkeit einschätzen, Opfer von Straftaten zu werden. Wir können anhand eines experimentellen Surveys in Leipzig einen Framing-Effekt zeigen: Da ein Verlust schwerer wiegt, werden im Verlust-Frame niedrigere subjektive Viktimisierungswahrscheinlichkeiten angegeben als im Gewinn-Frame. Weiterhin variiert die Zuordnung von prozentualen Wahrscheinlichkeitsangaben zu verbalen Antwortkategorien mit der Häufigkeit eines Ereignisses: Das häufige Ereignis ist ‚von einer Straftat verschont zu bleiben‘, das seltene Ereignis ist ‚einer Straftat zum Opfer fallen‘. Wir können zeigen, dass demselben verbalen Wahrscheinlichkeitslabel (z.B. ‚unwahrscheinlich‘ oder ‚sehr wahrschein-

Abstract

What is the meaning of 'likely', when respondents estimate the risk to become victim of crime? On the one hand, such risks can be presented and interpreted as gains, in terms of being spared of becoming a victim of crime. On the other hand, such risks can be presented and interpreted as losses, in terms of becoming a victim of crime. We demonstrate a framing-effect with data from an experimental survey in Leipzig. Because losses are perceived as more severe, respondents state lower subjective victimization probabilities in the loss-frame, compared to the gain-frame. In addition, the meaning of vague quantifiers varies with the frequency of an event: The frequent event is 'being spared of a crime'; the infrequent event is 'becoming a victim of crime'. We show that the same verbal likelihood labels (e.g. 'unlikely' or 'very unlikely') are more often chosen (in terms of percentages) for frequent than for infrequent events. Finally, with increasing severity of offences, respondents assign for their likeli-

lich') beim häufigen Ereignis höhere prozentuale Wahrscheinlichkeiten zugeordnet werden als beim seltenen Ereignis. Schließlich sinken mit zunehmender Schwere einer Straftat die verbalen Antwortkategorien zugeschriebenen prozentualen Wahrscheinlichkeitswerte. Unsere Ergebnisse zeigen, dass verbale Wahrscheinlichkeitslabel (Vague Quantifiers') nicht wörtlich zu interpretieren sind. Da kategoriale Antwortskalen für verschiedene Viktimisierungen nicht einheitlich verwendet werden, ist ein direkter Vergleich von Anteilswerten über verschiedene Viktimisierungen hinweg als problematisch einzustufen.

hood estimate lower percentages to the same verbal label. Our evidence shows that vague quantifiers cannot be interpreted literally. Because respondents do not use vague quantifiers consistently for different victimizations, it is problematic to compare subjective victimization risks across different victimizations.

1 Einleitung¹

"I thought you'd never come here again," I said.

'Every time I see you, you say the same thing,' she said, laughing. As always, she sat down next to me at the bar and rested both hands on the counter. 'But I did write you a note saying I wouldn't be back for a while, didn't I?'

'For a while is a phrase whose length can't be measured. At least by the person who's waiting,' I said.

'But there must be times when the word's necessary. Situations when that's the only possible word you can use,' she said.

'And probably is a word whose weight is incalculable.'"

"South of the border, west of the sun", Haruki Murakami (2000)

Das Problem, welches der populäre japanische Schriftsteller Haruki Murakami hier in einer poetischen Sprache beschreibt, beschäftigt empirisch arbeitende Sozialwissenschaftler ebenso. Allerdings begnügt sich die Wissenschaft nicht damit, den Wert von ‚for a while‘ oder ‚probably‘ als unmessbar zu bezeichnen. Es gibt bereits einige Versuche, die Bedeutung von solchen unbestimmten verbalen Mengenbegriffen – *Vague Quantifiers* – zu quantifizieren. Simpsons (1944) Aufsatz über Häufig-

1 Wir danken C. Gross, M. Jungbauer-Gans, F. Kreuter, P. Kriwy, M. Trappmann und T. Voss für anregende Diskussionen und wertvolle Hinweise. Wir danken auch den beiden anonymen Gutachtern für ihre konstruktiven Verbesserungsvorschläge. Zudem danken wir T. Köhler und J. Scharf für ihre Unterstützung bei der Datenerhebung. Der Beitrag ist an der Universität Leipzig entstanden.

keitsausdrücke und ihre Bedeutung ist einer der ersten Ansätze. Mittlerweile ist es Konsens, dass nicht jedem verbalen Ausdruck eine feste Zahl zugeordnet werden kann, sondern dass deren Interpretation sowohl intra- als auch interindividuell variiert (vgl. Reyna 1981; Nakao/Axelrod 1983; Wright et al. 1994; Budescu/Wallsten 1995; Kuhn 1997; Karelitz/Budescu 2004).

„One man’s *,rarely*’ is another man’s *,hardly ever*“ (Hakel 1969: 533) ist ein anschauliches und wohl auch eines der meistzitierten Beispiele hierfür. Auch für die intraindividuell variierende Verwendung von *Vague Quantifiers* lässt sich ein Beispiel anführen:

„The mean response for *,sometimes*’ in the shooting in Hollywood westerns context is higher than the mean response for *,very often*’ in the context of earthquakes.“ (Bradburn 1979: 94).

In der empirischen Sozialforschung ist die Frage nach der Verwendung von *Vague Quantifiers* besonders wichtig, da diese häufig bei Antwortskalen von Standard-items in Fragebögen eingesetzt werden.²

Wir untersuchen Effekte der Frageformulierung sowie der Häufigkeit und Schwere von Delikten auf die subjektive Einschätzung und Kommunikation von Viktimisierungswahrscheinlichkeiten. Zunächst werden wir zeigen, dass bei der Beantwortung von deliktspezifischen Fragen zur eigenen subjektiven Viktimisierungswahrscheinlichkeit mittels prozentualen Wahrscheinlichkeiten Framing-Effekte auftreten. Bereits frühere Studien konnten solche Effekte zeigen (vgl. Tversky/Kahneman 1981). Wir weisen nach, dass derartige Effekte auch im Feld im Rahmen von Surveys auftreten. In unserer zweiten Analyse wird dann der Einfluss des Framings auf die Zuordnung („Mapping“) von *Vague Quantifiers* zu prozentualen Wahrscheinlichkeitswerten untersucht. Es ist festzustellen, dass jedem *Vague Quantifier* im Gewinn-Frame ein höherer prozentualer Wert zugeordnet wird als im Verlust-Frame. Dieser Effekt lässt sich hauptsächlich durch die unterschiedliche Häufigkeit der zu bewertenden Ereignisse erklären („Opfer werden‘ ist seltener als ‚nicht Opfer werden‘). Bei einem seltenen

2 „Bitte sagen Sie mir nun für jede der Situationen, die ich Ihnen gleich nenne [z. B. angepöbelt oder bedroht zu werden, zu sexuellen Handlungen gezwungen zu werden, etc.], ob Sie es für *sehr wahrscheinlich*, *wahrscheinlich*, *weniger wahrscheinlich*, oder für *unwahrscheinlich* halten, dass Ihnen persönlich innerhalb der nächsten 12 Monate einmal so etwas passiert?“ (Wohlfahrtssurvey 1998).
 „First, how likely do you think your home is to be burgled in the next year? (...) And how likely do you think you are mugged or robbed in the next year?“ Die möglichen Antwortkategorien waren hierbei: *Very likely*, *fairly likely*, *fairly unlikely* and *very unlikely* (British Crime Survey 2004/2005).

Ereignis wird beispielsweise der verbale Ausdruck ‚oft‘ mit einem kleineren numerischen Wert assoziiert als bei einem sehr häufigen Ereignis (vgl. Bradburn 1979: 94). Allerdings sind auch andere Einflussfaktoren auf das Mapping möglich, so etwa die Schwere eines Deliktes. Wir diskutieren Alternativen zur Verwendung von *Vague Quantifiers*. Wir erörtern, in welchen Forschungskontexten numerische Antwortskalen gegenüber verbalen Antwortskalen vorteilhafter sind.

Unsere Fragestellung ist zudem nicht nur für die Sozialwissenschaften und hier insbesondere für die Fragebogenkonstruktion von Interesse, sondern erhält ihre Relevanz ebenso durch ihre praktische Anwendung in der Medizin. Wie die umfangreiche Literatur aus diesem Bereich zeigt (für einen Überblick vgl. Edwards et al. 2001; Ghosh/Ghosh 2005), wird hier vor allem untersucht, welchen Einfluss der gewählte Frame bei der Kommunikation des Gesundheitsrisikos (beispielsweise die Wahrscheinlichkeit von Nebenwirkungen oder die Wahrscheinlichkeit von Krebs) auf die Entscheidung des Patienten für oder gegen eine Vorsorgeuntersuchung hat (vgl. Gurm/Litaker 2000; Welkenhuysen et al. 2001).³

2 Literaturdiskussion

2.1 *Vague Quantifiers*

Die Verwendung von *Vague Quantifiers* stellt die Vergleichbarkeit von Antworten zwischen verschiedenen Items häufig in Frage (vgl. Kreuter 2000; Schnell/Kreuter 2000; Kreuter 2002: 226-229). Dennoch werden sie, selbst wenn es Alternativen gibt, weiterhin verwendet. Warum aber benutzt man *Vague Quantifiers* und fragt nicht nach dem genauen Wert? Grundsätzlich sind *Vague Quantifiers* die natürliche Antwort auf viele Fragen (vgl. Wallsten et al. 1993). Die Frage, wie oft man in der letzten Woche glücklich war, lässt sich nur schwer mit einer genauen Angabe in Prozent beantworten. Befragte bevorzugen es, Informationen verbal und nicht numerisch zu geben (vgl. Moxey/Sanford 2000), andererseits erhält man Informationen lieber numerisch (vgl. Budescu/Wallsten 1995: 298). Werden Befragte aufgefordert, ihre Antworten quantitativ zu geben, so löst dies in der Regel zunächst

3 Je nach dem, ob man das an sich gleiche Risiko von Nebenwirkungen im Gewinn-Frame oder Verlust-Frame kommuniziert, entscheidet sich ein Patient eher gegen bzw. für eine Vorsorgeuntersuchung.

ein Nachdenken aus, nur die wenigsten verweigern jedoch eine Antwort (vgl. Bradburn 1979: 95). Präzise quantitative Schätzungen erfordern eine hohe kognitive Leistung (vgl. Peters et al. 2006: 46). So geht in der Fragesituation viel Zeit und Konzentration für die Antwortfindung verloren. Gleichzeitig besteht das Risiko, dass die genaue Zahl gar nicht vorhanden ist („Non-Attitude“), sondern erst vom Befragten situativ generiert wird (vgl. Conrad et al. 1998: 363).

Da numerische Angaben häufig eine Genauigkeit suggerieren, die für den Befragten so nicht vorliegt, erscheint die Verwendung von *Vague Quantifiers* adäquater – wenn man nur wüsste, was sie bedeuten. Wenn bei der Auswertung bekannt wäre, welchem *Vague Quantifier* welcher exakte numerische Wert zugeordnet ist, könnte man weiterhin die kognitiv einfacher zu verstehenden *Vague Quantifier* in Befragungen nutzen. Das sogenannte ‚Mapping‘ realisiert diese Zuordnung von exakten numerischen Werten zu *Vague Quantifiers* (vgl. Wright et al. 1994: 481).

Theoretische Grundlage für das Mapping ist die Annahme von ‚Membership Functions‘ (Hammerton 1976). Es wird davon ausgegangen, dass Wahrscheinlichkeitsausdrücke ungenaue Konzepte sind, und dass verschiedene numerische Wahrscheinlichkeiten zwischen 0 und 1 diese Ausdrücke repräsentieren. Die Membership Function eines verbalen Ausdrucks ordnet jedem Wert zwischen 0 und 1 eine Zahl zu, die angibt, wie stark diese Wahrscheinlichkeit mit der verbal ausgedrückten Häufigkeit korrespondiert.

„Probability phrases are vague concepts and (...) different numerical probabilities in the [0, 1] range are represented to various degrees in these concepts. (...) The membership function of any given phrase assigns a number to each value on the probability line [0, 1] that represents its degree of membership in the concept defined by the phrase.“ (Karelitz/Budescu 2004: 27).

Erste Studien nahmen an, dass jeder Ausdruck nur eine Membership Function hat. Es wurde versucht, jedem *Vague Quantifier* genau einen numerischen Wert zuzuordnen (vgl. Hammerton 1976). Relativ schnell wurde aber klar, dass die Membership Functions und damit auch das Mapping von vielen Faktoren beeinflusst werden.

Vor allem psychologische Studien heben die Wichtigkeit des Kontextes hervor. Goocher (1965) betont, dass es eine Rolle spielt, ob man das Ereignis oder das Objekt mag oder nicht. Je mehr die Respondenten ein Ereignis oder ein Objekt mochten, desto kleinere Häufigkeitsausdrücke wurden verwendet. Daran anknüpfend zeigen Pepper/Prytulak (1974) sowie Moxey/Sanford (1993, 2000) den Einfluss der erwarteten Häufigkeit:

„In particular, if an event has a high base-rate expectation, such as people enjoying parties, then the values assigned to (say) *many* in *many* people enjoyed the party is higher than it is for a low baserate expectation (as in *many* of the doctors in the hospital were female).“ (Moxey/Sanford 2000: 241).

Schwarz et al. (1985) verweisen auf den Einfluss der Skalierung der Antwortvorgaben. Wenn man die niedrigeren Kategorien einer Antwortskala ausdifferenziert, geben die Befragten einen niedrigeren Fernsehkonsum an. Im höheren Bereich ausdifferenzierte Antwortkategorien ergeben dagegen höhere Häufigkeitsangaben des abgefragten Verhaltens.⁴ Hörmann (1983) und Newstead (2000) wiederum beschäftigen sich mit der Variation, die durch die veränderte Größe von Objekten ausgelöst wird.

„[...] *a few* people in front of a building meant more than *a few* people standing in front of a hut“ (Newstead 2000: 244).

Neben individuellen Variationen der Assoziationen von *Vague Quantifier* mit numerischen Angaben bzw. neben Manipulationen, die alle Befragten im gleichen Maße betreffen, ist es für die sozialwissenschaftliche Forschung problematischer, wenn Unterschiede im Verständnis von *Vague Quantifiers* gruppenspezifisch auftreten. Festgestellte Gruppenunterschiede beruhen dann möglicherweise nicht auf tatsächlichen Unterschieden, sondern auf einer unterschiedlichen Interpretation von *Vague Quantifiers*. Eine Studie von Schaeffer (1991) verdeutlicht dies: Sie befragte schwarze und weiße Teilnehmer nach dem Ausmaß ihrer Langeweile. Wurden numerische Werte verwendet, waren keine Unterschiede festzustellen. Bei der Verwendung von *Vague Quantifiers* gaben schwarze Teilnehmer allerdings an, häufiger gelangweilt zu sein. Nakao/Axelrod (1983) weisen Gruppenunterschiede zwischen Ärzten und Laien im Gebrauch von verbalen Häufigkeitsausdrücken wie ‚infrequent‘ oder ‚not infrequently‘ nach. King et al. (2004) zeigen, dass die Verwendung von Antwortskalen interkulturell variiert.

Wänke (2002) zeigt, dass es wichtig ist, mit welcher Referenzgruppe man sich vergleicht – die gesamte Bevölkerung oder die eigene soziale Gruppe. So schätzen Studenten, die sich mit der gesamten Bevölkerung vergleichen, ihre Kino-

4 Neben der Frage nach dem Fernsehkonsum konnten vergleichbare Effekte auch bei der Abfrage anderer Verhaltensweisen, wie etwa sexuelles Verhalten oder Konsumverhalten, festgestellt werden (vgl. Schwarz/Scheuring 1988; Menon et al. 1995). Ähnliche Effekte konnten auch bei der Frage nach Häufigkeiten körperlicher Symptome bei Patienten beobachtet werden (vgl. Schwarz/Scheuring 1992).

frequenz höher ein als Studenten, die sich mit anderen Studenten vergleichen. Auch spielt es eine Rolle welches Referenzverhalten man wählt:

„For example, when judging how often one goes to the movies, the response may depend on whether one compares this behaviour to the frequency one attends poetry readings or watches TV.“ (Wänke 2002: 302).

Wright et al. (1994) fassen die unterschiedlichen Faktoren, welche das Mapping beeinflussen, zu zwei Kategorien zusammen. Zum einen hängt das Mapping von Informationen ab, die der eigenen Erfahrung entspringen. Sollen den unbestimmten Adjektiven ‚jung‘, ‚alt‘, ‚sehr alt‘ ... numerische Altersangaben zugeordnet werden, dann spielt das eigene Alter eine Rolle. Diese Annahme wird als ‚Self-Information-Hypothese‘ bezeichnet. Zum anderen wird das Mapping durch Eigenschaften (Verhalten, Sprachgebrauch etc.) der sozialen Gruppe bestimmt, zu der man gehört. Ein Beispiel, das einen dieser Aspekte verdeutlicht, findet sich ebenfalls bei Wright et al. (1994): In zwei isolierten Dörfern werden die Bewohner befragt, wie groß sie seien. Als Antwortmöglichkeiten sind *Vague Quantifiers* vorgegeben. Angenommen, in dem einen Dorf sind alle Personen 1,80 m groß und im anderen Dorf sind alle Personen 1,90 m groß. Soll der Befragte nun seine Größe einschätzen, so denkt er an die anderen in seinem Dorf und kommt zu der Annahme, er sei normal groß. Entsprechend ordnet er sich in der Mitte der vorgegebenen Skala ein. Bei der Erfassung der Größe mittels *Vague Quantifiers* kommt man in diesem Fall also zu dem Ergebnis, dass die Bewohner beider Dörfer genau gleich groß sind – was aber nicht der Fall ist. Die Annahme, dass die Zugehörigkeit zu einer sozialen Gruppe das Mapping beeinflusst, wird als ‚Group-Norm-Hypothese‘ bezeichnet. Wright et al. (1994) bestätigen in ihrer Studie zum Fernsehkonsum⁵ die Gültigkeit beider Hypothesen, regen aber an, dass weitere Untersuchungen zu Gruppenunterschieden nötig seien, um ihre Schlussfolgerungen zu untermauern.

2.2 Effekte des Framings

Die subjektive Einschätzung von Wahrscheinlichkeiten hängt davon ab, wie solche Risiken präsentiert werden. Tversky/Kahneman (1981) beschreiben in ihrer mittlerweile klassischen Studie zum Framing von Entscheidungen zwei unterschiedliche

5 Wenn erst nach dem Fernsehkonsum des ‚typisch Anderen‘ gefragt wird, geben Respondenten in der folgenden Frage nach dem eigenen Konsum mehr Stunden an, als wenn sie zuerst nach ihrem eigenen Konsum gefragt werden.

Szenarien der Seuchenbekämpfung, ein sicheres und eine risikoreiches. Einmal wird das Ergebnis mit der Anzahl der Überlebenden beschrieben, das andere Mal mit der Anzahl derer, die sterben. Obwohl das Entscheidungsproblem in beiden Formulierungen den gleichen statistischen Erwartungswert hat, und somit formal äquivalent ist, entscheiden sich die Befragten für die risikoreiche Variante, wenn sie mit dem Verlust-Frame (Anzahl der Toten) konfrontiert werden, und für die sichere Variante, wenn sie mit dem Gewinn-Frame konfrontiert werden.

Laut Prospect Theorie (Kahneman/Tversky 1979) erzeugt die Darstellung eines Entscheidungsproblems unter Unsicherheit in Form von Gewinnen im Vergleich zu Verlusten eine Verschiebung des Referenzpunktes, von dem aus die Befragten ihre Evaluation durchführen. So ist der Referenzpunkt ein anderer, wenn man bei einem Risiko einerseits den drohenden Verlust in den Vordergrund stellt oder andererseits den Gewinn, von diesem Risiko verschont zu bleiben. Die Annahme der Prospect Theorie ist, dass erstens die Bewertungsfunktion für Gewinne konkav und für Verluste konvex verläuft und zweitens, dass der Anstieg der Bewertungsfunktion für Verluste steiler ist als für Gewinne. Diese Eigenschaften führen dazu, dass die gleiche Wahrscheinlichkeit des Eintretens eines bestimmten Ereignisses im Verlust-Frame stärker psychologisch wirksam ist als im Gewinn-Frame. In anderen Worten: zu 5% einen Verlust zu erleiden wiegt stärker, als zu 95% von diesem Verlust verschont zu bleiben.

McNeil et al. (1982) konnten diesen Effekt im klinischen Bereich bestätigen. Patienten hatten zu entscheiden, ob sie eine sicherere Strahlentherapie oder eine risikoreichere Operation bevorzugen. Einmal wurden die Risiken beider Behandlungen in Form von Sterberisiken, einmal in Form von Überlebenswahrscheinlichkeiten dargestellt. Wie erwartet zeigte es sich, dass die Darstellung in Form von Sterberisiken weniger Befragte dazu bewegte, sich für die risikoreichere Operation zu entscheiden, als dies bei der Darstellung in Form von Überlebenswahrscheinlichkeiten der Fall war. Für eine Übersicht dieser Darstellungseffekte in der Medizin siehe Marteau (1989), Banks et al. (1995) und Edwards et al. (2001).

In eine ähnliche Richtung für den Bereich von Konsumententscheidungen gehen die Überlegungen von Moxey/Sanford (2000). Den Befragten werden zwei logisch äquivalente Möglichkeiten vorgegeben, zwischen denen sie sich entscheiden sollen, beispielsweise einem Joghurt, der 95% fettfrei ist und einem, der 5% Fett enthält. Die Entscheidung fällt meist auf den 95% fettfreien Joghurt. Als entscheidendes Kriterium für diese Entscheidung betonen sie die Perspektive, die durch die Formulierung gesetzt wird – das eine Mal wird der positive Aspekt (fettfrei) betont, das andere Mal wird die Aufmerksamkeit auf den negativen Aspekt

gelenkt. Im ‚Joghurtproblem‘ ist die Bevorzugung des positiven Frames (95% fettfrei) gegenüber dem negativen Frame (5% Fett) auf eine innerlich höhere Bewertung im Gewinn-Frame zurückzuführen. Man müsste folglich im Gewinn-Frame den Fettgehalt des Joghurts heraufsetzen (z. B. 90% fettfrei), um eine subjektive Äquivalenz mit der Darstellung im Verlust-Frame (5% Fett) zu erhalten. In diesem Fall wäre die Versuchsperson indifferent zwischen den beiden Formulierungen.

Das Ergebnis, dass logisch äquivalente Ausdrücke unterschiedlich bewertet werden, wird ebenso von einer früheren Studie von Reyna (1981) bestätigt. Sie beschäftigt sich mit Modaladjektiven (wie ‚wahrscheinlich‘ oder ‚möglichlicherweise‘) und der Wirkung, die Verneinungen haben. Dabei stellt sie fest, dass Verneinungen nicht automatisch zu komplementären numerischen Schätzungen führen.

3 Subjektive Einschätzung von Viktimisierungswahrscheinlichkeiten

3.1 Hypothesen

Wenn wir die bisherigen Analysen auf subjektive Viktimisierungswahrscheinlichkeiten übertragen, ist Folgendes zu erwarten: Es wird vorgezogen ‚zu 95% von einem Verbrechen verschont zu bleiben‘ anstelle ‚zu 5% Opfer zu werden‘. Der Optimismus im Gewinn-Frame würde dabei auf einer, im Vergleich zum Verlust-Frame, niedrigeren subjektiven Viktimisierungswahrscheinlichkeit beruhen. Folglich wird erwartet, dass die Wahrscheinlichkeit, verschont zu bleiben (Gewinn-Frame) herabgesetzt wird, damit sie mit der Wahrscheinlichkeit, Opfer zu werden, subjektiv korrespondiert. Ein Beispiel wäre, dass ‚zu 90% verschont bleiben‘ als äquivalent aufgefasst wird wie ‚zu 5% Opfer werden‘. Unsere erste Hypothese lautet damit wie folgt: Im Gewinn-Frame (‚verschont werden‘) erwarten wir höhere subjektive Viktimisierungswahrscheinlichkeiten als im Verlust-Frame (‚Opfer werden‘).

3.2 Methode

Im Rahmen eines experimentellen Surveys zum Thema ‚Kriminalitätsfurcht‘, befragten wir im August 2006 in Leipzig 95 Taxifahrerinnen bzw. Taxifahrer und 96 Kellnerinnen bzw. Kellner (N=191). Erhoben wurden u. a. Kriminalitätsfurcht, Opferwerdung, subjektive Kosten einer Viktimisierung und Wehrhaftigkeit. Außerdem wurden die Befragten gebeten, ihre Wahrscheinlichkeit einzuschätzen, dass

sie selbst Opfer eines Verbrechens werden. Insgesamt befragten wir zu fünf Delikten, die wir in kurzen Szenarien vorstellten: Verkehrsunfall, Zechprellerei, Diebstahl, Raub, Mord. Wir wählten zwei Gruppen von Personen, die ihren Beruf in der Öffentlichkeit ausüben und somit einem höheren Viktimisierungsrisiko ausgesetzt sind. Die Ziehung der Probanden erfolgte mittels einer zweistufigen Klumpenstichprobe, wobei Taxistände bzw. Gastronomiebetriebe jeweils die Klumpen bildeten, in denen in einem zweiten Schritt eine oder mehrere Personen befragt wurden: Zur Auswahl der Taxifahrerinnen und Taxifahrer verwendeten wir eine Liste aller Taxistände der Taxigenossenschaft, aus der wir zunächst zufällig Taxistände zogen. An den einzelnen Ständen fingten wir jeweils mit der Fahrerin bzw. dem Fahrer des zweiten Taxis in der Reihe an und befragten weiter bei jedem zweiten Wagen. Die Anzahl der realisierten Interviews pro Taxistand variierte, je nachdem, wie frequentiert der Taxistand war. Zum anderen lag uns eine Liste aller eingetragenen Gastronomiebetriebe der IHK Leipzig vor, aus der wir eine Zufallsstichprobe zogen. In den einzelnen Betrieben gingen wir nach der Geburtstagsauswahl („last-birthday-method“) vor. Die Analyse beruht also auf einer direkten Erhebung aus dem Feld, die jederzeit reproduziert werden kann.

Es wurden von vier Interviewern mündliche Face-to-face Interviews mit einem standardisierten Fragebogen durchgeführt. Dabei gab es zwei Versionen mit unterschiedlichen Frames. Die Zufallsallokation zu den beiden Fragebogenversionen erfolgte mittels Zufallszahlen. Die Randomisierung fand sowohl innerhalb der beiden Berufsgruppen als auch innerhalb der Interviewer statt. Jeweils die Hälfte der TaxifahrerInnen und KellnerInnen fragten wir nach ihrer Schätzung, Opfer eines Verbrechens zu werden (Verlust-Frame), die andere Hälfte beider Berufsgruppen, für wie wahrscheinlich sie es halten, von einem bestimmten Verbrechen verschont zu bleiben (Gewinn-Frame). So wurden 41 KellnerInnen und 49 TaxifahrerInnen im Verlust-Frame und 55 KellnerInnen und 46 TaxifahrerInnen im Gewinn-Frame befragt. Der Fragebogen beinhaltete außerdem eine Kontrollfrage zum Delikt Raub (intraindividuelle Variation des Frames). So bekam die eine Hälfte der Befragten die Frage nach der Viktimisierungswahrscheinlichkeit zweimal im Verlust-Frame gestellt, die andere Hälfte der Befragten sollte erst im Verlust- und dann im Gewinn-Frame antworten. Die Befragten sollten ihre Einschätzung zuerst auf einer verbalen Skala von *Vague Quantifiers* und direkt anschließend noch einmal als numerische Prozentangabe angeben:

Tabelle 1 Die Fragen des Fragebogens

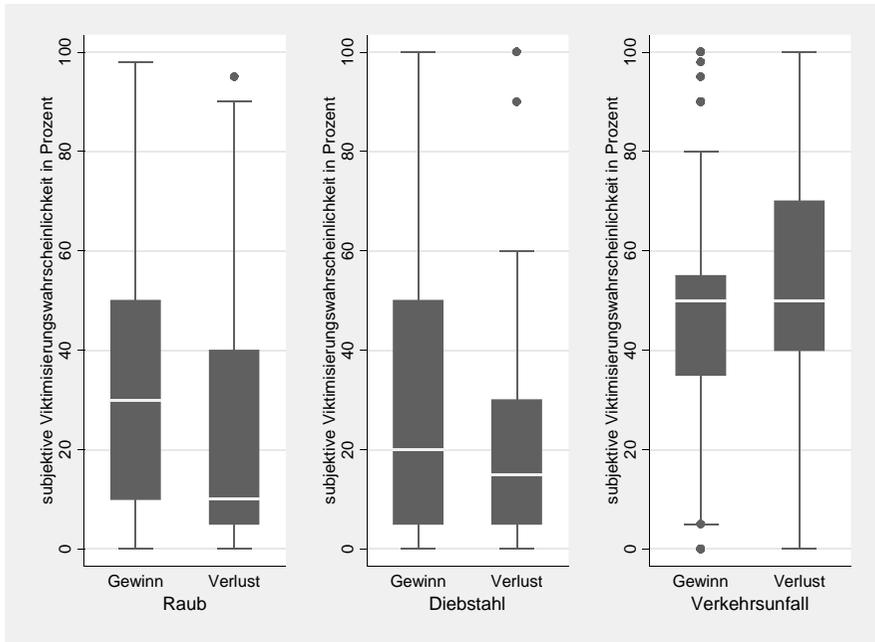
Frage 1:	<p>Ich möchte als Erstes gern über Ihre persönliche Gefährdung sprechen. Dazu nenne ich Ihnen verschiedene Situationen und würde Sie bitten, mir anhand dieser Karte zu sagen, für wie wahrscheinlich Sie es halten, dass Sie in den nächsten 12 Monaten in eine solche Situation geraten.</p> <p>Für wie wahrscheinlich halten Sie es, dass in den nächsten 12 Monaten ...</p> <p>... ein Kunde abhaut ohne zu zahlen?</p> <p>... Ihnen jemand Gewalt androht, um an Ihr Geld oder Ihre Wertgegenstände zu kommen?</p> <p>... Sie jemand umbringt?</p>
<hr/>	
Frage 9:	
<i>Version 1</i>	<p>Für wie wahrscheinlich halten Sie es, in den nächsten 12 Monaten von folgender Situation ... verschont zu bleiben? (...) Für wie wahrscheinlich halten Sie es, dass Ihnen folgendes <u>nicht</u> passiert:</p>
<i>Version 2</i>	<p>Für wie wahrscheinlich halten Sie es, in den nächsten 12 Monaten in folgende Situation zu geraten? (...) Für wie wahrscheinlich halten Sie es, dass Ihnen folgendes passiert:</p> <p>... Ein Kunde stiehlt Ihr Portemonnaie und verschwindet.</p> <p>... Jemand droht Ihnen Gewalt an, um an Ihr Geld oder Ihre Wertgegenstände zu kommen.</p> <p>... Ein anderer Verkehrsteilnehmer nimmt Ihnen die Vorfahrt und es kommt zu einem Zusammenstoß.</p>
<hr/>	
Antwort:	<p>1 – auf keinen Fall, 2 – sehr unwahrscheinlich, 3 – unwahrscheinlich, 4 – teils/teils, 5 – wahrscheinlich, 6 – sehr wahrscheinlich</p> <p>Geben Sie dazu bitte auch noch die Wahrscheinlichkeit in Prozent an.</p>

Die Fragen wurden nicht direkt nacheinander gestellt, es lagen etwa fünf Minuten Zeit zwischen der Beantwortung.

3.3 Ergebnisse

Zunächst betrachten wir nur die Ergebnisse auf der prozentualen Skala und überprüfen den Effekt des Framings auf die subjektive Viktimisierungswahrscheinlichkeit in der numerischen Einschätzung. Mit Boxplots vergleichen wir die Verteilungen pro Frame für die Delikte Diebstahl, Raub und Verkehrsunfall.

Abbildung 1 Boxplots der Verteilungen der subjektiven Viktimisierungswahrscheinlichkeit in beiden Frames pro Delikt. Im Gewinn-Frame wird die Gegenwahrscheinlichkeit dargestellt.



Die Daten zeigen, dass bei den Delikten ‚Raub‘ und ‚Diebstahl‘ wie erwartet eine höhere subjektive Viktimisierungswahrscheinlichkeit im Gewinn-Frame angegeben wurde. So liegen bei beiden Delikten der Median der empirischen Verteilungen im Gewinn-Frame über dem Median des Verlust-Frames. Im Delikt ‚Verkehrsunfall‘ liegen die beiden Mediane dagegen auf gleicher Höhe. Für einen besseren Vergleich der Wahrscheinlichkeiten betrachten wir die Mittelwerte (vgl. Tabelle 2). Die mittleren subjektiven Viktimisierungswahrscheinlichkeiten sind bei den Delikten ‚Raub‘ und ‚Diebstahl‘ in den beiden Frames unterschiedlich. Die Wahrscheinlichkeit verschont zu bleiben wird, wie theoretisch vermutet, herabgesetzt. Die subjektive Viktimisierungswahrscheinlichkeit im Gewinn-Frame wird somit höher eingeschätzt. So beträgt beim Delikt ‚Raub‘ der Mittelwert der eigenen subjektiven Wahrscheinlichkeit 35,5% im Gewinn-Frame (‚Verschont-Bleiben‘). Demgegenüber wird im Verlust-Frame (‚Opfer-Werden‘) eine geringere mittlere subjektive Viktimisierungswahrscheinlichkeit von 21,9% angegeben ($p < 0,01$). Auch beim Delikt ‚Diebstahl‘ ist die Differenz signifikant in Richtung unserer Hypothese ($p < 0,01$). In der folgenden Tabelle sind alle Mittelwerte aufgelistet.

Tabelle 2 Arithmetische Mittel der subjektiven Viktimisierungswahrscheinlichkeit (in Klammern: N)

	Raub	Diebstahl	Verkehrsunfall
Verlust-Frame	21,9 (87)	20,6 (87)	51,4 (85)
Gewinn-Frame (100-p)	35,5 (99)	31,0 (99)	48,8 (96)
p-Wert (einseitiger t-Test)	0,0003	0,0042	0,7560

Im Delikt ‚Verkehrsunfall‘ liegen dagegen die beiden Wahrscheinlichkeiten sehr nahe beieinander (51,4% im Verlust-Frame versus 48,8% im Gewinn-Frame).⁶ Wir können mit den beiden Delikten ‚Raub‘ und ‚Diebstahl‘ im Sinne unserer Hypothese zeigen, dass es einen Einfluss des Framings in der Fragestellung auf die numerischen Angaben in der Antwort gibt.

4 Kommunikation von Viktimisierungswahrscheinlichkeiten

4.1 Hypothesen

Während in der ersten Analyse gezeigt wurde, dass die subjektive Einschätzung von Viktimisierungswahrscheinlichkeiten durch die Frageformulierung beeinflusst wird, untersuchen wir in einer weiteren Analyse, inwiefern subjektive Viktimisierungswahrscheinlichkeiten je nach Situation unterschiedlich kommuniziert werden. Eine unterschiedliche Kommunikation von Risiken drückt sich darin aus, dass die Ausdrücke ‚wahrscheinlich‘ oder ‚unwahrscheinlich‘ in verschiedenen Situationen unterschiedlich verwendet werden. Hierzu analysieren wir die Zuordnung dieser *Vague Quantifiers* zu prozentualen Wahrscheinlichkeitswerten.⁷ Dieses Vorgehen wird als ‚Mapping‘ bezeichnet (vgl. Wright et al. 1994). Wir demonstrieren mit zwei Mapping-Analysen, dass verbale Wahrscheinlichkeitsausdrücke zu Viktimisierungsrisiken je nach Situation eine unterschiedliche Bedeutung haben. *Erstens* vergleichen wir

6 Dies kann mitunter darauf zurückzuführen sein, dass es sich bei einem Verkehrsunfall um ein relativ häufiges Ereignis handelt.

7 Einen Vergleich von verbalen und prozentualen Antwortskalen zu subjektiven Entdeckungswahrscheinlichkeiten bei selbstberichteter Delinquenz anhand von ALLBUS-Indikatoren der Jahre 1990 und 2000 führte unlängst Reuband (2002) durch.

die Interpretation einer gegebenen verbalen Viktimisierungswahrscheinlichkeit (z. B. ‚wahrscheinlich‘) im Gewinn-Frame mit derjenigen im Verlust-Frame. Hierbei wird über das Framing die Häufigkeit des zu bewertenden Ereignisses randomisiert: Das Ereignis ‚Opfer werden‘ ist seltener als das Ereignis ‚nicht Opfer werden‘. Es wird angenommen, dass bei seltenen Ereignissen (z. B. ‚Raub‘) eine gegebene verbale Antwortkategorie mit kleineren numerischen Wahrscheinlichkeiten in Verbindung gebracht wird als bei häufigen Ereignissen (z. B. ‚nicht Raub‘).⁸ Die zu testende Hypothese lautet somit wie folgt: Im Gewinn-Frame („nicht Raub“) werden demselben verbalen Antwortlabel (z. B. ‚sehr wahrscheinlich‘) höhere prozentuale Wahrscheinlichkeiten zugeordnet als im Verlust-Frame („Raub“). *Zweitens* analysieren wir, ob mit zunehmender Schwere eines Deliktes, die prozentualen Wahrscheinlichkeiten, die einem gegebenen verbalen Antwortlabel zugeschrieben werden, sinken. So ist etwa zu erwarten, dass Befragte mit einer ‚sehr wahrscheinlichen‘ Viktimisierung beim Delikt ‚Mord‘ ein geringeres Risiko kommunizieren möchten als bei einer ‚sehr wahrscheinlichen‘ Viktimisierung beim Delikt ‚Zechprellerei‘.⁹

8 „In short ‚often‘ for an improbable event is less than often for a highly probable event.“ (Bradburn 1979: 94). Der Häufigkeitseffekt kann anhand der Konversationslogik von Grice (1993) begründet werden. Hierbei wird eine Orientierung des Befragten am Relevanzprinzip und an der Kategorie der Relation angenommen. Wenn der Befragte die Wahrscheinlichkeit einschätzt, im nächsten Jahr Opfer eines Raubs zu werden, und die entsprechende Frage in Form einer absoluten Wahrscheinlichkeit beantwortet, ist die vermutete Antwort ‚sehr unwahrscheinlich‘. Da andere Befragte eine ähnliche Antwort geben werden, wäre diese Information nicht relevant. Den Überlegungen von Grice folgend vermutet der Befragte vielmehr, dass die Wahrscheinlichkeit in Relation zu allen anderen Befragten interessiert. Deshalb wird er die Antwortskala an der vermuteten durchschnittlichen Wahrscheinlichkeit verankern und sich entsprechend einordnen. Für eine detaillierte Beschreibung der „anchoring and adjustment heuristic“ vgl. Peters et al. 2006. Dieser Antwortprozess ist zunächst unabhängig von der Häufigkeit und findet Anwendung sowohl für numerische, als auch für verbale Skalen. Bei der Verwendung der verbalen Skala stehen jedoch qualitative Überlegungen im Vordergrund (vgl. Zimmer 1983; Budescu/Wallsten 1995: 303). Sowohl beim häufigen („nicht Raub“) als auch beim seltenen Ereignis („Raub“) ist eine relevante und relationale Antwort möglich, wenn die eigene Wahrscheinlichkeit in Relation zur vermuteten durchschnittlichen Wahrscheinlichkeit der anderen evaluiert, und diese durchschnittliche Wahrscheinlichkeit mit der Mittelposition der Antwortskala als Anker gleichgesetzt wird. Bei der numerischen Skala stehen hingegen in erster Linie quantitative Überlegungen im Vordergrund. So wird bei häufigen Ereignissen („nicht Raub“) der Anker höher gesetzt als bei seltenen Ereignissen („Raub“). Dies begründet den Häufigkeitseffekt beim Mapping.

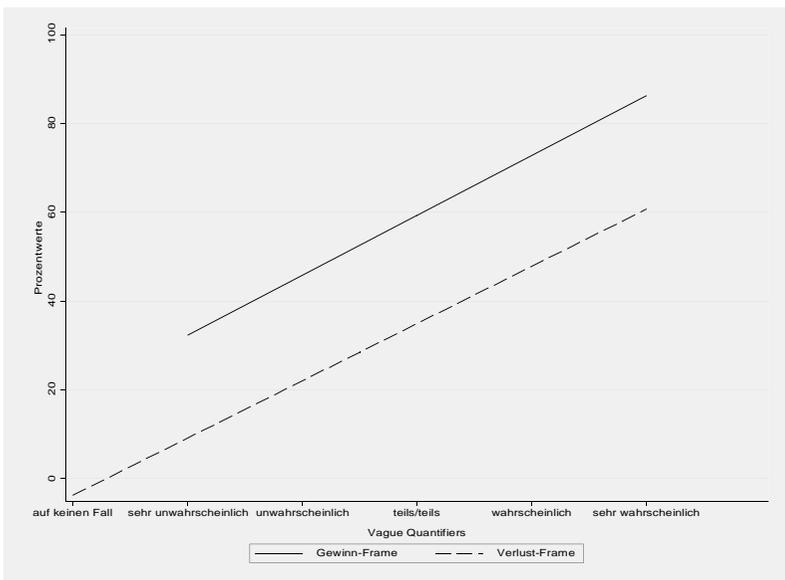
9 Die beiden Dimensionen ‚Häufigkeit‘ und ‚Schwere‘ eines Deliktes konnten hier nicht unabhängig voneinander manipuliert werden. Um eine mögliche Konfundierung in zukünftigen Studien auszuschließen, sollten Anwendungen entwickelt werden, bei denen dies möglich ist. Eine denkbare Möglichkeit in der Medizin wäre, bei der Untersuchung unterschiedlicher Krebskrankungen ‚Häufigkeit‘ (Erkrankungsrisiko) und ‚Schwere‘ (Sterberisiko) unabhängig voneinander zu manipulieren. Zu einer vertieften Diskussion dieses Sachverhaltes im medizinischen Kontext vgl. Weber/Hilton (1990); Merz et al. (1991); Sutherland et al. (1991); Budescu/Wallsten (1995: 294f.).

4.2 Einfluss der Häufigkeit eines Ereignisses

Wir rechnen sechs einfache lineare OLS-Regressionen. Für die Delikte ‚Raub‘, ‚Diebstahl‘ und ‚Verkehrsunfall‘ (vgl. Tabelle 1, Frage 9) regressieren wir die Wahrscheinlichkeitsangaben der verbalen Antwortskala auf die zugehörigen Prozentangaben. Wir berechnen jeweils drei Regressionen für den Gewinn-Frame und drei Regressionen für den Verlust-Frame. Die daraus resultierenden Response Functions werden für jedes Delikt graphisch abgebildet (vgl. Wright et al. 1994).¹⁰ Auf der X-Achse werden die Werte der verbalen Antwortskala und auf der Y-Achse die Werte der prozentualen Wahrscheinlichkeitsangaben abgetragen.

Unsere Vorhersage, wonach den *Vague Quantifiers* im Gewinn-Frame (also dem häufigeren Ereignis des Nicht-Eintretens einer Opferwerdung) höhere prozentuale Wahrscheinlichkeiten zugeschrieben werden, lässt sich bestätigen (vgl. Abbildungen 2-4).

Abbildung 2 Inter-personelles Framing, Delikt: Raub (Frage 9)



10 Wright et al. (1994) berechnen in ihrer Studie zum Mapping von *Vague Quantifiers* ebenfalls OLS-Regressionen und interpretieren die daraus gebildeten Regressionsgeraden als ‚Response Functions‘.

Abbildung 3 Inter-personelles Framing, Delikt: Diebstahl (Frage 9)

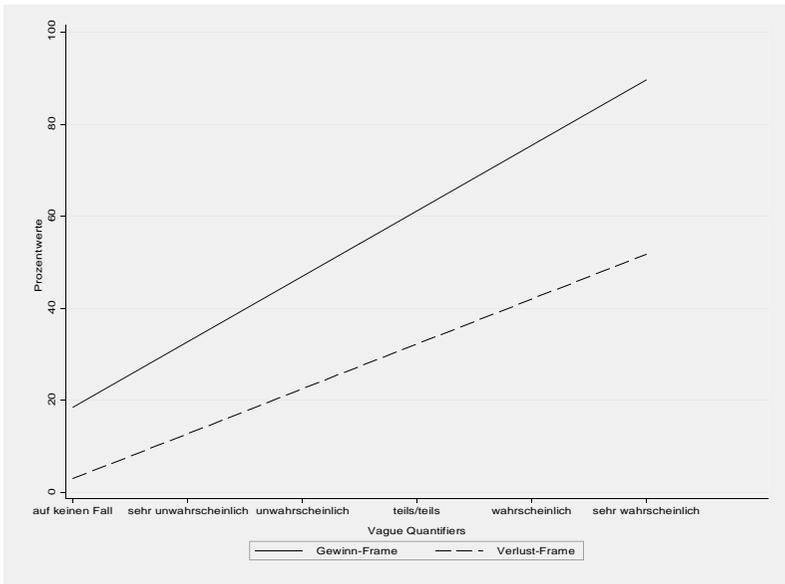
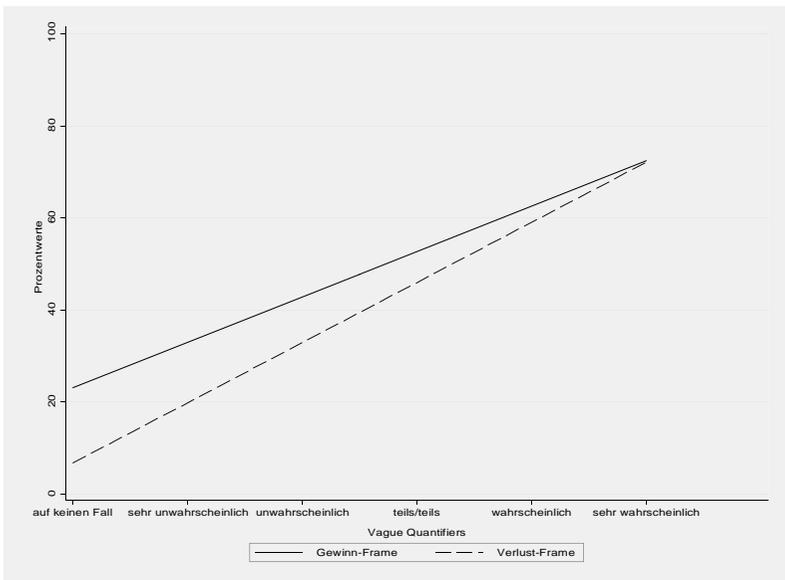


Abbildung 4 Inter-personelles Framing, Delikt: Verkehrsunfall (Frage 9)



Die in den Grafiken abgebildeten Response Functions der Befragten des Gewinn-Frames liegen bei allen Delikten über den Response Functions des Verlust-Frames. Bei den schwereren Delikten ‚Diebstahl‘ und ‚Raub‘ beträgt der Abstand zwischen den beiden Response Functions deutliche 20–25%. Somit liegt bei diesen beiden Delikten ein insgesamt recht starker Effekt im Sinne unserer Hypothese vor. Die Response Functions beim Delikt ‚Verkehrsunfall‘ liegen dagegen näher beieinander.¹¹

Ein noch stärkerer Effekt zeigt sich beim *intra-personellen Test*. Dabei schätzte dieselbe Person ihre subjektive Wahrscheinlichkeit, Opfer eines Raubs zu werden, jeweils zweimal ein: Durch Randomisierung wurden 2 Gruppen gebildet. Dabei beantworteten die Befragten in der ersten Gruppe zunächst die Frage nach der Wahrscheinlichkeit ‚ausgeraubt zu werden‘ (Frage 1, Verlust-Frame). Nach einigen weiteren Fragen wurden sie schließlich aufgefordert, die Frage nach ihrer subjektiven Wahrscheinlichkeit ‚nicht ausgeraubt zu werden‘ zu beantworten (Frage 9, Gewinn-Frame). Die Befragten der zweiten Gruppe schätzten dagegen die subjektive Wahrscheinlichkeit ‚ausgeraubt zu werden‘ zweimal ein (Frage 1 und 9 jeweils im Verlust-Frame). Die Ergebnisse dieser Analysen sind in Abbildung 5 und 6 visualisiert.

Im *intra-personellen Test* bewertete dieselbe Person die *Vague Quantifiers* im Gewinn-Frame höher als zuvor im Verlust-Frame. Das bestärkt die Ergebnisse des *inter-personellen Tests*. Dagegen gaben Respondenten in der Kontrollkondition, die zweimal dieselbe Frage beantworteten, relativ konsistente Antworten.

Gemäß unseren Erwartungen halten wir fest, dass bei häufigen Ereignissen den *Vague Quantifiers* höhere Prozentwerte zugeordnet werden. Umgekehrt werden bei seltenen Ereignissen geringere Werte auf der prozentualen Skala angegeben. Die Response Functions verschieben sich somit mit steigender Häufigkeit nach oben.

11 Eine detaillierte Darstellung der Kennwerte der OLS-Regressionen (Regressionskoeffizienten, Standardfehler, etc.) kann bei Bedarf von den Autoren angefordert werden.

Abbildung 5 Intra-personelles Framing, Delikt: Raub
(Frage 1 im Verlust-Frame, Frage 9 im Gewinn-Frame)

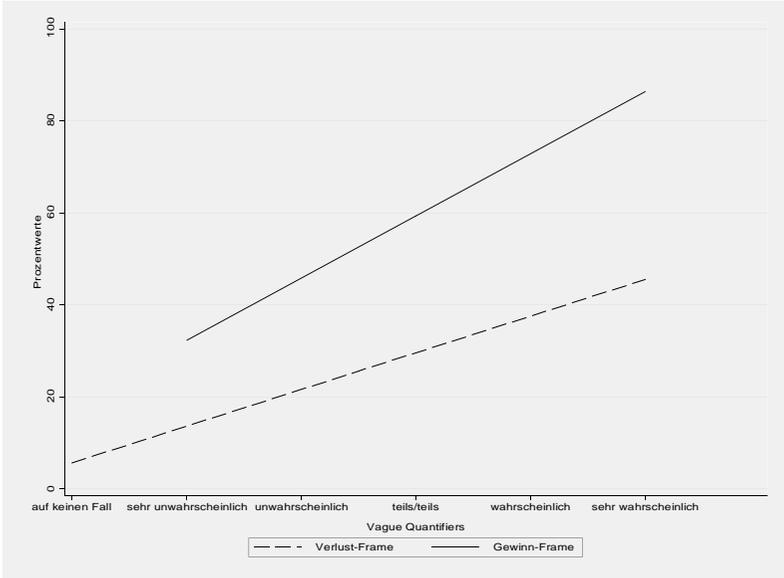
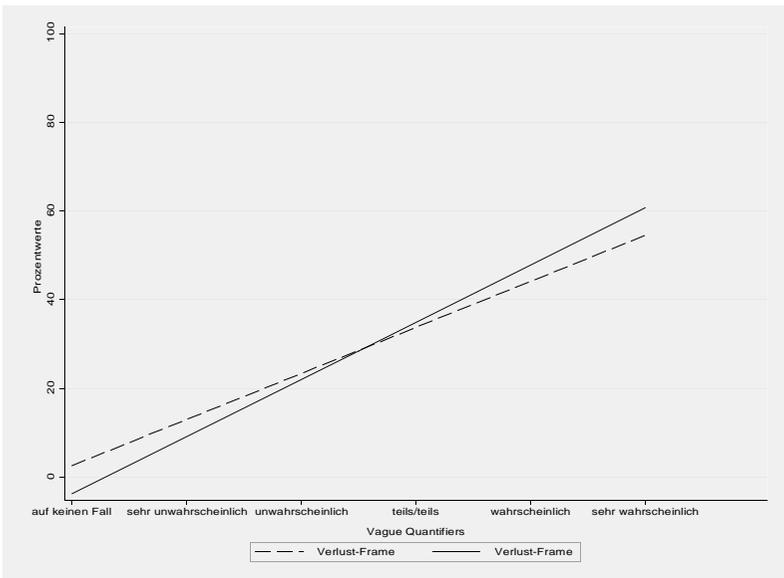


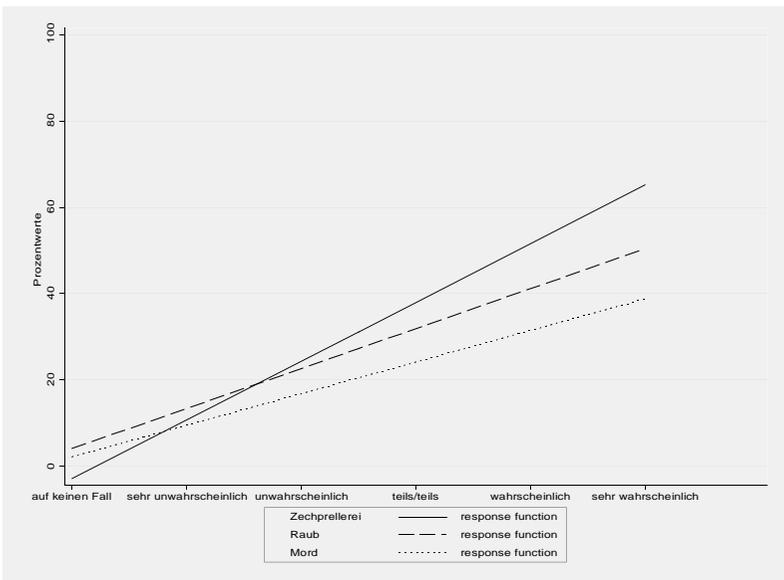
Abbildung 6 Intra-personelles Framing, Delikt: Raub
(Kontrollkondition: Frage 1 und 9 im Verlust-Frame)



4.3 Einfluss der Schwere einer Straftat

Hinsichtlich des Einflusses der Schwere einer Straftat auf das Mapping der verbalen mit der numerischen Wahrscheinlichkeitsskala bestätigt sich unsere Prognose: Mit zunehmender Deliktschwere (aufsteigend: Zechprellerei, Raub, Mord) sinken die prozentualen Wahrscheinlichkeiten, die einem gegebenen *Vague Quantifier* zugeschrieben werden (vgl. Abbildung 7).

Abbildung 7 Deliktspezifisches Mapping: Zechprellerei, Raub, Mord (Frage 1)



Aus Abbildung 7 wird ersichtlich, dass mit zunehmender Schwere eines Delikts die Steigungen der Response Functions abnehmen.¹² Vergleicht man die Wahrscheinlichkeiten zwischen verschiedenen Delikten, so wird deutlich, dass die verbalen Antwortkategorien nicht einheitlich verwendet werden. Unter ‚sehr wahrscheinlich‘ ermordet zu werden subsumieren Befragte eine kleinere Wahrscheinlichkeit,

12 Zechprellerei: $b = 13,6$; Raub: $b = 9,2$; Mord: $b = 7,3$. Weitere Kennwerte der Regressionen können bei Bedarf von den Autoren angefordert werden.

als unter ‚sehr wahrscheinlich‘ Opfer einer Zechprellerei zu werden. So liegt für das verbale Antwortlabel ‚sehr wahrscheinlich‘ beim Delikt ‚Mord‘ die mittlere prozentuale Wahrscheinlichkeit unter 40%. Dagegen liegt beim Delikt ‚Zechprellerei‘ für dasselbe verbale Antwortlabel die durchschnittliche zugeschriebene Wahrscheinlichkeit deutlich über 60%.

5 Diskussion

Bei der Interpretation von Daten, welche mit kategorialen Antwortskalen generiert wurden, sollte beachtet werden, dass der Fragekontext die Antworten beeinflusst. So können wir in einem randomisierten Experiment einen Effekt der Frageformulierung auf die subjektive Wahrscheinlichkeitsschätzung nachweisen: Respondenten schätzen ihre Viktimisierungswahrscheinlichkeit je nach Fragekontext bei formal gleichen Situationen unterschiedlich ein. Die subjektiven Wahrscheinlichkeiten unterscheiden sich, je nachdem, ob das Risiko eingeschätzt wird, von einer Straftat verschont zu bleiben oder einer Straftat zum Opfer zu fallen. Zudem zeigen wir, dass die gleichen verbalen Wahrscheinlichkeitsangaben in verschiedenen Situationen unterschiedliche Bedeutungen aufweisen. Einerseits ist mit ‚wahrscheinlich‘ bei seltenen Ereignissen ein geringeres Risiko gemeint als bei häufigeren Ereignissen. Darüber hinaus bedeutet eine ‚wahrscheinliche‘ Viktimisierung bei schweren Delikten ein geringeres Risiko als bei leichteren Delikten.

Ein direkter Vergleich von Anteilswerten über verschiedene Viktimisierungen hinweg ist somit als problematisch einzustufen, da die Interpretation der Antwortkategorien abhängig ist von der Frageformulierung, der Delikthäufigkeit und der Deliktsschwere. Wenn Befragte beispielsweise mit der Kategorie ‚sehr wahrscheinlich‘ ihr subjektives Risiko kommunizieren, Opfer einer bestimmten Straftat zu werden, dann meinen sie beim Delikt Mord eine geringere Wahrscheinlichkeit, als beim Delikt Raub. Um jedoch verbale Wahrscheinlichkeitsangaben zu verschiedenen Viktimisierungen direkt vergleichen zu können, müssten Befragte die verbalen Antwortkategorien für alle Delikte einheitlich verwenden: ‚Sehr wahrscheinlich‘ ausgeraubt zu werden sollte demnach genauso wahrscheinlich sein wie ‚sehr wahrscheinlich‘ ermordet zu werden. Da diese Annahme aber nicht zutrifft, lassen sich individuelle Unterschiede in der Risikoeinschätzung häufig nicht erkennen, wenn man lediglich Antworten auf den verbalen Antwortskalen vergleicht.

Da Befragte sich in ihren Antworten durch den Fragekontext beeinflussen lassen, besteht berechtigter Zweifel an der Annahme, dass die untersuchten Items

zur Risikoeinschätzung ein eindimensionales Konstrukt erfassen.¹³ Zu einer ähnlichen Schlussfolgerung kommt auch Kreuter (2002: 228):

„So nannten Befragte bei den Überlegungen zu ihren eigenen Risiken zunächst die allgemeinen Risiken, besannen sich auf ihre üblichen Alltagsroutinen und leiteten daraus ihre Antwort auf das eigene Risiko ab. Die explizite Erfassung allgemeiner Risiken vor der Abfrage subjektiver Risiken kann die Befragten möglicherweise um diese kognitive ‚Kalkulation‘ entlasten.“

Einen viel versprechenden Ausgangspunkt zur Erfassung dieser allgemeinen Risiken stellen frequentistische Ansätze zum Umgang mit Wahrscheinlichkeiten dar (vgl. Gigerenzer/Hoffrage 1995; Hoffrage et al. 2000).¹⁴

Neben diesen Überlegungen zu geeigneten Erfassungsstrategien von subjektiven Viktimisierungswahrscheinlichkeiten könnte als weiterer Anknüpfungspunkt an unsere Studie die Stärke der Framing-Effekte auf die verbale Skala und die numerische Skala verglichen werden. Einige Studien gehen davon aus, dass Kontext-Effekte auf der verbalen Skala stärker wirken als auf der prozentualen Skala (vgl. Moxey/Sanford 2000: 238; Welkenhuysen et al. 2001). Dagegen wird in anderen Studien berichtet, dass Framing-Effekte nur auf der numerischen Skala signifikant sind (vgl. Jasper et al. 2001). Eine vertiefte Untersuchung dieses Interaktionseffektes würde zu einem besseren Verständnis der Nutzung unterschiedlicher Antwortskalen beitragen.

- 13 Wir konnten in unseren Untersuchungen der Antwortstrategien von Befragten im Rahmen von qualitativen Interviews (N=12) im Pretest unseres Instruments feststellen, dass Viktimisierungswahrscheinlichkeiten bei häufigen Delikten (z. B. Zechprellerei) für das kommende Jahr aus Viktimisierungen des vergangenen Jahres abgeleitet wurden. Dagegen orientierten sich Befragte bei selteneren Delikten eher an der Viktimisierung anderer Personen. So wurden beispielsweise beim Delikt Mord zum Vergleich Zeitungsberichte herangezogen. Zukünftige Studien könnten diese Antwortstrategien noch differenzierter untersuchen.
- 14 „Frequentistische Ansätze“ basieren auf Erkenntnissen der Kognitionspsychologie, wonach die Darstellung und Kommunikation von statistischen Informationen in Form von Häufigkeiten intuitiver und verständlicher ist als alternative Darstellungsformen in Wahrscheinlichkeiten und Prozenten: „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). Ein bekanntes Problem ist die subjektive Überschätzung von Risiken im Zusammenhang mit seltenen Ereignissen (fiktives Beispiel: „Wie wahrscheinlich ist es, dass eine zufällig ausgewählte Person aus ihrer Wohngegend innerhalb der nächsten 12 Monaten Opfer eines Raubs wird?“). Werden Items zur Messung subjektiver Risiken in Häufigkeiten formuliert, so fördert dies ein Denken in Verteilungen und führt in vielen Fällen zu realistischeren Schätzungen der Base Rate (fiktives Beispiel: „Stellen Sie sich 1000 Personen aus ihrer Wohngegend vor. Wie viele von diesen Personen werden innerhalb der nächsten 12 Monaten Opfer eines Raubs werden?“). Teigen (1974) zeigt, dass die subjektive Überschätzung von kleinen Wahrscheinlichkeiten abnimmt, wenn Befragte Risiken in Häufigkeiten anstatt Wahrscheinlichkeiten schätzen. Coutts (2002) zeigt, dass Häufigkeitsformate („distributional perspective“) zu realistischeren, niedrigeren Einschätzungen von deliktsspezifischen Viktimisierungsrisiken führen als andere Formate („singular perspective“).

Dieses bessere Verständnis würde dann eine begründete Auswahl einer verbalen oder numerischen Skala ermöglichen. So wird in der Medizin die Verwendung von verbalen Häufigkeitsausdrücken größtenteils abgelehnt (vgl. Nakao/Axelrod 1983). Entsprechend beschäftigen sich die meisten Fachartikel in diesem Bereich mit den Vor- und Nachteilen verschiedener Varianten von numerischen Skalen. Ghosh/Ghosh (2005) haben in einer umfassenden Review Ergebnisse von insgesamt 52 Studien zusammengestellt. Hierbei wurde in 17 Studien untersucht, ob Medizinstudenten und Ärzte numerische Wahrscheinlichkeitsformate richtig deuten und korrekt kommunizieren können:

„Physicians demonstrate widely varying understanding of probability terms and NNT [number needed to treat]“ (Ghosh/Ghosh 2005: 178).

Noch größere Probleme hinsichtlich des Verständnisses von numerischen Angaben wurden dabei in den restlichen 35 Studien auf Seiten der Patienten festgestellt.

Der direkte Vergleich von verbalen und numerischen Skalen führt allerdings in einigen Studien zu einer positiveren Bewertung der Ersteren. So schätzen Frauen ihr eigenes Brustkrebsrisiko auf einer verbalen Skala realistischer ein als auf einer numerischen Skala, bei der es zu einer deutlichen Überschätzung des objektiven Krebsrisikos kam (vgl. Woloshin et al. 1999).

Windschitl/Wells (1996: 359) kommen schließlich zu dem Ergebnis, dass sowohl numerische als auch verbale Maße ihre Berechtigung haben. Sie stellen fest:

„Results suggest that numeric measures tend to elicit deliberate and rule-based reasoning from respondents, whereas verbal measures allow for more associative and intuitive thinking“ (Windschitl/Wells 1996: 343).

Literatur

- ALLBUS, 1990 u. 2000: Codebücher (Mannheim: Zentrum für Umfragen, Methoden und Analysen, ZUMA). <http://www.gesis.org/Datenservice/ALLBUS/index.htm> (08.05.2007).
- Banks, S. M., P. Salovey, S. Greener, A. J. Rothman, A. Moyer, J. Beauvais und E. Epel, 1995: The effects of message framing on mammography utilization. *Health Psychology* 14: 178-184.
- Bradburn, N. und C. Miles, 1979: Vague Quantifiers. *Public Opinion Quarterly* 43: 92-101.
- British Crime Survey 2004 u. 2005: Fragebogen (London: Research development statistics, rds). <http://www.data-archive.ac.uk/doc/5347/mrdoc/pdf/5347userguide2.pdf> (08.05.2007).
- Budescu, D. und T. S. Wallsten, 1995: Processing linguistic probabilities. General principles and empirical evidence. S. 275-318 in: J. Busenmeyer, D. L. Medin und R. Hastie (Hg.): *Decision making from a cognitive perspective*. San Diego, CA: Academic Press.
- Conrad F. G., N. Brown und E. Cashman, 1998: Strategies for Estimating Behavioural Frequency in Survey Interviews. *Memory* 6: 339-366.

- Coutts, E., 2002: Context effects in the measurement of subjective probabilities in surveys. Universität Konstanz: Diplomarbeit Fachbereich Psychologie.
- Edwards, A., G. Elwyn, J. Covey, E. Matthews und R. Pill, 2001: Presenting Risk Information – A Review of the Effects of "Framing" and other Manipulations on Patient Outcomes. *Journal of Health Communication* 6: 61-82.
- Ghosh, A. K. und K. Ghosh, 2005: Translating evidence-based information into effective risk communication. Current challenges and opportunities. *Journal of Laboratory and Clinical Medicine* 145: 171-180.
- Gigerenzer, G. und U. Hoffrage, 1995: How to improve Bayesian reasoning without instruction. Frequency formats. *Psychological Review* 102: 684-704.
- Goocher, B. E., 1965: Effects of attitude and experience on the selection of frequency adverbs. *Journal of Verbal Learning and Verbal Behaviour* 4: 193-195.
- Grice, H. P., 1993: Logik und Konversation. S. 243-265 in: G. Meggle (Hg.): *Handlung, Kommunikation, Bedeutung*. Frankfurt am Main: Suhrkamp-Verlag.
- Gurm, H. S. und D. G. Litaker, 2000: Framing procedural risks to patients. Is 99% safe the same as a risk of 1 in 100? *Academic Medicine* 75(8): 840-842.
- Hakel, M. D., 1969: How often is often? *American Psychologist* 25: 533-534.
- Hammerton, M., 1976: How much is a large part? *Applied Ergonomics* 7: 10-12.
- Hoffrage, U., S. Lindsey, R. Hertwig und G. Gigerenzer, 2000: Communicating Statistical Information. *Science* 290: 2261-2262.
- Hörmann, H. 1983: The calculating listener, or how many are „einige, mehrere und ein paar“. S. 221-234 in: R. Bauerle, C. Schwarze und A. von Stechow (Hg.): *Meaning, use and interpretation of language*. Berlin: De Gruyter.
- Jasper, J., R. Goel, A. Einarnson, M. Gallo und G. Koren, 2001: Effects of framing on teratogenic risk perception in pregnant women. *The Lancet* 358: 1237-1238.
- Kahneman, D. und K. A. Tversky, 1979: Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263-291.
- Karelitz, T. M. und D. Budescu, 2004: You Say "Probable" and I Say "Likely". Improving Interpersonal Communication with Verbal Probability Phrases. *Journal of Experimental Psychology Applied* 10: 25-41.
- King, G., J. L. Murray Christopher, J. A. Salomon und A. Tandon, 2004: Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review* 98: 191-207.
- Kreuter, F., 2000: Uncertainty in capturing uncertainty. Toward measurement of fear of crime. Part I – Measuring subjective probability. (Girona, Annual Conference of the International Society of Quality of Life Studies). www.uni-konstanz.de/FuF/Verwiss/Schnell/Kreuter/girona.pdf (08.05.2007).
- Kreuter, F., 2002: *Kriminalitätsfurcht. Messung und methodische Probleme*. Opladen: Leske & Budrich.
- Kuhn, K. M., 1997: Communicating uncertainty. Framing effects on responses to vague probabilities. *Organizational Behavior and Human Decision Processes* 71: 55-83.
- Marteau, T. M., 1989: Framing of information. Its influence upon decisions of doctors and patients. *British Journal of Social Psychology* 28: 89-94.
- McNeil, B. J., S. G. Pauker, H. C. Sox und K. A. Tversky, 1982: On the elicitation of preferences for alternative therapies. *New England Journal of Medicine* 306: 1259-1262.
- Menon, G., P. Raghubir und N. Schwarz, 1995: Behavioral frequency judgments. An accessibility-diagnosticity framework. *Journal of Consumer Research* 22: 212-228.
- Merz, J. F., M. J. Druzdzel und D. J. Mazur, 1991: Verbal expressions of probability in informed consent litigation. *Medical Decision Making* 11(4): 273-281.

- Moxey, L. M. und A. J. Sanford, 1993: Prior expectation and the interpretation of natural language quantifiers. *European Journal of Cognitive Psychology* 5: 73-91.
- Moxey, L. M. und A. J. Sanford, 2000: Communicating quantities: A Review of Psycholinguistic Evidence of How Expressions Determine Perspective. *Applied Cognitive Psychology* 14: 237-255.
- Murakami, H., 2000: *South of the border, west of the sun*. New York: Vintage.
- Nakao, M. A. und S. Axelrod, 1983: Numbers are better than words – Verbal specifications of frequency have no place in medicine. *American Journal of Medicine* 74(6): 1061-1065.
- Newstead, S. E. und K. R. Coventry, 2000: The role of context and functionality in the interpretation of quantifiers. *European Journal of Cognitive Psychology* 12: 243-259.
- Pepper, S. und L. S. Prytulak, 1974: Sometimes frequently means seldom. Context effects in the interpretation of quantitative expressions. *Journal of Research in Personality* 8: 95-101.
- Peters, E., K. McCaul, M. Stefanek und W. Nelson, 2006: A Heuristic Approach to Understanding Cancer Risk Perception. Contributions from Judgement and Decision-Making Research. *Annals of Behavioral Medicine* 31: 45-52.
- Reuband, K. H., 2002: Subjektive Wahrscheinlichkeiten und Antwortmuster: Der Einfluss von Personenbezug und Skalierungsart. *ZA-Information* 50: 46-58.
- Reyna, V., 1981: The language of possibility and probability. Effects of negation on meaning. *Memory & Cognition* 9: 642-650.
- Schaeffer, N.C. 1991: Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly* 55: 395-423.
- Schnell, R. und F. Kreuter, 2000: Das DEFECT-Projekt. Sampling-Errors und Nonsampling-Errors in komplexen Bevölkerungsstichproben. *ZUMA-Nachrichten* 47: 89-101.
- Schwarz, N., H. J. Hippler, B. Deutsch und F. Strack, 1985: Response categories. Effects on behavioural reports and comparative judgments. *Public Opinion Quarterly* 49: 388-395.
- Schwarz, N. und B. Scheuring, 1988: Judgments of relationship satisfaction. Inter- and intraindividual comparison strategies as a function of questionnaire structure. *European Journal of Social Psychology* 18: 485-496.
- Schwarz, N. und B. Scheuring, 1992: Selbstberichtete Verhaltens- und Symptommhäufigkeiten. Was Befragte aus Antwortvorgaben des Fragebogens lernen. *Zeitschrift für Klinische Psychologie* 22: 197-208.
- Simpson, R. H. 1944: The specific meanings of certain terms indicating differing degrees of frequency. *Quarterly Journal of Speech* 30: 328-330.
- Sutherland, H. J., G. A. Lockwood, D. L. Trichtler, F. Sem, L. Brooks und J. E. Till, 1991: Communicating probabilistic information to cancer patients – is there "noise" on the line? *Social Science & Medicine* 32(6): 725-731.
- Teigen, K. H., 1974: Overestimation of subjective probabilities. *Scandinavian Journal of Psychology* 15: 56-62.
- Tversky, K. A. und D. Kahneman, 1981: The framing of decisions and the psychology of Choice. *Science* 211: 453-458.
- Wänke, M., 2002: Conversational Norms and the Interpretation of Vague Quantifiers. *Applied Cognitive Psychology* 16: 301-307.
- Wallsten, T. S., D. Budescu, R. Zwick und S. M. Kemp, 1993: Preferences and reasons for communicating probabilistic information in numerical or verbal terms. *Bulletin of the Psychonomic Society* 31: 135-138.
- Weber, E. U. und D. J. Hilton, 1990: Contextual effects in the interpretation of probability words – perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance* 16(4): 781-789.

- Welkenhuysen, M., G. Evers-Kiebooms und G. D'Ydewalle, 2001: The language of uncertainty in genetic risk communication. Framing and verbal versus numerical information. *Patient Education and Counseling* 43: 179-187.
- Windschitl, P. D. und G. L. Wells, 1996: Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology-Applied* 2(4): 343-364.
- Wohlfahrtssurvey, 1998: Fragebogen (Mannheim: Zentrum für Umfragen, Methoden und Analysen, ZUMA). http://www.gesis.org/Dauerbeobachtung/Sozialindikatoren/Publikationen/Datenreport/pdf_files/ws1998.pdf (08.05.2007).
- Woloshin, S., L. M. Schwartz, W. C. Black und H. G. Welch, 1999: Women's perceptions of breast cancer risk. How you ask matters. *Medical Decision Making* 19(3): 221-229.
- Wright, D., G. Gaskell und C. O'Muirheartaigh, 1994: How much is "Quite a bit"? Mapping between numerical values and vague quantifiers. *Applied Cognitive Psychology* 8: 479-496.
- Zimmer, A. C., 1983: Verbal versus numerical processing of subjective probabilities. S. 159-182 in: R. W. Scholz (Hg.): *Decision making under uncertainty*. Amsterdam: Elsevier.

Korrespondenzadressen:

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

Heiko Rauhut
ETH Zürich
Professur für Soziologie
Universitätsstraße 41
CH-8092 Zürich
rauhut@gess.ethz.ch

Space and Time in Comparative Political Research

*Pooled Time-series
Cross-section Analysis
and Multilevel Designs
Compared*

Raum und Zeit in der vergleichenden Politikwissenschaft

*Die gepoolte
Zeitreihenanalyse und
die Mehrebenenanalyse
im Vergleich*

Isabelle Stadelmann-Steffen und Marc Bühlmann

Abstract

The combination of cross-section and time dimension is a central issue in current comparative political research. The state-of-the-art procedure in this context is pooled time-series cross-section analysis (PTSCS), which is en vogue in today's relevant literature but not uncontested. An interesting option are multilevel designs, which allow the combination of time and space by considering observations over time nested within country-specific contexts. The purpose of this paper is to illustrate the advantages of multilevel designs in comparative political research, which mainly concern the modeling of time-invariant variables, the possible distinction between cross-sectional and time related variance in the data, and the possibility to model heterogeneity instead of just correcting it. Using the example of an analysis of public education expenditure in the 26 Swiss cantons between 1978 and 2003, it can be shown that multilevel analysis – mainly due to its statistical and conceptual advantages – is indeed a promising alternative to PTSCS.

Zusammenfassung

Die Kombination von Längs- und Querschnittsvergleich ist ein zentrales Thema in der vergleichenden Politikwissenschaft. In aller Regel wird ein gepooltes Zeitreihen-Design angewandt, um verschiedene politische Einheiten über die Zeit und miteinander zu vergleichen. Diese Methode, obwohl sehr in Mode, ist allerdings nicht unumstritten. Eine interessante Alternative stellt die Mehrebenenanalyse dar, welche ebenfalls die Kombination von Zeit und Raum zulässt, indem sie Beobachtungen in der Zeitachse quasi als Eigenschaften verschiedener länderspezifischer Kontexte betrachtet. Das Ziel dieses Beitrags ist es, die Vorteile eines Mehrebenen-Designs für die vergleichende Politikforschung zu illustrieren. Diese ergeben sich v. a. in Bezug auf die Modellierung zeit-invarianter Variablen, die Unterscheidung zwischen querschnitt- und zeitbedingter Varianz sowie die Möglichkeit, Heterogenität zu modellieren statt lediglich zu korrigieren. Als praktisches Anwendungsbeispiel werden die öffentlichen Bildungsausgaben in den 26 Schweizer Kantonen zwischen 1978 und 2003 analysiert. Es zeigt sich, dass die Mehrebenenanalyse sowohl aus statistischen Gründen, aber auch aufgrund ihrer konzeptuellen Vorteile tatsächlich eine viel versprechende Alternative zu gepoolten Zeitreihen-Designs darstellt.

1 Introduction¹

Combining cross-sectional and time-dimensional data is one of the central issues in the present methodological discussion in comparative political research. Since data sets in social science are often limited to about 20 to 30 cases, and as dynamics over time are an important aspect of typical political and societal phenomena, the simultaneous examination of time and space seems to be indispensable in quantitative research. Pooled time-series cross-sectional analysis (PTSCS) is the currently prevailing method for analysing this type of data and is correspondingly en vogue in the relevant literature.²

However, together with the design's growing popularity a discussion on its methodological pitfalls has arisen. By pooling the time and cross-sectional dimensions the data obtains a specific structure, which violates important condition for ordinary least squares regression (OLS). Although the techniques to deal with these problems are quite sophisticated by now, concerns and criticism on the one hand and the search for better alternatives on the other remain.

An interesting option in this discussion is multilevel analysis (MLA), which allows for the combination of different levels of analysis. Prevalent in educational science for a long time, multilevel applications have more recently been used in political science, too. Although in most cases the method is used for the simultaneous analysis of individual and contextual data, it also offers the possibility of combining time and space dimensions in one single model: time and space are considered as two levels of analysis, whereas observations over time are nested within country-specific contexts.

While there is an extensive debate on the strengths and limits of pooled time-series cross-section analysis, multilevel designs as a possible alternative have been much less discussed. An exception is Western (1998: 1234) who claims that a multilevel specification of pooled time-series models has both substantial and statistical advantages. He stresses that MLA allows for greater causal complexity which provides a closer fit between typical comparative theory and model specifi-

1 We thank two anonymous referees for helpful remarks on an earlier version of this paper, and K. Gilland for linguistic assistance.

2 If we refer to the PTSCS method in the following we mean the "de facto Beck-Katz standard" which has become the "accepted econometric technique in comparative political economy" (Plümper et al. 2005: 327), even though PTSCS actually only describes the type of data structure, which would actually allow for a wide variety of model specifications.

cation. Hence, multilevel models not only control for, but model heterogeneity in the data. However, only recently multilevel models for time-series cross-section data have been subject to a more systematic evaluation. Shor et al. (2007) employing Monte Carlo simulations show that multilevel models indeed perform better than OLS estimators. They stress that MLA is more flexible offering the researcher additional advantages that can be summarized as follows:

1. MLA performs better than other common estimators in Monte Carlo Simulations. When the number of observations is low (either at the time- or country-level) MLA does better, since each "estimated parameter has the potential to borrow strength from other parameters in the model" (Shor et al. 2007: 5).
2. As a consequence, MLA generally provides a better model fit (Shor et al. 2007: 4f.). While complete pooling ignores important parts of contextual variance, no pooling – i.e. separate intercepts for single time series – is affected by effects of outliers and does not allow the sharing of important information from across units. Multilevel analysis which can be seen as a process of "partial pooling" (Shor et al. 2007: 5) offers an advantageous middle way. Multilevel models thereby allow distinguishing and analyzing the variances at both the cross-sectional and the time level.
3. Another advantage of MLA is the handling of slowly moving or completely time-invariant variables: While such variables often produce estimation problems in PTSCS, they are not a problem in multilevel models: Partial pooling as done in MLA "allows estimates of units to borrow strength from the whole sample and shrink toward a common mean" (Shor et al. 2007: 4). Moreover, theoretically and conceptionally it makes more sense to view such time-invariant variables as contextual effects which influence all observations of a given unit.
4. The basic random intercept multilevel model can easily be expanded by also including varying slopes, i.e. different effects of a predictor variable for different units.

Starting from these considerations the purpose of this paper is to demonstrate the differences between pooled time-series cross-sectional models and multilevel analysis. Its main aim is to illustrate the advantages of multilevel models in practical applications that are typical in comparative political research. First and foremost, we focus on the differences between the two methods in terms of the modeling of time-invariant variables and model fit. Time-invariant or slowly moving variables such as institutions or cultural settings are crucial elements of theories explaining differences between political entities. The simultaneous modeling of time-invariant variables and dynamic determinants within comparative politics therefore is an important task. Drawing on Shor et al. (2007) we suggest that the estimation of the

influence of time-invariant variables is more appropriately done by MLA than by PTSCS. Furthermore, it can be expected that due to its conceptual advantages MLA provides a better model fit than PTSCS.

As an example of a practical application we use a comparative analysis of public education expenditure in the 26 Swiss cantons between 1978 and 2003 (in so doing we particularly refer to Freitag/Bühlmann 2003). Following the theories of welfare state development (Schmidt 2000) we distinguish institutional, socio-economic, party-political, and historic-cultural variables, and derive respective hypotheses to explain the differences in cantonal education expenditure.

The paper unfolds as follows. First the theoretical background of our illustrative example is shortly demonstrated. Afterwards the main part of this contribution consists of two analyses of cantonal education expenditure, one using pooled time series and the other applying a multilevel design. The analyses are followed by a discussion of the results and the practical implications of this methodological comparison. The paper closes with concluding remarks.

2 Pooled Time-series Cross-section Analysis versus Multilevel Design – Two Comparative Analyses of the Public Education Expenditure in the Swiss Cantons

The aim of this contribution is to compare a pooled times-series cross-section design and multilevel analysis by means of an empirical example. We will therefore present two analyses of cantonal public education expenditure, one applying pooled times-series cross-sectional analysis and the other using a multilevel design. Both analyses will start with a short description of the method and the state-of-the-art procedures used in practical applications. Afterwards, the hypotheses on the differences in cantonal education expenditure, briefly outlined in the following section, will be empirically tested by applying the two methods. After having illustrated both methods, we will compare the results and try to demonstrate the advantages of a multilevel design.

2.1 Public education expenditure in the Swiss cantons – theory and hypotheses

In the Swiss federal system there are various areas in which the cantons have substantial autonomy in policy-making. This is especially true for the education system, which has been the responsibility of the cantons and the communes since the beginning of the Swiss nation-state. As a consequence, the education systems and the financial investments of public authorities in human resources substantially vary among the cantons. In the following we will therefore analyze the determinants of cantonal differences in public education expenditure.

Even though this paper focuses on the methodological implications rather than on substantial results, we will shortly lay out the theoretical background of our analysis in this paragraph.

Most of the quantitative studies done in this field (e.g. Freitag/Bühlmann 2003; Boix 1997; Castles 1999; Cameron/Hofferbert 1974) refer to theories of comparative welfare state research: education policy is seen as a part of a comprehensive welfare state concept (Castles 1999: 10). Theories of welfare state development (Schmidt 2000) can thus help to explain the variances in public education expenditure. Following this literature we distinguish four different approaches: socio-economic determinants, the partisan theory of public policy, institutional explanations, and historical and cultural legacies.

The first approach hypothesizes that public policy is substantially shaped by existing *socio-economic challenges* (Schmidt 2000: 23f.). In this sense public education expenditure depends on the demand for public education and the possibility to provide a corresponding supply. First, it can be hypothesized that public education expenditure is high, if the share of children of school age and thus the demand for investments in public education is considerable (Schmidt 2002: 10; Hega 1999: 62; Poterba 1999; Fernandez/Rogerson 1997). Similarly, a substantial share of foreigners in schools is expected to raise the need for public investments in education in order to finance necessary integration measures (Freitag/Bühlmann 2003: 146). However, theoretically, only those parts of the population support an extensive education system, which in some way can profit from the educational supply. All others will try to restrict public investment in education. Boix (1997: 837) finds that especially high employment shares in the agricultural sector tend to reduce public education expenditure since this group has only very limited interest in the public education system. In contrast, it can be expected that a large service sector demands a well-educated labor force and therefore has

a positive effect on public education expenditure. Similarly and according to Castles (1999: 180), the demand for human resources is higher in urban areas. On the other hand it can nevertheless be argued that the provision and organization of public education is cheaper in densely populated regions, which in turn leads to lower expenditure.

A second explanation for differences in public education expenditure is provided by the *partisan theory* of public policy (Schmidt 2000: 26f.; Hibbs 1977). It develops the idea that the party-political composition of government influences public investments in public policy in general and in education in particular. According to the findings in the literature a substantial share of leftist parties in government lead to higher (education) expenditure, as these parties attribute the responsibility for education to the state rather than to the private sector (Schmidt 2002: 13f.; Hega 1999: 63f.; Castles 1999: 53, 1989: 438; Boix 1997: 815). This effect should be even more pronounced, if strong leftist parties are supported by powerful unions (Schmidt 2002: 12). Conversely, Christian democratic parties disapprove of government interference in the area of education for two reasons: On the one hand education had been in the hands of the church for a long time, on the other hand public intervention in education policy goes against the principle of subsidiarity, which is an important element of Christian democratic politics (Schmidt 2002: 14; Wilensky 1981).

Thirdly, in the analysis of public policy provision *institutional explanations* are of particular importance. Formal or informal rules build the framework in which state activities take place. Institutional factors therefore constrain but also allow for governmental action (Schmidt 2000: 28f.). In Switzerland the three typical elements of the political system are of overriding interest: direct democracy, the amount of decentralization, and the character of decision making (i.e. whether decision making is rather consensual or follows the majority rule) (Vatter/Freitag 2002). With respect to direct democracy previous studies have found that the extensive use of direct-democratic rights have a restricting effect on public expenditure (Vatter 2002; Vatter/Freitag 2002; Hega 1999: 67). Generally, voters tend to avoid additional taxes (Hega 1999: 66). Direct democracy provides them with a strong instrument to bring these preferences into politics and, thus, to prevent further investment in public policy (Schaltegger 2001: 6).

A similar effect can be attributed to decentralized decision-making structures and federalism. Federalist power-sharing and decentralized fiscal power in particular narrows the scope for state intervention (Vatter/Freitag 2002). For reasons of possible migration, decentralized structures additionally lead to competi-

tion for the most attractive location – not least in terms of taxes – among the decentralized entities (Freitag/Bühlmann 2003: 149). In sum, these tendencies lead to the hypothesis that decentralized decision-making structures (in our case: high local autonomy) result in lower public education expenditure (Cameron/Hofferbert 1974).

An inverse effect is generally attributed to consensus democracy, which is expected to foster welfare state expansion (Lijphart 1999). The integration of various political, social or cultural minorities into the decision-making process tends to produce cost-intensive barter transactions and thereby a higher degree of state intervention. It can therefore be assumed that a high degree of consensual alignment in the government leads to more substantial public education expenditure (Freitag/Bühlmann 2003: 150).

Finally, *historical and cultural legacies* can help explaining variances in public education expenditure. As mentioned above, the relationship between church and state has played an important role in the development of educational systems. In societies with strong conservative-catholic background state intervention in education has been limited, since catholic forces tried to maintain their traditional competence in this field. In contrast, dominant protestant forces facilitated cooperation between church and state as Protestantism stood for a fast education expansion lead by the state. It can actually be shown that the power balance between catholic forces and the state at the end of the 19th century is an important factor for the path chosen in education policy (Archer 1979), and also for the degree to which the state engages in education even today. In Schmidt's (2000: 31) argumentation present education policy is thus path-dependent, which means that it is a result of earlier decisions and their intended or unintended consequences.

Cultural differences exist in Switzerland not only for religious but also for linguistic reasons. While the French- and Italian-speaking parts of Switzerland share a lot of institutional and cultural elements with the neighboring countries France and Italy, the German-speaking part is much more oriented towards Germany and Austria. This is also reflected in different education systems. Contrary to the French- and Italian-speaking parts of Switzerland, in the German-speaking cantons we find a strong tradition of vocational education like in Germany and Austria, and accordingly higher education and thus financial investments in public schooling are lower (Schmidt 2002: 9; Castles 1999: 175; Hega 1999: 63).

In the following, these hypotheses will be tested in a comparative quantitative analysis. For the operationalization of the variables see the appendix.³

The four approaches provide us with different variables. While some variables (mainly the socio-economic and the party political determinants) vary over time, this is not the case for others. Hence, the historic-cultural factors as well as most of the institutional variables do not or only slightly change over time. In what follows we will see that the two methods differently handle and interpret these two groups of explanatory variables.

2.2 The dependent variable and methodological challenges

The dependent variable is the public education expenditure in the Swiss cantons between 1978 and 2003. This period of investigation has been chosen mainly for data related reasons: The year 1978 is the first year (after the foundation of the 26th Swiss canton, Jura) for which data for all 26 cantons is available. Starting in 1978 thus provides us with enough data for applying a multilevel analysis (Bullen et al. 1994; Jones 1997). Since the cantons vary substantially in size and population we analyze the education expenditure per capita.

The cantonal public education expenditure has been steadily growing during the last 26 years, which lets one assume that the data follows a non-stationary development. The statistical tests for unit-root (Hadri 2000; Im et al. 1997; Levin/Lin 1993) indeed all conclude that public education expenditure is non-stationary.⁴ In the methodological discussion of PTSCS this phenomenon is an important issue. If the data follows a permanent growing development over time the mean of the residuals is not constant over time either, which violates the Gauss-Markov-assumption for the execution of OLS. As a consequence OLS-results are wrong. Additionally, there often result so-called "spurious regressions", which display relationships only resting on a common development of variables over time (Kittel 1999: 249). Generally, the best way to handle non-stationary data is the examination of first differences. However, in this case conceptual characteristics of the model change: It is the short-term changes that are now modeled, while the differences in the level between units and long-term developments have been

3 To allow for a comparison of the results all data were standardised (mean = 0, standard deviation = 1).

4 Further analyses not presented here show that this finding also applies for the logarithmized education expenditure per capita.

eliminated from the model. Thus, such a model is not suited to sensibly evaluate the effect of the time-invariant variables. To sum up, the analysis of first differences may solve the statistical problem of non-stationarity, but it is not a valuable alternative for most applications in political science where differences between countries and effects of (almost) time-invariant variables (in our case: institutions and historical-cultural characteristics) are of interest.

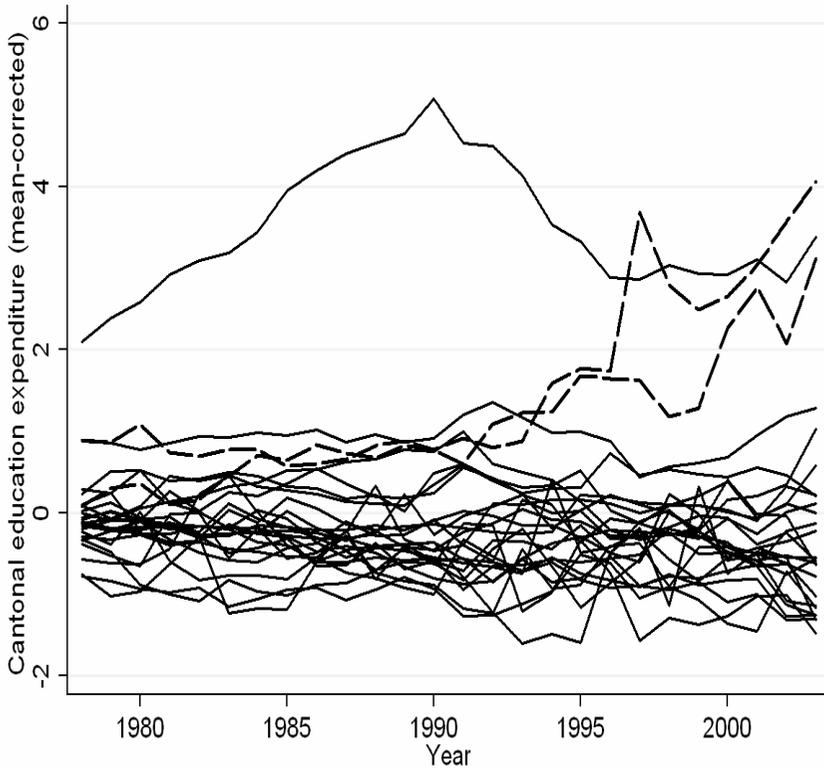
Therefore we proceed in another way that in some cases can also eliminate the problem of non-stationarity. We transform the education expenditure variable by sorting out the annual average values, which means that we remove the steadily increasing development in the time dimension.⁵ Using this transformation a big part of the development over time cannot be analyzed either. However, even though the model indeed focuses on cross-sectional variation, it also contains a substantial part of the time dimension. That is, the relative differences in the development of education expenditure between cantons can still be analyzed.

As shown in figure 1, this procedure largely solves the problem. Two cantons, however, possibly still follow a non-stationary development. For the purpose of this paper these two cantons – Zug and Basle Town – are excluded from the analysis. In so doing both the graphical examination of the data (figure 1) as well as the statistical tests indicate that the time-series under investigation are stationary.⁶ Thus, for the following analysis we use the mean-corrected public education expenditure for 24 cantons.

5 In the language of PTSCS this is actually equivalent to a model with fixed time effects.

6 The Levin-Lin-Chu test (lag[1]) tests the null hypothesis of Non-stationarity. In the present case the test produces a coefficient of -0.28 (p=0.02), thus implying that the time-series are stationary.

Figure 1 Standardized education expenditure in the Swiss cantons (mean-corrected)



Notes: The lines for Zug and Basle-Town (dashed, bold lines) indicate a non-stationary development.

2.3 Pooled time-series cross-section analysis (PTSCS)

In comparative political research the prevailing way of simultaneously analyzing variances over time and across countries is to pool cross-sectional data for different points in time:

$$(1) \quad Y_{ij} = \beta_0 + \beta X_{ij} + \varepsilon_{ij}$$

In our example Y_{ij} specifies the predicted educational investments of year i in canton j . This predicted value is explained by the overall mean (β_0) and the explanatory variables (the X variables and their respective β) measured for each

year and unit separately (subscripts i [year] and j [canton] respectively). β has no subscript since one coefficient per variable for all years and cantons is estimated.

However, this produces a data structure which practically by definition violates important conditions for OLS regression. In particular, PTSCS models suffer from temporally and spatially correlated errors as well as from panel heteroscedasticity (Beck/Katz 1995: 636; Kittel 1999: 228ff.; Podestà 2002: 10; Plümper et al. 2005). The consequence of these data characteristics is that the standard errors are systematically biased. Hence, conclusions concerning the statistical significance of the estimations become meaningless (Gujarati 1995: 366). In order to successfully use a PTSCS design it is therefore necessary to correct for these problems. To this end, the following procedures have won recognition, and represent the state of the art in this field. First, for the correction of heteroscedasticity and the spatial autocorrelation so-called panel-corrected standard errors (PCSE) are applied. This procedure draws on the fact that OLS still produces unbiased, but inefficient estimates. The OLS residuals allow for a reliable estimation of the covariances, which then are used to correct the standard errors (Beck/Katz 1995: 638). The PCSE procedure requires that any serial correlation has been eliminated before calculating the corrected standard errors (Beck/Katz 1995: 638). This can be done either by introducing a lagged dependent variable or by estimating the autocorrelation coefficient and correct for the correlation in the residuals (Kittel 1999: 230). The two proceedings represent different approaches: On the one hand, by modeling one or more lagged dependent variables autocorrelation is not seen as a problem, but is like other explanatory factors used for the explanation of the dependent variable. This procedure corresponds to the theoretical proposition that the economy is a self-regulating system which is only marginally influenced by political intervention (Kittel 1999: 230). The so-called AR1-approach, on the other hand, treats autocorrelation as a problem which has to be eliminated (Kittel 1999: 231). The corresponding correction of the standard errors is based on the proposition that the residuals follow a first order autoregressive process:

$$(2) \quad \varepsilon_{ij} = \rho\varepsilon_{i-1j} + v_{ij}$$

In this function the autocorrelation coefficient ρ (Rho), which takes on a value between 0 and 1, is estimated and is then used to correct the standard errors by applying a so-called Prais-Winsten-transformation (Kittel/Winner 2002: 16). This second approach is often preferred in comparative political research (Plümpfer et al. 2005: 342f.).⁷ Thus, the model takes the following form:

$$(3) \quad Y_{ij} = \beta_0 + \rho\gamma_{i-1j} + \beta X_{ij} + \varepsilon_{ij}$$

A further lively debate has evolved on whether unit fixed effects (FE) should be controlled for in order to account for unobserved unit-specific developments (equation 4). Shor et al. (2007: 4) call this option 'no pooling' and the 'opposite extreme' of complete pooling, since a separate intercept is calculated for each unit. The main disadvantage of this procedure from a political science point of view is, however, that such unit dummies "completely absorb differences in the level of independent variables across units" (Plümpfer et al. 2007: 331) and thus the time-invariant variables cannot be analyzed anymore. As a FE model is the 'opposite' of complete pooling⁸ and since a FE is conceptionally even more closely related to a multilevel design than a completely pooled model (Shor et al. 2007: 4) we will also present a FE model based on a "fixed effects vector composition", which should be a better way to estimate the effect of timeinvariant variables in a FE-context (Plümpfer/Troeger 2007).

$$(4) \quad Y_{ij} = \beta_{0j} + \rho\gamma_{i-1j} + \beta X_{ij} + \varepsilon_{ij}$$

In what follows, these procedures are applied to the analysis of the cantonal public education expenditure.

7 A discussion of the disadvantages of the lagged-variable approach can be found in Plümpfer et al. (2005).

8 A multilevel model would lie in between these two options (see section 2.4 and Shor et al. 2007: 4).

Table 1 TSCS-models for the explanation of public education expenditure (mean-corrected)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 (xtfevd)
Pupils per cantonal population	.10*** (.04)				.10*** (.03)	.02 (.09)
Proportion of foreign pupils	.15*** (.05)				.17*** (.05)	.00 (.03)
Sector 1	-.08 (.07)					
Urbanization	.29*** (.08)				.20*** (.07)	.24*** (.02)
Leftists in government		.09** (.05)			-.01 (.04)	.02 (.03)
CVP in government		-.07 (.05)				
Union		.07 (.05)				
Direct democracy			-.15*** (.05)		-.08 (.05)	-.21*** (.02)
Governmental concordance			-.03 (.03)			
Local autonomy			-.44*** (.06)		-.39*** (.06)	-.41*** (.02)
Strength of Catholics 19 th century				-.25*** (.06)	.00 (.05)	-.07*** (.02)
Efforts in education policy 1880 language				.06 (.05)		
				-.32*** (.08)	.08* (.06)	.13*** (.02)
eta						.93*** (.04)
constant	-.09** (.05)	-.12*** (.03)	-.11*** (.03)	-.13*** (.04)	-.08** (.04)	-.08*** (.01)
N	624	624	624	624	624	600
R2	.13	.03	.13	.05	.26	.87
AIC	189.5	322.3	206.2	179.6	232.5	415.4
Rho	.83	.79	.82	.86	.76	.67

Notes: Non-standardized OLS-coefficients; standard errors in brackets. Heteroscedasticity and autocorrelation is treated as follows: To correct for heteroscedasticity and spatial autocorrelation panel corrected standard errors (PCSE) are calculated as suggested by Beck/Katz (1995: 638). Beforehand, serial autocorrelation was eliminated by using a Prais-Winsten-transformation to correct the standard errors (in model 6 the Cochran-Orcutt method is used for the correction of the standard errors which reduces the number of observations by one in each panel). Thereby it is assumed that the residuals follow an AR1-process (Kittel/Winner 2002: 16). *** significant at the 1%-level; ** significant at the 5%-level; * significant at the 10%-level. In model 6 the following variables are used to explain unit fixed effects: urbanization, direct democracy, autonomy of the municipalities, strength of Catholics, language. The language variable and the strength of Catholics are typical examples for variables that per definition do not change over time, while urbanization is time-invariant due to the limited period of investigation and data availability. The same applies to local autonomy that can be seen as a constant institutional feature during our period of investigation. In contrast direct democracy is an example of a slowly-moving variable, since we have three different measures of the direct democracy index indicating only minor changes over time.

Table 1 reports the results for the multivariate models with the mean-corrected public education expenditure as dependent variable. In models 1 to 4 the explanatory power of the four theoretical approaches (socio-economic, party political, institutional, and historic-cultural approach) is tested separately. Afterwards a fifth model is estimated which incorporates the most important (i.e., significant) variables of the four approaches. Finally, a FE model is calculated on the basis of model 5 allowing for separate cantonal intercepts.

It can be shown that socio-economic determinants on the one hand and institutional factors on the other hand are the most important variables in the model. However, only for two variables a consistent and significant influence on cantonal education expenditure can be demonstrated: While a high degree of urbanization increases public investments in education, highly decentralized structures within the Swiss cantons (i.e., high autonomy of municipalities) decrease cantonal education expenditure. Moreover, held everything constant, the German speaking cantons spend more on public education, which is contrary to our hypothesis. This result can be seen against the background that the dependent variable incorporates expenditure for vocational training, but not for higher education. The latter is particularly important in the Latin cantons, and thus their total investments in education are underestimated by the variable chosen here. In contrast, the size of the agricultural sector, the share of Christian-democratic parties in government, the union strength as well as our indicator for consensus democracy seem not to influence public education expenditure. For some other variables, i.e. the number of (foreign) pupils, direct democracy, and the historical strength of the Catholic Church the PTSCS models are inconclusive.⁹

The Akaike Information Criterion (AIC)¹⁰ shows that neither model 5 nor model 6 are satisfactory in terms of model fit. Actually, even though the two models incorporate more (significant) variables they are not better in explaining

9 In what concerns direct democracy these results actually reflect a standard finding in Swiss political science in that this variable is highly correlated with the language region and therefore is highly sensitive to the introduction of the latter. This is because cantonal differences with regard to direct democracy strongly coincide with the language regions. It is mainly the German speaking cantons that show very extensive direct-democratic procedures (Linder 2005: 272, 1994; Vatter 2002: 401ff.).

10 The Akaike Information Criterion tests the model fit by penalizing for the addition of parameters, and thus selects a model that fits well but has a minimum number of parameters. By means of the AIC models based on different estimation methods and specifications can be compared, the sample however needs to be identical (Burnham/Anderson 2004: 267f.). Therefore, the AIC of model 6 cannot strictly be compared with the others.

the data than model 4 which includes the cultural variables only. Finally, even though we use the mean-corrected data, the high values of Rho show that autocorrelation in the model is still substantial. The investigation of the residuals however indicates that the first order autocorrelated error terms (AR1) in the model do satisfactorily account for these processes.

2.4 Multilevel analysis

"Multilevel analysis (MLA) is a methodology for the analysis of data with complex patterns of variability, with a focus on nested sources of variability" (Snijders/Bosker 1999: 1). In most of the empirical analyses with MLA, individuals are nested in contexts. Mainly applied in education science, a typical multilevel design consists of data on pupils nested within classes and/or schools, whereby it is argued that their school achievements, for instance, not only depend on individual capacities but also on class- and school-related factors.

In political science, the use of MLA is increasing. While the idea that context influences individual behavior has been crucial at least since the suicide-analysis by Durkheim (1897), corresponding quantitative analyses in political science in the past century have lacked the methodological base to properly analyze contextual effects. With MLA the simultaneous modeling of individual and contextual determinants to explain individual behavior is now possible (Bühlmann 2006).¹¹

The underlying principle of multilevel modeling is quite simple. Intercepts of common OLS regression analysis are allowed to vary:

$$(5) \quad Y_{ij} = \beta_{0j} + \beta X_{ij} + \varepsilon_{ij}, \text{ where}$$

$$(6) \quad \beta_{0j} = \beta_0 + \mu_{0j} \text{ (} \mu_{0j} \text{ stands for the residuals at level 2).}$$

11 With pure individual or pure ecological analyses, we cannot explain contextual influence on individual behaviour. For more detailed discussion see Bühlmann 2006; Goldstein 1987: 19; Hox/Kreft 1994: 285; Rasbash et al. 2002; Snijders/Bosker 1999; Steenbergen/Jones 2002; Stoker/Bowers 2002; Teachman/Crowder 2002: 284.

If individual behaviour is modelled as it is most often the case in MLA, such a model implies that individual behaviour can vary between contexts. In other words, unlike standard regression analysis this model does not assume that the constant (β_0) is the same in all units; but rather it varies from context to context. In our case we adapt the common multilevel model as follows: instead of modeling individuals within contexts, we apply the model to multiple observations nested within a unit. More precisely: the data for different points in time is seen as nested within cantons. Thereby, we suppose that the development of educational expenditure over time differs from canton to canton.¹² In so doing, we are able to analyze the effects of time-dependent developments (level 1) and of time-invariant cantonal characteristics (i.e., institutions or historical cultural characteristics) (level 2) separately. While, the number of pupils and of foreign pupils, or the strength of parties in government change over time and are, thus, level-1 variables, most of the institutional and historical cultural variables (e.g., local autonomy, the strength of the Catholic church at the end of the 19th century, or the language region) are constant over time. They are now explicitly modeled as context factors which differently influence all observations of a given canton over time. Furthermore, we can distinguish between the cross-sectional variance (between the cantons) on the one hand, and the variance over time within cantons on the other. To sum up, such a model enables us to model simultaneously different effects due to time (time-dependent variables) and due to cantonal characteristics (time-independent cantonal properties) on cantonal education expenditure.¹³ Our standard model, hence, takes the following form.

$$(7) \quad Y_{ij} = \beta_{0j} + \beta X_{ij} + \alpha W_j + \mu_{0j} + \varepsilon_{ij}$$

Education expenditure in canton j at time point i can be explained by an overall mean (β_0), time-dependent variables (the X variables and their respective β), time-independent cantonal properties (the W variables and their respective α), cantonal variation (μ_{0j} with an assumed mean of 0 and a total between-canton

12 It would also be possible to change the two levels, i.e., to consider the cantons as level 1 and the time dimension as level 2.

13 We do not go further in this article. It is however worth mentioning that MLA makes it possible to also model random slopes or even cross-level interactions. In other words: One can model canton-specific effects on the strength of time effects on education expenditure.

variance of σ_{μ}^2), and time variation (ε_{ij} with an assumed mean of 0 and a total within-canton variance of σ^2). Hence, the overall variation ($\sigma_{\mu}^2 + \sigma^2$) is divided into differences at the time level (level 1 variance), which shall be explained by time-dependent variables, and differences between contexts (level 2 variance), which shall be explained by cantonal, cross-sectional factors.¹⁴

However, as with PTSCS the issue of autocorrelation has to be discussed: In the multilevel model, too, correlated errors have to be expected. This is first of all true for level 1, which consists of 26 subsequent and thus related observations for the cantons. Just as in the PTSCS model, in the multilevel model, too, we therefore model the level-1 error term to follow a first-order autocorrelated process ($\rho_i \varepsilon_{ij}$).¹⁵

$$(8) \quad Y_{ij} = \beta_{0j} + \beta X_{ij} + \alpha W_j + \mu_{0j} + \rho_i \varepsilon_{ij}$$

As we did in the PTSCS, we use this multilevel model for the analysis of the cantonal public education expenditure.

Table 2 presents the results of the multilevel models with mean-corrected public education expenditure as dependent variable. As in the PTSCS analysis, we first estimate four models for the four theoretical approaches. Again, we calculate a joint fifth model by including the significant variables from the previous analyses. To further compare the MLA and the PTSCS design, we estimate a sixth model containing the variables of the PTSCS-model 5 from table 1.

Before calculating the multivariate models we estimate empty models without any specifications in order to assess the variance components. With MLA we can distinguish between the effects of cantonal characteristics and time-related changes. In the empty model we estimate the variance of the education expenditure which can be separated in a level 1 part (differences between time points) and a level 2 part (differences between cantons). The estimation shows

14 We refer to the relevant literature on MLA for a more thorough discussion of the method and expansions of the basic model presented here (Bullen et al. 1994; Ditton 1998; Goldstein 1987, 1995; Hox 1995; Jones 1997; Jones/Duncan 1996; Snijders/Bosker 1999; Teachman/Crowder 2002; for a critic see Rohwer 1998).

15 We use MLwiN to estimate our multilevel models (Rasbash et al. 2002). In this software program it is possible to model autocorrelated errors terms by applying time-series macros (cp. Yang et al. 2004).

that both, differences between the cantons and differences between time points within a canton contribute to the model's total variance: 78% of total variance in the empty model is found at level 2, while 22% can be attributed to level 1. This can be interpreted to mean that differences between countries are more pronounced than differences over time (which may be also due to fact that the data is mean corrected).

It can be seen from table 2 that the same procedure of model building does not result in the same final model as in the PTSCS context (model 5, table 2). Surely, like in the PTSCS model the most crucial variables are again the degree of urbanization and local autonomy. Compared to PTSCS, however, both the language region and the share of (foreign) pupils (complete pooled model 5, table 1), as well as direct democracy and the strength of Catholics at the end of the 19th century (FE model 6, table 1) are not significantly associated with public education expenditure. This is also the case, if the final model identified by the PTSCS method (model 5 and 6 respectively in table 1) is calculated (model 6, table 2). Hence, the results of different model specification are highly consistent in the multilevel context. Finally, and similar to PTSCS, we find high autocorrelation in the model, which is corrected for in the error terms at level 1. As for the model fit, the joint model 5 is now best in parsimoniously explaining cantonal education expenditure.

2.5 Comparison of the results and methodological implications

Formally, the PTSCS- and the multilevel model are quite similar (cf. equations 3 and 4 respectively, and equation 8). Actually, the main difference between the models can be found in the structure of the error term. First and foremost, the multilevel model incorporates an additional residual component which is μ_{0j} and accounts for country-specific random intercepts. This means that compared to PTSCS the multilevel model allows time-series coefficients "an extra stochastic component, an additional uncertainty which influences inferences" (Western 1998: 1234). Hence, as the basic idea of multilevel analysis suggests, the difference between PTSCS and MLA is that the latter allows for cantonal heterogeneity, which does not exist in the PTSCS setup.

Table 2 Multilevel models for the explanation of public education expenditure (mean-corrected)

	Empty Model	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6 (PTSCS)
Pupils per cantonal population		.06 (.07)					.06 (.07)
Proportion of foreign pupils		.05 (.04)					.05 (.04)
Sector 1		-.04 (.05)					
Urbanization		.40*** (.14)				.32** (.15)	.26* (.15)
Leftists in government			.04 (.03)				.03 (.03)
CVP in government			.01 (.06)				
Union			.13** (.05)			.11** (.05)	
Direct democracy				.02 (.09)			.04 (.09)
Governmental concordance				.03 (.02)			
Local autonomy				-.59*** (.17)		-.43** (.17)	-.42** (.17)
Strength of Catholics 19 th cent.					-.26* (.15)	-.04 (.14)	-.05 (.13)
Efforts in education policy 1880					.05 (.14)		
Language					-.32** (.15)	.04 (.15)	-.03 (.16)
constant	-.12 (.17)	-.08 (.13)	-.12 (.16)	-.10 (.13)	-.12 (.15)	-.07 (.12)	-.08 (.11)
<i>Between-variance of cantons (σ_{μ}^2)</i>	.64*** (.20)	.37*** (.14)	.62*** (.19)	.40*** (.13)	.48*** (.15)	.31** (.10)	.26** (.09)
<i>Between-variance of between time-points within cantons (σ^2)</i>	.18*** (.03)	.18*** (.03)	.17*** (.03)	.17*** (.03)	.18*** (.03)	.17*** (.03)	.18*** (.03)
<i>N</i>	624	624	624	624	624	624	624
<i>Autocorrelation level 1</i>	.798	.798	.798	.798	.798	.798	.798
<i>-2loglikelihood</i>	148.1	136.3	140.00	135.7	141.6	126.6	127.5
<i>AIC</i>	154.1	150.3	152.0	147.7	153.6	142.6	149.5

Notes: Non-standardized coefficients; standard-errors in brackets; time-independent variables, i.e. variables explaining contextual level variance, are highlighted (see Notes Table 1). Estimation with iterative generalized least squares (Goldsmith 1995). A first-order autocorrelation process was modeled in order to correct for autocorrelation at level 1 (MLwiN TS macro).

From a substantial point of view both methods show that structural and institutional characteristics are most important for the explanation of public education expenditure in the cantons (see tables 1 and 2). Particularly, cantonal education expenditure is high, where an urbanized structure demands high investments in education. In contrast, a high degree of decentralized decision making structures (e.g. substantial local autonomy in education) hinders extensive public education expenditure. Other factors such as the number of (foreign) pupils, direct democracy, the strength of Catholicism at the end of the 19th century, and the linguistic region tend to somehow influence cantonal education expenditure, the results are however not consistent and obviously depend on the model specification and estimation method. These inconsistent results may be due to fact that the two approaches handle heterogeneity differently. While in PTSCS heteroscedasticity and spatial autocorrelation is corrected for by calculating panel-corrected standard errors, in multilevel analysis heteroscedasticity is not corrected for, but modelled by distinguishing two levels of analysis. In other words: It is seen as "substance" and not as "nuisance" (Jones 1997).

Moreover, our results indicate that the multilevel design provides some clear advantages. The first aspect concerns the cross-sectional and time-related variance respectively. Most of our explanatory variables do not change much over time or are even constants, as it is typical for institutional and cultural variables in political research. This means that, logically, our model should be better in explaining cross-sectional variance than the development over time. Empirically, this can be best seen in the multilevel models: While the final models explain a substantial part of the variance between the cantons, namely about 60 percent, the variance within the cantons, i.e. the development over time, is not reduced at all. This means that the model is actually not able to explain variations in public education expenditure over time. Against the background that our model includes a quite typical set of explanatory variables for comparative political research these findings imply that one of the main advantages of pooling, i.e. the modeling of "*dynamics* inherent in the panel" (Plümper et al. 2005: 334), must be indeed questioned. Second, our results confirm the claim made by Shor et al. (2007) that multilevel models generally provide a better model fit. A glance at the AIC for the different models confirms that the multilevel models do better in explaining the differences in cantonal education expenditure and should therefore be preferred. This is also the case if we model fixed unit effects in the PTSCS context. Moreover,

different to PTSCS MLA produces a final model which exhibits the best model fit, which again speaks for the higher sensibility and consistency of multilevel models.¹⁶

3 Conclusion

The aim of this paper was to compare multilevel analysis and pooled time-series cross-section designs in the analysis of time and space in comparative political research. It has to be mentioned that by means of our analyses we are not able to actually *prove* whether MLA performs better than PTSCS, since such a comparison would require that the real error structures of the data are known which is normally not the case in empirical data. However, while such a systematic evaluation has recently been provided by Shor et al. (2007), this was not the aim of this paper. Rather, our analyses allow for some implications in terms of practical application of PTSCS and MLA respectively. From the discussion of the two approaches and the empirical examples we can make the following remarks that may be important for empirical analyses of time-serial cross-section data in comparative political research:

Generally, a simple two-level model with repeated measures (level 1) nested within units (level 2) and correlated error terms roughly corresponds to a basic pooled time-series cross-sectional model. The main difference is that the multilevel model allows for random intercepts. This can be seen as "partial pooling where the estimated parameter has the potential to borrow strength from other parameters in the model." (Shor et al. 2007: 4f.). As a result, both methods produce quite similar results for our data, implying that urbanized structures and high local autonomy are crucial for explaining cantonal education expenditure. For the rest, multilevel analysis is more conservative, but also more consistent in that it finds no further effects where some PTSCS-specifications do, thus, providing inconclusive results. Goodness of fit measures (AIC) thereby indicate that multilevel models better explain the variance in the data and should therefore be preferred.

16 It has to be mentioned that in this paper very simple models have been chosen. Both approaches would offer some more sophisticated procedures that would probably lead to better model fits (but probably also to more problems).

Comparing the two designs, MLA moreover offers some advantages particularly with regard to conceptual and theoretical aspects.

Firstly, a multilevel design is theoretically better suitable for modeling the effect of time-invariant variables (e.g., institutions) which often take center stage in comparative political research. If (almost) time-invariant variables are integrated into a PTSCS design the corresponding hypothesis is that a particular (not or very slowly changing) variable parameter at time t (or $t-x$ if a time-lag is incorporated) influences the dependent variable at time t . However, such a theoretical link between (time-invariant) explanatory factors and the dependent variables is often questionable. This is most pronounced in the presence of serial autocorrelation, when such a model implies that specific time-invariant characteristics explain some short-term changes in the dependent variable. In such a context a multilevel design is theoretically more sensible because it is based on the assumption that time-invariant factors build a unit-specific context that influences all observations of this unit in a particular way.

Secondly, a multilevel design allows for a clear distinction between cross-sectional effects and time effects.¹⁷ It is possible to analyze to what extent cross-sectional differences and developments over time each account for the overall variance in the data. Concerning substantial interpretation of the estimation results this feature is very useful: In our case it shows that the models do quite well in explaining cross-sectional variance, while they are not able to explain developments over time.

Thirdly, heterogeneity in the data is not just corrected for as in the pooled time-series cross-section analysis, but is modeled. Conceptually this is a more elegant way to handle the statistical problem, because heterogeneity is not just seen as a problem that has to be eliminated, but as an important (realistic) characteristic of the data that should be adequately modeled.

A further advantage is that while in this paper we applied a very simple model in order to make comparison easier, MLA would allow for various model extensions, be it the introduction of an additional level (e.g., an intervening municipality level), the estimation of random slopes or even cross-level interactions (e.g., the effect of urbanization on the strength of relationship between for-

17 In the case presented in this paper time effects correspond to the relative changes in the cantons' position over time.

eign pupils and education expenditure), or the calculation of growth curves in order to model particular developments over time (Bryk/Raudenbush 1987; Cnaan et al. 1997; Hox 2000; Huggins/Loesch 1998; Luke 2004; Yang/Goldstein 1996). In a multilevel design more sophisticated models are relatively easy to apply, which could be especially helpful when working with data reflecting changes over time.

However, multilevel analysis also has its shortcomings in practical applications. First and foremost, the necessity of large and particularly structured data sets has to be mentioned. Indeed, in order to use MLA, data has not only to be measured at different levels, but additionally each group at all levels should consist of at least 25 observations in order to get valid estimations for multivariate analyses.¹⁸ Obviously, in comparative political research such data is still quite rare. Furthermore, the analysis and handling of crucial statistical pitfalls is much better developed in the methodological discussion of PTSCS than it is in the multilevel context. Mainly in relation to serial autocorrelation, which is a crucial aspect in case of repeated measurements, procedures to correct for the correlated errors are much better elaborated and standardized in pooled time-series cross-section designs than in multilevel analysis. When carrying out a MLA with data that includes developments over time, it is useful to consider the methodological debate on non-stationarity and serial autocorrelation in the PTSCS-context.

To conclude, we would propose the following procedure for the analysis of time and space in comparative political research: If a large data set satisfying the requirements for multilevel analysis is available, the implementation of a multilevel design is recommended due to its statistical, conceptual and theoretical strengths discussed above. In smaller data samples a PTSCS design might generally be more appropriate. In both situations, however, we think a triangulation of methods worthwhile. In the multilevel case additional analyses from the PTSCS context help clarifying and understanding problems of autocorrelation and non-stationarity. In smaller data samples the calculation of empty multilevel models allows at least for an approximate distinction between longitudinal and cross-sectional variances in the data. Hence, the combination of pooled time-series cross-

18 If one is only interested in an analysis of variances at different levels the case-restriction is less problematic. If we have only few cases, but many time points within these cases it makes sense to control for the effects of the level 2-variance with MLA. However, in this case we would not specify level 2 with independent variables.

section analysis and multilevel analysis offers the possibility of profiting from the advantages of both methods, but also to compensate for their weaknesses.

References

- Année politique Suisse, various years: Schweizerische Politik. Bern: Institut für Politikwissenschaft.
- Archer, M. S., 1979: *Social Origins of Educational Systems*. London/Beverly Hills: Sage.
- Bassand, M., 1988: *Les enjeux de l'urbanisation*. Bern/Frankfurt a. M./Paris: Lang.
- Beck, N. and J. Katz, 1995: What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89(3): 634-647.
- Boix, C., 1997: Political Parties and the Supply Side of the Economy. The Provision of Physical and Human Capital in Advances Economies 1960-90. *American Journal of Political Science* 41(3): 814-845.
- Bryk, A. S. and S. W. Raudenbush, 1987: Application of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin* 101(1): 147-158.
- Bühlmann, M., 2006: Politische Partizipation im kommunalen Kontext. Der Einfluss lokaler Kontexteigenschaften auf individuelles politisches Partizipationsverhalten. Bern: Haupt.
- Bullen, N., K. Jones and C. Duncan, 1994: Modelling complexity-analysing between-individual and between-place variation. A multilevel tutorial. Mimeo. Portsmouth: University of Portsmouth, Department of Geography.
- Burnham, K. P. and D. R. Anderson, 2004: Multimodel Inference. Understanding AIC and BIC in Model Selection. *Sociological Methods Research* 33: 261-304.
- Cameron, D. R. and R. I. Hofferbert, 1974: The Impact of Federalism on Education Finance. A Comparative Analysis. *European Journal of Political Research* 2: 225-258.
- Castles, F. G., 1989: Explaining public education expenditure in OECD nations. *European Journal of Political Research* 17: 431-448.
- Castles, F. G., 1999: *Comparative Public Policy. Patterns of Post-war Transformation*. Cheltenham/Northampton: Edward Elgar Publishing Limited.
- Cnaan, A., N. M. Laird and P. Slasor, 1997: Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine* 16: 2349-2380.
- Ditton, H., 1998: *Mehrebenenanalyse. Grundlagen und Anwendungen des Hierarchisch Linearen Modells*. Weinheim/München: Juventa.
- Eidgenössische Finanzverwaltung (EFV), various years: *Öffentliche Finanzen der Schweiz*. Neuchâtel: Bundesamt für Statistik.
- Fernandez, R. and R. Rogerson, 1997: *The Determinants of Public Education Expenditure. Evidence from the States 1950-1990*. NBER-Working Paper 5995. Cambridge: National Bureau of Economic Research.
- Freitag, M. and M. Bühlmann, 2003: Die Bildungsfinanzen der Schweizer Kantone. Der Einfluss sozioökonomischer Bedingungen, organisierter Interessen und politischer Institutionen auf die Bildungsausgaben im kantonalen Vergleich. *Swiss Political Science Review* 9(1): 139-168.
- Goldstein, H., 1987: *Multilevel Models in Educational and Social Research*. London/New York: Oxford University Press.
- Goldstein, H., 1995: *Multilevel Statistical Models*. New York: Halstead Press.

- Gujarati, D. N., 1995: *Basic Econometrics*. Third Edition. New York: McGraw-Hill International Editions.
- Hadri, K., 2000: Testing for stationarity in heterogeneous panel data. *Econometrics Journal* 3: 148-161.
- Hega, G. M., 1999: *Consensus Democracy? Swiss Education Policy between Federalism and Subsidiarity*. New York: Peter Lang.
- Heidenheimer, A. J., 1999: External and Domestic Determinants of Education Expansion. How Germany, Japan, and Switzerland Have Varied. *Governance* 6(2): 194-219.
- Hibbs, D., 1977: Political Parties and Macroeconomic Policy. *American Political Science Review* 71: 1467-1487.
- Hox, J. J., 1995: *Applied Multilevel Analysis*. Den Haag: CIP-Gegevens Koninklijke Bibliotheek.
- Hox, J. J., 2000: Multilevel Analyses of Grouped and Longitudinal Data. Pp. 15-32 in: T. D. Little, K. U. Schnabel and J. Baumert (Eds.): *Modeling longitudinal and multilevel data*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hox, J. J., and I. G. Kreft, 1994: Multilevel Analysis Methods. *Sociological Methods & Research* 22(3): 283-299.
- Hsiao, C., 2003: *Analysis of Panel Data*. 2. Edition. Cambridge: Cambridge University Press. <http://assets.cambridge.org/0521818559/sample/0521818559WS.pdf> (21.8.2003).
- Huggins, R. M. and D. Z. Loesch, 1998: On the analysis of mixed longitudinal growth data. *Biometrics* 54: 583-95.
- Hunziker, O., 1882: *Geschichte der Schweizerischen Volksschule, in gedrängter Darstellung*. Vol. 3. Zürich: Friedrich Schulthess.
- Im, K. S., H. M. Pesaran and Y. Shin (Eds.): Testing for unit roots in heterogeneous panels. *Journal of Econometrics* 115(1): 53-74.
- Jones, K., 1997: Multilevel Approaches to Modelling Contextuality. From Nuisance to Substance in the Analysis of Voting Behaviour. Pp. 19-43 in: G. P. Westert and R. N. Verhoeff (Eds.): *Places and people. Multilevel modelling in geographical research*. Utrecht: The Royal Dutch Geographical Society.
- Jones, K. and C. Duncan, 1996: People and places. The multilevel model as a general framework for the quantitative analysis of geographical data. Pp. 79-104 in: P. Longley and M. Batty (Eds.): *Spatial Analysis: Modelling in a GIS Environment*. New York: John Wiley and Sons.
- Jones, K. and C. Duncan, 1996: People and places. The multilevel model as a general framework for the quantitative analysis of geographical data. In: P. Longley and M. Batty (Eds.): *Spatial Analysis: Modelling in a GIS Environment*. New York: John Wiley and Sons.
- Kittel, B., 1999: Sense and Sensitiveness in Pooled Analysis of Political Data. *European Journal of Political Research* 35(2), 225-253.
- Kittel, B. and H. Winner, 2002: How Reliable is Pooled Analysis in Political Economy? The Globalization-Welfare State Nexus Revisited. Max-Planck-Institut für Gesellschaftsforschung: Discussion Paper 2(3).
- Ladner, A., 1994: Finanzkompetenzen der Gemeinden – ein Überblick über die Praxis. Pp. 64-85 in: F. Eng, A. Glatthard and B. H. Koenig (Eds.): *Finanzföderalismus*. Zürich: Emissionszentrale der Schweizer Gemeinden.
- Levin, A. T. and C.-F. Lin, 1993: Unit Root Tests in Panel Data. New Results. University of California at San Diego: San Diego Discussion Paper (93-56).
- Linder, W., 1994: *Swiss Democracy. Possible Solutions to Conflict in Multicultural Societies*. New York: St. Martin's Press.

- Linder, W., 2005: Schweizerische Demokratie. Institutionen, Prozesse, Perspektiven. 2nd Edition. Bern: Paul Haupt.
- Lijphart, A., 1999: Patterns of Democracy. Government forms and Performance in Thirty-Six Countries. New Haven/London: Yale University Press.
- Luke, D. A., 2004: Multilevel Modeling. Thousand Oaks: Sage.
- Pennings, P., H. Keman and J. Kleinnijenhuis, 1999: Doing Research in Political Science. An Introduction to Comparative Methods and Statistics. London/Thousand Oaks/New Delhi: Sage Publications.
- Plümper, T. and V. E. Troeger, 2007: Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis* 15: 124-139.
- Plümper, T., V. E. Troeger and P. Manow, 2005: Panel data analysis in comparative politics. Linking method to theory. *European Journal of Political Research* 44: 327-354.
- Podestà, F., 2002: Recent Developments in Quantitative Comparative Methodology. The Case of Pooled Time Series Cross-Section Analysis. Universität Brescia: Discussion Paper (3-02). <http://www.eco.unibs.it/segdss/paper/pode202.pdf> (21.8.2003).
- Popper, K. R., 1992: Die offene Gesellschaft und ihre Feinde. Vol. 2. Tübingen: J.C.B Mohr.
- Poterba, J. M., 1999: Demographic Structure and the Political Economy of Public Education. Cambridge: Working Paper 5677 of National Bureau of Economic Research.
- Rasbash, J., W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford and T. Lewis, 2002: A user's guide to MLwiN. Version 2d for use with MLwiN 1.10. London: Centre for Multilevel Modelling, Institute of Education, University of London.
- Rohwer, G., 1998: Mehrebenenmodelle und das Mikro-Makro-Problem in der empirischen Sozialforschung. Bochum: Ruhr Universität, Fakultät für Sozialwissenschaften.
- Schaltegger, C., 2001: Ist der Schweizer Föderalismus kleinräumig? *Swiss Political Science Review* 7(1): 1-18.
- Schmidt, M. G., 2000: Die sozialpolitischen Nachzüglerstaaten und die Theorien der vergleichenden Staatstätigkeitsforschung, Pp. 22-36 in: H. Obinger and U. Wagschal (Eds.): Der gezügelte Wohlfahrtsstaat. Sozialpolitik in reichen Industrienationen. Frankfurt am Main/New York: Campus.
- Schmidt, M. G., 2002: Warum Mittelmaß? Deutschlands Bildungsausgaben im internationalen Vergleich. *Politische Vierteljahresschrift* 43 (1): 3-19.
- Shor B., J. Bafumi, L. Keele and D. Park, 2007: A Bayesian Multilevel Modeling Approach to Time Series Cross-Sectional Data. *Political Analysis* 15: 165-181.
- Steenbergen, M. R. and B. S. Jones, 2002: Modelling Multilevel Data Structures. *American Journal of Political Science* 46(1): 218-237.
- Stoker, L. and J. Bowers, 2002: Designing multilevel studies. Sampling voters and electoral contexts. Pp. 77-109 in: Mark N. Franklin and C. Wlezien (Eds.): *The Future of Election Studies*. Amsterdam: Pergamon.
- Stutzer, A., 1999: Demokratieindizes für die Kantone der Schweiz. Zürich: Working Paper 23 of Institutes for Empirical Research in Economic.
- Teachman, J. and K. Crowder, 2002: Multilevel Models in Family Research. Some Conceptual and Methodological Issues. *Journal of Marriage and Family* 64(2): 280-294.
- Vatter, A., 2002: Kantonale Demokratien im Vergleich. Entstehungsgründe, Interaktionen und Wirkungen politischer Institutionen in den Schweizer Kantonen. Opladen: Leske & Budrich.

- Vatter, A. and M. Freitag, 2002: Die Janusköpfigkeit von Verhandlungsdemokratien. Zur Wirkung von Konkordanz, direkter Demokratie und dezentralen Entscheidungsstrukturen auf den öffentlichen Sektor der Schweizer Kantone. *Swiss Political Science Review* 8(2): 53-80.
- Vatter, A. et al., 2004: Politische, soziale und ökonomische Daten zu den Schweizer Kantonen, 1983-2000. Bern: Institut für Politikwissenschaft.
- Western, B., 1998: Causal Heterogeneity in Comparative Research. A Bayesian Hierarchical Modelling Approach. *American Journal of Political Science* 42 (4): 1233-1259.
- Wilensky, H., 1981: Leftism, Catholicism, and Democratic Corporatism. The Role of Political Parties in Recent Welfare State Development. Pp. 345-382 in: P. Flora and A. J. Heidenheimer (Eds.): *The Development of Welfare State in Europe and America*. New Brunswick: Transaction.
- Yang, M., and H. Goldstein, 1996. Multilevel Models for longitudinal data. Pp. 191-220 in: U. Engel and J. Reinecke (Eds.): *Analysis of Change, Advanced Techniques in Panel data Analysis*. Berlin: Walter de Gruyter.

Contact Addresses:

Isabelle Stadelmann-Steffen
University of Konstanz
Department of Politics and Management
P.O. Box D84
D-78457 Konstanz
Isabelle.Stadelmann-Steffen@uni-konstanz.de

Marc Bühlmann
Universität Zürich
Zentrum für Demokratie
Küttigerstraße 21
CH-5000 Aarau
Marc.buehlmann@zda.uzh.ch

Appendix Variables, Hypotheses, and Operationalization

Variable	Hypothesis (ceteris paribus)	Expected Relationship	Operationalization / Source
<i>Dependent variable</i>			
Public education expenditure per capita, mean corrected			Cantonal education expenditure in CHF per capita (for elementary school and vocational education, less federal contributions), according to the functional classification of the expenditure of cantons and communes, annual mean corrected. Source: EFV (various years).
<i>Socio-economic variables</i>			
Pupils per cantonal population	The more pupils a canton has, the higher is cantonal public education expenditure.	+	Number of pupils in elementary school per cantonal population. Source: Statistisches Jahrbuch (various years).
Proportion of foreign pupils	The higher the share of foreigners in school, the higher is expenditure for educational integration.	+	Share of foreigners in elementary schools, in percent. Source: Statistisches Jahrbuch (various years).
Employment in the third sector	The higher the employment in the service sector, the higher is cantonal education expenditure.	+	Share of employed in the third sector according to the federal census of enterprises 1991, in percent. Source: Swiss Federal Statistical Office.
Urbanization	The higher urbanization, the higher is cantonal public education expenditure.	+	Share of residents living in urban areas, in percent. Source: Own calculation on the basis of Bassard (1988).
<i>Party-political variables</i>			
Strength of leftist parties in the government	The higher the share of leftist parties in the government, the higher is public education expenditure in a canton.	+	Share of seats in cantonal government held by SPS, DSP and the Green parties. Source: Année Politique Suisse (various years).
Strength of Christian-democratic parties in the government	The higher the share of Christian democratic parties in the cantonal government, the lower the public education expenditure turn out to be.	-	Share of seats in cantonal government held by Catholic parties (CVP and CSP). Source: Année Politique Suisse (various years).
Strength of unions	The better unions in a canton are organized, the more this canton spends for public education.	+	Union-membership in percent of total employment. Source: Own calculation on the basis of the Swiss federation of trade unions and BAK Basel.

Variable	Hypothesis (ceteris paribus)	Expected Relationship	Operationalization / Source
<i>Institutional variables</i>			
Local autonomy	The more decentralized the decision making structures in a canton are, the lower is public education expenditure.	-	Degree of communal autonomy in relation to the Federation and the Canton (index) according to Ladner (1994: 81) (0 = no autonomy at all; 10 = very substantial autonomy).
Direct democracy	The more extensive direct democratic rights in a canton are, the lower is public education expenditure.	-	Index of direct democracy according to Stutzer (1999): 1978-1991: value for 1970; 1992-1995: value for 1992; 1996-2003: values for 1996.
Consensus democracy	The higher the consensual alignment of the government is, the more a canton spends for public education.	+	Summarized percentages of votes of the governing parties (if the majority rule is used for parliamentary elections (Grisons, Uri, Appenzell Inner- and Outer-Rhodes) the summarized share of parliamentary seats of the governing parties is calculated instead). Source: <i>Année Politique Suisse</i> (various years).
<i>Historical and cultural variables</i>			
Strength of the Catholic Church	The stronger the influence of the Catholic Church on education was at the beginning of the education expansion (end 19th century), the lower is public education expenditure today.	-	Estimation of the influence of the catholic church when the cantonal educational system was built up. 1 = weak influence; 2 = medium influence; 3 = strong influence. Source: Hunziker (1882: 304-367).
Efforts in education policy 1880	The higher the efforts of the government in education were around 1880, the higher are investments in public education still today.	+	Efforts in public education of the cantonal governments around 1880. Index of ranking points based on five indicators: expenditure per capita, pupils per capita, pupil-teacher-quota, the introduction of an education law, number of school weeks. 1=little effort; 2=medium effort; 3=strong effort. Source: Hunziker (1882: 304-367).
Language	German-speaking cantons spend less for public education than the Latin cantons.	-	Dummy-variable. 1 = majority of German-speaking inhabitants, otherwise =0. Source: Swiss federal statistical office, <i>Statistisches Jahrbuch der Schweiz</i> .

Note: The data used is based on the data set of Vatter et al. (2004).

Persönliche Codes „reloaded“

Personal Codes “reloaded”

Andreas Pöge

Zusammenfassung

Bei Längsschnittuntersuchungen mit sensiblen Fragestellungen werden die Fragebogenzuordnungen zwischen den einzelnen Erhebungswellen aus Datenschutzgründen oftmals mit persönlichen und selbstgenerierten Codes vorgenommen – so auch in dem DFG-Projekt „Kriminalität in der modernen Stadt“. Die Ergebnisse mit dieser Zuordnungsmethode, insbesondere bezüglich der Probleme bei der Durchführung, der Ausschöpfungsquote und der Verzerrung der resultierenden Paneldaten, waren Gegenstand der Ausführungen in dem Artikel „Persönliche Codes bei Längsschnittstudien: Ein Erfahrungsbericht“ (Pöge 2005b), der sich auf die Münsteraner Teilstudie bezog. Mittlerweile wurde die Studie auf den Erhebungsort Duisburg ausgedehnt, wobei die Erkenntnisse aus Münster zu Modifikationen des Verfahrens führten. Hier sollen die teils sehr deutlichen Verbesserungen der Datenqualität aufgezeigt werden, die damit in Duisburg erreicht werden konnten.

Abstract

In panel studies the assignments of questionnaires are often carried out between the waves using personal codes – also in the DFG project “Juvenile delinquency in modern town”. The results with this allocation method, in particular with regard to the problems by the realisation, the exhaustion rate and the bias of the resultant panel data, have been published in the article “personal codes with profile studies: A report of experience” (Pöge 2005b) which refers to the study in Münster. Meanwhile the study was expanded to Duisburg and the findings from Münster led to modifications of the procedure. Here the very clear improvements in the exhaustion rate and smaller bias in Duisburg shall be described.

1 Einleitung

Eine grundlegende Problematik bei Panelstudien stellt die Zuordnung der Fragebögen aus den einzelnen Erhebungswellen dar, die von denselben Personen ausgefüllt wurden. Ist aus Gründen des sensiblen Befragungsgegenstandes (beispielsweise Krankheitsverläufe, kriminelle Verhaltensweisen etc.) oder des besonderen Persönlichkeitsschutzes der Befragten (Studien mit minderjährigen Teilnehmerinnen und -nehmern) besonderes Augenmerk auf Datenschutzbelange zu richten und daher die Erhebung und Speicherung beispielsweise von Adressen, Versicherungs- oder Matrikelnummern unerwünscht, bietet sich ein Verfahren mit persönlichen und selbstgenerierten Codes an (Schnell/Bachteler et al. 2006: 129). Die einzelnen Codefragen sollten dabei zeitstabile Merkmale mit persönlichem Bezug zu den einzelnen Respondenten aufweisen – aus diesem Grund wird oftmals auf Buchstaben des eigenen Namens oder dem Verwandter, auf Geburtsdatumsziffern oder andere persönliche Merkmale zurückgegriffen (Grube/Morgan et al. 1989: 159). Die Länge der Codes umfasst dabei meist sechs bis zehn Zeichen bzw. Ziffern (Schnell/Bachteler et al. 2006: 129). Problematisch an solch einem Verfahren sind Fehler bei der Erzeugung der Codes, da dann die Fragebögen nur noch unter erschwerten Bedingungen, nämlich mithilfe von fehlertoleranten Verfahren einander zugeordnet werden können und somit teils erheblicher Datenverlust droht.

Die folgenden Ausführungen stellen eine Fortschreibung des in der ZA-Information 56 (S. 50–69) veröffentlichten Artikels „Persönliche Codes bei Längsschnittstudien: Ein Erfahrungsbericht“ dar. Wie dort vorgestellt, wurden bzw. werden im Zuge der Panelstudie *Kriminalität in der modernen Stadt*¹ mithilfe von solchen persönlichen und selbstgenerierten Codes, die über Codeblätter abgefragt wurden, Paneldatensätze erstellt. In der Münsteraner Teilstudie kam es im Hinblick auf die Fragebogenzuordnung der einzelnen Jahre zu Problemen, die zusammen mit der Problemlösestrategie und den Ergebnissen im genannten Artikel dargestellt wurden. Parallel wurde ab dem Jahr 2002 begonnen, in *Duisburg* einen Paneldatensatz aufzubauen, wobei die Erfahrungen mit den Münsteraner Codes berücksichtigt wurden. Auf Grundlage des Duisburger Vier-Wellen-Datensatzes 2002 bis 2005 sollen die Modifikationen im Codeverfahren aufgezeigt werden, die zu einer deutlich besseren Datenqualität in den Duisburger Paneldaten führen.

1 Projektleitung: Prof. Dr. Klaus Boers, Institut für Kriminalwissenschaften, WWU Münster und Prof. Dr. Jost Reinecke, Fakultät für Soziologie, Universität Bielefeld.

2 Schülerbefragungen in Duisburg

Bei den im Jahr 2002 begonnenen Untersuchungen in Duisburg wurden Schülerinnen und Schüler der siebten Jahrgangsstufe befragt, die eine Sonder-, Haupt-, Real-, Gesamtschule oder ein Gymnasium besuchten.² In den Wellen zwei bis drei wurden die Probanden der nun achten und neunten Jahrgangsstufe wiederbefragt, in der vierten Welle zusätzlich zu den Jugendlichen der nunmehr zehnten Klassen auch diejenigen, die nicht versetzt worden waren, also im Jahr 2005 die neunte Jahrgangsstufe wiederholten.

Die in den vier Querschnittsbefragungen gewonnenen Daten zeigen – nach den durchgeführten Plausibilitätskontrollen – als Stichprobengrößen 3.243 bis 3.411 verwertbare Fälle. Vergleicht man diese realisierten Fallzahlen mit den offiziellen Belegungszahlen der an den Befragungen *teilnehmenden Schulen* aus den jeweiligen Schulstatistiken, so ergeben sich sehr gute Rücklaufquoten zwischen 85 und 92%³.

3 Gestaltung des Codeblattes

Um den für die Fragebogenzuordnung zwischen den Wellen erforderlichen Code zu erzeugen, sollte jede befragte Schülerin bzw. jeder Schüler in jedem Jahr ein Codeblatt mit verschiedenen Codefragen ausfüllen, so dass bei einem stabilen Antwortverhalten die Codeblätter und die über eine eindeutige Nummer mit ihnen verknüpften Fragebögen einander zugeordnet werden könnten. Als grundlegende Problembereiche waren in den früheren Münsteraner Befragungen die Probleme der Identifizierung und der Reproduktion aufgetreten (Pöge 2005a: 7ff.; Pöge 2005b). Das heißt, dass zum einen nicht alle in einem Querschnitt aufgetretenen Codes eindeutig waren und zum anderen, dass sich ein nicht unerheblicher Teil der Codes, die zwischen den Jahren hätten gleich sein müssen, in der Praxis unter-

2 Die Befragungen fanden als schriftliche Fragebogeninterviews im Klassenverband statt. Sie wurden von Interviewerinnen und Interviewern begleitet, die den Jugendlichen zu Beginn der doppelstündigen Befragungsphase Ausfüllanweisungen und Informationen zur Studie, zu Datenschutzfragen und zum Ablauf gaben.

3 Zur Dokumentation der Duisburger Querschnittsstudien der Jahre 2002 bis 2005 siehe Motzke/Brondies (2004); Brondies (2004); Hilfert (2005); Kunadt (2006), zu der Panelstudie siehe Pöge (2007). Die Befragungen werden seit dem Jahr 2005 im jährlichen Rhythmus weiter fortgeführt, wobei (zunächst nur teilweise) auf eine postalische Befragungsform umgestellt wurde, da die Jugendlichen nicht mehr alle eine allgemeinbildende Schule besuchen.

schied. Schon im letzten Befragungsjahr in Münster erfolgte daher eine grundlegende Überarbeitung des Codeblattes, die für die Duisburger Studie ab der zweiten Welle übernommen und weitergeführt wurde. Es wurde insbesondere versucht, die Schwierigkeit der Codefragen noch weiter herabzusetzen und Layoutmängel zu beseitigen.

Bei den Befragungen in Duisburg wurde zunächst wie in Münster ein fünfstelliger Code verwendet, bei dem allerdings aufgrund der Erfahrungen zum Teil bereits einfachere Fragen mit größerer Antwortvarianz zum Einsatz kamen.⁴ So mussten zum Beispiel nicht mehr der letzte Buchstabe der Haarfarbe des Vaters, sondern der erste Buchstabe des Vornamens des Vaters angegeben werden.⁵ Im Jahr 2003 wurde auf Grundlage der bisher gewonnenen Erkenntnisse der Duisburger Code um eine Frage erweitert und das Layout des gesamten Codeblattes überarbeitet. Im Zuge dieser Umstellung musste kein handschriftliches Ausfüllen mehr erfolgen, sondern alle Antwortvorgaben wurden zum Ankreuzen aufgeführt. Ab dem Befragungsjahr 2003 wurden außerdem zusätzlich Fragen nach einer Befragungsteilnahme im Vorjahr, einem Schulwechsel sowie einer eventuellen Nichtversetzung im vergangenen Jahr gestellt. Da im letzten Erhebungsjahr 2005 auch die nicht versetzten Schülerinnen und Schüler befragt wurden, musste auf dem Codeblatt eine Zusatzfrage nach der entsprechenden Jahrgangsstufe gestellt werden (siehe Abbildung 1). In Ergänzung zu diesen acht bzw. neun Fragen standen für die Fragebogenzuordnungen das Geschlecht der Befragten sowie die Schule, die mit einer Kennnummer erhoben wurde, zur Verfügung. Zusätzlich zu den Hinweisen auf dem Codeblatt selber wurden von den jeweiligen Interviewerinnen und Interviewern Ausfüllanweisungen gegeben. Zu nennenswerten Problemen oder Schwierigkeiten kam es beim Ausfüllen der Blätter nicht.

Zum Problembereich der Identifizierung ist mit den abgeänderten Fragen und deren größerer Antwortvarianz sowie der zusätzlichen Codestelle festzuhalten, dass in Duisburg ab dem Jahr 2003 fast alle Codes eindeutig sind. In Münster waren in allen vier Erhebungsjahren nur bei rund drei Vierteln der Schülerinnen

4 Probleme mit bewussten Falschangaben oder unbewussten Fehlern bei der Ausfüllung des Codeblattes tauchen selbstverständlich auch in Duisburg auf. Allerdings zeigten sich zumindest die unbewussten Fehler durch die leichtere Beantwortbarkeit der Fragen als positiv beeinflussbar.

5 Durch die wesentlich größere Datengrundlage in Duisburg war es möglich, wie ursprünglich auch in Münster geplant, oftmals die *ersten* Buchstaben der Antworten auf die einzelnen Codefragen notieren zu lassen ohne eine Verletzung der Anonymität durch eventuell mögliche Identifizierung einzelner Personen befürchten zu müssen. Das Notieren dieser jeweils ersten Buchstaben scheint in der Tat eine wesentliche Erleichterung für die ausfüllenden Personen zu sein.

und Schülern eindeutige Codes aufgetreten – bei rund einem Viertel der Befragten wiesen unterschiedliche Personen einen gleichen Code auf und waren somit ohne Hinzunahme weiterer Informationen nicht eindeutig zu identifizieren. In Duisburg hingegen haben bei dem modifizierten fünfstelligen Code aus dem Jahr 2002 immerhin rund 80%, bei dem sechststelligen Code, der in allen weiteren Befragungsjahren Anwendung fand, fast alle Befragten einen eindeutigen Code (siehe Tabelle 1). Insofern kann das Problem der Identifizierung durch die geschilderten Codemodifizierungen als gelöst betrachtet werden.

Abbildung 1 Das Codeblatt in Duisburg 2005

*Bitte kreuze bei jeder der sechs Fragen immer nur ein Feld an!
Wenn du eine der Fragen überhaupt nicht beantworten kannst, kreuze bitte kein Feld an!*

Hier nun die sechs Fragen zur Erstellung deines persönlichen Codes:

1	Bitte kreuze den ersten Buchstaben des Vornamens deines Vaters (oder einer Person, die für dich einen Vater am nächsten kommt) an. (z. B. <u>A</u> nton, <u>B</u> ernd, <u>H</u> ans-Peter usw.) <input type="checkbox"/> a <input type="checkbox"/> b <input type="checkbox"/> c <input type="checkbox"/> d <input type="checkbox"/> e <input type="checkbox"/> f <input type="checkbox"/> g <input type="checkbox"/> h <input type="checkbox"/> i <input type="checkbox"/> j <input type="checkbox"/> k <input type="checkbox"/> l <input type="checkbox"/> m <input type="checkbox"/> n <input type="checkbox"/> o <input type="checkbox"/> p <input type="checkbox"/> q <input type="checkbox"/> r <input type="checkbox"/> s <input type="checkbox"/> t <input type="checkbox"/> u <input type="checkbox"/> v <input type="checkbox"/> w <input type="checkbox"/> x <input type="checkbox"/> y <input type="checkbox"/> z <input type="checkbox"/> ä <input type="checkbox"/> ö <input type="checkbox"/> ü <input type="checkbox"/> ß
2	Bitte kreuze den ersten Buchstaben des Vornamens deiner Mutter (oder einer Person, die für dich eine Mutter am nächsten kommt) an. (z. B. <u>A</u> нна, <u>B</u> eate, <u>J</u> utta, <u>M</u> aria, usw.) <input type="checkbox"/> a <input type="checkbox"/> b <input type="checkbox"/> c <input type="checkbox"/> d <input type="checkbox"/> e <input type="checkbox"/> f <input type="checkbox"/> g <input type="checkbox"/> h <input type="checkbox"/> i <input type="checkbox"/> j <input type="checkbox"/> k <input type="checkbox"/> l <input type="checkbox"/> m <input type="checkbox"/> n <input type="checkbox"/> o <input type="checkbox"/> p <input type="checkbox"/> q <input type="checkbox"/> r <input type="checkbox"/> s <input type="checkbox"/> t <input type="checkbox"/> u <input type="checkbox"/> v <input type="checkbox"/> w <input type="checkbox"/> x <input type="checkbox"/> y <input type="checkbox"/> z <input type="checkbox"/> ä <input type="checkbox"/> ö <input type="checkbox"/> ü <input type="checkbox"/> ß
3	Bitte kreuze den ersten Buchstaben deines Vornamens an (z. B. <u>M</u> ichael, <u>T</u> homas, <u>U</u> te usw.) <input type="checkbox"/> a <input type="checkbox"/> b <input type="checkbox"/> c <input type="checkbox"/> d <input type="checkbox"/> e <input type="checkbox"/> f <input type="checkbox"/> g <input type="checkbox"/> h <input type="checkbox"/> i <input type="checkbox"/> j <input type="checkbox"/> k <input type="checkbox"/> l <input type="checkbox"/> m <input type="checkbox"/> n <input type="checkbox"/> o <input type="checkbox"/> p <input type="checkbox"/> q <input type="checkbox"/> r <input type="checkbox"/> s <input type="checkbox"/> t <input type="checkbox"/> u <input type="checkbox"/> v <input type="checkbox"/> w <input type="checkbox"/> x <input type="checkbox"/> y <input type="checkbox"/> z <input type="checkbox"/> ä <input type="checkbox"/> ö <input type="checkbox"/> ü <input type="checkbox"/> ß
4	Bitte kreuze den Tag deines Geburtsdatums an (z. B. Geburtstag am 7. Januar = <u>7</u> , am 12. Mai = <u>12</u> , am 31. Oktober = <u>31</u>) <input type="checkbox"/> 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> 6 <input type="checkbox"/> 7 <input type="checkbox"/> 8 <input type="checkbox"/> 9 <input type="checkbox"/> 10 <input type="checkbox"/> 11 <input type="checkbox"/> 12 <input type="checkbox"/> 13 <input type="checkbox"/> 14 <input type="checkbox"/> 15 <input type="checkbox"/> 16 <input type="checkbox"/> 17 <input type="checkbox"/> 18 <input type="checkbox"/> 19 <input type="checkbox"/> 20 <input type="checkbox"/> 21 <input type="checkbox"/> 22 <input type="checkbox"/> 23 <input type="checkbox"/> 24 <input type="checkbox"/> 25 <input type="checkbox"/> 26 <input type="checkbox"/> 27 <input type="checkbox"/> 28 <input type="checkbox"/> 29 <input type="checkbox"/> 30 <input type="checkbox"/> 31
5	Bitte kreuze den letzten Buchstaben deiner natürlichen Haarfarbe an. (z. B. braun, <u>G</u> latz, schwarz, usw.) <input type="checkbox"/> a <input type="checkbox"/> b <input type="checkbox"/> c <input type="checkbox"/> d <input type="checkbox"/> e <input type="checkbox"/> f <input type="checkbox"/> g <input type="checkbox"/> h <input type="checkbox"/> i <input type="checkbox"/> j <input type="checkbox"/> k <input type="checkbox"/> l <input type="checkbox"/> m <input type="checkbox"/> n <input type="checkbox"/> o <input type="checkbox"/> p <input type="checkbox"/> q <input type="checkbox"/> r <input type="checkbox"/> s <input type="checkbox"/> t <input type="checkbox"/> u <input type="checkbox"/> v <input type="checkbox"/> w <input type="checkbox"/> x <input type="checkbox"/> y <input type="checkbox"/> z <input type="checkbox"/> ä <input type="checkbox"/> ö <input type="checkbox"/> ü <input type="checkbox"/> ß
6	Bitte kreuze den letzten Buchstaben deiner Augenfarbe an. (z. B. braun, grün, grau, usw.) <input type="checkbox"/> a <input type="checkbox"/> b <input type="checkbox"/> c <input type="checkbox"/> d <input type="checkbox"/> e <input type="checkbox"/> f <input type="checkbox"/> g <input type="checkbox"/> h <input type="checkbox"/> i <input type="checkbox"/> j <input type="checkbox"/> k <input type="checkbox"/> l <input type="checkbox"/> m <input type="checkbox"/> n <input type="checkbox"/> o <input type="checkbox"/> p <input type="checkbox"/> q <input type="checkbox"/> r <input type="checkbox"/> s <input type="checkbox"/> t <input type="checkbox"/> u <input type="checkbox"/> v <input type="checkbox"/> w <input type="checkbox"/> x <input type="checkbox"/> y <input type="checkbox"/> z <input type="checkbox"/> ä <input type="checkbox"/> ö <input type="checkbox"/> ü <input type="checkbox"/> ß

Hast du im letzten Jahr an der Befragung teilgenommen?
 Hast du im letzten Jahr die Schule gewechselt?
 Bist du im letzten Jahr sitzen geblieben?
 In welcher Klasse bist du?

<input type="checkbox"/> ja	<input type="checkbox"/> nein
<input type="checkbox"/> ja	<input type="checkbox"/> nein
<input type="checkbox"/> ja	<input type="checkbox"/> nein
<input type="radio"/> 9	<input type="radio"/> 10

Tabelle 1 Häufigkeiten der Codes

Häufigkeit	2002		2003		2004		2005	
	Anzahl Codes	Gesamt						
1	3064	3064	3352	3352	3350	3350	3358	3358
2	324	648	32	64	40	80	56	112
3	18	54	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-
5	5	25	-	-	-	-	-	-
Summe ^a	3411	3791	3384	3384	3390	3430	3414	3470

^a Die Anzahlen für die Codeblätter können von den Stichprobengrößen abweichen. Zum einen wurden die Codeblätter von Schulen, die nicht an allen Befragungszeitpunkten teilnahmen, zum anderen Codeblätter, deren Fragebögen durch die Plausibilitätskontrollen fielen, verwendet.

Das Problem der Reproduktion, dass also Befragte in den einzelnen Wellen einen nicht übereinstimmenden Code angeben, existiert allerdings auch in Duisburg noch in einem nicht zu vernachlässigendem Maße, konnte aber gleichwohl reduziert werden. In Münster fanden sich zwischen jeweils zwei Erhebungszeitpunkten lediglich rund 500 bis 600 Zuordnungen mit exakt übereinstimmendem Code, was bei allen realisierten Zuordnungen einen Prozentsatz von 40 bis 50% ausmachte (Pöge 2005a: 7f; Pöge 2005b: 66). In Duisburg hingegen liegt die Quote recht stabil bei rund 60% aller gefundenen Zuordnungen (siehe exemplarisch für die Zuordnungen zwischen 2003 und 2004 Tabelle 2). Die Reproduktion des Codes hat also mithilfe des verbesserten Layouts und der vereinfachten Fragen deutlich besser funktioniert, obwohl eine Codestelle mehr fehlerfrei angegeben werden musste als in Münster.

4 Das angewendete Zuordnungsverfahren

Um die Fragebögen der einzelnen Erhebungsjahre über die Codes einander zuzuordnen, wurde auch in Duisburg das im genannten Artikel für Münster vorgestellte *fehlertolerante Verfahren mit manuellem Handschriftenvergleich* angewendet. Es bestand zunächst aus drei, später aus vier Schritten: In einem *ersten Schritt* wurden maschinell alle exakt übereinstimmenden Codes aus zwei Erhebungswellen heraus-

gefunden.⁶ Die zusammengehörigen Fragebögen und Codeblätter wurden daraufhin einer manuellen Handschriftenkontrolle⁷ unterzogen, wobei die offensichtlich nicht passenden Zuordnungen aussortiert wurden. Die passenden Fragebogennummern wurden daraufhin aus dem Datensatz genommen, so dass sie für die nachfolgenden Schritte nicht mehr zur Verfügung standen. Im *zweiten Schritt* wurde nach Codeübereinstimmungen unter Zulassung eines Fehlers und im *dritten Schritt* unter Zulassung von zwei Fehlern gesucht und die zugehörigen Bogennummern herausgeschrieben. Seit dem Jahr 2003 war es mit dem eingesetzten erweiterten Code möglich, in einem *vierten Schritt* drei Fehler zuzulassen. Auch in diesen Schritten wurden als Validierung der Zuordnungen Handschriftenvergleiche durchgeführt, die offensichtlich nicht passenden Zuordnungen verworfen und vor der Durchführung des nächsten Schrittes die erfolgreich zugeordneten Nummern aus dem Datensatz entfernt. Um die im weiteren Verlauf vorgestellten, realisierten Fallzahlen erreichen zu können, waren pro Datensatz rund 4.000 manuelle Überprüfungen per Handschriftenvergleich nötig (exemplarisch siehe hierzu Tabelle 2).

Tabelle 2 Zugelassene Fehler, Anzahl der handschriftlichen Kontrollen und deren Ergebnisse (vor Plausibilitätskontrollen) 2003/2004

Fehler	passt	passt nicht	fehlt ^a	Kontrollen
0	1528	42	16	1586
1	735	175	7	917
2	292	853	5	1150
3	67	119	1	187
Summe	2622	1189	29	3840

^a In der Spalte „fehlt“ sind die Anzahlen der unauffindbaren Fragebögen ausgewiesen.

Es stellte sich heraus, dass die Fragebogenpärchen auf Grundlage der Codezuordnung ohne Fehler meist tatsächlich von derselben Person ausgefüllt wurden. Bei den Zuordnungen mit zugelassenen Fehlern ist der Grad der falsch positiven Über-

6 Das Heraussuchen der Zuordnungsvorschläge wurde mithilfe des Office-Programms Access und SQL-Skripten durchgeführt.

7 Hierbei wurden vor allem die offenen Fragen im Fragebogen mit den handschriftlichen Eintragungen verwendet.

einstimmungen deutlich höher. Hier kommt wiederum zum Tragen, dass unter Auslassung einer oder mehrerer Codestellen der Code uneindeutiger wird. So passt dann ein Fragebogen mit reduziertem Code vermeintlich zu mehreren Fragebögen der jeweils anderen Welle. Dieser Umstand lässt noch längere Codes ratsam erscheinen.

Unter Berücksichtigung der gültigen Fälle – derjenigen also, die auch den Plausibilitätskontrollen der einzelnen Querschnitte standhielten, liegen die mit dem vorgestellten Zuordnungsverfahren erreichten Fallzahlen zwischen 2.472 und 2.640 für die Zwei-Wellen-, bei 2.012 und 2.186 für die Drei-Wellen-Panels und bei 1.715 für das Vier-Wellen-Panel (Tabelle 3).⁸

Tabelle 3 Zuordnungsgüte der Paneldaten in Duisburg und Münster im Vergleich (N_e bedeutet erwartetes N ; N_b tatsächlich beobachtetes N)

	(a) Münster			(b) Duisburg			
	N_e	N_b	Quote (%)	N_e	N_b	Quote (%)	
$P_{t_{1,2}}$	1683	1271	75,5	$P_{t_{1,2}}$	3075	2472	80,4
$P_{t_{2,3}}$	1710	1373	80,3	$P_{t_{2,3}}$	3010	2596	86,2
$P_{t_{3,4}}$	1705	1406	82,2	$P_{t_{3,4}^*}$	3106	2640	85,0
$P_{t_{1,2,3}}$	1502	997	66,4	$P_{t_{1,2,3}}$	2626	2012	76,6
$P_{t_{2,3,4}}$	1497	1075	71,8	$P_{t_{2,3,4}^*}$	2755	2186	79,3
$P_{t_{1,2,3,4}}$	1316	813	61,8	$P_{t_{1,2,3,4}^*}$	2403	1715	71,4

Die Duisburger Paneldatensätze sind somit weitaus umfangreicher als die Münsteraner (siehe ebenfalls Tabelle 3), was vor allem auf die größere Querschnittsdatengrundlage zurückzuführen ist – allerdings leistet auch das deutlich besser funktionierende Zuordnungsverfahren seinen Beitrag zu den guten Ergebnissen. Die in

8 Zum vierten Zeitpunkt (2005) wurden in Duisburg zusätzlich zum zehnten Jahrgang auch Schülerinnen und Schüler der neunten Jahrgangsstufe befragt (Wiederholer). Die hochgestellten Sternchen bezeichnen diejenigen Datensätze, bei denen, um einen Vergleich mit den Münsteraner Daten zu ermöglichen, zum vierten Zeitpunkt *nur Jugendliche der zehnten Jahrgangsstufe* enthalten sind.

Duisburg mit dem beschriebenen Verfahren erreichten Zuordnungsquoten⁹ für die Zwei-Wellen-Panels aus jeweils zwei aufeinanderfolgenden Zeitpunkten liegen zwischen 80 und 86%. Die Quote ist dabei zwischen den Jahren 2002 (t_1) und 2003 (t_2) deutlich niedriger als zwischen den späteren Zeitpunkten. Auch die absolute Fallzahl steigt im Laufe der Jahre an und ist beim Panel 2004/2005 am höchsten. Dies mag zum einen am steigenden Alter der Befragten und der möglicherweise damit verbundenen höheren Fähigkeit, die Codefragen richtig zu beantworten bzw. sich zu konzentrieren, liegen. Zum anderen mag auch ein positiver Gewöhnungseffekt bei den Probanden eine Rolle spielen. Die Quoten der beiden lückenlosen Drei-Wellen-Panels liegen bei 77 und 79%, die des Vier-Wellen-Panels erwartungsgemäß niedriger bei 71%. Im Vergleich zu den Quoten aus der Münsteraner Panelstudie zeigen sich damit deutlich bessere Ergebnisse (siehe Tabelle 3). Es zeigt sich also, dass durch den verbesserten Code ein deutlich besseres Ergebnis bei der Ausschöpfung erzielt werden konnte.

Auch die Duisburger Paneldaten wurden auf Verzerrungen geprüft, mit dem Ergebnis, dass die Anzahl der Fehler bei Beantwortung des Codeblattes mit dem Geschlecht und der Schulbildung der Befragten korreliert. Während die Verzerrungen auf Ebene des gesamten Paneldatensatzes im Vergleich mit den Querschnittsdaten und der Schulstatistik nur leicht sind, lassen sich bei der Aufschlüsselung nach Fehlern deutliche Verschiebungen erkennen: Der Anteil der Mädchen ist bei den Zuordnungen mit keinem Fehler deutlich größer als derjenige der Jungen. Auch die Schulbildung hat einen (erwartungsgemäßen) Effekt. Je höher die Schulbildung, desto weniger Fehler wurden bei der Beantwortung gemacht. Vermutlich spielen kognitive Fähigkeiten – bei aller Einfachheit der Fragen – ebenso eine Rolle wie Konzentrationsfähigkeit und -wille. Im Vergleich zu den Verzerrungen in den Münsteraner Daten ist auf Ebene des Gesamtpanels festzustellen, dass der verbesserte Code hier zu deutlich geringer verzerrten Daten führt. Zwar unterscheiden sich die Abweichungen in der Verteilung nach Geschlecht kaum voneinander, die Verbesserung im Bereich der Verteilung nach Schulform ist jedoch erheblich.

9 Hier wird dieselbe Art der Quotenberechnung zugrundegelegt, wie für die Münsteraner Daten vorgestellt (Pöge 2005a: 62ff.).

5 Empfehlungen für die Praxis

Resümiert man nun nach acht Jahren die Erfahrungen, die in unserem Forschungsprojekt mit der Zuordnung von Fragebögen gemacht wurden, so ist festzuhalten, dass nach anfänglichen Schwierigkeiten ein durchaus gangbarer und erfolgreicher Weg gefunden wurde, die Paneldatensätze zu erstellen.

Aus vorwiegend datenschutzpsychologischen Gründen wurde ein Zuordnungsverfahren über persönliche Codes gewählt, damit gar nicht erst Bedenken im Hinblick auf die Anonymität bei den Befragten auftreten sollten. Dies erschien insbesondere wegen des jungen Alters der Befragungsteilnehmerinnen und -teilnehmer, welches sie besonders schützenswert macht, und des sehr sensiblen Datenmaterials (unter anderem Angabe von verübten Straftaten) geboten. Das Alternativverfahren eines Adresspanels mit Einwilligung zur Adressspeicherung kam daher zunächst nicht in Frage. Allerdings zeigen vergleichbare Studien auch mit diesem Verfahren keine deutlich besseren Ausschöpfungsquoten. So tritt erfahrungsgemäß teils schon bei der ersten Adresserhebung ein Ausfall von ca. 50% auf (exemplarisch hierzu siehe Böttger/Ehret et al. 2003: 35ff.), der im Laufe solcher Adresspanels deutlich größer wird. Mithilfe von selbstgenerierten Codes können oftmals bessere Quoten erreicht werden, wobei hier den Erhebungsabständen eine bedeutende Rolle zukommt. So liegen die Zuordnungsquoten bei monatlichem Befragungsintervall deutlich höher als bei jährlichem (siehe Kearney/Hopkins et al. 1984; Grube/Morgan et al. 1989).

In den beiden hier veröffentlichten Artikeln wurde vorgestellt, mit welchen Problemen in der Praxis mit selbstgenerierten persönlichen Codes zu rechnen ist. Im Vergleich der Erhebungsorte Münster und Duisburg wurde aufgezeigt, wie es möglich ist, durch Codemodifikationen deutliche Verbesserungen in den Zuordnungsquoten zu erreichen. Hierzu kann man zusammenfassend festhalten, dass bei der Planung eines solchen Zuordnungsverfahrens zuallererst die Variationsmöglichkeit des Codes geprüft werden muss. Es muss gewährleistet sein, dass jede Person mit hinreichender Wahrscheinlichkeit einen eindeutigen Code aufweist. In diese Überlegungen muss außerdem einbezogen werden, dass sich möglicherweise später einzelne Codestellen als fehlerhaft erweisen und bei der späteren Zuordnung nicht zur Verfügung stehen. Der Code sollte auch in diesem Fall noch eindeutig sein. Als Daumenregel stellte sich in unserer Untersuchung heraus, dass der Code auch unter Auslassung zweier Stellen eindeutig sein sollte. Als Stellschrauben für die Anforderungen an die Eindeutigkeit sind zum einen die Länge des Codes bzw. die Anzahl der Codestellen und zum anderen die Art der einzelnen

Codefragen im Hinblick auf deren Antwortvariation zu nennen. Es wurde hier gezeigt, wie abgeschätzt werden kann, ob der Code die genannten Anforderungen erfüllt (Pöge 2005a: 63; Pöge 2005b: 60).

Der zweite Punkt, den es zu beachten gilt, ist die Einfachheit der Codefragen. Wie hier aufgezeigt werden konnte, hat diese einen deutlichen Effekt auf die zu erzielenden Zuordnungsquoten. Zwar zeigen sich Unterschiede zwischen Geschlecht und Schulform der Befragten in dem Sinne, dass Mädchen sowie Jugendliche mit höherer Schulbildung weniger Probleme mit dem korrekten Beantworten der Fragen haben, aber um es auf eine kurze Formel zu bringen: Je einfacher die Codefragen sind, desto besser. Dies gilt dabei sowohl für den Inhalt der Fragen als auch für den zu notierenden Teil der Antwort. So fiel es den Befragten leichter, den ersten Buchstaben einer Antwort zu notieren als den letzten.

Neben diesen beiden grundlegenden Bereichen spielte in unserer Untersuchung auch das Layout der auszufüllenden Codeblätter eine Rolle. Es erwies sich als sehr hilfreich, handschriftliches Ausfüllen zu vermeiden und alle Antwortmöglichkeiten (Buchstaben und Zahlen) zum Ankreuzen aufzuführen. Hierdurch wurde die spätere Datenerfassung wesentlich erleichtert und deutlich weniger fehleranfällig.

In Bezug auf das Zuordnungsverfahren, welches mit den selbstgenerierten Codes arbeitet, ist zu sagen, dass es fehlerhafte Zuordnungen ermöglichen sollte, wenn nicht gar muss, um hinreichend gute Ergebnisse zu liefern. In beiden Erhebungen wäre die Ausschöpfungsquote mit fehlerfreien Codes unbefriedigend gewesen. In der Praxis stellte sich mit den vorgestellten Codes in unseren Stichproben die Zulassung von zwei Fehlern als guter Richtwert heraus. Mit dem erweiterten Code in Duisburg war auch die Zulassung von drei Fehlern möglich, dies erwies sich durch den relativ geringen Ertrag jedoch nur noch bedingt als sinnvoll.

In unserer Untersuchung wurde die Validierung der elektronisch gefundenen Fragebogenpärchen mithilfe von manuellen Handschriftenvergleichen durchgeführt, wobei in Duisburg pro Welle ein Aufwand von rund 4.000 Kontrollen aufgewendet wurde. Hierbei wurden mit erheblichem finanziellen Aufwand studentische Hilfskräfte eingesetzt. Schnell/Bachteler et al. haben ein fehlertolerantes Record-Linkage-Verfahren vorgestellt, welches das Heraussuchen der fehlerhaften Zuordnungsvorschläge automatisieren kann. Somit könnte möglicherweise auf die semi-automatische Vorgehensweise mit SQL-Skripten verzichtet werden. Auch sie schlagen bei ihrer Methode allerdings längere und komplexere Codes vor als bislang üblich (Schnell/Bachteler et al. 2006).

Berücksichtigt man die aufgeführten Punkte und setzt bei Paneluntersuchungen durchdachte, wohlkonzeptionierte Codes ein, können – wie vorgestellt –

durchaus beachtliche Zuordnungsquoten erreicht werden. Besonders in Studien, in denen selbstgenerierte persönliche Codes ratsam erscheinen, können sie sich trotz aller Probleme als probates Mittel der Fragebogenzuordnung erweisen.

Literatur

- Böttger, A., B. Ehret und F. Othold et al., 2003: Methoden der Untersuchung. S. 35-66 in: K. F. Schumann (Hg.): Delinquenz im Lebensverlauf. Bremer Längsschnittstudie zum Übergang von der Schule in den Beruf bei ehemaligen Hauptschülern (Bd. 2). Weinheim und München: Juventa.
- Brondies, M., 2004: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2003. Münster und Trier: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden (6).
- Grube, J. W., M. Morgan und K. A. Kearney, 1989: Using Self-Generated Identification Codes to Match Questionnaires in Panel Studies of Adolescent Substant Use. *Addictive Behaviors* 14: 159-171.
- Hilfert, N., 2005: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2004. Münster und Bielefeld: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden (10).
- Kearney, K. A., R. H. Hopkins und A. L. Mauss et al., 1984: Self-Generated Identification Codes for Anonymous Collection of Longitudinal Questionnaire Data. *The Public Opinion Quarterly* 48: 370-378.
- Kunadt, S., 2006: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2005. Münster und Bielefeld: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden (11).
- Motzke, K. und M. Brondies, 2004: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002. Münster und Trier. Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden (5).
- Pöge, A., 2005a: Methodendokumentation der kriminologischen Schülerbefragung in Münster 2000-2003 (Vier-Wellen-Panel). Münster und Trier: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden (9).
- Pöge, A., 2005b: Persönliche Codes bei Längsschnittstudien. Ein Erfahrungsbericht. *ZA-Information* 56: 50-69.
- Pöge, A., 2007: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2005 (Vier-Wellen-Panel). Münster und Bielefeld: Schriftenreihe Jugendkriminalität in der modernen Stadt – Methoden (13).
- Schnell, R., T. Bachteler und J. Reiher, 2006: Die Anwendung statistischer Record-Linkage-Methoden auf selbstgenerierte Codes bei Längsschnitterhebungen. *ZA-Information* 59: 128-142.

Korrespondenzadresse: Andreas Pöge
 Universität Bielefeld
 Fakultät für Soziologie
 Postfach 10 01 31
 D-33501 Bielefeld
 Andreas.Poege@uni-bielefeld.de

Rezensionen



UDO KELLE, 2007:
Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Theoretische Grundlagen und methodologische Konzepte. Wiesbaden: VS Verlag.
ISBN-13: 9783531153124,
ISBN-10: 3531153129,
331 Seiten,
34,90 EUR.

Udo Kelle arbeitet seit langer Zeit an einem (extrem) wichtigen und schwierigen Thema: Die Integration qualitativer und quantitativer Methoden in der empirischen Sozialforschung. Das dazu jetzt unter diesem Titel erschienene Buch fasst die über die Zeit entstandenen Positionierungen und (praktischen) Umsetzungen zusammen und gießt sie in ein „integratives Programm empirischer Sozialforschung“. Das Buch enthält 12 Kapitel, die sich in fünf thematische Punkte gruppieren lassen. Die Kapitel 1 und 2 eröffnen die Diskussion um die beiden Schulen von „quantitativer“ und „qualitativer“ Sozialforschung, die sich (zunächst) leicht auch unterschiedlichen theoretischen Orientierungen und der Dichotomie von „erklärenden“ und „verstehenden“ Ansätzen zuordnen lassen. Das Kapitel 3 geht auf einen Kernbereich der gesamten Argumentation ein: Es gebe im Bereich der gesellschaftlichen Phänomene immer nur „Strukturen begrenzter Reichweite“. Damit ist gemeint: bestimmte Zusammenhänge gelten nur bei einer gewissen Stabilität eines (Reproduktions-)Gleichgewichts von Handeln und Strukturbildung. Zu Änderungen kann es jederzeit kommen, insbesondere

durch die „agency“ der Akteure: ihre Fähigkeit, neue Wege zu gehen, aber auch durch exogen verursachte neue Problemlagen. Daraus wird ein Zentralargument für die Integration von quantitativen und qualitativen Methoden abgeleitet: Quantitative Studien sind nötig, um die in den (begrenzten) Gleichgewichten vorhandenen Verteilungen von Eigenschaften verlässlich zu ermitteln, qualitative, um herauszufinden, über welche (Mikro-)Mechanismen sich die (begrenzten) Gleichgewichte von Handeln und Strukturen herausbilden und reproduzieren. Das läuft (zwingend) auch auf eine Auflösung des anderen Dualismus hinaus, dem zwischen Erklären und Verstehen, zwischen „Kausalität“ und „Sinn“ also. Dieser Problematik sind die dann folgenden Kapitel 4 bis 9 gewidmet. Kapitel 4 nimmt sich die Grenzen des „deduktiv-nomologischen Modells sozialwissenschaftlicher Erklärung“ vor. Gemeint ist das (inzwischen) bekannte Badewannen-Modell, wie es in der Variante von Coleman auch als „Rational-Choice“-Ansatz bekannt geworden ist. Der zentrale Punkt bei den „Grenzen“ ist, dass die in diesem Modell „zur Erklärung herangezogenen Theorien fast immer durch Brückenhypothesen und Zusatzannahmen ergänzt werden müssen, um empirischen Gehalt zu gewinnen“, aber weil die betreffenden „Strategien hypothetisch-deduktiver Theoriebildung und Forschung keine Verfahren zur Verfügung stellen, um solche Brückenhypothesen empirisch begründet zu formulieren“, fehlt ihnen grundsätzlich etwas. Das macht die Verwendung vor allem der qualitativen Verfahren „unumgänglich“ (S. 82), etwa zum Rückgriff auf eine „Gewohnheitsheuristik des Alltagswissens“, die die Akteure, etwa durch kulturell festgelegte und inhaltlich oft sehr spezifische „Handlungsmaximen“, leitet und ohne deren Einbezug man allein mit einer abstrak-

ten Theorie, etwa der RC-Theorie, gar nichts sagen könne. In Kapitel 5 wird sozusagen die andere Seite aufs Korn genommen: Die Vorstellung, dass die Intentionalität des menschlichen Handelns einen Gegensatz zur kausalen Erklärbarkeit konstituiere, weil, so die dualistische These, der Begriff des Handelns die Intentionalität schon logisch impliziere und sie daher auch keine eigenständige Ursache sein könne. Diese (lange) Debatte beruht (wie so oft bei bloß philosophischen Denküben) auf einer eigentlich leicht klärbaren definitorischen Unterscheidung: Begriffsnominalistisch kann man das ruhig so sehen („Handeln“ ist definiert als ein „Verhalten“ mit Intentionen, warum nicht?), aber theoretisch (und empirisch) gibt es keinen Grund, Intentionen nicht als „Ursachen“ anzusehen, wie das etwa die Wert-Erwartungstheorie oder die Einstellungstheorie tut, die die Intentionen oder die Einstellungen als den Reaktionen kausal vorgeordnet ansieht. Kurz: Sinnverstehen und kausale Erklärung sind keine Gegensätze, sondern Teil ein und desselben Zusammenhangs, dem der „doppelten Hermeneutik“ nämlich, wie das Anthony Giddens einmal ausgedrückt hat. Wichtig wird dann, dass bei solchen Intentionen-Erklärungen auch die „Alltagstheorien“ der Akteure (ihre Erwartungen über kausale Zusammenhänge also) bzw. gewisse „Handlungsmaximen“ (also für sozial gültig angenommene Regeln oder gar „Gesetze“ ihres Tuns) in den Mittelpunkt rücken – und empirisch erfasst werden müssen. Hier stellt sich dann aber gerade für eine (nahe liegende) Verwendung qualitativer Verfahren die Frage nach der Fallauswahl und der Geltungsreichweite, und die (ebenso nahe liegende) Empfehlung ist der Einsatz auch quantitativer Verfahren zur Kontrolle dieser beiden Probleme. Die Kapitel 6 bis 9 greifen dann verschiedene Aspekte und Teilprobleme der Berücksichtigung von Kausalität in der sozialwissenschaftlichen Theoriebildung und empirischen Sozialforschung auf. In Kapitel 6 wird das Problem

der (meist zahllosen) Hintergrundbedingungen für die kausale Wirkung bestimmter „Faktoren“ angesprochen und welche dieser Hintergrundbedingungen bei einer „akteursorientierten Sozialforschung“ speziell zu beachten sind: externe Umstände, die als „weil-Motive“ wirksam werden können, die Intentionen der Akteure als ihre „um-zu“-Motive und gewisse, „mehr oder weniger lokale Handlungsregeln bzw. –maximen, die Handlungsziele und Handlungsbedingungen zu sinnvollen Handlungsstrategien verknüpfen“ (S. 164), die „Frames“ und die „Skripte“ also, die es für typische Situationen gibt. Die wohl wichtigste Botschaft an die „verstehende“ (qualitative) Methodologie ist dabei, dass auch hier, wenigstens implizit, stets kausale Argumentationen vorgenommen werden – ganz gegen die dort oft herrschende Vorstellung, dass das eine ganz unangemessene Kategorie sei, wenn es um den „Sinn“ gehe. Das Kapitel 7 behandelt dann die Frage der Kausalität bei kleinen Fallzahlen, wie sie in der vergleichenden Sozialforschung oft vorkommt (und mit dem Konzept der „Qualitativen Komparativen Analyse“ nach Ragin eine systematische Fassung gefunden hat). In Kapitel 8 geht es um das Problem der (Nicht-)Determinierung des Handelns, also des in den Sozialwissenschaften üblichen Falls großer Anteile nicht erklärter Varianzen, und in Kapitel 9 um die Frage der Aufdeckung von Scheinbeziehungen durch die (bekannteren) Verfahren der statistischen oder der (quasi-)experimentellen Kontrolle von Hintergrundvariablen. Das Fazit ist für alle diese Fälle und Problemlagen: Eigentlich sucht man nicht nach fixen Kausalfaktoren und –beziehungen, sondern nach den diesen Zusammenhängen „underlying, generative processes“, bei denen die „statistischen Aggregatphänomene“ durch generative Mechanismen erklärt werden, die auf „Mikromodelle(n) sozialen Handelns“ beruhen. Die Kausalfaktoren sind also nicht selbst schon Teile des Explanans für die Erklärung gewisser Sachverhalte, sondern

ihrerseits Explananda. Zu Recht wird in diesem Zusammenhang auf das Makro-Mikro-Makro-Modell verwiesen, das aber, wie in Kapitel 4 dargelegt wird, nur eine abstrakte Theorie anbieten und daher (leider) aus sich heraus keinen empirischen Gehalt vorweisen könne. Für die Ermittlung der nun nötigen Brückenhypothesen müsse man über diesen Ansatz hinausgehen. Die Kapitel 10 bis 12 skizzieren schließlich das Programm einer integrativen Sozialforschung. In Kapitel 10 wird an verschiedenen Beispielen die Fruchtbarkeit der komplementären Kombination von qualitativen und quantitativen Verfahren demonstriert: die Erklärung „überraschender“ statistischer Ergebnisse oder die Aufdeckung von Validitätsproblemen bei Interviews etwa. Das Kapitel 11 bildet sozusagen das kanonisierte Konzentrat des Programms. Der „harte“ Kern des Vorschlags besteht aus vier Elementen (S. 264): ein „nicht-nomothetisches Kausalitätskonzept“, die Überwindung des Dualismus von Verstehen und Erklären, ein Konzept von gleichzeitig „theoriegeleiteter“ und „empirisch begründeter Theoriebildung“, das den Dualismus von explorativer und hypothesengeleiteter Forschung überwinden soll, und die Überwindung der Schwächen sowohl des einseitig „hypothetiko-deduktiven“ Modells, wie das eines ebenso einseitigen theorielosen „Induktivismus“. Vor diesem Hintergrund wird das Modell eines Forschungsablaufs skizziert, bei dem in den verschiedenen Phasen die eigenen Leistungen und Vorzüge jeweils der qualitativen wie der quantitativen Methoden zum Zuge kommen sollen: Festlegung und Feststellung eines Explanandums; die „Konstruktion des Explanans“ über einen als Heuristik fungierenden, abstrakten, universellen, aber empirisch leeren Theorierahmen, der dann auf verschiedene Weise mit empirischem Gehalt gefüllt wird; und schließlich die Überprüfung der Geltungsreichweite der so theoretisch begründeten und empirisch gefundenen Beziehungen von Kausalfaktoren (begrenzter Reichweite)

durch „zusätzliches empirisches Material“. Den Abschluss bilden zwei Modelle der Kombination von qualitativen und quantitativen Verfahren und in Kapitel 12 eine Art Epilog über die „Empirische Sozialforschung jenseits des Methodendualismus“.

Es kann gar keinen Zweifel geben, dass Udo Kelle einen außerordentlich wichtigen, informativen, durchdachten, lesens- und beherzigenswerten Vorschlag zur Überwindung einer der ärgerlichsten Spaltungen in der (empirischen) Sozialwissenschaft vorgelegt hat. Dass dabei eine in weiten Teilen sehr informative, manchmal fast lehrbuchreife Sammlung von wichtigen Einzelheiten, etwa in den Kapiteln über die Kategorie der Kausalität abfällt, sollte nicht unerwähnt bleiben.

Gäbe es denn dann überhaupt etwas zu bemängeln? Wir glauben schon, und auch, dass es kein peripherer Punkt ist (und viele kleinere kritische Punkte seien belassen), wengleich letztlich das alles für die praktische Arbeit keinen sonderlichen Unterschied macht. Udo Kelle kritisiert in Kapitel 4 und immer einmal wieder über das ganze Buch hinweg das Modell der soziologischen Erklärung, welches er für unzureichend hält. Dabei konzentriert sich seine Kritik auf ein spezielles Problem: die Verbindung der universal-abstrakten Terme der Handlungstheorie bei der Logik der Selektion mit den stets spezifischen und (in der Tat: meist) begrenzten Strukturen in der Logik der Situation, den Brückenhypothesen also. Dem Modell der soziologischen Erklärung wird dabei unterstellt, dass diese Verbindung etwas sei, was ihm eigentlich ganz fremd wäre und dass es daher immer einer „Ergänzung“ bedürfe, die nur von außen kommen könne (etwa dann ggf. auch über qualitative Verfahren). Vor diesem Hintergrund sieht Kelle zwei, jeweils für sich unvollständige „Typen theoretischer Aussagen“: Empirisch gehaltvolle Aussagen mit raum-zeitlich begrenztem Geltungsbereich

einerseits und Konzepte und Aussagen mit großem oder gar universellem Geltungsbereich mit geringem oder ganz fehlendem empirischen Gehalt. Die abstrakten Konzepte sind die an sich „leeren“ Heuristiken der diversen Theorieangebote, die es im Forschungsprozess mit „gegenstandsbezogenen empirischen Aussagen“ zu verbinden gilt, damit sie (raum-zeitlich stets begrenzten) „empirischen Gehalt“ bekommen. Und dafür ist die Kombination quantitativer und qualitativer Verfahren das geeignete Mittel.

Was ist dazu zu sagen? Vier Punkte erscheinen wichtig, um den Vorschlag von Udo Kelle, vor dem Hintergrund der im Modell der soziologischen Erklärung bereits seit langem ausgearbeiteten Möglichkeiten und Vorgaben, einordnen zu können. Erstens: Es ist einfach unzutreffend, die Formulierung und Validierung von Brückenhypothesen als eine dem Modell der soziologischen Erklärung externe, fremde oder gar widersprechende „Zusatzaufgabe“ zu deklarieren. Das hat schon den allgemeinen methodologischen Grund, dass keine Theorie mit latenten, abstrakten Variablen ohne irgendeine empirische Interpretation prüfbar ist und es das Problem der „Hilfstheorien“ also sowieso immer schon gibt. Allein deshalb stehen die Brückenhypothesen im Zentrum des ersten Schrittes einer jeden soziologischen Erklärung, und es ist kaum nachzuvollziehen, wieso das als ein innerhalb der Konzeption eigentlich nicht vorgesehenes oder gar nicht zu lösendes Problem angesehen wird. Damit wird zweitens auch sofort erkennbar, dass die abstrakte Handlungstheorie, auf die sich die Brückenhypothesen beziehen, mehr sein muss als eine bloße „Heuristik“: Es muss (möglichst) genau spezifiziert werden, wie denn bestimmte „raum-zeitlich begrenzte“ empirische Situationsparameter mit welchen Variablen der universal-abstrakten Handlungstheorie variieren, etwa: je größer die Gruppe, umso kleiner ist die Erfolgserwartung für den Effekt des eigenen Tuns. Die abstrakte Variable ist hier die eine

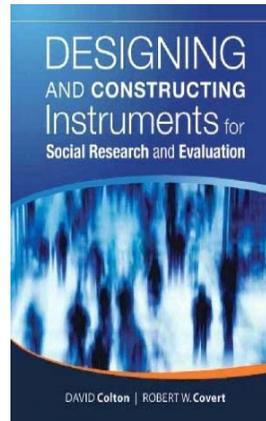
„Erwartung“, die konkret-empirische Strukturbedingung die Gruppengröße, und der einfachste Fall wäre eine lineare Verbindung, die, wenn sie nicht stimmt, genauer zu betrachten wäre und sich auch wieder ändern kann (was praktisch sehr lästig ist). Diese Verbindung folgt sicher nicht schon aus der Handlungstheorie, aber sie valide zu bestimmen, ist im Rahmen des Modells der soziologischen Erklärung auch kein besonderes methodologisches Problem (wenngleich meist ein schwieriges praktisches Geschäft), sondern genuiner Teil der Bestimmung der „Logik der Situation“. Schon hier kann es – drittens – um den „Gehalt“ der ganzen Theorie gehen: den logischen Gehalt nämlich, dass möglichst viel an (potentiellen) Falsifikatoren in der abstrakten Handlungstheorie vorkommt und nicht, wie bei den lockeren Heuristiken üblich, eigentlich schon logisch und begrifflich so gut wie alles möglich ist und damit die Unwiderlegbarkeit schon formal erzwingt – noch vor jeder empirischen Interpretation. Noch der vierte Punkt. Nur auf der Ebene der abstrakten Handlungstheorie liegt dann auch der (letzte) Rest an allgemeinem Hempel-Oppenheim-Gesetz, das für den Anspruch einer über den Fall und das jeweilige Strukturgleichgewicht hinausweisenden Erklärung unerlässlich ist, und ansonsten ist in der Tat alles mehr oder weniger von „begrenzter Reichweite“. Das aber ist für das Modell der soziologischen Erklärung auch die Grundlage für den Anspruch, weiterhin eine „universale“ und „nomothetische“ Angelegenheit zu sein, wobei sich das eben nicht auf Strukturzusammenhänge, sondern nur (noch) auf die Selektionstheorien bezieht, denen, so die Annahme, die Menschen so ziemlich universell folgen. Wenn es das nicht gäbe, stünde man außerhalb der Gleichgewichte begrenzter Reichweiten theoretisch mit recht leeren Händen da.

Das Hauptergebnis der Studie kann sich jedoch sehen lassen und die Hardliner beider Seiten sollten es sich genau ansehen –

und hinter die Ohren schreiben: Es gibt keinen Grund für irgendeinen „qualitativen“ induktivistisch-theoriefreien Dualismus auf der einen Seite und es gibt ebenso keinen Grund für die Annahme eines „quantitativen“ deduktivistischen Universalismus kausaler Gesetze auf der anderen. Es gibt einige Besonderheiten der Sozialwissenschaften, die es erzwingen, sich oftmals sehr genau die historisch und sozial sehr spezifischen (Mikro-)Verhältnisse anzusehen, aber das heißt noch lange nicht, dass damit auch die allgemeinen Regeln der Theoriebildung und -prüfung, wie sie für alle Wissenschaften gelten, etwa die Überprüfung der Geltungreichweite eines Zusammenhangs oder die Annahme von Kausalbeziehungen auch bei der Erklärung von Handlungen, außer Kraft gesetzt sind. Dem ist nichts hinzuzufügen, und man muss Udo Kelle sehr dankbar sein, dass er die Grenzen der beiden Extrempositionen und die Unhaltbarkeit der damit verbundenen Dualismen überzeugend begründet und unvoreingenommen und ausbalanciert vorträgt. Die Integration der qualitativen und der quantitativen Verfahren fällt dann wie eine reife Frucht vom Baume – und mancher wird sich fragen: warum gab es denn eigentlich darüber mal einen Streit?

HARTMUT ESSER, MANNHEIM

* * * * *



DAVID COLTON and ROBERT W. COVERT, 2007: *Designing and Constructing Instruments for Social Research and Evaluation*. 1. Auflage. San Francisco: Jossey-Bass, 412. Seiten, ISBN: 978-0-7879-8784-8, 44,99 EUR.

Der Band tritt mit einem ambitionierten Titel an, der einen an der Umfrageforschung interessierten Leser zunächst zu der Annahme führt, es handele sich um ein Buch, in dem – zumindest auch – ausführlich über die Fragebogenkonstruktion informiert wird. Bei genauerer Inspektion zeigt sich jedoch, dass es sich eher um ein Einführungslehrbuch in die Methoden der empirischen Sozialforschung mit Schwerpunkt Instrumentenentwicklung (Beobachtungsbögen, diagnostische Instrumente, Fragebögen, Check-Listen, Tests, Beurteilungsbögen) handelt. Das Buch gliedert sich in 3 Hauptabschnitte, die mit „Grundlagen“, „Anwendung“ und „Organisation und Durchführung“ überschrieben sind, die intern in 15 Unterkapitel untergliedert sind:

Das einleitende Kapitel 1 liefert zunächst einige konzeptionelle Überlegungen zu den charakteristischen Differenzierungen verschiedener Arten von Instrumenten und ihrer Bestandteile. Hier wird auch der Prozess der Konstruktion eines Instruments entworfen, der im Verlauf des Bandes immer wieder aufgegriffen wird. In dieser Betonung des Prozesshaften liegt sicher eine der Stärken des Bandes, indem nicht nur die aus wissenschaftstheoretischer Sicht zu lösenden Aufgaben diskutiert, sondern auch die notwendigen Arbeitsschritte – und

dabei auftretende Fallstricke – thematisiert werden. Das zweite Kapitel stellt die Rolle der Instrumente und der Instrumentenentwicklung in den Kontext eines Forschungsprojekts. Dabei werden sehr kursorisch die Grundstrukturen quantitativer und qualitativer Forschungsmethodologie präsentiert, bevor im anschließenden Kapitel 3 die Skalenniveaus vermittelt werden.

Auf dem Niveau eines Einführungslehrbuchs wird im anschließenden Kapitel 4 ein Überblick über Validitäts- und Reliabilitätskonzepte sowie über Methoden zur Überprüfung und Sicherstellung der beiden Gütekriterien gegeben. Einzelne Verweise auf weiterführende Literatur und Ressourcen reichen das Kapitel an, bevor es im darauffolgenden Kapitel 5 um die Explikation der Ziele der Studie geht, in deren Rahmen ein Instrument zu entwickeln ist. Diese Explikation soll den Forscher zu einer Vergegenwärtigung der zu messenden Konzepte und der Untersuchungseinheiten/Merkmalsträger anregen. Außerdem werden Entscheidungen über die Art der Administration nahegelegt. Um diese Explikation der Ziele anzuregen und herbeizuführen, werden Techniken für eine systematische Generierung von kreativen Ideen diskutiert, unter anderem: Brainstorming, Snowballing und Pyramiding, Delphie-Methode, was wiederum die sozialen Entwicklungsprozesse eines Instruments betont.

Im sechsten Kapitel werden Methoden des Pre-Testings diskutiert – was systematisch eigentlich nach den im Weiteren beschriebenen Arbeitsschritten liegt. Zudem bleiben die meisten Standardmethoden unberücksichtigt und werden z. T. nicht einmal genannt (Behavior-Coding, Interaction-Coding, Random-Probe usw.). Auch die einschlägige Standardliteratur wird nur spärlich erwähnt. Dafür unternehmen die Autoren einen interessanten Versuch, die Pre-Test-Methoden nach ihrer Stellung im Entwicklungsprozess eines Instruments zu ordnen: In der

Vorbereitungsphase, in der sich die Zielsetzung der Studie herauschält und präzisiert wird, in der eigentlichen Entwicklungsphase des Instruments, in der verschiedene Entwurfsfassungen schrittweise verfeinert und optimiert werden, und schließlich in der Testphase des vorläufig fertigen Instruments. Von daher erklärt sich dann auch, dass diese Methoden relativ früh in den Entwicklungsprozess eines Instruments eingeführt werden.

Jedoch krankt dieses Kapitel an der theoretischen Einbettung der Pre-Test-Methoden: Zwar wird zunächst auf im Prozess der Instrumententwicklung unentdeckte Validitätsprobleme verwiesen, dann aber folgt eine stichpunktartige Zusammenstellung von bekannten „Antwortfehlern“ (Context-Effekte, Antwortreihenfolgeneffekte, Effekte von sensitiven Items, Non-Attitude, Antwortinkonsistenzen und viele andere bekannte Effekte), deren Auftreten als Begründung für die Notwendigkeit eines Pre-Tests angeführt wird. Hier wird offensichtlich, dass die „kognitionspsychologische Wende“, die seit etwa 20 Jahren die methodische Forschung zu den verschiedenen Facetten des Measurement-Error wesentlich vorangebracht hat, in diesem Buch nur wenig Berücksichtigung gefunden hat.

Das anschließende Kapitel 7 bietet eine schöne, pragmatische Einführung in die Gestaltung von geschlossenen Antwortvorgaben. Einige basale Regeln, eine Systematisierung in Anlehnung an den Band von Fink und viele Beispiele von Standardskalen werden geboten. Allerdings fehlen wiederum die umfangreich existierenden Belege aus der experimentellen Forschung für die gegebenen Empfehlungen. Außerdem reicht manche „Guideline“ einfach zu kurz oder ist nur schwer nachvollziehbar: „By convention, response scales are ordered from low to high“ (S. 156). Weiter wird die Frage der numerischen Relative diskutiert, aber ohne Bezug zu den einschlägigen Texten, die z. B

belegen, dass diese die vom Befragten vorgenommene Verankerung der Skala zu verändern vermögen. Entscheidend ist aber, dass die Konstruktion der Antwortkategorien bei geschlossenen Fragen auf das Kriterium reduziert wird, welche Kategorien „most effective for capturing the data you need“ sind. Damit bleibt der zentrale Ansatz der sozialwissenschaftlichen Methodenforschung, der die Qualität einer Frage am Umfang der auftretenden Antwortfehler misst, unberücksichtigt. In Kapitel 8 folgt erneut eine Zusammenstellung typischer Probleme bei der Formulierung von Fragebogenfragen und Antwortkategorien. Dabei wird auf das Problem der sozialen Erwünschtheit und der sensitiven Fragen sowie auf Probleme der Bedeutungsäquivalenz bei interkulturellen Befragungen eingegangen. Hinzu kommen einige Regeln für die Formulierung von Antwortvorgaben bei Rating-Skalen. Daran schließt sich im Kapitel 9 eine Darstellung von Mehrfachnennungen und besonderen Formen der Einfachauswahl an, sowie eine Diskussion der Exklusivität der Antwortvorgaben.

Zwar treten an verschiedenen Stellen im Buch Redundanzen auf, die mit Blick auf selektive Leser, die nicht das ganze Buch zur Kenntnis nehmen, durchaus sinnvoll sein können. In den Kapiteln 7, 8 und 9 nehmen diese aber überhand. Eine systematischere Aufbereitung und Gliederung dieser drei für den Fragebogenkonstrukteur zentralen Kapitel hätten sicher zu einer gelungeneren, weniger redundanten Präsentation der an sich kenntnisreichen Darstellung führen können.

Im Kapitel 10 werden offenen Fragen und ihre Auswertung diskutiert, was zu einer relativ ausführlichen Passage über inhaltsanalytische Verfahren führt. Die Formulierung der offenen Frage wird kurz dargestellt; die aktuelle Diskussion über die „Visual Design Language“ der offenen Frage (u. a. Dillman, Smyth, Christian) wird hinge-

gen nicht aufgegriffen, auch klassische Studien zur unterschiedlichen Mächtigkeit von geschlossenen und offenen Items bleiben unberücksichtigt. Gelungener ist demgegenüber das anschließende Kapitel 11, in dem eine Beschreibung der Funktionsweise von verschiedenen Skalentypen geleistet wird: Neben dem semantischen Differenzial und der allfälligen Likert-Skala werden zwei Varianten von Thurstone und die Goal-Attainment-Scale (GAS) diskutiert. Die Darstellung gibt einen Einblick in die Grundzüge der Funktionsweise und der Konstruktion der jeweiligen Skala, der für den Leser dieses Einführungsbuches durchaus aufschlussreich ist.

Der abschließende dritte Teil des Buches beginnt mit einem Kapitel über die Zusammenführung verschiedener Items und Skalen zu einem Gesamtinstrument und zu seiner formalen Gestaltung. Hier werden Titel, Einleitungsstatement und Formulierungen zum Datenschutz diskutiert, aber auch – kurz – das Design von Filterinstruktionen in selbst-administrierten Fragebögen. Ausführlich gehen die Autoren auf die sozialstatistischen Fragen ein, wobei vor allem auch Ethnizität und ‚Race‘ behandelt werden. Außerdem wird auf die Reihenfolge von Fragetypen und Themen im Fragebogen eingegangen sowie – erneut – auf die Länge des Fragebogens. Die Frage des visuellen Designs wird nur auf wenigen Seiten abgehandelt und vor allem mit Bezug zur Schriftart und Schriftgröße.

Schließlich werden in Kapitel 13 die Administration der Instrumente getrennt für Beobachter-/Rater-/Interviewer-administrierte und selbst-administrierte Instrumente diskutiert. Dabei werden im Bereich der nicht selbst-administrierten Instrumente vor allem Beobachtungsbögen und das mit der Beobachtung einhergehende soziale Setting besprochen – interviewer-administrierte Befragungen kommen nur am Rande vor, eine ausführliche Diskussion des Modus der Admi-

nistration entsprechend ebenfalls nicht. Dafür wird über die Auswahl der Orte für eine Beobachtung, die Auswahl der Beobachteten und die Schulung der Beobachter gesprochen.

Im Weiteren wird dann die Durchführung einer selbst-administrierten Befragung thematisiert, wobei Stichprobenverfahren und Nonresponse problematisiert aber keinesfalls erschöpfend abgehandelt werden. In diesen Passagen wird deutlich, dass das Buch eben nicht nur den Konstruktionsprozess von Instrumenten beschreiben, sondern eher einen umfassenden Blick auf den Forschungsprozess in der empirischen Sozialforschung liefern will. Diesem überambitionierten Ziel kann der Band auf knapp 400 Seiten jedoch kaum gerecht werden. Diese Ausführungen über Stichproben und Feldarbeit eignen sich daher eher als Ausblick und Problemaufriss, allerdings fehlt für eine sinnvolle Weiterführung der Verweis auf die einschlägige Literatur, mit deren Hilfe die Leser die Probleme von Coverage, Sampling und Nonresponse gründlicher aufarbeiten könnte. Ebenfalls im Schlussteil des Bandes findet sich ein Kapitel (Kapitel 14) zum Einsatz von Computern bei der Erstellung, der Administration und Auswertung von Instrumenten. Diese sehr knappe Darstellung – Web Surveys zum Beispiel werden auf 1-2 Seiten abgehandelt – vermag immerhin eine gewisse Sensibilität für die Folgen der Computerisierung zu wecken; einige Vor- und Nachteile der Computer-Administration werden diskutiert. Allerdings fehlt der Verweis auf die Standardterminologie (CASI, CSAQ usw.) und auf die entsprechende Literatur. Das abschließende Kapitel 15 diskutiert Fragen des Datenzugangs und Copyright-Fragen bei von abhängig beschäftigten Mitarbeitern erstellen Instrumenten und gesammelten Daten. Schließlic werden einige Seiten auf die Berichterstattung und die Dokumentation der Studie verwendet. Auch hier gilt, dass dieses Kapitel mehr einem Pro-

blemaufriss gleicht, als dass es Antworten liefert.

Eine Gesamtwürdigung kommt zu einem uneinheitlichen Bild: Sehr erfreulich sind die vielen Beispielinstrumente im Text. Am Ende jedes Kapitels finden sich meist mehrere Instrumente aus allen Bereichen der empirischen Forschung (Beobachtung, psychologische Diagnose, Befragung), an denen jeweils verschiedene Aspekte oder Schritte der Instrumentenentwicklung verdeutlicht werden. In diesem Anwendungsbezug, der dem Leser die Möglichkeit gibt, die diskutierten Aspekte der Instrumententwicklung nachzuvollziehen, sind sicher die Stärken des Bandes zu verbuchen. Doch auch wenn diese Beispielfragebögen recht gut demonstrieren, was im Text erläutert wird, enthalten einige von ihnen Unzulänglichkeiten oder zumindest zweifelhafte Elemente, die bei einem am kognitionspsychologischen Konzept orientierten Vorgehen aufgefallen wären (z. B. die „not sure“ Option im Fragebogen auf Seite 124, unterschiedliche Spaltenbreite bei Matrixfragen auf den Seiten 294/295, alternierende Graustufen in aufeinanderfolgenden Spalten einer Matrixfrage auf Seite 339 – um nur einige dieser Stellen zu nennen). Mehrere kleine Flüchtigkeitsfehler (S. 174, 212, 209) z. B. bei der Exklusivität der Antwortkategorien, die ein erfahrener Fragebogenkonstrukteur als solche übersehen kann, bzw. die ihn nicht daran hindern, das, was im Beispiel verdeutlicht werden soll wahrzunehmen, sind nicht nur ärgerlich, sondern bedenklich, weil unerfahrene Leser diese Ungenauigkeiten möglicher Weise als intendiert missverstehen können.

Doch abgesehen von diesen Kleinigkeiten fällt die Würdigung des Bandes noch aus einem anderen Grund nicht rundum positiv aus: Während die internationale Methodenforschung die Fragebogenentwicklung in den letzten beiden Jahrzehnten definitiv von einer ‚Kunst‘ zu einer ‚Wissenschaft‘

weiterentwickelt hat, die auf experimentell gewonnenen Aussagen und theoretischen Modellen beruht, wird die Entwicklung eines Instruments von Colton und Covert eher als Gleichklang von ‚Kunst‘ und ‚Wissenschaft‘ diskutiert – mit einem beträchtlichen Schwerpunkt auf ‚Kunst‘. Auf dem Weg von der ‚Kunst‘ der Instrumententwicklung zur ‚Wissenschaft‘ ist dieser Text daher nur einen ersten Schritt gegangen. Die in fast allen Kapiteln bemühte Analogie zur Malerei und zur Komposition eines Gemäldes bekräftigt den Eindruck, dass die Instrumentenentwicklung eben noch wesentlich vom Genius des Konstrukteurs mitbedingt wird. Entsprechend finden sich explizit die Formulierungen „Purposeful Creativity“ (S. 97) und „The Art of Instrument Construction“ (S. 43), die nicht nur der manchmal überstrapazierten Gemälde-Analogie geschuldet sind, sondern tatsächlich Programm sind.

Die dem Buch zugrunde liegende Literatur enthält zwar einzelne Titel aus den Jahren 2006 und 2007; ein systematisches Review des State of the „Art“ scheint dem Buch aber nicht vorausgegangen bzw. bereits vor einigen Jahren durchgeführt worden zu sein. Z. T. werden einzelne wichtige Autoren genannt (Dillman, Fowler, Krosnick, Schwarz ...) andere werden übersehen (Tourangeau, Conrad, Couper ...). Daher entsteht der Eindruck, dass dieses Buch vor allem auf dem Erfahrungsschatz der Autoren basiert, der durch ausgewählte Publikationen abgesichert wurde. Positiv zu würdigen ist die Tatsache, dass manche abgelegene Quelle benannt und verarbeitet wird, die im Kanon der Instrumentenentwicklung und insbesondere der Fragebogenforschung bisher nicht berücksichtigt wird, aber durchaus interessante Teilbefunde – allerdings für den Experten – liefert (etwa zu den offenen Fragen).

Nach allem eignet sich das Buch nicht unbedingt als vertiefende Einführung in die Fragebogenkonstruktion. Es scheint eher für

eine Leserschaft geschrieben zu sein, die noch gar keine Erfahrungen mit der empirischen Sozialforschung hat. Das Fehlen der Bezüge zu der immens angewachsenen Literatur zum Total Survey Error, zum Measurement Error im Besonderen lässt einem mit dem Gefühl zurück, dass die Konstruktion eines Instruments zwar nicht mehr freihändig „aus dem Bauch heraus“ vorzunehmen sei, aber eben auch nicht auf der Basis der umfangreichen Fachliteratur zu verschiedenen Facetten des auf das Messinstrument zurückgehenden Messfehlers sowie der vorhandenen theoretischen Konzepte zu Integration und Erklärung dieser Effekte.

Positiv zu würdigen ist hingegen die Betonung des Prozesscharakters der Instrumentenentwicklung, die das iterative Vorgehen bei der Erstellung eines Fragebogens hervorhebt und betont. In diesem Zusammenhang soll explizit die allgegenwärtige Aufforderung zu reflektieren, zu überdenken und Transparenz herzustellen, gelobt werden. Damit kann dem Buch als Verdienst die Betonung des sozialen Entstehungsprozesses eines Instruments zugeschrieben werden, die in einer stärker wissenschaftstheoretisch oder kognitionspsychologisch zugeschnittenen Einführung sicher weniger sichtbar geworden wäre.

Die Stärken des Buchs liegen aber eindeutig darin, dass es den Anfänger in einer verständlichen Sprache einige Grundprobleme der Instrumentenentwicklung nahe bringt. Dadurch werden diese für die auftretenden Untiefen bei der Planung und Durchführung einer quantitativen empirischen Studie sensibilisiert. Die gegebenen Antworten bleiben aber hinter dem zurück, was man Personen, die erstmals eine Befragung durchführen, mit auf den Weg geben möchte.

MAREK FUCHS, KASSEL



JOACHIM BEHNKE,
THOMAS GSCHWEND,
DELIA SCHINDLER und
KAI-UWE SCHNAPP
(Hg.), 2006:
Methoden der Politik-
wissenschaft. Neuere
qualitative und quanti-
tative Analysever-
fahren. Baden-Baden:
Nomos. ISBN:
978-3-8329-2225-2,
364 Seiten, 29 EUR.

Die Herausgeber des Bandes „Methoden der Politikwissenschaft. Neuere qualitative und quantitative Analyseverfahren“ haben sich in mehrerer Hinsicht viel vorgenommen. Wie der Untertitel bereits andeutet, sollen qualitative und quantitative Zugänge zur Analyse empirischer Materialien in einer Publikation vereint vorgestellt werden. Ihnen liegt ebenfalls daran, der Leserin/dem Leser einen Eindruck von neueren Ansätzen zur Analyse sozialwissenschaftlicher Fragestellungen zu vermitteln. Ihre Zielsetzung konkretisierend schreiben sie: „Das erste Problem war, dass es insbesondere für die Politikwissenschaft keinen aktuellen Überblick über methodische Neuentwicklungen in deutscher Sprache gibt. [...] Zweitens wurde [...] festgestellt, dass nach wie vor nur selten qualitative und quantitative Zugänge zur Analyse empirischen Materials gemeinsam in einer Publikation vorgestellt werden“ (S. 9). Ausdrücklich vertreten sie dabei den Anspruch, dass ihr Werk denjenigen, die vor allem passiv mit der Rezeption wissenschaftlicher Verfahren befasst sind, eine „allgemeine Lesefähigkeit“ (S. 9) vermitteln soll. Anders formuliert: Die Beiträge sollen vor allem einen einführenden Charakter haben. Schließlich bieten eine ausführliche Einleitung der Herausgeber sowie die folgenden 29 Kapitel des Buches ausreichend Lesestoff.

Das erste, von den Herausgebern verfasste Kapitel, ist dem „Methoden-Schisma“ in den empirischen Sozialwissenschaften gewidmet und versucht, qualitative und quantitative Verfahren aus einer „integrativen Perspektive“ heraus zu beschreiben. Die Herausgeber führen hier ihre wissenschaftstheoretischen Überlegungen aus, die das Buch methodologisch begleiten. Ihre diesbezüglichen Ausführungen sind lesenswert und ihre (allerdings wenig überraschenden) Schlussfolgerungen im Hinblick auf die angemessene Wahl eines Analyseverfahrens lassen sich wie folgt zusammenfassen: „Je nach Art des Forschungsgegenstandes werden sich dabei der Forscherin im einen Fall eher quantitative Verfahren, im anderen Fall eher qualitative Verfahren als ‚natürlich‘ erscheinende Lösungen zur Behandlung ihres Forschungsproblems anbieten“ (S. 21). Am Ende dieses einführenden Kapitels werden sämtliche Analyseverfahren des Buches anhand von sieben Dimensionen verortet und in Form einer tabellarischen Synopse dargestellt. Im Einzelnen wird der Theorie-, Kausal- sowie Zeitbezug unterschieden; die weiteren Dimensionen beziehen sich auf das Aggregationsniveau der Analyseeinheiten, die Fallzahl, die methodische Reichweite sowie die Art der Datenerhebung.

Wie oben angesprochen, umfasst die Darstellung neuerer qualitativer und quantitativer Verfahren insgesamt 29 Kapitel. Die Auswahl möglicher Verfahren gründet sich auf der von Lewis-Beck et al. (2003) herausgegebenen „Encyclopedia of Social Science Research Methods“. „Neuer“ ist ein Verfahren, wenn es entweder innerhalb der letzten zehn bis fünfzehn Jahren entwickelt wurde oder in der deutschsprachigen Politikwissenschaft erst in „jüngster Zeit“ (S. 22) Anwendung fand. Auf Grundlage der Einträge in der „Encyclopedia of Social Science Research Methods“ und der genannten Kriterien wurde eine vorläufige Liste möglicher Themen erstellt und versucht, Autorinnen und Autoren für jeden Beitrag zu gewinnen – was

nach Angabe der Herausgeber nicht in allen Fällen gelang. Auf quantitativer Seite lassen sich exemplarisch Kapitel zu „Bayesianische Datenanalyse“ (Andreas Broscheid), „Item-Response Modellierung“ (Daniel Finke), „Nichtparametrische und semiparametrische Verfahren“ (Martin Elff), „Ökologische Inferenz“ (Thomas Gschwend), „Propensity Score Matching“ (Markus Gangl) nennen. Die Auswahl qualitativer Verfahren umfasst etwa „Biographieforschung“ (Bettina Dausien), „Prozessanalyse“ (Frank Schimmelfenig) oder „Qualitative Comparative Analysis and Fuzzy Sets“ (Carsten Q. Schneider). Eine Aufzählung aller Analyseverfahren würde an dieser Stelle zu weit führen und stattdessen wird auf die folgende Website verwiesen, die genau das leistet: <http://www.nomos-extra.de/ne/Behnke_2225-2_start.lasso>.

Dem Ziel, eine Einführung in den jeweiligen Sachverhalt zu geben, wird durch den standardisierten Aufbau der meisten Einzelbeiträge Rechnung getragen. Zunächst erfolgt eine Kurzdarstellung des Verfahrens, gefolgt von einer ausführlicheren Beschreibung. Schwächen des Verfahrens werden in einem separaten Abschnitt zur „innermethodischen Kritik“ behandelt. Die Vorstellung eines typischen Anwendungsbeispiels sowie kommentierte Literaturempfehlungen beschließen die Ausführungen eines jeden Kapitels. Für ein Herausgeberwerk eher untypisch, aber hervorhebenswert ist der ausführliche Index am Ende des Bandes.

Eine Kritik an der Auswahl der Verfahren ist (eigentlich) müßig, denn immer finden sich Analyseverfahren, die einer Aufnahme Wert gewesen wären. So hätten beispielsweise noch Beiträge zur Meta-Analyse, dem sogenannten „Data-Mining“ oder Strukturgleichungsmodelle gut in das Konzept des Bandes gepasst. Oben wurde darauf hingewiesen, dass die Herausgeber nicht für alle geeigneten Verfahren Autoren gewinnen konnten. Ob das auch für die genannten drei Verfahren gilt, lässt sich leider nicht

sagen. Insofern wäre eine Auflistung der relevanten, aber unbearbeiteten Themen interessant gewesen. Bereits hinreichend etabliert und damit verzichtbar sind dagegen Beiträge wie etwa zu „Modelle diskreter Entscheidungen“, „Grounded Theory“ oder „Ereignisanalyse“ – ohne mit der Nennung dieser Beiträge anzudeuten, sie seien schlecht geschrieben. An der Autorenauswahl lässt sich kaum Kritik üben. Im Gegenteil, die Wahl der in vielen Fällen einschlägig bekannten Autoren und die damit einhergehende Qualität der Beiträge ist ein starkes Argument für das Buch.

Mit (wenigen) Einschränkungen können vor allem Einsteiger in ein bestimmtes Analyseverfahren Gewinn aus dem Band ziehen. Gewähr dafür bietet der in den meisten Kapiteln gut umgesetzte einheitliche Aufbau. Die Beiträge, die das nicht leisten, tun das zumeist zu ihrem Nachteil – vor allem dann, wenn etwa keinerlei Ausführungen zu innermethodischer Kritik zu finden sind. Die vereinzelt praktizierte Neigung, im Rahmen der kommentierten Literaturliste vor allem auf die eigenen Werke abzuheben, hat einen etwas eitlen Beigeschmack.

Viele der vorgestellten Verfahren und Techniken setzen den Einsatz von Software voraus. Dem lehrbuchartigen Charakter entsprechend, sollten deshalb Beispielcodes verfügbar gemacht werden. Für die Mehrzahl der Beiträge wird auch auf sogenannte „Online Materialien“ verwiesen, doch es wird leider keine konkrete WWW-Adresse genannt. Das ist umso ärgerlicher, als dass in einigen Fällen dort nicht nur Programmcodes, sondern weiteres und die jeweiligen Beiträge ergänzendes Textmaterial zu finden ist – die einschlägige URL wurde oben genannt.

Abschließend sei noch die Frage erlaubt, ob die Titelwahl „Methoden der Politikwissenschaft“ dem Buch gerecht wird. Es gab möglicherweise gute Gründe dafür, nur die Politikwissenschaft anzusprechen, doch ein nicht

unerheblicher Teil der Autoren wurde aus anderen sozialwissenschaftlichen Disziplinen rekrutiert. Der Hinweis auf die Politikwissenschaft engt den Adressatenkreis unnötig ein, denn auch Vertreter anderer Fächer können Gewinn aus der Lektüre ziehen.

Trotz der genannten Kritikpunkte ist es den Herausgebern gelungen, eine gut lesbare und informative Sammlung von Beiträgen zu neueren qualitativen und quantitativen Analyseverfahren der empirischen Sozialwissenschaften zusammenzustellen. Personen mit einem gezielten Interesse an einem der beschriebenen Verfahren wird ein guter Einstieg vermittelt. An Methoden interessierte Leserinnen und Leser ohne spezielle Fragestellungen lädt der Band zum Schmökern ein und erlaubt auf angenehme Art und Weise, den eigenen Horizont zu erweitern.

BERND WEIß, KÖLN

* * * * *



NICOLE BURZAN, 2008: Quantitative Forschung in der Sozialstrukturanalyse. Anwendungsbeispiele aus methodischer Perspektive. Wiesbaden: VS-Verlag. ISBN: 978-3-531-15550-0, 184 Seiten, 19,90 EUR.

Wer Einführungen in die Sozialstrukturanalyse lehrt, kennt das Problem. Die neuere Literatur ist methodisch recht anspruchsvoll, aber viele Studierende haben zu diesem Zeitpunkt nur eine geringe Kenntnis im Bereich sozialwissenschaftlicher Methoden. Einerseits werden Lehrveranstaltungen zur

Sozialstruktur traditionell für angehende Sozialkundeführer angeboten. Andererseits hat sich die empirische Sozialstrukturforschung in den letzten Jahren immer stärker zu einem methodisch höchst anspruchsvollen Kernbereich der Soziologie entwickelt, der mit den traditionellen, auf die Lehramts-Zielgruppe ausgerichteten Curricula wenig kompatibel ist.

Mit ihrem Buch versucht Nicole Burzan eine Lösung für dieses Problem. Das Buch ist so konzipiert, dass es für Studierende ohne oder mit nur geringen Kenntnissen in Erhebungsmethodik, Statistik und Datenanalyse lesbar sein soll, während es – im scheinbaren Gegensatz dazu – die Bedeutung gerade dieser Verfahren für die angewandte Sozialstrukturforschung thematisiert. Anhand von Beispielen werden die Wege der Forschung erläutert, die Abhängigkeit der Ergebnisse von methodischen Entscheidungen wird dargestellt.

Nach der Einleitung gibt die Autorin eine kurze Einführung in den Forschungsprozess, wobei sie die Schwerpunkte auf Finanz- und Zeitplanung legt und auch die Drittmittelbeschaffung thematisiert. Diese wichtigen Dinge verdienen sicher ihre Abhandlung in einem solchen Buch, vermutlich sind sie aber nicht gerade von prominenter Bedeutung bei der Motivation von Studierenden zum Verständnis quantitativer Sozialstrukturanalyse.

Es folgen sechs Kapitel mit inhaltlichem Schwerpunkt, in denen jeweils die Methodenabhängigkeit der Ergebnisse zur Diskussion gestellt wird. Gegenstand des ersten Kapitels ist die Armutsforschung. Hier referiert die Autorin diverse Studien und thematisiert unter anderem Einkommenskonzepte, Einkommenserhebung, die Kostendegression im Haushaltsverbund, verschiedene Arten der Mittelwertbildung, zeitliche Dimensionen der Armut und das Verhältnis von Armut und Reichtum. Meines Erachtens unterliegt sie aber in der Zustimmung zum

Zitat der Bundesregierung (S. 23), Armut sei im „streng wissenschaftlichen Sinne“ nicht messbar, einem Irrtum, weil jede Operationalisierung – wie eben auch die von Armut – Annahmen unterliegt. Die Behauptung, die Beschreibung des Reichtums habe gegenüber der Armutserhebung die Besonderheit, dass neben Einkommen Vermögen zu berücksichtigen sei (S. 43), ist nicht ganz richtig, wenn die Autorin auf der Seite davor in Anlehnung an das IAB korrekterweise Überschuldung (also gewissermaßen negatives Vermögen) als einen „Indikator zur Armutsbestimmung“ aufführt.

Das nächste Kapitel befasst sich mit Problemen der Altersforschung. Ausführlich werden zunächst Abgrenzungsprobleme bei der Definition der Älteren diskutiert, ebenso für diese Personengruppe typische Probleme bei der Erhebung. Dann werden drei publizierte Studien exemplarisch dargestellt, nämlich eine zum Einfluss des Alters auf die Mediennutzung, eine zweite zu Handlungstypen im Alter und schließlich als drittes eine ältere, auf Beobachtungen gestützte Untersuchung von Bewohnern eines Altersheims. Das APK-Problem wird in knapper Form zwar angesprochen, ohne jedoch den Studierenden wirklich das Problem analytisch deutlich zu machen. Der Index zur Nutzung von Massenmedien erscheint mir auch nicht gerade konsistent, hier wäre die Angabe einer Korrelationsmatrix der Items sinnvoll gewesen.

Das folgende Kapitel befasst sich mit der Bildungshomogamie; zwei retrospektive, auf Deutschland bezogene Kohortenstudien und ein internationaler Querschnittsvergleich werden dargestellt. In der durchaus ausführlichen Diskussion zur Problematik der Bildungsklassifikation wird leider nicht erwähnt, dass mit CASMIN und ISCED Standardverfahren existieren, stattdessen wird eine wenig überzeugende ad hoc Klassifikation auf Basis einer Kombination allgemeiner und beruflicher Bildung dargestellt, die mir mit der

starken Berücksichtigung der beruflichen Bildung zwar in Hinblick auf Erwerbsperspektiven einleuchtet, nicht jedoch bezogen auf kulturelle Korrelate der Bildungsabschlüsse. Der eingeschobene Exkurs zu multivariaten Analysemethoden (S. 80f.) wird wohl eher für eine Mischung aus Staunen und Verwirrung sorgen.

Schließlich befasst sich die Autorin in einem weiteren Kapitel ausführlich mit der empirischen Überprüfung der Individualisierungsthese. Dabei stellt sie eine deutsche Studie mit US-Daten und eine Studie zur Entwicklung der Statusinkonsistenz in Deutschland ab 1980 vor. Der durchaus gelungenen Problemdarstellung – einschließlich der hier erforderlichen Kritik am geringen Präzisionsgrad der Theorie – korrespondiert eine ordentliche Darstellung der zwei Studien, die allerdings nur sehr bedingt repräsentativ für die umfangreiche Literatur zu diesem Thema sind.

Zwei weitere Kapitel wurden von Mitarbeiterinnen Burzans verfasst (was allerdings nur auf der inneren Titelseite und nicht auf dem äußeren Einband sichtbar wird). Die Zusammenfassung der Magisterarbeit Rückerts beschäftigt sich mit Gerechtigkeitsvorstellungen im internationalen Vergleich. Zentrale Fragestellung ist die Legitimation sozialer Ungleichheit, vier Dimensionen (Funktionalismus, meritokratische Orientierung, soziales Kapital, Egalitarismus) werden theoretisch vorgestellt und in einer Faktorenanalyse auf Basis von ISSP-Daten abgebildet. Um den Lehrbuchcharakter zu wahren, hätten allerdings in Tabelle 14 auch erklärte Varianzen oder Eigenwerte der Faktoren angegeben werden müssen.

Schließlich gibt Lökenhoff eine kurze Einführung in die Netzwerkanalyse, die sie – in der egozentrierten Variante – anhand einer eigenen, zuvor nicht veröffentlichten Studie über die Kinderbetreuungs-Netzwerke von 30 Müttern in Deutschland vorstellt. Dabei zeigt sie anschaulich das Potential von

Netzwerkanalysen, weniger schön ist die Ausweisung von Nachkommastellen bei Prozenten (S. 161) bei dieser Fallzahl. Aufgrund der Willkür der Stichprobe ist es ohnehin nur mit großen Einschränkungen sinnvoll, Prozentanteile anzugeben.

Eine kritische Besprechung dieses Buchs steht vor dem Problem, dass die einzelnen Studien nur als Beispiel dienen. Studien, die zum Zeitpunkt der Erstellung des Buchs noch nicht veröffentlicht waren, werden so gegen Kritik immunisiert, denn sie sollen ja nur als Beispiel dienen. (Ob unveröffentlichte Studien überhaupt als Lehrbuchbeispiele – jedenfalls als solche für erfolgreiche Forschung – taugen, sei einmal dahingestellt.) Wenn dann aber ohne Beleg Trends wie ein Abbau „institutioneller Kinderbetreuungsangebote“ postuliert werden (S. 141), aus denen sich dann ein „allgemeines Interesse an den informellen Hilfsressourcen von Müttern“ „ableiten“ soll, stellt sich doch die Frage nach der empirischen wie argumentativen Legitimierung solcher Behauptungen.

Die Hagener Studientexte sollen – ihrem eigenen Anspruch zufolge – „mit einer verständlichen Sprache und mit einer unaufdringlichen, aber lenkenden Didaktik zum eigenen Studium anregen“ (S. 2). Für Anfänger sicher ganz unverständliche Textteile zu „lorenz-konsistenten Maßen“ (S. 38) und „LAG-Sequenzanalysen“ (S. 67) erwecken aber den Eindruck des Versuchs über Rätselhaftigkeit zu motivieren, einer im Kontext soziologischer Theorievermittlung traditionell durchaus erfolgreichen Strategie. Dass den Studierenden zum Nachschlagen ein Verzeichnis inhaltlicher Stichworte fehlen dürfte, stellt einen weiteren Mangel dieses gut gemeinten Bandes dar, der Studierende weder zur Durchführung eigener Sozialstrukturanalysen befähigen dürfte noch solide Grundlagen in der Interpretation sozialstruktureller Daten vermittelt.

PETER HARTMANN, DÜSSELDORF



CHRISTIAN FLECK, 2007:
Transatlantische
Bereicherungen.
Zur Erfindung der
empirischen Sozial-
forschung.
Frankfurt am Main:
Suhrkamp. ISBN-10:
3518294237 ISBN-13:
978-3518294239,
578 Seiten, 18 EUR.

Philanthropic foundations endowed by entrepreneurs who made enormous fortunes in business and industry are a peculiarity of the United States, and none played a greater role in the advancement of science than the Rockefeller Foundation (RF), established in 1913 by the enormously wealthy founder of the Standard Oil Corporation. The purpose of the RF was to "promote the well-being of mankind throughout the world" and among others, to examine the causes of social problems and cure them at their source. It supported generously higher education and research in the U.S., most notably the founding of the University of Chicago, established the first schools of public health (Harvard and Johns Hopkins), funded research for a vaccine to prevent yellow fever, programs in maternal health, contraception and sex education, the development of the social sciences, and later the 'green revolution'. Moreover, it was an international philanthropy, e.g. it funded the first modern medical school in China in 1921, the Peking Union Medical College. My aunt Ilma Oberschall who later became a founder of the Freedom Party in 1945 in Hungary headed a RF program in rural health in the late 1930s in Hungary and Transylvania. In 1921 the RF started an international fellowship program to train

scholars; by 2000 there had been 13,000 Rockefeller Fellows. In the 1930's the RF helped many German and Austrian scholars and intellectuals relocate, among others Thomas Mann, Claude Levi-Strauss and Leo Szilard.

Christian Fleck's topic are the contrast between German and U.S. universities' openness for social science research and innovation, the activities of the RF and especially the Laura Spellman Rockefeller Memorial Fund (which merged with the RF in 1928 and had a social science focus) on behalf of German and Austrian social science in the interwar years, and the opportunity for social scientists, in particular those that defined themselves as 'sociologists', to emigrate to the United States following the Nazi seizure of power, establish careers there, and innovate in empirical sociology. Although the relocation story has been told by prominent émigré scholars in their biographies and by historians focusing on distinguished scientists, Fleck believes that his archival research at foundations, universities and in the private papers of scholars provides a fuller and unbiased account of German and Austrian sociology in the interwar years and shortly after the war. He describes the social science program of the RF in Germany and Austria, puts together a 'collective biography' of some 800 German language social scientists, analyzes the accomplishments and career of the émigrés, contrasts the 'movers' with the 'stayers', and presents two case studies of émigré research in the U.S., both receiving some RF financing, Paul Lazarsfeld's Princeton Radio Research Project of 1937-41 and the Frankfurt Institute for Social Research's Studies in Prejudice, which culminated with the influential *The Authoritarian Personality* (1950).

At the turn of the 19/20 century many U.S. social scientists and historians studied at German universities and returned to estab-

lish graduate education and PhD programs on a scale unimagined in Europe. Because of the rapid expansion of youth attending colleges and the founding of new universities, there was steady demand for college teachers and post-doctoral programs for recruiting them to leading institutions. Several professors in the same field and department were established, and team work was expected between them and between senior and junior faculty. The model in the social sciences was the University of Chicago and its Local Community Research Committee, generously funded by the RF. Chicago was the immigrant megacity par excellence in the 1910s and 1920s and inspired Robert Park to refer to it as a 'social laboratory'. American sociology became an established social science because its subject matter was the adjustment, acculturation and assimilation of European migrants (many rural) to American cities, industry and way of life. Moreover, U.S. sociologists did not have a socialist political disposition. The American professorate accepted capitalism and democracy and shared the 'social progress through science' ideology of the business elite that endowed the foundations, and of their philanthropy administrators. It is because social science was expanding and open to newcomers that the European émigrés of the 1930's were successfully absorbed as social scientists, which they would not have been able to accomplish in Germany and Austria, even without Nazism.

At RF, Beardsley Ruml directed huge funds to European scholars and institutions through the European fellows program and funding books, journals, and staff for research centers and sent American fellows to study in Germany. The foundation administrators cultivated academic links in Europe and were persistently looking for promising scholars and projects to support. In Germany, recipients were the Hamburg Institut für Auswärtige Politik and the Kieler Institut für Weltwirtschaft. After the

Nazi regime most of the Kieler RF fellows emigrated and the RF trustees stopped funding the institutes. In Austria, the Buhlers and most of the other RF fellows emigrated after 1938. Altogether the RF helped 17 Austrian and 53 German former social science fellows between 1929 and 1941 emigrate for political reasons, including Albert Hirschman, Jakob Marschak, Fritz Machlup, Paul Lazarsfeld, and Leo Strauss. That was a huge gain for the U.S. and a loss to Germany and Austria.

Fleck devotes a long chapter and a lot of research to a collective portrait of some eight hundred German language social scientists, ranks them on productivity, visibility, and recognition based on citations and other criteria, and compares their career paths, émigrés versus stayers, Germans versus Austrians, birth cohorts, men versus women. I did not find this material particularly useful because Fleck does not connect it to the creation and establishment of social science. It is good to know that Karl Popper and Friedrich Hayek scored high (sixth and seventh) on reputation, but it in no way explains their extraordinary cross disciplinary and international influence on twentieth century social science. Similarly, the story of Lazarsfeld at the Office of Radio Research, which at some time or another involved the RF, Frank Stanton at CBS, Hadley Cantril at Princeton, public opinion polling, market research, Theodor Adorno in his guise as 'music expert' and some others, focuses on the minute details of their interpersonal relations without providing an understanding of how mass communications as a quantitative social science field got started. Fleck centers the story on how *Radio and The Printed Page* (1940) was finally published, yet the *Language of Social Research* (1955), the core text of the Lazarsfeld school of methodology, reprints only three pages out of 590 from it. Lazarsfeld and Adorno had a polar opposite conception about human choice and how to

study it that could not be reconciled. Lazarsfeld made choice central to his life-long ambition to create the empirical study of action, whether it be consumer choice, political choice, occupation choice, choice of residence, choice of spouse. The actor has dispositions which interact with influences, some from his social milieu, some from exposure to mass media, that solidifies intentions into a choice. The process could be studied quantitatively using repeated interviews and observations with the same persons, called the panel method. This was a genuine innovation in social science and first tried in the now classic *The People's Choice* (1944). Adorno on the other hand denied human agents real choice in a capitalist mass society that manipulated their mentality, and eleven years in the U.S. did not shake this conviction. One cannot build methodological individualism from such a foundation. Despite Lazarsfeld's multiple efforts to include Adorno in his research projects and get them funded by the RF, these two lived on different social science planets, as Fleck makes clear.

Similarly the chapter on the Frankfurt Institute and the Studies in Prejudice research in New York, Los Angeles and Berkeley in cooperation with some U.S. social scientists is full of interpersonal detail – intrigue is actually a better word to describe some of the material – but lacks a bridge to the book's subtitle "Zur Erfindung der Empirischen Sozialforschung." There are some intriguing loose ends. Horkheimer hypothesized that anti-semites were not a group with a distinct personality type, as many of the researchers thought, and that under some circumstances all persons were at some risk of becoming anti-semitic. This is similar to the contemporary controversy over what sorts of people become suicide bombers, terrorists, war criminals, and torturers. Why did not Horkheimer's views prevail? It would have resulted in a different research instrument to measure intolerance and anti-

semitism for the "Authoritarian Personality" and could have stimulated an innovative political psychology.

Although well written and researched, the book is too long with 580 pages. One ends up learning about a lot of people that are not particularly noteworthy. There are some omissions. For instance, table 6.1 on page 355 lists more than twenty prominent German writers and books on totalitarianism and related topics (Hannah Arendt, Joseph Schumpeter, Franz Neumann, etc). Most are political and philosophic treatises. Unmentioned is the classic *From Democracy to Nazism* (1945), by the émigré Rudolf Heberle, which was the first empirical (as opposed to philosophical) account of the rise of the Nazi party and regime and which I studied with great profit as a student, and later used in my lectures and writings on the rise of Nazism.

Whatever its shortcomings, the book tells a remarkable story, only partially known and somewhat sanitized in autobiographical accounts, of successful transnational cooperation in social science which, despite the human frailties of the protagonists, led to pioneering work and innovation.

ANTHONY OBERSCHALL, CHAPEL HILL

Ankündigungen

Call for Papers

**European Labour Force Survey
(EU-LFS) and
European Union Statistics on Income
and Living Conditions (EU-SILC):
European User Conference**

Organized by
German Microdata Lab, GESIS,
in cooperation with Eurostat

*Mannheim
March 5 – 6, 2009*

*European User
Conference
EU-LFS and
EU-SILC*

EU-LFS and EU-SILC are the most important official micro databases for comparative social research in Europe. With the availability of these surveys as scientific use files more and more researchers use these data for a wide range of economic and social analyses; topics addressed include e.g. intergenerational transmission of economic disadvantages, transition from school to work, gender gaps on the labour market. Against the backdrop of the rapidly growing use of EU-LFS and EU-SILC micro data there is a clear need to share experience between researchers, to provide feedback to producers, but also to learn more about the way these European statistics are developed, compiled and disseminated.

The first European User Conference for EU-LFS and EU-SILC, which is organized by the German Microdata Lab, GESIS, in cooperation with Eurostat, will provide researchers who use these data with the opportunity to present and discuss their work and share their experience. In addition to fostering the discussion within the research community on both substantive and methodological issues, the conference offers researchers the opportunity to give feedback to the European Statistical System. It is important that the potential and limitations of EU-LFS and EU-SILC are understood by researchers and, at the same time, that the needs of researchers become clear.

Researchers of all disciplines (e.g. economics, demography, geography, political science, public health and sociology) who use either EU-LFS or EU-SILC micro data are encouraged to participate and to submit an abstract. Substantive topics may include, among others, all aspects of the European labour market, living conditions, and static or dynamic aspects of income poverty and social exclusion. Methodological topics may include e.g. questions of data quality, cross-national and inter-temporal comparability, and statistical modeling. All presentations must be comparative and include data from at least three countries.

SUBMISSIONS:

The deadline for submissions of abstracts is October 31, 2008. Please send your submissions to the local organizers of the conference:

Christof Wolf (Christof.Wolf@gesis.org) and
Heike Wirth (Heike.Wirth@gesis.org).

Abstracts should not be longer than 350 words; the abstract should also indicate the dataset(s) used and the countries analyzed.

The submitters will be notified by December 1, 2008, whether their paper was accepted for presentation or not. For any further questions please contact the local organizers and visit the conference webpage at http://www.gesis.org/Veranstaltungen/Konferenzen/EU_User_Conference/index.htm

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 Tabellenummer 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 zu liefern; Auflösung mindestens 150 dpi bei Schwarz-Weiss-Grafiken, 300 dpi bei farbigen Grafiken.

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

Werden in einem Beitrag empirische Daten verwendet, 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 Normalfall 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:

Fuchs, M. und M. Sixt, 2007: Zur Nachhaltigkeit von Bildungsaufstiegen. Soziale Vererbung von Bildungserfolgen über mehrere Generationen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 59: 1-29.

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. ZUMA How-to-Reihe, Nr. 14 (Mannheim: Zentrum für Umfragen, Methoden und Analysen, ZUMA). http://www.gesis.org/Publikationen/Berichte/ZUMA_How_to/Dokumente/pdf/how-to14rp.pdf (29.06.2007).

ISSN 1864-6956

2. Jahrgang 2008 © GESIS, Mannheim, Juni 2008