methoden daten analysen
ZEITSCHRIFT FÜR EMPIRISCHE SOZIALFORSCHUNG

mda

2011, Jahrgang 5, Heft 1

Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung

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

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

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

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

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

http://www.gesis.org/publikationen/zeitschriften/mda/bestellung/

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

ISSN 1864-6956

5. Jahrgang 2011

Inhalt

ANKÜNDIGUNGEN

Kontextgestützte Abfrage in Retrospektiv-erhebungen

Ein kognitiver Pretest zu Erinnerungsprozessen bei Weiterbildungsereignissen

Context Aided Retrieval in Retrospective Surveys

A Cognitive Pretest of Memory Processes Related to Episodes of Further Education

Andrea Dürnberger, Katrin Drasch und Britta Matthes

Zusammenfassung
Weiterbildungsaktivitäten gelten als schwer erinnerbare Ereignisse. Daher wurde bislang auf die retrospektive Erhebung nicht-formaler und informeller Weiterbildungsaktivitäten weitgehend verzichtet. Im Rahmen der IAB-Studie „Arbeiten und Lernen im Wandel (ALWA)" stehen jedoch auch kurze und unbedeutende, oft weiter zurückliegende Weiterbildungsaktivitäten im Fokus. Ziel der vorliegenden Studie war es zu untersuchen, ob die Erinnerung an diese Aktivitäten mittels einer kontextgestützten Erfassung verbessert werden kann. Daher sind wir der Frage nachgegangen, ob es beim Erinnern von Weiterbildungsereignissen einen Zusammenhang zwischen den von den Befragten verwendeten Erinnerungsstrategien und der Anzahl der Ereignisse sowie dem zeitlichen Abstand zu den Erinnerungskontexten – in unserem Fall den Erwerbsepisoden – gibt. Dabei übertragen wir den Ansatz des kognitiven Pretests – einer qualitativen Methode, die ursprünglich zur Validierung

Abstract
Further education and training activities are considered to be recalled only with difficulty. Therefore, non-formal and informal training activities have rarely been collected in retrospective surveys. However, within the framework of the IAB-ALWA study "Working and Learning in a Changing World" short and relatively insignificant training activities that occurred sometime in the past are also of central interest. It was the goal of this study to examine whether recalling these activities is aided by providing context. In particular, we addressed the question whether there is a connection between the memory strategies used by respondents and the amount of events as well as the time-lag to the context of memorization when remembering further education episodes. In our study the context of memorization were employment episodes. Our approach involved applying cognitive pretesting – a qualitative method originally developed to validate the understanding of

des Verständnisses von Survey-Fragen entwickelt wurde – auf die Untersuchung der Art und Weise des Abrufens von an sich schwer erinnerbaren Ereignissen.
Die Ergebnisse des kognitiven Pretests zeigen, dass sich Befragte gut an Weiterbildungsaktivitäten erinnern, die weniger weit in der Vergangenheit liegen. Bei länger zurückliegenden Ereignissen werden meist kontextgestützte Erinnerungsstrategien angewandt. Insgesamt zeigt sich, dass die kontextspezifische Abfrage ein geeignetes Mittel für die retrospektive Erfassung von Weiterbildungsaktivitäten ist, die nur wenige Jahre zurückliegen. Mit einer Zunahme des retrospektiven Zeitraums wachsen die Erinnerungsprobleme allerdings so stark, dass diese auch durch eine kontextspezifische Abfrage nicht behoben werden können. Die aufgrund des kognitiven Pretests gewonnenen Erkenntnisse wurden anschließend in der Haupterhebung der IAB-ALWA Studie und in der Nachfolgestudie, der ersten Welle der Erwachsenenetappe des Nationalen Bildungspanels (NEPS) bei der Instrumentenentwicklung eingesetzt.

survey questions – to the analysis of recall processes that are related to events that are difficult to remember per se.
The results of the cognitive pretest show that respondents remember training activities from the recent past well. For events that are further back in the past, context-based memory strategies are frequently used. In sum, we demonstrate that context-based data collection is an appropriate instrument for the retrospective collection of training activities which occured only a few years ago. The longer the retrospective intervall is, however, the greater are the memory problems which cannot be compensated by context-based retrieval. The findings from the cognitive pretest were used for designing instruments of the main survey of the IAB-ALWA study and its follow-up study, the first wave of the adult stage of the National Educational Panel Study (NEPS).

1 Einleitung

Weiterbildungsaktivitäten stellen aufgrund ihrer kurzen Dauer und mangelnden Einbettung in Berufs- und Privatleben der Personen zumeist besonders schwierig zu erinnernde Inhalte dar. Für die retrospektive Erfassung von Weiterbildungen, die zeitlich nicht sehr weit in der Vergangenheit liegen, gibt es in verschiedenen Studien in Deutschland bereits bewährte Instrumente (z. B. Adult Education Survey (AES), Berichtssystem Weiterbildung (BSW), BIBB-IAB-Erwerbstätigenbefragung, Mikrozensus (MZ), SOEP (Sozio-oekonomisches Panel)).

Diese empirischen Erhebungen liefern jedoch sehr unterschiedliche, zum Teil sogar widersprüchliche Informationen über die Weiterbildungsbeteiligung in Deutschland (vgl. Wohn 2007). So nahmen entsprechend der Daten des Berichtssystems Weiterbildung im Jahr 2003 41 % der erwachsenen Bevölkerung an Weiterbildung teil (ebd. S. 9), während sich laut Mikrozensus im selben Jahr nur knapp 13 % der Erwachsenen weiterbildeten (ebd. S. 20). Diese starken Abweichungen erklärt

Wohn (2007) insbesondere durch große Unterschiede zwischen den Frageinstrumenten: Im Mikrozensus wird mit einer offenen Frage nach der Weiterbildungsbeteiligung einschließlich einer umfangreichen Liste an Beispielen, die formale nicht-berufliche, aber auch die formale berufliche Weiterbildung untererfasst. Dagegen wird die Weiterbildungsbeteiligung im Berichtssystem Weiterbildung durch die themengestützte Abfrage der Teilnahme an einem bestimmten Lehrgang, Kurs oder Vortrag ermittelt, was zu einer Überschätzung vor allem der formalen nicht-beruflichen Weiterbildung führt (ebd. S. 1f.). Offensichtlich wird der Begriff „Weiterbildung" in beiden Studien sehr unterschiedlich interpretiert.

Diese Abfragemodi sind also bereits für die Erfassung von Weiterbildungsaktivitäten problematisch, die nur ein oder drei Jahre zurückliegen. Die Erfassung von noch länger zurückliegenden Weiterbildungsereignissen scheint damit nahezu unmöglich. In den meisten Befragungen zur Weiterbildungsbeteiligung wurde daher auf die retrospektive Erfassung von länger zurückliegenden Weiterbildungsaktivitäten verzichtet. Will man dennoch unterschiedliche Weiterbildungsaktivitäten für einen länger zurückliegenden Zeitraum erheben, bleibt erstens festzuhalten, dass es nicht ausreicht, die Frage zu stellen, ob eine Person im letzten Jahr (oder in den letzten Jahren) an einer Weiterbildung teilgenommen hat. Es sind Erinnerungshilfen notwendig, die die Befragten beim Erinnern an Weiterbildungsaktivitäten unterstützen. Zweitens dürfen diese Erinnerungshilfen nicht allein themengestützt sein, da dadurch eine Übererfassung von Weiterbildungsereignissen möglich ist.

Die IAB-Erhebung „Arbeiten und Lernen im Wandel" (ALWA) (Antoni et al. 2010) hat sich unter anderem zum Ziel gesetzt, im Rahmen der retrospektiven Erhebung des gesamten Bildungs- und Erwerbsverlaufs auch alle Weiterbildungsaktivitäten, deren Umfang und Datierung zu erfassen (Kleinert/Matthes/Jacob 2008).[1] Viele Erfahrungen, die im Rahmen der ALWA-Erhebung gesammelt wurden, sind in das Erhebungs- und Fragebogendesign der Erwachsenenetappe des Nationalen Bildungspanels (NEPS) eingeflossen (vgl. Allmendinger et al. 2011).

Im vorliegenden Papier wird die zentrale Annahme überprüft, ob auch die Erfassung von kurzen und relativ unbedeutenden, weiter zurückliegenden Weiter-

1 ALWA ist ein Datensatz, der im Rahmen des Projektes „Qualifikationen, Kompetenzen und Erwerbsverläufe" am Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA), Nürnberg, erhoben wurde. Die ALWA-Daten enthalten detaillierte Informationen über die Bildungs- und Erwerbsverläufe, die Wohnort-, Partner- und Kindergeburtsgeschichte von 10.400 Personen der Geburtsjahrgänge 1956-1988 und erlauben Längsschnittanalysen insbesondere zum Schul- und Ausbildungsverhalten, zum Erwerbsverlauf sowie zu Prozessen der Familienbildung und der regionalen Mobilität. Für externe Wissenschaftler wurde ALWA als Scientific Use File aufbereitet und kann über das Forschungsdatenzentrum der BA (FDZ) bezogen werden (http://fdz.iab.de/de/FDZ_Individual_Data/ALWA.aspx).

bildungsaktivitäten möglich ist, wenn die Erfassung dieser Aktivitäten kontextgestützt erfolgt, d. h. im Zusammenhang mit anderen Ereignissen. Die Idee dabei ist, dass zunächst die einzelnen Lebensbereiche modulweise erfragt werden (Matthes/Reimer/Küster 2007; Reimer/Matthes 2007). Die einzelnen Phasen im Bildungs- und Erwerbsverlauf werden also nicht – wie sonst häufig bei Befragungen – sukzessive entlang einer chronologischen Zeitachse abgefragt, sondern die Befragten gehen mehrmals, d. h. für jeden einzelnen Lebensbereich (wie z. B. Schule, Ausbildung, Erwerbstätigkeit, Arbeitslosigkeit etc.) separat durch ihren Lebensverlauf. Für jeden dieser Lebensbereiche wird einzeln gefragt, ob es Weiterbildungsaktivitäten gab oder nicht. Dieses Design orientiert sich an der Struktur des autobiografischen Gedächtnisses, also daran, in welcher Art und Weise Individuen persönliche Erinnerungen speichern. Dadurch sollen retrospektive Erinnerungsprozesse unterstützt werden (vgl. Drasch/Matthes 2009).

Lebenslanges Lernen kann als „die Gesamtheit allen formalen, nicht-formalen und informellen Lernens über den gesamten Lebenszyklus eines Menschen hinweg" (Timmermann et al. 2004: 6) verstanden werden. In ALWA fassen wir als formale Weiterbildung solche Bildungsaktivitäten auf, die zu einem anerkannten Zertifikat führen (Kleinert/Matthes 2009). Das können Ausbildungen sein, die in formalisierten Bildungseinrichtungen stattfinden, aber auch Vorbereitungslehrgänge für Prüfungen (wie z. B. Meister- oder IHK-Prüfungen). Unter nicht-formalen Weiterbildungen verstehen wir kürzere Weiterbildungen wie Kurse und Lehrgänge, die zu keinem bzw. keinem anerkannten Zertifikat führen, wie z. B. Computerkurse (Kleinert/Matthes 2009). Mit informellen Weiterbildungsaktivitäten können einerseits intendierte Selbstlernaktivitäten, wie das Lesen von Fachbüchern oder das Einarbeiten in ein neues Computerprogramm, gemeint sein. Andererseits wird darunter auch nichtintendiertes informelles Lernen verstanden (Kleinert/Matthes 2009). Grundsätzlich lässt sich allerdings nichtintentionales Handeln in einer Personenbefragung nicht erfassen, da es ohne Absicht der Zielperson geschieht und deswegen nicht erinnert werden kann.

Im Fokus des vorliegenden Beitrages stehen die nicht-formalen und die intendierten informellen Weiterbildungen, weil sie gewöhnlich gemeint sind, wenn von Weiterbildung im Erwachsenenalter gesprochen wird. Die formalen Bildungsaktivitäten werden aufgrund ihrer starken Institutionalisierung in ALWA separat in einem eigenen Modul erfasst und sind deshalb nicht Bestandteil des vorliegenden kognitiven Pretests. Wir beschränken uns darüber hinaus auf die Untersuchung der Erfassung nicht-formaler und informeller Weiterbildungen, die im Zusammenhang mit Erwerbsepisoden erinnert werden. Hierbei setzt die berufliche Tätigkeit den Kontext für die Erfassung dieser Weiterbildungsaktivitäten. Wir gehen aber davon

aus, dass die daraus resultierenden Erkenntnisse auch auf die anderen Lebensbereiche übertragen werden können.

Ein kontextgestütztes Erhebungsverfahren von Weiterbildungsereignissen wurde bislang noch nicht bei Befragungen verwendet und muss daher empirisch getestet werden. Wir gehen dabei nicht wie üblich quantitativ vor, indem wir erhobenen Retrospektivdaten Paneldaten gegenüberstellen (Powers/Goudy/Keith 1978; Peters 1988; Dex/McCulloch 1997; Klein/Fischer-Kerli 2000; Solga 2001). Unser Ansatz ist ein qualitativer, der die Forderung von Blair und Burton (1987: 288) nach einer stärkeren Auseinandersetzung mit den Mechanismen des Erinnerns und den möglichen Gründen für gute oder schlechte Erinnerbarkeit von Ereignissen berücksichtigt. Mittels eines kognitiven Pretests, einem qualitativen Verfahren, das ursprünglich zur Testung des Frageverständnisses von Survey-Fragen entwickelt wurde, wird untersucht, ob ein kontextbezogenes Erhebungsverfahren das Erinnern von Weiterbildungsereignissen erleichtert. Im Mittelpunkt des kognitiven Pretests steht hier jedoch nicht die Validierung einer bestimmten Frageformulierung oder eines bestimmten Begriffes, sondern die Art und Weise des Erinnerns von Weiterbildungsereignissen.

Die vorliegende Arbeit ist folgendermaßen aufgebaut: Im nächsten Kapitel werden zentrale Befunde der Kognitionspsychologie zu den grundlegenden Erinnerungsprozessen bei autobiografischen Ereignissen dargestellt und erläutert, wieso eine kontextspezifische Abfrage für das Erinnerungsvermögen des Befragten förderlich sein kann. Im dritten Kapitel werden Hypothesen zur Erinnerung von Weiterbildungsaktivitäten im Kontext von Erwerbstätigkeiten formuliert. Anschließend wird in Kapitel vier das Design des kognitiven Pretests vorgestellt. Im fünften Kapitel werden die zentralen Ergebnisse hinsichtlich der Erinnerungsfähigkeit und des Vorgehens beim Erinnern in Abhängigkeit von der Anzahl der zu erinnernden Erwerbsepisoden und der Länge des retrospektiven Erinnerungszeitraums vorgestellt. Im folgenden Kapitel werden Implikationen dieser Ergebnisse für die retrospektive, kontextgestützte Erhebung von Weiterbildungsaktivitäten in der IAB-ALWA Studie dargestellt. In einem kritischen Ausblick wird auf den weiteren allgemeinen Forschungsbedarf im Bereich der retrospektiven Erhebung von schwer zu erinnernden Ereignissen hingewiesen.

2 Kognitionspsychologische Befunde zur kontextspezifischen Erinnerung

In diesem Kapitel wird der Frage nachgegangen, welche kognitionspsychologischen Befunde die These stützen, dass eine kontextspezifische Abfrage das Erinnerungsvermögen des Befragten fördert. Reimer (2001) unterscheidet in Anlehnung an Barsalou (1988), Conway (1996) und Conway/Pleydell-Pearce (2000) folgende grundlegenden Vorgänge beim Erinnerungsprozess:

- die Erinnerung, dass ein bestimmtes Ereignis im Leben einer Person stattgefunden hat („Erinnern, Dass"),
- die Erinnerung an wichtige Details und Zusammenhänge im Kontext des erinnerten Ereignisses („Erinnern, Wie/Wo/Warum"),
- die Erinnerung an die „korrekte zeitliche Verortung" des Ereignisses („Erinnern, Wann") (Reimer 2001: 16).

Diese drei Erinnerungsleistungen stellen einen hierarchischen Prozess dar: Erst wenn die Person über die Erinnerung verfügt, dass ein bestimmtes Ereignis stattgefunden hat, können im zweiten Schritt weitere Informationen zu diesem Ereignis rekonstruiert werden. Im dritten Schritt wird die Datierung eines Ereignisses abgerufen, da Ereignisse nicht „time-tagged" (Wagenaar 1986) sind, d. h. Ereignisse und deren Datierung im menschlichen Gedächtnis normalerweise nicht zusammen abgespeichert werden. Daher muss die Datierung eines Ereignisses abgekoppelt vom Erinnerungsprozess an das Ereignis selbst gesehen werden und bedarf eines eigenständigen Rekonstruktionsprozesses.

2.1 Erinnerung an Ereignisse

Die Tatsache, dass sich eine Person an bestimmte Ereignisse, Sachverhalte oder Fakten erinnert, hängt entscheidend von deren Eigenschaften ab. In ihrer Studie über Erinnerungsfehler arbeiten Dykema und Schaeffer (2000) folgende wichtigen Eigenschaften der zu erinnernden Ereignisse heraus, die sich mehr oder weniger stark auf die Lücken- und Fehlerhaftigkeit von Erinnerungen auswirken: (1) die Komplexität, die sich aus Häufigkeit, Regelmäßigkeit und Ähnlichkeit der Ereignisse zusammensetzt, (2) die Eindeutigkeit der Erinnerungsinhalte sowie (3) die Intensität, die über Emotionen Erinnerungsleistungen beeinflussen kann. Daher ist anzunehmen, dass nicht nur die Fehlerquote beim Erinnern mit zunehmender Komplexität, Uneindeutigkeit und sinkender Intensität steigt, sondern auch die Erinnerungsfähigkeit insgesamt abnimmt.

Blair und Burton (1987) zeigen, dass eine steigende Anzahl von Ereignissen die Wahrscheinlichkeit eines zuverlässigen Abrufens und Aufzählens von Ereignissen verringert. Als Grund nennen sie, dass mehr Zeit und Anstrengung auf das Aufzählen verwendet werden muss, was auch dazu führen kann, dass Zielpersonen das Nennen verweigern. Ein weiterer Grund ist, dass auch die Befragungsdauer und somit die Zeit, die für den einzelnen Abruf von Ereignissen zur Verfügung steht, zeitlich begrenzt ist. Weiterhin gehen die Autoren davon aus, dass mit einer Ausweitung des zeitlichen Erinnerungsrahmens die Wahrscheinlichkeit für einen zuverlässigen Abruf der Ereignisse abnimmt. Dabei weisen sie darauf hin, dass die Anzahl der relevanten Ereignisse stark mit der Länge des retrospektiven Intervalls zusammenhängt. Grund hierfür ist, dass ein längerer Zeitraum zumeist mehr und auch zeitlich weiter voneinander entfernte Ereignisse umfasst, dadurch der Zugang zu den Ereignissen erschwert wird.

Auch Reimer (2001) nennt eine Reihe wichtiger Eigenschaften, die sich je nach Ausprägung auf die Erinnerungsfähigkeit und vor allem auf die Erinnerungsgenauigkeit auswirken: die emotionale Bedeutsamkeit des Ereignisses, dessen Folgenschwere, die Einzigartigkeit und Nicht-Erwartung, die Dichte von Ereignissen, die Ereignisdauer sowie deren Positionen innerhalb verschiedener Ereignissequenzen (ebd. 2001: 43ff.).

Die kognitionspsychologischen Forschungsergebnisse legen nahe, dass eine Reihe von Eigenschaften von Ereignissen für eine zuverlässigere Erinnerungsleistung mitverantwortlich ist. Da wir uns von vornherein auf Weiterbildungsereignisse beschränken wollen, konzentrieren wir uns im kognitiven Pretest auf die aus unserer Sicht zwei wichtigsten und empirisch am besten zugänglichen Eigenschaften:

- die Häufigkeit von Ereignissen (Anzahl der zu erinnernden Ereignisse),
- das retrospektive Intervall (zeitlicher Abstand zwischen dem zu erinnernden Ereignis und dem Interviewzeitpunkt).

2.2 Erinnerung an die weiteren Umstände von Ereignissen

Nachdem zuerst das Ereignis als solches erinnert ist, findet sich beim Erinnern innerhalb thematischer Kontexte die Struktur des autobiografischen Gedächtnisses wieder (Belli 1998; Conway 1996). Conway (1996) und darauf aufbauend Van der Vaart (2004) beschreiben das autobiografische Gedächtnis als hierarchisches Netzwerk mentaler Repräsentationen aus Erinnerungen an Lebensabschnitte, allgemeinen Ereignissen und spezifischem Wissen zu diesen Ereignissen. Das autobiografische Gedächtnis erlaubt das Abrufen von vergangenen Ereignissen über drei

verschiedene Erinnerungspfade. Erstens existieren hierarchische Erinnerungspfade, die vom Ereignis selbst zu dessen Eigenschaften („top-down") verlaufen. Zweitens gibt es sequenzielle Erinnerungspfade, die entlang einer kausal-temporären Abfolge von Ereignissen innerhalb einer Lebensdomäne verlaufen und die oft länger andauernden Ereignisse miteinander verbinden. Drittens werden noch parallele Erinnerungspfade zwischen Lebensdomänen unterschieden, die sowohl gleichzeitige als auch unmittelbar aufeinander folgende Ereignisse miteinander verknüpfen.

Auch diese zweite Erinnerungsleistung, das Erinnern, wie, wo und warum sich ein Ereignis ereignet hat, nimmt im kognitiven Pretest eine zentrale Rolle ein. Es wird in Anlehnung an die vorgestellten Erkenntnisse zur Funktionsweise des autobiografischen Gedächtnisses (Belli 1998; Conway 1996) davon ausgegangen, dass die genauen Details des zu erinnernden Ereignisses dessen unmittelbaren Kontext bilden und in den inhaltlichen Kontext des Lebensverlaufes der Zielperson (z. B. den Erwerbsverlauf im Falle von Weiterbildungsereignissen) eingeordnet sind. Es kann daher vermutet werden, dass der Zielperson das Erinnern der näheren Umstände, wie, wo oder warum sich ein Ereignis ereignet hat, leichter fallen müsste, wenn die Abfrage kontextgestützt erfolgt.

2.3 Erinnerung an die Datierung von Ereignissen

Die Datierung des Ereignisses stellt die dritte Stufe beim Erinnern von Ereignissen dar. Van der Vaart (2004) stellt fest, dass eine Abfrage mit Hilfe eines chronologischen Zeitstrahls sich zwar positiv auf die Zahl und Art der Ereignisse auswirkt, nicht aber auf deren Datierung. Andere Forschungsergebnisse zum kontextspezifischen Erinnern zeigen, dass die zeitliche Verortung eines Ereignisses mit Rückgriff auf den Kontext zuverlässiger erfolgen konnte. So konnten die Befragten in einer Studie von Auriat (1993) über die Zeitpunkte wichtiger familiärer Ereignisse deutlich schlechter Auskunft geben, wenn diese Erinnerungen ohne jegliche Erinnerungsstütze verliefen. Auch einige ältere Studien (Rogoff-Ramsøy 1973; Balan/Browning/Jelin 1973; vgl. hierzu Tölke 1979) zeigen, dass es für Befragte hilfreich ist, ihre Lebensgeschichte anhand eines selbst gewählten Lebensbereiches zu rekonstruieren. Insgesamt zeigen bisherige Forschungsarbeiten, dass eine kontextgestützte Abfrage zu weniger Lücken und Fehlern in der biografischen Erinnerung führen (Drasch/Matthes 2009). Diese Erkenntnis ist ein weiterer Grund dafür, auch in der vorliegenden Untersuchung den kontextgebundenen Abfragemodus zu wählen.

2.4 Erinnerungsunterstützung und -strategien im kognitiven Pretest

Kontextspezifisches Erinnern kann auf unterschiedliche Art und Weise unterstützt werden (Van der Vaart 2004; Sudman/Bradburn 1974; Sudman/Bradburn/Schwarz 1996). Möglichkeiten sind zum einen der aided recall und zum anderen dessen Sonderform, der bounded recall (Van der Vaart 2004). Bei einem gestützten Abruf der Erinnerungsinhalte in Form eines aided recalls werden dem Befragten so genannte memory cues genannt, die als Schlüsselreiz fungieren sollen, um den Befragten in den entsprechenden inhaltlichen Kontext zu versetzen (z. B. Erwerbsepisoden). Bounded recalls bilden einen besonderen Schlüsselreiz in Form von zeitlichen Verortungen. Durch diese Art von Reizen soll der zeitliche Rahmen, also die Referenzperiode des abgefragten Ereignisses genau abgesteckt und als Raster definiert werden, wodurch das Telescoping (Blair/Burton 1987) von Ereignissen, also eine Falschdatierung zu nah am Erhebungszeitpunkt, vermieden werden soll.

Zeitliche und/oder inhaltliche Raster als Erinnerungskontext sollen den Befragten helfen, sich zu erinnern, dass überhaupt ein bestimmtes Ereignis stattgefunden hat und neben der Datierung auch Details der Episode abrufbar machen. Im vorliegenden kognitiven Pretest dient dem Befragten der Kontext der entsprechenden Erwerbsepisode als inhaltliches und/oder zeitliches Raster. Darüber hinaus soll jede Zielperson beim kontextspezifischen Erinnern weitere, für sich selbst als geeignet erscheinende Erinnerungsstrategien verwenden. Angenommen wird, dass diese sich unterschiedlich gestalten, je nachdem, welche Art von Erinnerung vom Befragten gefordert wird.

Im Hinblick auf die Erinnerung an Episoden und Ereignisse werden die drei vorgestellten Erinnerungspfade aufgegriffen. Für den hierarchischen („top-down") Abruf gilt, dass die zentralen Lebensabschnitte (z. B. Bildungs- und Erwerbsepisoden) der Zielperson als Raster dienen, an dem sich die Zielperson, zumeist mit dem frühesten Ereignis beginnend, chronologisch entlang der Zeitachse bewegt. Dies soll die Zielperson dazu nutzen, andere meist unwichtigere Ereignisse (z. B. Weiterbildungsereignisse) im Zusammenhang mit diesen Lebensereignissen zu verorten. Wichtige Lebensereignisse können aus dem privaten Bereich stammen, wie zum Beispiel der Auszug aus dem Elternhaus, die Heirat oder die Geburt des ersten Kindes. Sie können aber auch aus dem beruflichen Umfeld sein, wie beispielsweise der Beginn der Ausbildung oder die erste Vollzeitstelle. Außerdem soll die Erinnerung „entlang von Ereignisnetzwerken" erfolgen (Reimer 2001: 100), die die verschiedenen Lebensbereiche miteinander verbinden – also mittels sequentieller und paralleler Erinnerungspfade. Es wird erwartet, dass sich Zielpersonen besser an Kurse oder Lehrgänge in ihrem Erwerbszusammenhang erinnern, wenn sie gleichzeitig Parallelen zu eben diesen privaten Lebensereignissen ziehen.

Im Zusammenhang mit dem Abruf von Häufigkeiten werden ebenfalls drei Strategien genannt (Blair/Burton 1987). Eine Möglichkeit, die Häufigkeiten von Ereignissen zu erinnern, besteht darin, zunächst nur die relevanten Ereignisse abzurufen und aufzuzählen, um dann im Anschluss daran deren Häufigkeit schätzen zu lassen. Im Fall der Anzahl der Weiterbildungen ist das Vorgehen wie folgt: die Zielpersonen nennen die Weiterbildungen und ermitteln daraus die absolute Anzahl. Der zweite Weg besteht in der ad-hoc Schätzung der Häufigkeiten ohne die vorherige Aufzählung spezifischer Episoden. Hierbei geht die Zielperson so vor, dass sie zunächst die durchschnittliche Anzahl an Weiterbildungen pro Jahr schätzt, um diesen Wert dann auf einen Schätzwert für einen längeren Zeitraum hochzurechnen. Die dritte Alternative besteht darin, Subklassen und Untergruppen zu bilden. Im Falle der Erinnerung an Weiterbildungen würde der Befragte beispielsweise zunächst die Anzahl der Weiterbildungen im Bereich IT oder Sprachen ermitteln, um am Ende eine Gesamtzahl an Weiterbildungen durch Addieren von Weiterbildungsereignissen aus den verschiedenen Feldern zu berechnen (Blair/Burton 1987).

Die vorgestellten Erinnerungsstrategien stellen nur einen kleinen Ausschnitt aus der Vielzahl der möglichen Vorgehensweisen dar. Daher werden in diesem Papier auch verschiedenste Erinnerungsstrategien aufgezeigt und deren unterschiedliche Verwendung charakterisiert.

3 Fragestellung und Zielsetzung

Die vorliegende Studie stützt sich auf die Erkenntnis, dass eine kontextspezifische Abfrage das Erinnerungsvermögen des Befragten fördern kann (Auriat 1993; Balan/Broning/Jelin1973; Tölke 1979). Unsere Untersuchung trägt dazu bei, weitere Erkenntnisse über bewusste Strategien und unbewusste Rekonstruktionsprozesse beim kontextbezogenen Erinnern zu erlangen. Im Anschluss daran wird untersucht, ob sich durch die Verwendung einer doppelten Strategie, d. h. einer modularisierten Befragung zu Erwerbsepisoden und anschließenden modulinternen kontextbezogenen Fragen zu Weiterbildungsereignissen die Erinnerungsfähigkeit der Befragten verbessern lässt. Folgende Fragen sollen beantwortet werden: Wie gehen die Befragten bei ihren Erinnerungen vor bzw. welche Erinnerungsstrategien gibt es? Welche Schwierigkeiten treten dabei auf? Was erleichtert ihnen neben der kontextspezifischen Abfrage den Abruf von erfragten Sachverhalten? Ist es durch eine kontextgestützte Abfrage möglich, Weiterbildungsaktivitäten auch für einen längeren Zeitraum retrospektiv zu erheben?

Weiterhin ist interessant, wie sich die Anzahl der Ereignisse und die Länge des retrospektiven Intervalls auf den kontextgestützten Erinnerungsprozess auswirken. Dazu werden zwei Hypothesen aufgestellt, die im kognitiven Pretest überprüft werden sollen.

H1: *Je länger das retrospektive Intervall ist, umso schwieriger wird es für eine Person, sich an bestimmte Details zu erinnern.*

Weiter zurückliegende Ereignisse können deutlich schlechter erinnert werden, wonach das Ausmaß des Vergessens also „eine Funktion der verstreichenden Zeit" (Reimer 2001: 19) ist. Die erste Hypothese wurde von Kurz, Prüfer und Rexroth (1999) für andere Ereignistypen bereits als Ergebnis formuliert. Die Länge des retrospektiven Intervalls kann allerdings nicht allein für die Verbesserung oder Verschlechterung von Erinnerungsleistungen verantwortlich sein. Daher lautet die zweite Hypothese, dass sich auch die Anzahl der Ereignisse negativ auf die Erinnerungsfähigkeit der Zielperson auswirkt.

H2: *Je mehr Erwerbsepisoden, Weiterbildungen oder Arbeitgeber eine Person aufzuweisen hat, umso schlechter sind deren Erinnerungen an einzelne Weiterbildungen.*

4 Design des kognitiven Pretests

4.1 Allgemeines Verfahren

Kognitive Pretests sind entwickelt worden, um in der Entwicklungsphase eines Fragebogens einzelne Fragen auf ihre Verständlichkeit und Durchführbarkeit zu prüfen. Zur Identifikation von Problemen wird nach dem Stellen der eigentlichen Frage eine Reihe von Strategien angewandt (Kurz/Prüfer/Rexroth 1999; Prüfer/Rexroth 1996, 2000, 2005). Im vorliegenden kognitiven Pretest werden das *Probing* und die *Think Aloud* Methode verwendet.[2]

Das Probing-Verfahren nutzt gezielte Nachfragen zur Ermittlung von Problemen bei Survey-Fragen. Die in der Psychologie häufig verwendete *Think Aloud* Methode stützt sich auf das gleichzeitige (concurrent) oder rückblickende (retrospective) laute Mitdenken der Befragten und stellt damit eine große Anforderung

2 Weitere Verfahren sind das Response-Latency-Verfahren, das Sorting Verfahren sowie das Paraphrasing, auf die in diesem Papier nicht eingegangen werden soll (vgl. Prüfer/Rexroth 1996, 2000, 2005).

an die Interviewpartner. Beide Verfahren wurden im durchgeführten kognitiven Pretest miteinander kombiniert. Die Nachfragen, die sich an die eigentliche Survey-Frage anschließen, können dabei sowohl auf inhaltliche Probleme (*Comprehension Probing*) abzielen, als auch den Erinnerungsprozess selbst zum Thema haben (*Information Retrieval Probing*). Durch *Comprehension Probing* werden mittels gezielter Nachfragen missverständliche Formulierungen aufgedeckt, Begriffe definiert oder umschrieben, Sinnzusammenhänge der jeweiligen Frage erläutert oder Kontexteffekte aufgespürt. Der Befragte wird aufgefordert, sein Verständnis der Survey-Frage offenzulegen. Durch diese Art von *Probing* stößt man oftmals auf nicht vermutete Missverständnisse oder Verständnisprobleme. Diese Probleme auszuschalten, ist zur Sicherstellung der Bedeutungsäquivalenz unter allen Befragten notwendig, um eine sinnvolle Interpretation der Antworten überhaupt erst zu ermöglichen. Beim *Information Retrieval Probing* steht der Erinnerungsprozess selbst im Fokus, und es werden diesbezüglich gezielte Nachfragen zum Erinnerungsprozess und dem Vorgehen beim Erinnern gestellt. Hier kommt auch häufig die *Think Aloud* Methode zum Einsatz und der Befragte wird gebeten, seine Gedanken beim Erinnern laut auszusprechen. Weitere verwendete Techniken sind das *Confidence Ratin*g, um die Verlässlichkeit der Antworten durch die Befragten selbst einschätzen zu lassen und das *Behavior Coding* (Prüfer/Rexroth 1996, 2005). Das letztgenannte Verfahren klassifiziert non-verbales Verhalten wie z. B. das Zögern, Stirnrunzeln, Lachen, Stöhnen oder die Verweigerung beim Beantworten von Fragen und erlaubt somit eine vereinfachte Bewertung der Qualität von Fragen.

Die bisherigen Studien, die kognitive Pretests verwenden, sind auf die inhaltlichen Probleme bei Survey-Fragen fokussiert (Oksenberg/Cannell/Kalton 1991; Presser/Blair 1994; Foddy 1995, 1998). Die Bitte, eine bloße Begründung für ihre Antwort anzugeben und außerdem die Gedanken zu nennen, die dem Interviewpartner bei der Beantwortung der Fragen durch den Kopf gegangen waren (*Think Aloud*) (Kurz/Prüfer/Rexroth 1999: 105), erwies sich aber für eine Untersuchung von Erinnerungsprozessen als nicht ergiebig genug. Deshalb steht in dieser Studie das *Information Retrieval Probing* im Vordergrund. Den Befragten werden zunächst eher allgemeine, dann fokussierte Nachfragen zur Vorgehensweise und den Problemen beim Erinnern gestellt.

4.2 Umsetzung des kognitiven Pretests

Insgesamt wurden acht kognitive Interviews durchgeführt. Bei der Auswahl der Interviewpartner(innen) wurde darauf geachtet, das empirische Spektrum an möglichen Lebenszusammenhängen, Berufsverläufen und Bildungskarrieren zu erfas-

sen, um eine Vielzahl an denkbaren Erinnerungsstrategien, Problemen und Hilfen beim Erinnern abzubilden. Um eine persönliche Gesprächsatmosphäre zu gewährleisten, wurden die kognitiven Pretests persönlich durchgeführt, obwohl die IAB-ALWA Befragung als computergestützte telefonische Befragung (CATI) angelegt ist. Die Transkripte des kognitiven Pretests wurden anonymisiert. Die Teilnehmer(innen) am kognitiven Pretest haben der anonymisierten Veröffentlichung der Ergebnisse zugestimmt.

Im Rahmen des kognitiven Pretests wurde eine Reihe von Fragen zum Thema Weiterbildung getestet, also dem Besuch von Kursen, Lehrgängen, Seminaren oder Fachvorträgen. Diese bezogen sich zum einen auf nicht-formale Weiterbildungen mittels kürzerer, nicht-zertifizierter Kurse und zum anderen auf informelle Weiterbildungsaktivitäten mit Hilfe elektronischer Medien, wie das berufsbezogene Lernen mit Freunden, Bekannten und Verwandten.[3] Im Rahmen der Erfassung dieser Weiterbildungen wurden deren Dauer und die Unterstützung der Weiterbildung durch den Arbeitgeber abgefragt. Um eine kontextgestützte Erinnerung sicherzustellen, wurde die Zielperson vor den Fragen zu Weiterbildungsaktivitäten zu ihrer zum jeweiligen Zeitpunkt zutreffenden Erwerbsepisode und deren Datierung inklusive Wochenarbeitszeit und zugehöriger Einarbeitungsphase befragt. Durch dieses Vorgehen wird die Zielperson in den jeweiligen Kontext der Erwerbssituation zurückversetzt.

Folgende Survey-Fragen zum Thema Weiterbildung wurden getestet:

- Wurden Ihnen während dieser Zeit Kurse oder Lehrgänge vom Arbeitgeber angeboten?
- Haben Sie während dieser Zeit Kurse oder Lehrgänge besucht?
- Haben Sie an Seminaren oder Fachvorträgen teilgenommen?
- Wie viele Stunden verbrachten Sie während dieser Zeit insgesamt mit Kursen und Lehrgängen?
- Haben Sie sich während dieser Zeit mit Hilfe von elektronischen Medien fortgebildet?
- Haben Sie sich berufsbezogene Dinge von Freunden, Bekannten oder Verwandten beibringen lassen?

Diese Fragen wurden jeweils für mehrere Erwerbsepisoden, sowohl für Beschäftigungsverhältnisse in der Vergangenheit als auch für die aktuelle Erwerbstätigkeit, abgefragt.

3 Für eine detaillierte Erläuterung der hier verwendeten Unterscheidung von Weiterbildungen wird auf Kleinert und Matthes (2009) verwiesen.

Im Rahmen des vorliegenden kognitiven Pretests wurden folgende *Information Retrieval Probing-Fragen* angepasst an den jeweiligen Kontext gestellt:

- Wie sind Sie bei Ihren Erinnerungen vorgegangen?
- Fiel es Ihnen leicht, sich an diesen Sachverhalt zu erinnern? Wenn ja: warum? Wenn nein: Was bereitete Ihnen bei Ihren Erinnerungen Schwierigkeiten?
- Wie kam es, dass Sie die Antwort sofort wussten? Haben Sie sich an etwas Bestimmtes erinnert?
- Haben Sie die Antwort sofort gewusst oder mussten Sie länger überlegen? (Teilweise auch durch *Behavior Coding* ermittelt, so dass lediglich gezielt nachgefragt wurde: Warum konnten Sie diese Frage so schnell beantworten?)

In Abhängigkeit von den Gegebenheiten wurden weitere kognitive Pretestmethoden verwendet, so z. B. das gleichzeitige laute Mitdenken (*Concurrent Think Aloud*) mit den daran anschließenden Nachfragen (Probing) sowie die Erfassung relevanter Verhaltensweisen der Befragten (*Behavior Coding*). Diese Verfahren wurden durch das *Confidence Rating* ergänzt. Die Befragten sollten in einem weiteren Schritt beurteilen, wie sicher sie sich ihrer Aussagen und angegebenen Daten waren. Beispiele für derartige Nachfragen sind:

- Wie sicher sind Sie sich Ihrer Antwort? Ist das Ergebnis ziemlich genau oder eher geschätzt?
- Wie sind Sie bei der Ermittlung des Schätzwertes vorgegangen?

Ebenso wie bei Kurz, Prüfer und Rexroth (1999) erwies es sich als sinnvoll, die *Probing-Fragen* direkt im Anschluss an die Survey-Fragen zu stellen, da die Vorgehensweise beim Erinnern noch unmittelbar präsent ist. Ein Nachteil, der dabei in Kauf genommen werden muss, ist, dass Effekte im Zusammenhang mit der Fragereihenfolge nicht zuverlässig nachgewiesen werden können, da der Fragenkatalog durch die Nachfragen häufig unterbrochen wird.

4.3 Fallauswahl

In der vorliegenden Untersuchung ist nicht die einzelne Person, sondern die einzelne Erwerbsepisode Untersuchungsgegenstand. Dennoch dürfte von Interesse sein, welche Charakteristika die befragten Personen aufweisen. Bei der Fallauswahl wurde auf eine möglichst große Streuung der sozio-demografischen Charakteristika geachtet. In der Stichprobe befinden sich sechs Frauen und zwei Männer, die zum Erhebungszeitpunkt zwischen 23 und 58 Jahre alt waren. Hinsichtlich der Bildungsabschlüsse dominieren Personen mit abgeschlossener Berufsausbildung (N=5). Zwei Personen haben ein Studium erfolgreich abgeschlossen, ein Befragter

ist Beamter. Während eine Befragte lediglich zwei Erwerbsepisoden in ihrem Leben durchlaufen hat, liegt das Maximum der durchlaufenen Erwerbsepisoden bei über zehn. Für die vorliegende Untersuchung ist aber nicht von Interesse, welche Unterschiede die einzelnen Individuen im Hinblick auf ihre Erinnerungsfähigkeit aufweisen, sondern wie sich die Erinnerungen an Ereignisse im Zusammenhang mit verschiedenen Erwerbsepisoden unterscheiden. Ähnlich wie bei der Studie von Blair und Burton (1987), die sich in ihren Ausführungen an den Studien von Sudman und Bradburn (1974) orientieren, wird hier davon ausgegangen, dass die Eigenschaften der Erinnerungsaufgabe für den Erinnerungsprozess interessanter sind als die Charakteristiken der Befragten.

Die Erwerbsepisoden werden hinsichtlich zweier Dimensionen differenziert (vgl. Tabelle 1). Zum einen werden die einzelnen Episoden unter dem Gesichtspunkt der zeitlichen Einordnung betrachtet. Dafür werden drei Referenzzeitpunkte ausgewählt (15 Jahre vor dem Befragungszeitpunkt, fünf Jahre vor dem Befragungszeitpunkt sowie aktuell im Befragungsjahr, also maximal ein Jahr zurück). Diese bilden die Ankerpunkte für die Länge des retrospektiven Intervalls. Die Auswahl dieser Referenzzeitpunkte folgte der Überlegung, im Vergleich zum gesamten Erwerbsverlauf, lange, mittlere und kurze retrospektive Intervalle adäquat abzubilden. Zum anderen werden die Episoden entweder der Gruppe der geringen (bis zu 3) oder der hohen Episodenanzahl (mehr als 3) zugeordnet, je nachdem, wie viele Erwerbsphasen die entsprechende Person insgesamt aufzuweisen hat. Dabei werden Erwerbsphasen, die bei mehreren der betrachteten Erinnerungszeiträume eingeordnet werden können, dem am längsten zurückliegenden zugeordnet. Eine aktuell andauernde, bereits vor 15 Jahren bestehende Erwerbsepisode wird als „vor 15 Jahren" gelistet. Jede Untersuchungsperson kann also maximal drei Erwerbsepisoden beitragen. Insgesamt ergaben sich dadurch 18 Erwerbsepisoden, für die die acht Befragten Weiterbildungsaktivitäten berichten sollten.

Tabelle 1 Einordnung der Erwerbsepisoden hinsichtlich der beiden Dimensionen Anzahl der Episoden insgesamt und Erinnerungszeitraum

		Erinnerungszeitraum		
		III aktuell	II vor 5 Jahren	I vor 15 Jahren
Anzahl der Erwerbsepisoden	1-3	3	4	1
	über 3	2	4	4

4.4 Analysemethode

Nach der Durchführung der persönlichen Interviews wurden die Tonband-Mitschnitte transkribiert. Aus den Forschungsfragen ergaben sich drei grobe Kategorien, nach denen die Aussagen der Befragten geordnet wurden: Erinnerungsstrategien, Gründe für gute Erinnerbarkeit und Probleme beim Erinnern. Die transkribierten Mitschnitte wurden im Hinblick auf diese Kategorisierung analysiert und wichtige Aussagen den entsprechenden Gruppen zugeordnet. Anschließend wurden die Aussagen innerhalb der Gruppen zu weiteren Subkategorien zusammengefasst und inhaltlich benannt. Für die Analysen wurde ein Verfahren gewählt, das sich an der Inhaltsanalyse nach Mayring (2007) orientiert. Dem hier angewandten Verfahren liegt eine induktive Kategorienentwicklung zugrunde. Im Zuge der Zuordnung entsprechender Aussagen bestätigen sich die bisherigen Erkenntnisse (Reimer 2001; Wagenaar 1986), dass große Unterschiede zwischen der Erinnerung an Ereignisse bzw. deren Datierung bestehen, die im Folgenden dargestellt und diskutiert werden.

5 Ergebnisse des kognitiven Pretests

5.1 Erinnerungsstrategien

Insgesamt können zehn Strategien identifiziert werden, mit deren Hilfe die Befragten ihre Erinnerungen rekonstruierten. Diese können in Strategien zur Erinnerung an Ereignisse und Vorgehensweisen beim Abrufen von Zahlen und Fakten sowie in Strategien zur Erinnerung von Datierungen unterschieden werden. Die Strategien sind nach der Häufigkeit ihrer Verwendung geordnet und mit jeweils einem Beispiel versehen. In den Tabellen 1 und 2 im Anhang sind die kompletten Transkripte eines oder mehrerer Beispiele für die jeweiligen Erinnerungsstrategien aufgeführt.
Erinnerungsstrategien beim Erinnern an das Ereignis:

- bewusstes Sich-Zurückversetzen in den situativen (beruflichen) Kontext, z. B. über den Arbeitsalltag („Das weiß ich noch, weil ich den ersten Tag da gearbeitet habe.“)
- Überprüfen spontan erinnerter Ereignisse auf ihre Passung hinsichtlich Referenzperiode und –menge („Ich war öfters mal weg in F. (...) es waren zwei reine Schulungen.“)
- personengebundene Erinnerungen („Da war ich mit dem H. und der K.“)
- chronologische Suche nach Erinnerungen innerhalb der Referenzperiode (meist vom Vergangenen zum Aktuellen) („Von Juni 72 bis Dezember 73. (...) Und danach vom April 75 bis November 87 als Fahrdienstleiter.“)

- Vorstellen möglicher Ereignisse aus der Referenzmenge mit anschließender Überprüfung, ob solche in der Referenzperiode stattgefunden haben („Ach da [während der Erwerbsepisode] habe ich ganz unterschiedliche Sachen gemacht. Da habe ich mal Telefontraining gemacht. Das waren so Kassetten zum Abhören. Dann habe ich mal Italienisch gelernt auch mit Kassetten.")
- Parallelisierung mit dem Privatleben („Ich habe angefangen [mit der Weiterbildung zur F.], da hatte meine Mutter ihren zweiten Schlaganfall und war ein Pflegefall.")

Erinnerungsstrategien beim Abrufen von Zahlen, Fakten und Datierungen:

- Nutzung des eigenen Grundwissens („Während dieser Zeit waren 38,5 Wochenarbeitsstunden üblich.")
- sukzessives Auflisten und Aufaddieren relevanter Daten (chronologisch oder vom wichtigen zum weniger wichtigen Ereignis) („Die zwei Tage waren vielleicht zwei mal acht Stunden oder zwei mal sieben Stunden, dann sind es 14.")
- „grobes" Schätzen („Ich würde jetzt mal sagen: 120 Stunden oder vielleicht so etwas. Das ist geraten.")
- Visualisierung in Form einer imaginären Liste/eines Lebenslaufs („Ich habe so die Liste im Kopf. Das ist eine DIN A4 Seite voll.")

Die Verwendung der Strategien unterscheidet sich insbesondere nach der Länge des retrospektiven Intervalls des erinnerten Ereignisses: Etwa jeweils zwei Fünftel der genannten Strategien bezogen sich auf Erinnerungen an 15 oder fünf Jahre zurückliegende Erwerbsepisoden und lediglich ein Fünftel auf die aktuelle Erwerbsphase. Diese Verteilung zeigt, dass die Anwendung von Erinnerungsstrategien für die stark präsente, aktuelle Beschäftigungsphase weniger bedeutsam ist. Vielmehr scheinen die Befragten die jüngsten Ereignisse ohne Systematik oder spezifische Vorgehensweise zu erinnern oder sich der angewandten Strategien nicht bewusst zu sein. Zudem kann mittels des kognitiven Pretests festgestellt werden, dass die Zielpersonen oft erst nach wiederholtem Nachfragen wiedergeben konnten, wie sie beim Abrufen ihrer Erinnerung vorgegangen sind.

5.2 Einflussfaktoren auf die Erinnerungsfähigkeit

In diesem Abschnitt werden zentrale Gründe erläutert, wieso Erinnern gelingen bzw. problematisch sein kann. Diese sind erneut nach Häufigkeit des Auftretens geordnet und mit einem Beispiel versehen. Ausführliche Beispiele finden sich in Tabelle 3 und 4 im Anhang.

Gründe für gelingendes Erinnern:

- Länge bzw. Dauer, Bedeutung oder Zeitintension zum Ereignis (Je länger, bedeutender oder zeitintensiver in Bezug zur verfügbaren Zeit ein Ereignis ist, desto besser wird es erinnert.) („Das [der Wechsel in einen anderen Betrieb] war ein ziemlicher Einschnitt in meinem Berufsleben.")
- Aktualität der verschiedenen Ereignisse („Weil die Stelle erst kurz ist, ganz neu ist, seit 3 Monaten.")
- Nutzung von Grundwissen („Das ist üblich, dass die um 8 beginnen und bis 16 Uhr gehen. Standard.")
- systematische Anordnung der verschiedenen Ereignisse (z. B. im Lebenslauf) („Ich habe meinen Werdegang schon so oft dargelegt bei Bewerbungen …")
- Erhalt von Zertifikaten oder Dokumenten („Weil man auch eine Urkunde bekommen hat.")
- Unkenntnis über die Referenzmenge (eine Person gibt an, nie an einer bestimmten Fortbildungsaktivität teilgenommen zu haben, weil ihm diese Art von Veranstaltungen unbekannt ist) („Weil ich gar nicht weiß, was das ist. So Qualitätsmanagement schon, aber Zirkel?")
- geringe bzw. große Anzahl an Ereignissen gleichen Typs inklusive regelmäßiger Wiederholung („Da habe ich mich leicht erinnern können, weil es eigentlich nur diese PC-Kurse waren.")
- kein Abweichen vom Erwarteten („So etwas merkt man sich, ob man das vom Arbeitgeber freigestellt kriegt oder nicht. Das merkt sich jeder Arbeitnehmer.")

Jüngere Personen erinnerten sich meist nicht – sie wussten. Nur selten verwendeten sie Erinnerungsstrategien und fühlten sich stattdessen überfordert, einen Grund für ihre gute Erinnerungsfähigkeit zu nennen. Die Tatsache der schnell präsenten Daten begründet sich nur indirekt mit dem Alter des Befragten und ist stattdessen hauptsächlich auf die Aktualität, also die geringe Zeitspanne zwischen Befragung und Ereignis zurückzuführen. Auffällig war außerdem, dass sich die Zielperson leichter an Ereignisse oder Zahlen erinnern konnte, wenn sie eine geringe Anzahl solcher Ereignisse gleichen Typs aufzuweisen hatte. Die Unkenntnis über die Referenzmenge erwies sich ebenfalls als positiver Einflussfaktor. Das ist dadurch zu erklären, dass der Zielperson unklar ist, um welche Ereignisse es überhaupt geht und daher eher willkürlich über ein Ereignis berichtet.
Gründe für Probleme beim Erinnern:

- unklare Abgrenzung der Referenzmenge („Jetzt ist die Frage: was ist ein Fachvortrag? Fachvortrag okay. Aber was ist ein Seminar im Gegensatz zum Lehrgang/Kurs?")

- unwichtige, alltägliche oder beiläufige Situationen („Also das ist ganz schwierig jetzt ohne, wie soll ich sagen, ohne Stütze zu beantworten (...) Teilweise kurze, teilweise lange.")
- zusätzliche Anforderung des Rechnens („Schwierig, ja. Das müsste man ausrechnen, was halt vier Tage in 15 Jahren sind prozentual.")
- fehlende systematische Auflistung („Wenn ich jetzt meinen Lebenslauf vor mir hätte, dann könnte ich es dir genau sagen.")
- Abstimmungsschwierigkeiten beim Vergleich des Ereignisses mit der Referenzperiode („Nein, nicht schwer, aber ich habe überlegt, ob das in meiner Lehre war, als ich da auf Fortbildung war oder ob das danach war.")
- hohe Zahl von Ereignissen oder langes retrospektives Intervall („Also das ist ganz schwierig (...) weil das relativ viele waren.")

Das erstgenannte Problem bezieht sich darauf, dass die Grundvoraussetzung des Erinnerns, die Kenntnis über die relevante Referenzmenge, nicht gegeben war. Dieses Phänomen zeigt nicht direkt eine Schwierigkeit im Erinnerungsprozess auf, sondern macht deutlich, dass oftmals bereits vorher Defizite auftauchen, die auch in der Frageformulierung begründet sein können. Dem Befragten war nicht zweifellos klar, an was er sich erinnern sollte. Der kognitive Pretest zeigt, dass derartige Missverständnisse und Fehlinterpretationen häufiger vorkamen. Daher ist anzuraten, schwierige Begriffe und Formulierungen vorab mit Hilfe inhaltlich fokussierter Pretest-Methoden zu testen. Nur so kann sichergestellt werden, dass eine fehlende Antwort tatsächlich auf Gedächtnislücken und nicht auf ein Nicht-Verstehen der Frage zurückzuführen ist. Eine weitere Schwierigkeit, die auch wiederum kein direktes Erinnerungsproblem darstellt, war die zusätzliche Anforderung des Rechnens. Oftmals schienen Zielpersonen durch Fragen, die eine Berechnung notwendig machten, überfordert. Blieben sie dann eine Antwort schuldig, so war nicht klar, ob diese an einer lückenhaften Erinnerung oder an der mangelnden Fähigkeit des Rechnens gescheitert war. Seltener wurden die hohe Anzahl an relevanten Ereignissen oder die lange Zeitspanne zwischen Befragungsdatum und zu erinnerndem Ereignis – also die Länge des retrospektiven Intervalls – explizit als Erinnerungsproblem genannt.

Die dieser Arbeit zugrunde liegende Hypothese, dass ein längeres retrospektives Intervall und eine erhöhte Anzahl an relevanten Ereignissen die Erinnerungsfähigkeit negativ beeinflussen, wurde von den Befragten nicht durch explizite Nennung bestätigt. Daher werden im Folgenden die anfänglich aufgestellten Hypothesen auf Grundlage der bisherigen Ergebnisse diskutiert. Unberücksichtigt bleiben hier erneut Eigenschaften der Befragten und deren Einfluss auf die Erinnerbarkeit von Weiterbildungsaktivitäten.

6 Zusammenfassung der Ergebnisse

6.1 Ergebnisse zum autobiografischen Gedächtnis

Die Hypothesen bezüglich des autobiografischen Gedächtnisses beziehen sich auf zwei wesentliche Merkmale der Erinnerungsinhalte: auf das retrospektive Intervall und die Anzahl der relevanten Erwerbsepisoden, die den Kontext für die Erinnerung darstellen. Erstens wurde angenommen, dass Weiterbildungsaktivitäten umso schwieriger erinnert werden können, je länger das retrospektive Intervall ist, also je länger das erinnerte Ereignis zurückliegt. Zweitens wurde die Hypothese aufgestellt, dass sich auch eine hohe Anzahl an relevanten Erwerbsepisoden negativ auf die Erinnerungsfähigkeit der Befragten auswirkt.

Zusammenfassend lässt sich sagen, dass eine kontextspezifische Abfrage vor allem für den zweiten Referenzzeitraum mit einem Erinnerungszeitraum von fünf Jahren sinnvoll erscheint. Lagen die Ereignisse weiter zurück, so schien ein bloßes „in den Kontext Zurückversetzen" nicht mehr auszureichen. In diesem Fall wurden weitere Strategien genutzt. Dies deutet darauf hin, dass die Methode der kontextspezifischen Abfrage in der Konzeption der Hauptuntersuchung vor allem für Episoden im mittleren retrospektiven Intervall zielführend ist. Immer wieder gingen hier die Befragten selbst auf die jeweiligen Kontexte der abgefragten Erwerbsphase ein und bestätigten damit die zugrunde liegende Vermutung, dass eine kontextspezifische Abfrage den Erinnerungsprozess für dieses retrospektive Intervall vereinfacht. Auch die Vorgehensweise des personengebundenen Erinnerns untermauert diese These. Hier versetzte sich der Befragte zurück in die jeweilige Situation, die er mit der entsprechenden Person erlebt hatte. Diese Strategie wurde ebenfalls häufiger für die zweite Referenzperiode verwendet.

Lagen die erfragten Inhalte dagegen weiter zurück, so beschränkten sich die Interviewpartner nicht nur auf die Strategie des kontextgestützten Erinnerns. Stattdessen verwendeten die Befragten für Ereignisse des Referenzzeitraums 15 Jahre vor dem Befragungszeitpunkt zusätzlich Strategien, denen ein bestimmter Suchalgorithmus zugrunde lag. So gingen Zielpersonen bei den weiter zurückliegenden Erwerbsepisoden meist so vor, dass sie die Zahl der Ereignisse entweder sukzessive aufaddierten oder chronologisch innerhalb der Referenzperiode suchten. Außerdem stellten sich die Zielpersonen mögliche Weiterbildungsereignisse vor, die sie im Anschluss daran mit ihrer eigenen Biografie und insbesondere mit Erwerbsphasen während der Referenzperiode verglichen.

Für die Referenzperiode mit dem kürzesten retrospektiven Intervall zeigte sich, dass die Befragten deutlich seltener oder zumindest weniger bewusst bestimmte

Erinnerungsstrategien verwendeten. Damit ist auch die durchschnittliche Anzahl an genutzten Strategien pro Episode etwas geringer als bei Episoden der mittleren Bezugszeit und deutlich niedriger als bei Episoden, die bereits 15 Jahre zurückliegen. Nutzten die Befragten für die aktuelle Referenzperiode dennoch eine bestimmte Strategie, so wählten sie meist eine Vorgehensweise, die sich aus der Aktualität der Situation ergab. Häufig erinnerten sie sich aufgrund von Schlüsselreizen in der Fragestellung spontan an ein bestimmtes Ereignis und überprüften dieses im Anschluss daran hinsichtlich seiner Passung in die Referenzperiode. Dieses Ergebnis deutet zumindest vorläufig auf die Gültigkeit der aufgestellten Hypothesen bezüglich der Länge des retrospektiven Intervalls hin.

Ähnliche Ergebnisse zeigten sich auch für die Anzahl der Erwerbsepisoden. Eine kontextspezifische Abfrage scheint vor allem für Personen sinnvoll zu sein, die nur wenige Erwerbsepisoden durchlebt haben. Hat ein Befragter mehrere Beschäftigungsverhältnisse zur Auswahl und demnach auch mehrere Ereignisse zu erinnern, so griff er auf andere Strategien zurück. Während Befragte mit nur wenigen Beschäftigungsverhältnissen sich vor allem personengestützt, also an den situativen Kontext in Zusammenhang mit einer relevanten Person erinnerten, gingen die Befragten mit mehr als drei Erwerbsepisoden viel schematischer vor. Sie suchten chronologisch innerhalb der Referenzperiode nach Ereignissen, stellten sich mögliche Ereignisse vor, um sie dann mit ihrer Biografie zu vergleichen, riefen die abgefragten Inhalte durch Parallelisieren mit dem Privatleben ab oder verwendeten Grundwissen. Diese anderen Strategien können jedoch die Vorgehensweise, bei der sich die Person in den situativen Kontext zurückversetzt, nicht ersetzen. Darauf weist die Anzahl der genannten Strategien pro Episode hin. Hatte eine Person mehr als drei Beschäftigungsverhältnisse, so nennt sie im Durchschnitt weniger Strategien, die sie zur Erinnerung nutzt. Umgekehrt werden insgesamt mehr Strategien pro Episode genannt, wenn die befragte Person nur wenige Erwerbsepisoden in ihrem Leben durchlaufen hat. Zusammenfassend kann vermutet werden, dass Personen mit mehr als drei Erwerbsepisoden, die ihre Erinnerung nicht durch den situativen Kontext gestützt abrufen (können), über nicht genügend alternative Erinnerungsstrategien verfügen. Das Zurückversetzen in den situativen Kontext ist demnach eine wichtige Vorgehensweise im Prozess des Erinnerns an Weiterbildungsaktivitäten. Bezieht man die beiden Aspekte *Länge des retrospektiven Intervalls* und *Anzahl der Erwerbsepisoden* mit ein, so zeigt sich für die Erwerbsepisoden, die 15 Jahre vor dem Befragungszeitpunkt stattfanden, dass die kontextgestützte Erinnerungsstrategie von den Befragten deutlich häufiger angewandt wurde, wenn maximal drei Erwerbsepisoden vorlagen.

Betrachtet man nicht nur die Strategien, welche die Befragten während ihres Erinnerungsprozesses verwendeten, sondern untersucht außerdem die explizit genannten Gründe für oder gegen eine gute Erinnerungsfähigkeit, so zeigte sich, dass viele Personen für den am längsten zurückliegenden Zeitraum die geringe Anzahl an relevanten Weiterbildungsereignissen als Grund für die gute Erinnerbarkeit der Inhalte nannten. Dies lässt darauf schließen, dass die Erinnerungsanforderung durch das längere retrospektive Intervall zusätzlich erhöht wird, wenn außerdem noch eine große Anzahl an Ereignissen vorliegt.

Hinsichtlich der Anzahl der Erwerbsepisoden lassen sich deutlichere Ergebnisse formulieren. Während für Personen mit weniger als drei Beschäftigungsverhältnissen vor allem die Intention der Ereignisse eine wichtige Rolle spielt, nannten Befragte mit mehr als drei Jobepisoden ganz explizit die Aktualität und die geringe Anzahl an abzurufenden Inhalten als Gründe für die gute Erinnerbarkeit. Des Weiteren gaben diese an, dass ihre schlechte Erinnerungsleistung durch eine fehlende Systematisierung der zahlreichen Ereignisse entstanden ist.

Außerdem sind sich die Befragten nur wenig bewusst darüber, warum sie sich an bestimmte Inhalte besser oder schlechter, leichter oder schwieriger, genauer oder ungenauer erinnern. Keine der Zielpersonen nannte auf die Nachfrage nach den Gründen für eine rasche oder genaue Erinnerung das Zurückversetzen in den situativen Kontext. Obwohl viele Befragte bewusst oder unbewusst diese Strategie auswählten und ihre Erinnerung scheinbar auch durch den kontextgestützten Aufbau der Befragung gefördert wurde, waren sie sich der erinnerungsfördernden Wirkung der kontextgestützten Abfrage offenbar nicht bewusst.

6.2 Ergebnisse hinsichtlich der Methode des kognitiven Pretests

Um Erinnerungsprozesse und Erinnerungsstrategien nachzeichnen zu können, benötigt man detaillierte Informationen über den Vorgang des Erinnerns, der nur durch Selbstreflexion der Befragten erlangt werden kann. Schon bei der Auswahl der Untersuchungspersonen waren Entscheidungen notwendig, die auch die Ergebnisse des kognitiven Pretests nachhaltig beeinflussten: die ersten Interviews zeigten bereits, dass die Fähigkeit, reflexiv zu handeln und die eigenen Gedanken zu kommunizieren, stark von individuellen Charakteristika wie dem Bildungsniveau oder dem Beruf etc. abhängt. Befragt man in kognitiven Pretests vornehmlich besser gebildete Personen, so erhält man zwar mehr Einblicke in die einzelnen Erinnerungsprozesse, wird sich jedoch nicht klar darüber, ob die (vielleicht gute) Erinnerungsfähigkeit nicht eher mit der relativ hohen Bildung in Verbindung steht. Auch im durchgeführten Pretest wird nicht deutlich, ob sich niedriger gebildete Personen

in gleicher Weise erinnern und sich dessen nur nicht bewusst werden (können) oder ob sie sich tatsächlich schlechter erinnern. Anders als in der Studie von Kurz, Prüfer und Rexroth (1999), in der niedriger gebildete Bevölkerungsgruppen bewusst überrepräsentiert waren, um derartige Erinnerungsprobleme zu verdeutlichen, wurde in der vorliegenden Untersuchung darauf geachtet, die verschiedenen Bildungsgruppen in ähnlichem Umfang zu repräsentieren. Jüngste Forschungsergebnisse weisen aber darauf hin, dass es keinen oder nur einen geringen Bildungsbias bei retrospektiven Erinnerungen gibt (Drasch/Matthes 2009). Die sehr persönliche Ebene in den Gesprächen spricht ebenfalls dafür, dass nicht das geringere Verbalisierungsvermögen des bewusst gewordenen Erinnerungsprozesses oder eine größere Scheu vor dem Interview verantwortlich für die mangelnden Informationen waren. Vielmehr schienen die niedriger qualifizierten Personen eher kein Bewusstsein darüber aufzuweisen, wie und wodurch sie sich erinnerten.

Im Laufe des kognitiven Pretests zeigte sich zudem, dass einige Survey-Fragen nicht problemlos bei der Untersuchung von kontextgestützten Erinnerungen herangezogen werden konnten. Hier wäre es sinnvoll gewesen, Fragen, bei denen häufiger Verständnisschwierigkeiten auftraten, in einem vorangestellten inhaltlichen Pretest z. B. durch *Paraphrasing* zu prüfen. Daher mussten unbekannte Begriffe während des kognitiven Pretests erklärt werden, um im Anschluss daran die Erinnerungsfähigkeit zu testen. Für die zukünftige Durchführung kognitiver Pretests zur Ermittlung von Erinnerungsprozessen und –problemen ist daher ein zweistufiges Verfahren anzuraten: im ersten Schritt der inhaltliche kognitive Pretest der Survey-Fragen und erst im zweiten Schritt die Analyse der Erinnerungsprozesse.

Ein Problem, das sich außerdem zeigte, war der Kontexteffekt der Fragepositionierung. So überprüften die Zielpersonen auf die Frage nach den angebotenen Kursen von Seiten des Arbeitgebers oft nur ziemlich rasch, ob die tatsächlich wahrgenommenen Kurse vom Arbeitgeber angeboten worden waren. Durch die vorangegangene Frage nach den Lehrgängen, an denen die Zielpersonen teilgenommen hatten, schlossen sie bei der Folgefrage ganz selbstverständlich auf die eingeschränkte Referenzmenge der tatsächlich besuchten Kurse. Auch die schematische Anordnung der Fragen in mehreren Schleifen, wodurch ähnliche Prozesse zu den drei verschiedenen Zeitpunkten und unter neuen Arbeitgebern abgefragt werden konnten, schien zu Problemen zu führen. Es traten Ermüdungseffekte und/oder Motivationsverluste bei den Befragten auf. Durch die sich ständig wiederholenden Fragen kam es nicht selten vor, dass die Interviewpartner bereits vor Beendigung der Frage ohne zu Überlegen die Antwort nannten.

Abschließend wurden die Befragten in jedem Gespräch über die genaue Intention der Interviews aufgeklärt. Erst zu diesem Zeitpunkt wurden ihnen explizit

Beispiele für Erinnerungsstrategien genannt. Auffallend war, dass daraufhin viele Befragte bestätigten, dass sie sich selbst auch über private Ereignisse oder zeitliche Raster erinnert hatten und dafür dann auch Beispiele nannten. Dieses Ergebnis bestätigte die im Vorfeld aufgestellte Vermutung, dass eine vorherige Nennung möglicher Vorgehensweisen nur reaktiv auf die Befragten wirken und deren Erinnerungsprozesse beeinflussen könnte. Daher entschieden wir uns dafür, den Befragten keine möglichen Strategien an die Hand zu geben, sondern stattdessen deren ganz unvoreingenommene Selbsteinschätzung abzurufen. Der Nachteil dieser Vorgehensweise war jedoch, dass sich viele Personen ihres Erinnerungsverhaltens nicht bewusst waren, auch wenn sie vor dem Gespräch aufgefordert wurden, laut zu denken und ihre Vorgehensweise beim Erinnern zu beschreiben. Erst wenn ihnen Beispiele für Erinnerungsprozesse genannt wurden, schien ihnen ihre Aufgabe deutlich zu werden und sie nannten ähnliche Vorgehensweisen.

7 Schlussfolgerungen und weiterer Forschungsbedarf

7.1 Konsequenzen für aktuelle Befragungen

Zusammenfassend zeigen die Ergebnisse, dass für ein retrospektives Intervall von etwa fünf Jahren eine kontextgestützte Abfrage die Erinnerungsfähigkeit unterstützt. Für länger zurückliegende Ereignisse reicht diese Art der Abfrage aber nicht aus. Konsequent zu Ende gedacht heißt das auch, dass selbst bei einer kontextgestützten Abfrage die Analyse nicht-formaler Weiterbildungsbeteiligung auf Ereignisse beschränkt bleiben muss, die maximal fünf Jahre vor dem Interview stattgefunden haben. Will man weiter zurückliegende nicht-formale Weiterbildungsereignisse in die Analysen einbeziehen, ist eine prospektive Erhebung im Rahmen eines Paneldatensatzes unabdingbar. Ist beabsichtigt, dass das retrospektive Intervall mehr als ein Jahr umfasst, sollte eine kontextgestützte Abfrage erfolgen, da Personen mit vielen Kontextereignissen bereits nach einem Jahr Schwierigkeiten haben, sich ohne Weiteres an ihre nicht-formalen Weiterbildungsereignisse zu erinnern.

Auf diesen Ergebnissen aufbauend und unter Berücksichtigung anderer Ergebnisse zum autobiografischen Gedächtnis, z. B. zur Bedeutung der Salienz von Ereignissen, wurden folgende Entscheidungen für die Erfassung von retrospektiven Weiterbildungsereignissen getroffen:

- Sowohl in ALWA als auch in der Erwachsenenetappe des NEPS wurden formale und somit zumeist längere zertifizierte Weiterbildungsaktivitäten im Rahmen des Ausbildungsmoduls als eigenes Untermodul in die Befragung aufgenommen,

um parallele Erinnerungspfade innerhalb eines Moduls zu stimulieren. Die Abfrage dieser Ereignisse erfolgte detailliert, es wurde die Datierung der Episode auf Monatsbasis sowie Angaben zur Abschlussart erfasst.

- Nicht-formale Weiterbildungsereignisse wie Kurse und Lehrgänge wurden in ALWA kumuliert für jede Erwerbs-, Arbeitslosigkeits-, Wehrdienst- und Erziehungsepisode erfasst. Für die Befragung der Erwachsenen im Nationalen Bildungspanel wurde aufgrund der Panelstruktur der Befragung entschieden, die Teilnahme an Lehrgängen oder Kursen jeweils für das vergangene Jahr zu erfassen. Für maximal zwei dieser per Zufall ausgewählten Lehrgänge oder Kurse wurden darüber hinaus detaillierte Angaben zu Inhalten, Kosten und Finanzierung erhoben.
- Nach informellen Weiterbildungsaktivitäten wie dem Besuch von Fachvorträgen wurde in ALWA nur für einen Zeitraum von zwei Jahren vor dem Interviewdatum gefragt. Der Referenzzeitraum für die Erfassung von informellen Weiterbildungsaktivitäten wurde in der Erwachsenenetappe des NEPS auf ein Jahr verkürzt.

7.2 Weiterer Forschungsbedarf

Aus den Ergebnissen des vorliegenden Pretests lassen sich keine allgemeingültigen Erkenntnisse zu den Grenzen der Einsetzbarkeit der kontextgestützten retrospektiven Abfrage für eine Vielzahl von verschiedenen Ereignissen ableiten. Vielmehr bedeuten die vorliegenden Ergebnisse, dass alle neu entwickelten Survey-Fragen mit retrospektivem Inhalt mittels eines zusätzlichen kognitiven Pretests, der auf die Erinnerungsfähigkeit fokussiert ist, überprüft werden sollten. Hier ist eine Reihe von möglichen Frageinhalten denkbar, die mittels dieser Verfahren getestet werden könnten, wie z. B. die retrospektive Erfassung von Einstellungen, familiärer Ereignisse oder Einkommensangaben.[4]

Eine wesentliche Neuerung dieses kognitiven Pretests war, sich auf Erinnerungsprozesse zu konzentrieren und nicht wie bisher üblich primär auf Nachfragen zum Verständnis von Fragen. Nicht im Fokus dieses Pretests standen dabei individuelle Unterschiede zwischen Erinnerungsleistungen von Personen. Diese Vorgehensweise erscheint insbesondere bei Befragungen, die die Erfassung von retrospektiven Lebensverlaufsdaten zum Ziel haben und sich auf bisher nicht erfasste Inhalte beziehen, sinnvoll. Nach eingehender Prüfung und Verbesserung

4 Insbesondere Einkommensangaben gelten als schwer erinnerbar, sind aber seit langem in sozialwissenschaftlicher Forschung von zentralem Interesse. Zur Überprüfung der Validität der Einkommensangaben bietet sich auch ein Vergleich mit prozessproduzierten Daten der Bundesagentur für Arbeit (BA) an.

der Frageformulierung von neuen Survey-Fragen sollte sich insbesondere bei der Erhebung von Retrospektivdaten ein zweiter Schritt, die Überprüfung von Erinnerungsprozessen mittels kognitiver Verfahren anschließen. Nur so kann gewährleistet werden, dass zum einen Survey-Fragen semantisch richtig verstanden werden und zum anderen auch adäquat beantwortet werden können, da sie allgemein als erinnerbar gelten. Sollten z. B. die Ergebnisse eines Pretests wie im vorliegenden Fall zeigen, dass es bestimmte Grenzen bei der Erinnerung von Ereignissen gibt, die mit der Dauer des retrospektiven Intervalls und/oder der Anzahl der zu erinnernden Ereignisse in Zusammenhang stehen, so muss das Befragungsinstrument entsprechend darauf abgestellt werden.

In diesem kognitiven Pretest lag der Schwerpunkt auf der Testung der Erinnerungsfähigkeit, abhängig von Eigenschaften des Ereignisses selbst, aber abgekoppelt von der individuellen Ebene. Faktoren, die dadurch keine Berücksichtigung fanden, waren Persönlichkeitseigenschaften der Befragten, deren Geschlecht, Alter, familiäre Situation, Bildung oder Einstellungen. Würde man hier Unterschiede feststellen, müsste das Befragungsinstrument an die Bedürfnisse und Besonderheiten der jeweiligen Zielgruppe angepasst werden. Eine solche Anpassung der Erhebungsinstrumente auf die Bedürfnisse verschiedener Befragtengruppen beim Erinnern ginge jedoch zu Lasten der Standardisierung und wäre somit nur eingeschränkt umsetzbar.

Abschließend bleibt zu bemerken, dass in der Survey-Methodologie mittlerweile eine Reihe von Strategien existiert, um die Erinnerung an länger zurückliegende und weniger bedeutende Ereignisse zu stimulieren. Diese können auch im Rahmen der Erfassung insbesondere von formalen, aber auch wie der vorliegende Bericht zeigt, für die Erfassung von non-formalen Weiterbildungsaktivitäten erfolgreich genutzt werden. Allerdings sind der Erinnerungsfähigkeit des Menschen Grenzen gesetzt, was dazu führt, dass manche Ereignisse, wie etwa im vorliegenden Fall sehr lange zurückliegende informelle Weiterbildungsereignisse, nicht bzw. nur unzureichend retrospektiv abgefragt werden können. Hier bietet sich als Alternative eine langjährige Panelstudie wie das Nationale Bildungspanel an, das Weiterbildungsereignisse zwischen den Erhebungszeitpunkten erfasst.

Literatur

Allmendinger, J., M. Antoni, B. Christoph, K. Drasch, F. Janik, C. Kleinert, K. Leuze, B. Matthes, R. Pollak, M. Ruland und A. Trahms, 2011 (im Erscheinen): Adult Education and Lifelong Learning. In: H.-P. Blossfeld, H.-G. Roßbach und J. von Maurice (Hg.): Education as a Lifelong Process. The German National Educational Panel Study (NEPS). Sonderheft der Zeitschrift für Erziehungswissenschaft.

Antoni, M., K. Drasch, C. Kleinert, B. Matthes, M. Ruland und A. Trahms, 2010: Arbeiten und Lernen im Wandel. Teil I: Überblick über die Studie. FDZ Methodenreport 05/2010. Nürnberg: IAB. http://doku.iab.de/fdz/reporte/2010/MR_05-10.pdf (03.02.2011).

Auriat, N., 1993: "My Wife Knows Best". A Comparison of Event Dating Accuracy Between the Wife, the Husband, the Couple and the Belgium Population Register. Public Opinion Quarterly 57: 165-190.

Balan, J., H. Browning und E. Jelin, 1973: Men in a Developing Society. Geographic and Social Mobility in Monterrey, Mexico. Austin: University of Texas Press.

Barsalou, L., 1988: The Content and Organization of Autobiographical Memories. S. 193-243 in: U. Neisser und E. Winograd (Hg.): Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory. Cambridge: Cambridge University Press.

Belli, R., 1998: The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. Memory 6: 383-406.

Blair, E. und S. Burton, 1987: Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions. The Journal of Consumer Research 14: 280-288.

Conway, M., 1996: Autobiographical Knowledge and Autobiographical Memories. S. 67-93 in: D. Rubin (Hg.): Remembering our Past: Studies in Autobiographical Memory. Cambridge/England: Cambridge University Press.

Conway, M. und C. Pleydell-Pearce, 2000: The Construction of Autobiographical Memories in the Self Memory System. Psychological Review 107: 261-288.

Dex, S. und A. McCulloch, 1997: The Reliability of Retrospective Unemployment History Data. Working Papers of the ESRC Research Centre on Micro-social Change. Paper 97-117. Colchester: University of Essex.

Drasch, K. und B. Matthes, 2009: Improving Retrospective Life Course Data by Combining Modularized Self-reports and Event History Calendars. Experiences from a Large Scale Survey. IAB Discussion Paper 21/2009. Nürnberg: IAB. http://doku.iab.de/discussionpapers/2009/dp2109.pdf (03.02.2011).

Dykema, J. und N. Schaeffer, 2000: Events, Instruments and Reporting Errors. American Sociological Review 65: 619-629.

Foddy, W., 1995: Probing: A Dangerous Practice in Social Surveys? Quality & Quantity. International Journal of Methodology 29: 73-86.

Foddy, W., 1998: An Empirical Evaluation of In-Depth-Probes Used to Pretest Survey Questions. Sociological Methods & Research 27: 103-133.

Klein, T. und D. Fischer-Kerli, 2000: Die Zuverlässigkeit retrospektiv erhobener Lebensverlaufsdaten. Analysen zur Partnerschaftsbiographie des Familiensurveys. Zeitschrift für Soziologie 29: 294-312.

Kleinert, C., B. Matthes und M. Jacob, 2008: Die Befragung „Arbeiten und Lernen im Wandel". Theoretischer Hintergrund und Konzeption. IAB Forschungsbericht, 5/2008. Nürnberg: IAB. http://doku.iab.de/forschungsbericht/2008/fb0508.pdf (03.02.2011).

Kleinert, C. und B. Matthes, 2009: Data in the Field of Adult Education and Lifelong Learning. Working Paper Series of the German Council for Social and Economic Data 91. Berlin: RatSWD. http://www.ratswd.de/download/RatSWD_WP_2009 /RatSWD_WP_91.pdf (03.02.2011).

Kurz, K., P. Prüfer und M. Rexroth, 1999: Zur Validität von Fragen in standardisierten Erhebungen. Ergebnisse des Einsatzes eines kognitiven Pretestinterviews. ZUMA-Nachrichten 44: 83-107. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_44.pdf (09.02.2011).

Matthes, B., M. Reimer und R. Künster, 2007: Techniken und Werkzeuge zur Unterstützung der Erinnerungsarbeit bei der computergestützten Erhebung retrospektiver Längsschnittdaten. Methoden – Daten – Analysen 1 (1): 69-92. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.1_Heft_1/MDA1_Matthes_Reimer_Kuenster.pdf (09.02.2011).

Mayring, P., 2007: Qualitative Inhaltsanalyse. Grundlagen und Techniken. Weinheim: Deutscher Studienverlag.
Oksenberg, L., C. Cannell und G. Kalton, 1991: New Strategies for Pretesting Survey Questions. Journal of Official Statistics 7: 349-365.
Peters, E., 1988: Retrospective versus Panel Data in Analyzing Lifecycle Events. The Journal of Human Resources 23: 488-513.
Powers, E., W. Goudy und P. Keith, 1978: Congruence between Panel and Recall Data in Longitudinal Research. Public Opinion Quarterly 42: 380-401.
Presser, S. und J. Blair, 1994: Survey Pretesting: Do Different Methods Produce Different Results? Sociological Methodology 24: 73-104.
Prüfer, P. und M. Rexroth, 1996: Verfahren zur Evaluation von Survey-Fragen: Ein Überblick. ZUMA-Nachrichten 39: 95-116. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_39.pdf (09.02.2011).
Prüfer, P. und M. Rexroth, 2000: Zwei-Phasen-Pretesting. ZUMA-Arbeitsbericht 2000/08. Mannheim: ZUMA. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/zuma_arbeitsberichte/00_08.pdf (03.02.2011).
Prüfer, P. und M. Rexroth, 2005. Kognitive Interviews. ZUMA How-to-Reihe 15. Mannheim: ZUMA. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/howto/How_to15PP_MR.pdf (03.02.2011).
Reimer, M., 2001: Die Zuverlässigkeit des autobiographischen Gedächtnisses und die Validität retrospektiv erhobener Lebensverlaufsdaten. Kognitive und erhebungspragmatische Aspekte. Materialien aus der Bildungsforschung 71. Max-Planck-Institut für Bildungsforschung.
Reimer, M. und B. Matthes, 2007: Collecting Event Histories with True Tales. Techniques to Improve Autobiographical Recall Problems in Standardized Interviews. Quality & Quantity. International Journal of Methodology 41: 711-735.
Rogoff Ramsøy, N., 1973: The Norwegian Occupational Life History Study. INAS, Memorandum from the Occupational History Study. Oslo.
Solga, H., 2001: Longitudinal Surveys and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison. Quality & Quantity. International Journal of Methodology 35: 291-309.
Sudman, S. und N. Bradburn, 1974: Response Effects in Surveys. A Review and Synthesis. Chicago: Aldine Publishing Co.
Sudman, S., N. Bradburn und N. Schwarz, 1996: Thinking about Anwers: The Application of Cognitive Processes to Survey Methodology. San Francisco: Jossey-Bass.
Timmermann, D., G. Färber, U. Backes-Gellner, G. Bosch und B. Bernhard, 2004: Finanzierung Lebenslangen Lernens: Der Weg in die Zukunft. Abschlussbericht der Expertenkommission Finanzierung Lebenslangen Lernens. Bielefeld.
Tölke, A., 1979: Literaturbericht zu methodischen Problemen und Varianten von Retrospektivbefragungen bei der Erfassung von Lebensgeschichten. Arbeitspapier Nr. 10, Sonderforschungsbereich 3, Mikroanalytische Grundlagen der Gesellschaftspolitik, J.W. Goethe-Universität Frankfurt und Universität Mannheim.
Van der Vaart, W., 2004: The Time-line as a Device to Enhance Recall in Standardized Research Interviews: A Split Ballot Study. Journal of Official Statistics 20: 301-317.
Wagenaar, W., 1986: My Memory: A Study of Autobiographical Memory over Six Years. Cognitive Psychology 18: 225-252.
Wohn, K., 2007: Effizienz von Weiterbildungsmessung. Research Notes Series of the German Council for Social and Economic Data 15. Berlin: RatSWD. http://www.ratswd.de/download/RatSWD_RN_2007/RatSWD_RN_15.pdf (03.02.2011).

Anschrift der Autorinnen

Andrea Dürnberger
Staatsinstitut für Familienforschung an der Universität Bamberg (ifb)
Heinrichsdamm 4
96047 Bamberg
andrea.duernberger@ifb.uni-bamberg.de

Katrin Drasch
Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA)
Regensburger Straße 104
90478 Nürnberg
katrin.drasch@iab.de

Dr. Britta Matthes
Institut für Arbeitsmarkt- und Berufsforschung (IAB) der Bundesagentur für Arbeit (BA)
Regensburger Straße 104
90478 Nürnberg
britta.matthes@iab.de

Anhang[5]

Tabelle 1 Strategien bei der Erinnerung an Ereignisse

Strategie	Transkribierte Beispiele
bewusstes Sich-Zurückversetzen in den situativen Kontext	*Gab es auf dieser Stelle eine Einarbeitungszeit?* – Nee, da gab es keine Einarbeitungszeit, ich bin da ins kalte Wasser geschmissen worden. Ja, das weiß ich noch, weil ich den ersten Tag da gearbeitet habe und mir dann gedacht habe, da geh ich morgen nicht mehr hin. Das war so heftig. Aber ich habe mich dann doch eines Besseren belehren lassen und bin dann doch wieder hin und hab es dann tatsächlich vier Jahre ausgehalten.
Überprüfen spontan erinnerter Ereignisse auf ihre Passung hinsichtlich Referenzperiode und -menge	*Haben Sie während dieser Zeit Kurse oder Lehrgänge besucht?* – Ich war öfters mal weg in F. oder K., aber ich wusste jetzt nicht mehr genau, ob es Schulungen oder Meetings waren. Aber es waren zwei reine Schulungen. *Haben Sie an Seminaren oder Fachvorträgen teilgenommen?* – Lass mich mal überlegen: Seminare (denkt laut, überlegt lange). Wir waren ein paar Mal in F. Da haben wir schon ein paar Mal etwas gehabt. Ich weiß jetzt gar nicht mehr genau. (...) Ich war ein paar Mal in F., aber ob das jetzt so etwas war oder irgendwelche Besprechungen, da bin ich mir nicht mehr sicher.
personengebundene Erinnerungen	*Warum konnten Sie sich so gut an diese Messe erinnern?* – Da war ich mit dem H. und der K.. An das Drumherum. Da waren wir halt in Köln und Köln war eine tolle Stadt. Und überhaupt war auch die Messe super-schön. Es war halt sauinteressant. *Haben Sie sich berufsbezogene Dinge von Freunden, Bekannten oder Verwandten beibringen lassen?* – Eigentlich war immer so das Gegenteil der Fall. Ich bin eigentlich immer ziemlich nieder gemacht worden. Also mein Mann zum Beispiel konnte das überhaupt nicht verstehen. Also wie man sich jetzt nach der Arbeit noch da rein hocken kann. Und wozu das Ganze, das bringt doch nichts. Das war immer sein Standardsatz. Das bringt doch nichts, für was denn?
chronologische Suche nach Erinnerungen innerhalb der Referenzperiode	Ich war immer bei der Eisenbahn und dort halt in verschiedenen Bereichen. – *Was war dort deine erste Tätigkeit nach der Ausbildung?* – Nach der Ausbildung Fahrkartenschalter. – *Können Sie auch sagen von wann bis wann?* – Von Juni 72 bis Dezember 73. Und dann war ich bei der Bundeswehr 15 Monate. Und danach vom April 75 bis November 87 als Fahrdienstleiter.
mögliche Ereignisse der Referenzmenge vorstellen, dann Überprüfung, ob diese im betreffenden Zeitraum stattgefunden haben	*Haben Sie sich während dieser Zeit mit Hilfe von elektronischen Medien fortgebildet?* – Ja. Ach da habe ich ganz unterschiedliche Sachen gemacht. Da habe ich mal Telefontraining gemacht. Das waren so Kassetten zum Abhören. Dann habe ich mal Italienisch gelernt auch mit Kassetten. Was gibt es da noch? (überlegt laut) DVDs? Na ja, ab und an kommt dann so etwas. Irgendwelche Demosachen, wenn du ein neues Programm auf den Rechner kriegst, mit dem du dich dann einarbeiten kannst.
Parallelisierung mit dem Privatleben	Das mit der Stelle von 2003 habe ich aus dem Grund so genau gewusst, weil ich habe die Heizung 2003 umgebaut und da weiß ich, da hab ich im Heizraum unten rumgewerkelt und da hat der K. angerufen. Das war sein erster Kontakt, ob ich denn nicht Interesse hätte, so etwas zu machen. (...) Und das weiß ich, das war im Juni/Juli 2003, als es so warm war und im September habe ich dann angefangen. Sicher, das ganze läuft ja schon konform mit meinem Privatleben. Dieser Englischkorrespondent, das war der Hammer hoch drei. Da war dann auch meine Mutter. Ich habe angefangen [mit der Weiterbildung zur Fremdsprachenkorrespondentin], da hatte meine Mutter ihren zweiten Schlaganfall und war ein Pflegefall. Ich habe sie dann auch noch mit gepflegt. Das heißt, ich war zweimal die Woche in dem Englischkurs, zwei Tage die Woche bei meiner Mutter und der Rest war dann für die Familie noch. Und gearbeitet auch noch.

5 Zur Übersichtlichkeit wurde die „Du"-Form, die bei einigen Interviews auf Wunsch der Zielperson verwendet wurde, in die „Sie"-Form umgewandelt. Nicht relevante Textteile sind nicht dargestellt und mittels (...) als Auslassung gekennzeichnet.

Tabelle 2 Erinnerungsstrategie beim Abrufen von Zahlen und Daten

Strategien	Transkribierte Beispiele
Nutzung des eigenen Grundwissens	Wie war zu dieser Zeit Ihre wöchentliche Arbeitszeit? – Während dieser Zeit waren 38,5 Wochenarbeitsstunden üblich, also gehe ich davon aus, dass auch wir damals in diesem Umfang gearbeitet haben.
	Wenn Sie alle Kurse zusammen nehmen, wie viele Stunden dauerten diese? – (lacht) Die Stunden? Na ja, 20 Wochen, das waren auch meistens immer so Vollzeit. Was hat so ein Lehrgang? Wobei dann wieder die Frage ist, Schulstunden oder Zeitstunden? Normalerweise sind das immer acht Schulstunden. Das sind also dann auch 40 Stunden die Woche, das mal 20, 800.
Auflisten und Aufaddieren relevanter Daten	Wie viele Stunden verbrachten Sie während dieser Zeit insgesamt mit Kursen und Lehrgängen? – Die zwei Tage waren vielleicht zwei mal acht Stunden oder zwei mal sieben Stunden, dann sind es 14 und dann noch mal zwei sind 16 und dann noch mal 4, also 20 Stunden.
	Wie sind Sie denn auf die absolute Wochenzahl Ihrer Kurse gekommen? – Also in erster Linie muss ich sagen, die langen Kurse. Ich hatte so einen PC-Programmierkurs drei Wochen, dann hatte ich mal fünf Wochen am Stück zwei Kurse, zwei Wochen, drei Wochen. Dann sind das ja schon, sagen wir mal, acht Wochen gewesen. Und dann kommen noch viele andere kleine, wo man mal eine Woche da ist oder zwei Wochen hier.
„Grobes" Schätzen	Über wie viele Stunden erstreckten sich alle Kurse und Lehrgänge insgesamt während dieser Zeit? – Also wie gesagt, ich könnte nachschauen. Ich habe das schwarz auf weiß zu Hause. Ich würde jetzt mal sagen: 120 Stunden oder vielleicht so etwas. Das ist geraten. – Wie sicher sind Sie sich dabei? – Fifty-fifty.
Visualisierung in Form einer imaginären Liste/eines Lebenslaufes	Wie viele Stunden verbrachten Sie während dieser Zeit mit Kursen und Lehrgängen? – Während dieser Stelle. Oh. (kurzes Überlegen) 30. Das waren halt von Eintageslehrgängen bis Dreiwochenlehrgängen oder so. – Fiel es Ihnen schwer, diese Frage zu beantworten. – Nee, eigentlich nicht. – Sind es genau 30 oder haben Sie geschätzt? – Das ist geschätzt. – Wie sind Sie bei Ihrer Schätzung vorgegangen? – Ich habe so die Liste im Kopf. Das ist eine DIN A4 Seite voll.

Tabelle 3 Gründe für gelingendes Erinnern

Gründe	Transkribierte Beispiele
Länge, Bedeutung oder zeitlicher Bezug zum Ereignis	Mir fiel es leicht, weil das eine ganz harte Zeit war und so etwas merkt man sich. Ich habe ja gearbeitet, Vollzeit und nach dem Geschäft dann abends noch bis um neun die Abendschule besucht.
	Also sagen wir mal so, die großen Sachen weiß ich, weil sie zum einen sehr teuer, zum anderen sehr aufwendig, sehr langfristig und auch mit einem hohen Lernaufwand verbunden waren. Der Fremdsprachenkorrespondent, der hat mich wirklich an meine Grenzen gebracht, muss ich sagen. (...) Und das vergisst man dann auch nicht, vor allem, wenn du es dann auch gut abgeschlossen hast (schmunzelt).
	Das [der Wechsel in einen anderen Betrieb] war ein ziemlicher Einschnitt in meinem Berufsleben, als ich vom Schichtdienst in den Tagesdienst gekommen bin. Und solche Dinger, die merkt man sich einfach. Ich weiß sogar, dass es der 17. November war, weil das war mehr oder weniger, warte mal, der Buß- und Bettag... (denkt laut) Ich weiß genau, der 17. November war's.
Aktualität der verschiedenen Ereignisse	Haben Sie im Rahmen Ihrer aktuellen Stelle an Kursen oder Lehrgängen teilgenommen? – Nein – Woher wussten Sie, dass Sie bisher noch keine Kurse besucht haben? – Weil die Stelle erst kurz ist, ganz neu ist, seit 3 Monaten. (überlegt kurz) Wie gesagt: der eine Kurs war 10-stündig, nein 10 mal je 2 Stunden und der andere Kurs war einen Samstag, ich glaube von 9 bis zwei oder bis 14 Uhr halt. – Wie konnten Sie sich an die Stundenzahl erinnern? Haben Sie sich an etwas Besonderes erinnert? – Nö, das weiß ich.
Nutzung von Grundwissen	Zweimal zwei Tage von in der Früh um acht bis 16 Uhr. Also ganze Tage. – Ist die Stundenzahl also wieder einfacher? – Genau. – Wie haben Sie sich an die Stundenzahl erinnert? – Das ist üblich, dass die um 8 beginnen und bis 16 Uhr gehen. Standard.
systematische Anordnung der verschiedenen Ereignisse	Ich habe meinen Werdegang schon so oft dargelegt bei Bewerbungen bei B. jetzt speziell. Bis ich jetzt in diesen Innendienst hinein gekommen bin. Da habe ich das schon hunderttausend Mal niederschreiben müssen. Das macht schon auch etwas aus.
Erhalt von Zertifikaten oder Dokumenten	Warum ist Ihnen die Erinnerung an die Anzahl der Kurse so präsent bzw. warum fiel es Ihnen so leicht, die Frage zu beantworten? – Wahrscheinlich weil es so viele waren. Weil man auch eine Urkunde bekommen hat. Ich habe da eine ganze Latte. Bei mir ist es wahrscheinlich auch noch so, man gibt ja, wenn man sich hier im Haus bewirbt, dann macht man ja auch so eine Übersicht von den Lehrgängen. Da gibt es auch so schriftliche Übersichten. Und die hat man dann halt, wenn man sie ein paar Mal gemacht hat, dann merkt man sich die.
Unkenntnis über die Referenzmenge	Haben Sie zu dieser Zeit an einer Beteiligungsgruppe oder einem Werkstattzirkel teilgenommen? – Nein, habe ich nicht. (lacht) – Wie konnten Sie sich daran erinnern? – Weil ich gar nicht weiß, was das ist. So Qualitätsmanagement schon, aber Zirkel?
geringe bzw. große Anzahl an Ereignissen gleichen Typs inklusive regelmäßiger Wiederholung	Da habe ich mich leicht erinnern können, weil es eigentlich nur diese PC-Kurse waren. Ansonsten war da nichts.
keine Abweichung vom Erwarteten	Woher wussten Sie so schnell, dass Ihre Kurse bezahlt wurden? – So etwas merkt man sich, ob man das vom Arbeitgeber freigestellt kriegt oder nicht. Das merkt sich jeder Arbeitnehmer.

Tabelle 4 Gründe für Probleme beim Erinnern

Gründe	Transkribierte Beispiele
unklare Abgrenzung der Referenzmenge	Jetzt ist die Frage: was ist ein Fachvortrag? Fachvortrag okay. Aber was ist ein Seminar im Gegensatz zum Lehrgang/Kurs? Was ist da der Unterschied? Manche sagen, sie haben ein Seminar zu dem und dem Thema besucht und andere würden wahrscheinlich sagen, das ist ein Lehrgang zu dem und dem Thema.
	Hat Sie Ihr Arbeitgeber damals finanziell unterstützt? – Nein. Halt, jetzt passen Sie mal auf: was heißt finanziell unterstützt? Mein Gehalt ist halt weiter gelaufen. Ich bin praktisch da gezahlt worden in der Zeit. – Macht das Schwierigkeiten, diese Frage zu beantworten? – Nein. Ist das jetzt finanzielle Unterstützung, wenn das Gehalt weiter läuft?
unwichtige, alltägliche oder beiläufige Situationen	Warum ist das schwierig? – Also sagen wir mal so, die großen Sachen weiß ich, weil sie zum einen sehr teuer zum anderen sehr aufwendig, sehr langfristig und auch mit einem hohen Lernaufwand verbunden sind. Der Fremdsprachenkorrespondent der hat mich wirklich an meine Grenzen gebracht, muss ich sagen. Das war in Englisch. Dieser Kurs, man kann sagen, 1/3 ist das, wo man drin sitzt im Kurs und 2/3 musst du daheim machen, sonst packst du das nicht. Und da war ich dann halt wirklich ganze Wochenenden dran gesessen und habe gelernt und habe Übersetzungen gemacht. Ich war dann auch auf einer Sprachschule, so zur Vorbereitung auf die mündliche Prüfung. Und das vergisst man dann auch nicht, vor allem wenn es dann auch gut abgeschlossen hat? (schmunzelt)
	Also das ist ganz schwierig jetzt ohne, wie soll ich sagen, ohne Stütze zu beantworten, weil das relativ viele waren. Teilweise kurze, teilweise lange und von daher: ich schreibe mir das zu Hause immer auf. Ich habe den Zettel nicht im Kopf. Ich kann das schlicht und einfach nicht sagen.
zusätzliche Anforderung des Rechnens	Wie viel Prozent Ihrer Arbeits- und Freizeit haben Sie mit Kursen verbracht? – Na ja, von der ganzen Arbeitszeit von 1996 bis 2003 wären es wahrscheinlich 0,1%. Das geht so nicht. Das wäre jetzt etwas anderes, wenn ich jedes Jahr einen Kurs belegen würde, dann könnte ich das natürlich prozentual sagen, aber zwei Kurse.
	Fällt es Ihnen schwer, hier einen Prozentwert anzugeben? – Schwierig, ja. Das müsste man ausrechnen, was halt vier Tage in 15 Jahren sind prozentual.
fehlende systematische Auflistung	An wie vielen Kursen oder Lehrgängen haben Sie während dieser Zeit teilgenommen? – Ich sag mal, so kleinere Sachen, die jetzt vielleicht nur ein Wochenende waren, dann könnten es auch mehr gewesen sein. Schwierig zu beantworten. Wenn ich jetzt meinen Lebenslauf vor mir hätte, dann könnte ich es dir genau sagen. Aber mehr als 20 waren es jetzt nicht. Ich will jetzt nicht so auf den Putz hauen. Aber 15 bis 20 waren es schon.
Abstimmungsschwierigkeiten beim Vergleich des Ereignisses mit der Referenzperiode	Fiel Ihnen die Beantwortung der Frage schwer? – Nein, nicht schwer, aber ich habe überlegt, ob das in meiner Lehre war, als ich da auf Fortbildung war oder ob das danach war. Ich war schon öfters auf Fortbildung. Das war danach erst.
hohe Zahl von Ereignissen oder langes retrospektives Intervall	Also das ist ganz schwierig jetzt ohne, wie soll ich sagen, ohne Stütze zu beantworten, weil das relativ viele waren. Teilweise kurze, teilweise lange und von daher...

Das Stichprobendesign des registergestützten Zensus 2011

The Sample Design for the Register-Assisted Census 2011

Ralf Münnich, Siegfried Gabler, Matthias Ganninger, Jan Pablo Burgard und Jan-Philipp Kolb

Zusammenfassung

Im Rahmen der europaweiten Zensus-Erhebungsrunde im Jahr 2011 wird zum ersten Mal seit 1987 auch im vereinigten Deutschland wieder eine Volkszählung stattfinden, diesmal allerdings nicht in Form einer Vollerhebung, sondern in Form einer kosten- und ressourcenschonenden registergestützten Erhebung. Diese wird flankiert durch eine Haushaltsstichprobe, aus der erstens in den Registern nicht erfasste Informationen gewonnen werden sollen und zweitens eine Abschätzung der Zahl der Karteileichen (KAL) und Fehlbestände (FEB) in den Melderegistern erfolgen soll. Aus den Register- und Stichprobendaten sollen möglichst verlässliche und genaue Schätzungen der Totalwerte vorgenommen werden. Ziel des von DESTATIS eingesetzten Stichprobenforschungsprojektes ist es, Antworten auf die Frage zu geben, welches Stichprobendesign unter den gegebenen Restriktionen empfohlen werden kann. Darüber hinaus sollen Schätzstrategien entwickelt werden, die zur Verwendung im Zensus 2011 vorgeschlagen werden können. Der vorliegende Aufsatz stellt einige wichtige Erkenntnisse aus dem Forschungsprojekt dar, wobei ein Schwerpunkt auf der Darstellung eines optimalen Stichprobendesigns liegt.

Abstract

Within the context of the Europe-wide census elicitation in 2011 there will be the first population census in reunified Germany. In contrast to the last German census in 1987, where all households were interviewed, the new census will be conducted by means of a cost- and resource-effective register-assisted census. In addition to the register information, a household sample will be drawn. On the one hand this sample will provide information that is not included in the register, on the other hand it will allow for the estimation of over- and undercounts in the register. Reliable estimates for total values of interest are to be derived from the register and sample data. The aim of the research project, which was initiated by DESTATIS, is to elaborate an efficient sample design as well as to develop estimation strategies which allow accurate estimates for the census 2011. This article presents some important findings from the research project. However, one focus is on the description of an optimal sample design.

1 Ausgangssituation

Im Rahmen der europaweiten Zensus-Erhebungsrunde im Jahr 2011 wird zum ersten Mal seit 1987 auch im vereinigten Deutschland wieder eine Volkszählung stattfinden. Diese wird jedoch nicht in Form einer Vollerhebung erfolgen, sondern wird in Form einer kosten- und ressourcenschonenden registergestützten Erhebung vollzogen. Zudem wird eine Haushaltsstichprobe gezogen, aus der zum einen eine Abschätzung der Zahl der Karteileichen (KAL) und Fehlbestände (FEB) in den Melderegistern erfolgen soll und zum anderen in den Registern nicht erfasste Informationen gewonnen werden.

Ziel ist es insbesondere, die amtliche Einwohnerzahl zu ermitteln, aber auch eine Schätzung von Totalwerten zusätzlicher in den Registern nicht enthaltener Merkmale zu erhalten. Zur Erarbeitung eines geeigneten Stichprobendesigns hat das Statistische Bundesamt (DESTATIS) ein Forschungsprojekt in Auftrag gegeben, das von einem Konsortium unter der Leitung von Ralf Münnich (Uni Trier) in Zusammenarbeit mit GESIS bearbeitet wird. Das Forschungsprojekt soll vor allem Antworten auf die Frage geben, welches Stichprobendesign unter den gegebenen Restriktionen empfohlen werden kann. Daneben sollen Schätzstrategien entwickelt werden, die zur Verwendung im Zensus 2011 vorgeschlagen werden können.

Um diese Ziele zu erreichen, wurde eine geeignete Simulationsumgebung erarbeitet, mit deren Hilfe der Beantwortung der Forschungsfragen nachgegangen wird. Ebenso wichtig ist es, eine geeignete Darstellungsform für die Visualisierung und Evaluation der hochdimensionalen Ergebnisse zu finden.

Der vorliegende Aufsatz gibt einige wichtige Erkenntnisse aus dem Forschungsprojekt wieder, wobei ein Schwerpunkt auf der Darstellung eines optimalen Stichprobendesigns liegt.

2 Methodische und statistische Grundlagen

Mit Hilfe der Daten aus der 2011 zu realisierenden Stichprobe des Zensus sollen nach Maßgabe von DESTATIS zwei Ziele erfüllt werden

Ziel 1: *Die Ermittlung der amtlichen Einwohnerzahl*

Ziel 2: *Die Schätzung von Kennzahlen bei Zusatzvariablen*

Um diese Ziele zu erreichen, werden geeignete Stichprobendesigns untersucht. Von Seiten des Auftraggebers (DESTATIS) existieren hierzu einige Vorgaben, die in jedem Fall eingehalten werden müssen. So werden im Zensus 2011 komplette Anschriften

aus dem Anschriften- und Gebäuderegister (AGR) gezogen, dessen Aufbau in Kleber et al. (2009) beschrieben ist. In einer ausgewählten Anschrift wird dann die Anzahl an Personen ermittelt, die eine bestimmte interessierende Eigenschaft aufweisen. Bei dieser Eigenschaft handelt es sich im Fall von Ziel 1 um die Existenz einer Person in der Anschrift *(Ziel 1 Variable)* und bei Ziel 2 um die konkrete Ausprägung einer interessierenden Variablen *(Ziel 2 Variable)*. Im AGR ist die Anschriftengröße, also die Anzahl der an der Anschrift registrierten Personen, enthalten. Es liegt nahe, diese Information durch eine Schichtung schon in das Auswahlverfahren einfließen zu lassen, da für etliche interessierende Merkmale, insbesondere Anzahl der Karteileichen und Fehlbestände, ein Zusammenhang mit der Anschriftengröße vermutet wird. Hierzu wird die Variable Anschriftengröße in Klassen eingeteilt und die so entstandene Anschriftengrößenklasse (ADK) als Schichtungsvariable verwendet. Allem Folgenden liegen die in Tabelle 1 dargestellten drei Schichtungsvarianten zugrunde, wobei SMP für *Sampling Point* steht (zur Definition der Sampling Points siehe Abschnitt 2.2).

Tabelle 1 Schichtungsvarianten

Bezeichner	Beschreibung
ADK1	6 Schichten mit Schichtgrenzen: 1, 2, 3, 4-6, 7-10, 11+ registrierte Personen in der Anschrift
ADK2	4 mit registrierten Personen gleich stark besetzte Schichten pro SMP
ADK3	8 mit registrierten Personen gleich stark besetzte Schichten pro SMP

Die Definitionen der Schichtgrenzen der Varianten ADK2 und ADK3 lassen sich auch verstehen als eine Bestimmung der Quartile (ADK2) beziehungsweise der Oktile (ADK3) in Bezug auf die Variable Anschriftengröße.

Neben der Festlegung der Schichten ist eine wesentliche Aufgabe des Forschungsprojekts, eine geeignete Aufteilung des Stichprobenumfangs auf die Schichten zu finden. Wünschenswert ist die Einhaltung von Präzisionsanforderungen an die Schätzungen (in unterschiedlichen Graden je nach Ziel und Untergliederungstiefe; siehe Aufstellung in Abschnitt 2.3) sowie das Nicht-Überschreiten eines vorab festgelegten Gesamtstichprobenumfangs bezogen auf in Deutschland registrierte Personen.

2.1 Simulationsumgebung

Bei der Entwicklung eines geeigneten Stichprobendesigns und der Bewertung von Schätzverfahren wird auf unterschiedliche Datenbestände zurückgegriffen. Dem

zeitlichen Vorliegen dieser Daten folgt die chronologische Einteilung des Projektes in drei Phasen. In der Phase 0 liegen rein synthetische Daten zugrunde, die sich aber an der Realität orientieren. Diese synthetische Grundgesamtheit wurde im Rahmen des DACSEIS-Projektes auf Basis des Vorgehens von Devroye (1986) erzeugt.[1] In Phase 1 wurden zum einen ausgewählte anonymisierte Melderegisterdaten aus vier Bundesländern um synthetisch generierte Variablen angereichert sowie erste gesamtdeutsche aggregierte Melderegisterdaten zur Planung herangezogen. In Phase 2 stehen die anonymisierten Melderegisterdaten für Gesamtdeutschland (im Folgenden MR-Daten) zur Verfügung.

Die Simulationsgesamtheiten, die der ersten und zweiten Phase zugrunde liegen, setzen sich aus verschiedenen Teilen zusammen.[2] Da die MR-Daten als deterministischer Block angesehen werden können, dienen die 85.790.381 Einträge als Rahmen der Simulation. Es werden also keine neuen Personen hinzu simuliert, sondern lediglich weitere Variablen für diesen schon vorhandenen Block synthetisch generiert. Zudem werden einzelne Personen verschiedenen Modellen zufolge als nicht im Melderegister erfasste (Fehlbestände) oder fälschlicherweise im Melderegister erfasste Personen (Karteileichen) ausgewiesen (Münnich et al. 2009; Burgard 2009; Burgard/Münnich 2010). Die Grundlage dieser Karteileichen- und Fehlbestandsmodelle liefert ein anonymisierter Datensatz, der aus dem Zensustest 2001 resultiert.[3]

Als dritte wichtige Datenquelle zur Erzeugung der synthetischen Variablen dient der anonymisierte Mikrozensus aus dem Jahr 2006. Ziel der synthetischen Datengenerierung in den Phasen 1 und 2 war es, heterogenere Strukturen über die administrativen Einheiten zu erzeugen, als dies noch für die Phase 0 der Fall war.

Bei der synthetischen Datengenerierung werden nun aufbauend auf bereits vorhandenen Variablen weitere Variablen generiert, welche auch in Gruppen (blockweise) erzeugt werden können. Zunächst wird hier der Block der Registervariablen verwendet. Die Erzeugung erfolgt nun rekursiv auf Basis von Kreuztabellen oder Modellen. Die Kreuztabellen werden im Mikrozensus auf Kreis-Ebene berechnet, da tiefer gegliederte Ebenen bei der Auszählung kaum noch valide Ergebnisse liefern. Darüber hinaus müssen die Hochrechnungsfaktoren aus dem Mikrozensus 2006 (Variable EF951) bei der Auszählung berücksichtigt werden. Damit wird eine Kon-

1 Zum genauen Vorgehen zur Erzeugung des Phase 0 Datensatzes siehe Münnich und Schürle (2003).

2 Die beiden Simulationsgesamtheiten unterscheiden sich hinsichtlich des Umfangs (Phase 1 – anonymisierte Melderegisterauszüge aus vier Bundesländern, Phase 2 - anonymisierte Melderegisterdaten aus allen Bundesländern) und weniger hinsichtlich der verwendeten Methodik. Deshalb wird im Folgenden nur von einer Simulationsgesamtheit die Rede sein.

3 Für nähere Informationen zum Zensustest 2001 siehe Schäfer 2004.

sistenz zwischen den hochgerechneten Stichprobenergebnissen aus dem Mikrozensus und der synthetischen Population erzwungen. Wichtig ist hierbei, dass die Verteilungen der Variablen, die aus den MR-Daten resultieren, sich auch in der Kreuztabelle wiederfinden lassen.

Allgemein kann die gemeinsame Verteilung einer *n*-variaten Verteilung rekursiv über ihre bedingten Verteilungen erzeugt werden. Es gilt:

$$F(x_1,\ldots,x_n) = F(x_1)\cdot F(x_2 \mid x_1)\cdots F(x_n \mid x_1,\ldots,x_{n-1})$$

Speziell kann man hieraus für die Erzeugung eines neuen Blockes $x_{k+1},\ldots,x_n$ bei bereits gegebenen oder erzeugten Variablen $x_1,\ldots,x_k$ die Verteilungsfunktion

$$F(x_{k+1},\ldots,x_n \mid x_1,\ldots,x_k) = \frac{F(x_1,\ldots,x_n)}{F(x_1,\ldots,x_k)}$$

verwenden. Durch die blockweise Generierung der Daten werden innerhalb der Blöcke die beobachteten Strukturen erhalten, was beispielsweise bei Ausbildungsvariablen sinnvoll ist. Die Reihenfolge der Erzeugung der Variablen spielt damit keine Rolle, solange auf alle bereits erzeugten Variablen konditioniert wird.

Ein weiteres Problem bei der Datenerzeugung ist die Unterscheidung zwischen Stichprobennullen und strukturellen Nullen in den Kreuztabellen. Bei den Erstgenannten handelt es sich um vorhandene, aber durch die MZ-Stichprobe nicht beobachtete Ausprägungen. Strukturelle Nullstellen sind Ausprägungen, die tatsächlich nicht vorkommen. Insbesondere bei der Kombination von Alter, Bildungsvariablen sowie Erwerbsvariablen ergeben sich strukturelle Nullen. Verwendet man bei der Erzeugung synthetischer Daten Modelle oder benutzt nicht alle bereits vorhandenen Variablen für die Konditionierung, kann es sein, dass strukturelle Nullen nicht korrekt erkannt werden. In solchen Fällen muss nach der Erzeugung noch ein Editing durchgeführt werden.

2.2 Datenstruktur und Erhebungseinheiten

Die nachfolgend dargestellte hierarchische Struktur von Zusammenfassungen regionaler Einheiten ist derart gestaltet, dass die einzelnen Ebenen der Präzisionsanforderungen widerspruchsfrei und eindeutig berücksichtigt werden. Diese Struktur dient als Basis, um eine regionale Aufteilung des Gesamtstichprobenumfangs zu ermöglichen. Eine anschließende Schichtung zur Erhöhung der Präzision der Schätzungen bleibt davon unberührt.

Eine *Stichprobenbasiseinheit*[4] ist als regionale Einheit definiert, aus der eine Stichprobe gezogen wird, wobei der Stichprobenumfang noch festzulegen ist. Die Einteilung der SMPs in vier Typen erfolgt nach einem streng hierarchischen Schlüssel, bei dem ein Typ höherer Ordnung sich nur noch auf den Rest bezieht, der von dem Typ niederer Ordnung verblieben ist. Sind daher in einer Verbandsgemeinde eine oder mehrere Gemeinden mit mehr als 10.000 Einwohnern (EW) vorhanden, so sind diese großen Gemeinden Typ 1 zugehörig. Die größten deutschen Städte werden in Stadtteile unterteilt, diese Stadtteile sind dem Typ 0 zugehörig. Die restlichen kleinen Gemeinden der Verbandsgemeinde werden dem Typ 2 zugeordnet, wenn sie zusammen mehr als 10.000 Einwohner haben. Ansonsten gehören sie zum Typ 3.

Tabelle 2 Verteilung der SMP-Typen in den Bundesländern

	SMP-Typ (absolute Häufigkeiten)				
Bundesland	**SDT**	**GEM**	**VBG**	**KRS**	**Summe**
Baden-Württemberg	2	244	126	35	407
Bayern	8	216	30	71	325
Berlin	12	0	0	0	12
Brandenburg	0	71	5	14	90
Bremen	3	1	0	0	4
Hamburg	7	0	0	0	7
Hessen	3	168	0	21	192
Mecklenburg-Vorpommern	0	24	30	12	66
Niedersachsen	2	205	68	34	309
Nordrhein-Westfalen	12	339	0	17	368
Rheinland-Pfalz	0	46	122	20	188
Saarland	0	40	0	5	45
Sachsen	4	69	13	22	108
Sachsen-Anhalt	0	60	27	11	98
Schleswig-Holstein	0	53	52	11	116
Thüringen	0	33	6	17	56
Deutschland	53	1569	479	287	2391

4 Mit der Wahl dieser Bezeichnung soll vermieden werden, dass diese mit den so genannten *Small Areas* beziehungsweise den Schichten, die innerhalb dieser Einheiten gebildet werden, direkt in Verbindung gebracht werden.

Typ 0 (SDT): Stadtteile ab 200.000 EW aus Gemeinden mit mindestens 400.000 EW

Typ 1 (GEM): Gemeinden mit mindestens 10.000 EW, sofern sie nicht zum Typ 0 gehören

Typ 2 (VBG): Kleine Gemeinden (unter 10.000 EW) innerhalb eines Gemeindeverbands beziehungsweise einer Verbandsgemeinde werden zusammengefasst, sofern sie in der Summe mindestens 10.000 EW betragen

Typ 3 (KRS): Zusammenfassung aller Gemeinden eines Kreises, die bis dahin noch keinem Typ zugeordnet wurden

Insgesamt lassen sich auf diese Weise 2.391 SMPs bilden[5], die sich auf die Bundesländer wie in Tabelle 2 dargestellt verteilen.

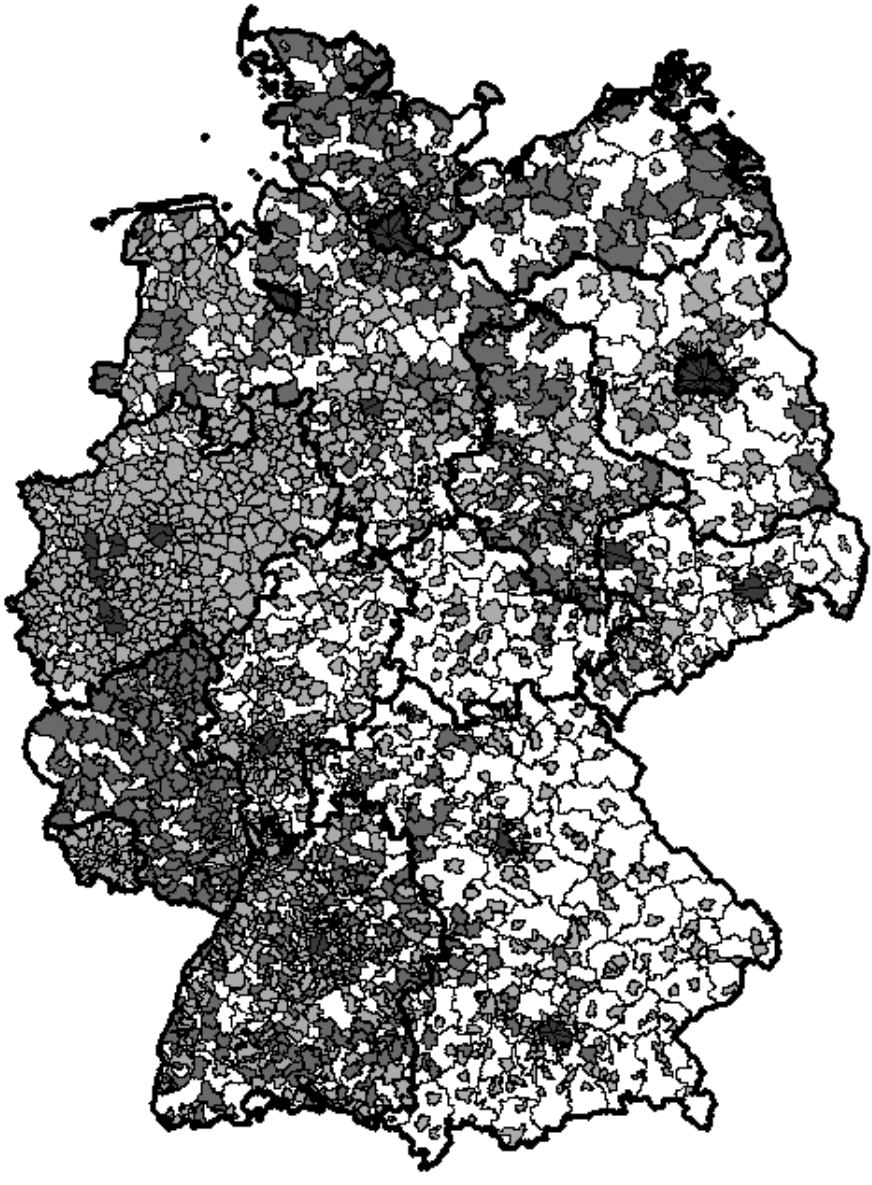
In Abbildung 1 ist eine Darstellung der geografischen Verteilung der SMPs zu finden.[6] Die Einfärbung in Abbildung 1 orientiert sich dabei an der Klassifikation der SMPs (GEM: dunkelstes Grau; SDT: dunkles Grau; KRS: helles Grau; VBG: hellstes Grau); diese sind durch feine schwarze Linien abgegrenzt.

Es sei noch erwähnt, dass von den in Tabelle 1 beschriebenen Schichtgrenzen jene von ADK1 in allen SMPs gleichermaßen definiert sind, wohingegen die Schichtgrenzen von ADK2 und ADK3 von der Verteilung der Personen innerhalb einer SMP abhängen. Da in jeder SMP eine Stichprobe gezogen wird und alle Anschriften des Bundesgebietes vollständig auf die SMPs aufgeteilt sind, muss auch die SMP als Schichtungsvariable betrachtet werden. Durch die Kreuzung von SMP und der Anschriftengrößenklassen ergeben sich schließlich insgesamt 14.346 SMPxADK1, 9.564 SMPxADK2 beziehungsweise 19.128 SMPxADK3 Schichten.

5 Es gilt zu beachten, dass diese Zahl auf den Melderegisterdaten der Phase 1 beruhen, die endgültige Zahl der SMPs zum Zensus-Stichtag kann hiervon noch geringfügig abweichen.

6 Die Abbildung stammt aus einer Forschungsarbeit des Forschungszentrums für Regional- und Umweltstatistik zum Thema *Wirkung der Verwendung von Verbandsgemeinden beim Zensus 2011 in Rheinland-Pfalz*. Die Darstellung wurde mit Hilfe des Datensatzes VG250 vom 31.12.2006 des Geodatenzentrums erstellt und kann kleinere Abweichungen zum aktuellen Registerauszug aufweisen, welche jedoch für die Aussagen keine Auswirkungen haben. Den verwendeten Karten liegen die Vektordaten in den Verwaltungsgrenzen 1:250.000 des Bundesamtes für Kartographie und Geodäsie zugrunde.

Abbildung 1 Darstellung der Stichprobenbasiseinheiten in Deutschland



2.3 Präzisionsanforderungen und Schätzer

Zentrales Ziel der Forschungsarbeiten ist es, Vorschläge für Stichprobendesign und Schätzer zu entwickeln, welche gegebene Präzisionsanforderungen erfüllen. Im Folgenden werden nun zunächst diese Präzisionsanforderungen genauer dargestellt. In einem weiteren Abschnitt wird schließlich auf das Referenzschätzverfahren sowie weitere verwendete Schätzer näher eingegangen. Neben dem Horvitz-Thompson (HT) Schätzer und dem modellunterstützten verallgemeinerten Regressionsschätzer (GREG) werden synthetische Schätzer und empirisch beste lineare unverzerrte Prädiktoren (EBLUP) verwendet.

Präzisionsanforderungen

Im Vordergrund des Forschungsprojekts steht die Frage, wie die folgenden drei zentralen Anforderungen an ein Auswahlverfahren erfüllt werden können

1. Einhaltung der gestellten Präzisionsanforderungen
2. Eingrenzung der Variation der Designgewichte
3. Berücksichtigung eines maximalen Stichprobenumfangs

Tatsächlich lassen sich diese Anforderungen nicht alle gleichzeitig erfüllen, da es sich um konkurrierende Ziele handelt. Vor der Beantwortung obiger Frage stehen Überlegungen zu einer geeigneten regionalen Gliederung von Gemeinden, Verbandsgemeinden und Stadtteilen in die oben beschriebenen Stichprobenbasiseinheiten (SMP) (im Folgenden indiziert durch <area>), aus denen die Teilstichproben von Anschriften mit zu bestimmenden Umfängen in den Schichten gezogen werden. Zentrales Bewertungskriterium der konkurrierenden Stichprobendesigns ist der relative root mean square error (RRMSE). Es wird also die Präzision der Schätzung von interessierenden Variablen für verschiedene Schätzverfahren und Stichprobendesigns verglichen. Im Rahmen des Forschungsprojekts wurden für die zwei oben genannten Ziele Präzisionsanforderungen formuliert, welche eingehalten werden müssen. Dabei bezeichnet $\hat{\tau}_{Z,<area>}$ den mit den Stichprobendaten geschätzten Totalwert der in $\mathrm{SMP}_{<area>}$ lebenden Personen und $\hat{\tau}_{Y,<area>}$ den mit den Stichprobendaten geschätzten Totalwert einer Untersuchungsvariable Y der in $\mathrm{SMP}_{<area>}$ tatsächlich lebenden Personen. Es gelten insbesondere die folgenden Anforderungen bei[7]

7 Siehe hierzu auch: Gesetz zur Anordnung des Zensus 2011 sowie zur Änderung von Statistikgesetzen, §7.

Ziel 1: Schätzungen ausschließlich in Gemeinden ab 10.000 EW

- Stadtteile von Großstädten (SMP-Typ 0):
  $\text{RRMSE}(\hat{\tau}_{Z,<\text{area}>}) \leq 0{,}5\%$
- Gemeinden ab 10.000 EW (SMP-Typ 1):
  $\text{RRMSE}(\hat{\tau}_{Z,<\text{area}>}) \leq 0{,}5\%$

Ziel 2: Betrachtet wird bei

$$\frac{\tau_{Y,<\text{area}>}}{\tau_{Z,<\text{area}>}} \approx p \qquad (\text{mit } p \geq 1/15) \text{ in \%}$$

Es gilt jeweils $\text{RRMSE}(\hat{\tau}_{Y,<\text{area}>}) \leq \frac{1}{p}$ in % für:

- Stadtteile von Großstädten (SMP-Typ 0)
- Gemeinden ab 10.000 EW (SMP-Typ 1)
- VBG in Rheinland-Pfalz (SMP-Typ 2)
- Kreise (SMP-Typ 3)

Bei der Ermittlung der amtlichen Einwohnerzahl (Ziel 1) interessiert ausschließlich die Präzision der Schätzer in so genannten *großen Gemeinden*, also Gemeinden und Stadtteilen von Großstädten mit 10.000 und mehr Einwohnern. Wird die Präzision der Schätzung von Ziel 2 Variablen betrachtet, so unterscheiden sich die Präzisionsanforderungen danach, wie hoch der Anteil p der interessierenden Merkmalsausprägung ist. Bei Merkmalsausprägungen von einem Anteil unter $p = 1/15 \approx 6{,}67$ % wird keine explizite Anforderung an die Präzision der Schätzer gestellt. Liegt der Anteil einer Ziel 2 Variablen über $p = 1/15$, so muss der RRMSE kleiner sein als $1/p$ in % und zwar bezogen auf die Schätzung in den vier oben genannten Gebietstypen. Bei der Schätzung einer Ziel 2 Variablen, bei der eine Merkmalsausprägung beispielsweise mit 25 % von Interesse ist, müsste der RRMSE also kleiner oder gleich 4 % sein, wird eine Merkmalsausprägung von 50 % betrachtet, so muss der RRMSE kleiner gleich 2 % sein.

In den bisherigen Untersuchungen hat sich gezeigt, dass die Präzisionsanforderungen für Ziel 1 von vielen Auswahlverfahren und Schätzern eingehalten werden können. Als komplexer zu beurteilen hat sich die Einhaltung der für Ziel 2 formulierten Präzisionsanforderungen erwiesen, da hier sowohl die Anteile als auch deren Verteilung auf die Schichten zu erheblich unterschiedlichen Ergebnissen führen können. In Abschnitt 4 wird an einem Beispiel verdeutlicht, wie unterschiedlich dann auch die Schätzergebnisse beurteilt werden müssen.

Schätzer

Im Rahmen des Forschungsprojekts werden acht Schätzer für den Totalwert interessierender Variablen untersucht. Eine ausführlichere Diskussion dieser Schätzer findet sich in Münnich et al. (2007). Allgemein bezeichnet d_k das Designgewicht, also die Inverse der Inklusionswahrscheinlichkeit für Einheit *k*. Bei dem verwendeten geschichteten Auswahlverfahren ist die Inklusionswahrscheinlichkeit des *i*-ten Elements in Schicht *h* gegeben durch

$$\pi_k = \pi_{h,i} = \frac{n_{A,h}}{N_{A,h}} \quad ,$$

wobei $n_{A,h}$ die Anzahl der zu ziehenden und $N_{A,h}$ die Gesamtzahl der Anschriften in Schicht *h* bezeichnet. Allgemein verzichten wir im Folgenden auf das Subskript *A* und verwenden es nur, wenn eine explizite Abgrenzung der Anzahl der Anschriften von der Anzahl der Personen *(*Subskript *P)* notwendig ist. Somit ist das Designgewicht des *i*-ten Elements in Schicht *h* gegeben durch

$$w_k = w_{h,i} = \frac{1}{\pi_{h,k}} = \frac{N_h}{n_h} \quad .$$

Die untersuchten Schätzer lassen sich unterteilen in *designbasierte* sowie *modellbasierte* Schätzer. Aus der Klasse der designbasierten Schätzer werden der HT Schätzer und drei Varianten des verallgemeinerten Regressionsschätzers (GREG-Schätzer) betrachtet.

Es bezeichne *d* die SMP, y_i die Ausprägung der *i*-ten Einheit auf der Untersuchungs- und x_i die Ausprägung der *i*-ten Einheit auf der Hilfsvariable. Als Hilfsvariable wird in der Regel die Anschriftengröße aus dem AGR verwendet. Die Qualität der Schätzer hängt natürlich von den verfügbaren Hilfsmerkmalen und deren Korrelationen mit der Untersuchungsvariable ab (siehe Abbildung 2).

Darüber hinaus ist $\tau_{X,d}$ der bekannte Totalwert der Hilfsvariable in der SMP *d* und $\hat{\tau}_{X,d}$ der aus der Stichprobe *s* geschätzte Totalwert der Hilfsvariable in der SMP *d*. Außerdem ist $\hat{\tau}_{Y,d}$ der aus der Stichprobe *s* geschätzte Totalwert der Untersuchungsvariable. Damit ist der HT- und der GREG-Schätzer für den Totalwert gegeben durch

HT Schätzer $$\hat{\tau}_Y^{\text{HT}} = \sum_{i \in s} w_k \cdot y_k$$

GREG (Small Area, SA) $$\hat{\tau}_{Y,d}^{\text{GREG,SA}} = \hat{\tau}_{Y,d}^{\text{HT}} + \left(\tau_{X,d} - \hat{\tau}_{X,d}\right)' \hat{\beta} \,,$$

mit $$\hat{\beta} = \left(\sum_{i \in s} w_i x_i x_i^T \right)^{-1} \sum_{i \in s} w_i x_i y_i \,.$$

Bei den modellbasierten Schätzern lassen sich bei Small Area-Schätzern synthetische und zusammengesetzte Schätzer unterscheiden. Letztere sind in unserem Fall empirisch beste lineare unverzerrte Prädiktoren (EBLUP). Innerhalb dieser zwei Schätzer-Familien kann weiterhin zwischen den den Schätzern zugrunde liegenden Modellen unterschieden werden. Die folgende Aufstellung gibt eine Übersicht über die untersuchten modellbasierten Schätzer.

Synthetische Schätzer

Modell A:
Angenommen, unit-level Daten von Hilfsvariablen $x_{di} = (x_{di1}, \ldots, x_{dip})^T$ seien für jede Einheit i in der Small Area Gesamtheit d verfügbar. Weiter sei die Zielvariable y_{di} durch ein lineares Regressionsmodell mit x_{di} und hierarchischen Fehlertermen verbunden

$$y_{di} = x_{di}^T \beta + u_d + \varepsilon_{di} \,,$$

wobei β der Regressionskoeffizientenvektor ist, u_d der Area-spezifische Effekt mit $E(u_d) = 0$, $\text{var}(u_d) = \sigma_u^2$ und $\text{var}(\varepsilon_{di})$ der unabhängige Zufallsfehler mit $E(\varepsilon_{di}) = 0$ und $\text{var}(\varepsilon_{di}) = \sigma_\varepsilon^2$ ist. Der synthetische Schätzer ist durch

$$\hat{\mu}_{Yd}^{\text{SYNA}} = \mu_{Xd}^T \hat{\beta}$$

definiert mit bekanntem Area-level Kovariatenvektor $\mu_{Xd} = (\mu_{Xd1}, \ldots, \mu_{Xdp})^T$, dem Vektor der wahren Mittelwerte von p Kovariaten (x_i ist p-dimensional) in der Area d. $\hat{\beta}$ ist wie beim GREG-Schätzer definiert. Beispiele für Kovariate finden sich in Kapitel 4.

Modell B:
Der synthetische Schätzer B verwendet ein lineares (normalverteiltes) Modell mit Area-level Kovariaten und einer gepoolten Schätzung der Varianz innerhalb der Areas. Das Modell ist

$$\overline{y}_{d.} = \mu_{Xd}^T \beta + \xi_d \ ,$$

wobei ξ_d der Area-spezifische Effekt mit $E(\xi_d) = 0$ ist und $\text{var}(\xi_d) = \sigma_u^2 + \psi_d$ gilt. Die Stichprobenvarianz ist $\psi_d = \sigma_\varepsilon^2 / n_d$, wobei n_d den Stichprobenumfang in der Area *d* bezeichnet.

Empirische beste linear unverzerrte Prädiktoren (EBLUP)

Modell A:

$$\hat{\mu}_{Yd}^{\text{EBLUPA}} = \hat{\gamma}_d \hat{\mu}_{Yd}^{\text{GREG}} + (1-\hat{\gamma}_d)\hat{\mu}_{Yd}^{\text{SYNA}} = \hat{\gamma}_d (\hat{\mu}_{Yd}^{\text{GREG}} - \hat{\mu}_{Xd}^T \hat{\beta}) + \mu_{Xd}^T \hat{\beta} \ ,$$

mit

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2 / n_d} \ .$$

Modell B:

$$\hat{\mu}_{Yd}^{\text{EBLUPB}} = \hat{\gamma}_d \hat{\mu}_{Yd}^{\text{GREG}} + (1-\hat{\gamma}_d)\hat{\mu}_{Yd}^{\text{SYNB}} = \hat{\gamma}_d \hat{\mu}_{Yd}^{\text{GREG}} + (1-\hat{\gamma}_d)\mu_{Xd}^T \hat{\beta} \ ,$$

mit

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\psi}_d} \ .$$

Für alle Modelle gilt, dass $\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2$ und $\hat{\psi}_d$ adäquate Schätzer für σ_u^2, σ_ε^2 und ψ_d sind.

Eine eingehende Übersicht über Small Area-Verfahren kann der Monografie von Rao (2003) oder Münnich, Burgard und Vogt (2011) entnommen werden. Sowohl bei den modellunterstützten als auch bei den modellbasierten Verfahren werden die Daten der Melderegister als Hilfsvariablen herangezogen, insbesondere die Anschriftengröße.

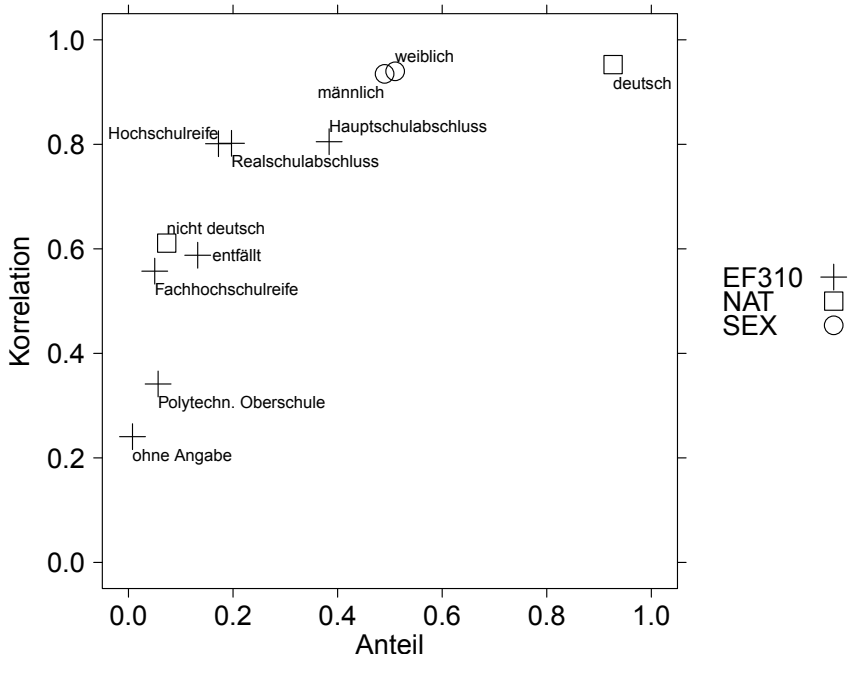
Von Interesse ist neben der Punktschätzung natürlich die Genauigkeit, welche im Rahmen dieser Untersuchungen durch den RRMSE der Schätzer quantifiziert wird. Die Schätzer hierfür lassen sich mitunter nicht in geschlossener Form

ausdrücken. Es existieren jedoch Approximationen, die zur Varianzschätzung benutzt werden können. Für eine Übersicht über einige wichtige Methoden sei auf Wolter (2007) und Münnich (2008) verwiesen. Als Benchmark dient die Varianz des GREG-Schätzers

$$\mathrm{var}\left(\hat{\tau}_Y^{\mathrm{GREG}}\right) = \sum_{h=1}^{H} N_h^2 \cdot \frac{S_{h,Y}^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \left(1 - \rho^2\right),$$

wobei $S_{h,Y}^2$ die Varianz der Untersuchungsvariable in der Schicht h ist und $\rho = 0{,}993$ die Korrelation zwischen der Hilfsvariable (hier: Anschriftengröße im Melderegister) und dem Totalwert einer Merkmalsausprägung in einer Anschrift einer interessierenden Ziel 1 oder Ziel 2 Variablen bezeichnet. Ist diese Korrelation hoch, so ist auch der Gewinn an Präzision durch Einbeziehen der Anschriftengröße als Hilfsvariable hoch. Voruntersuchungen im Zensustest haben gezeigt, dass selbst Ziel 2 Variablen mitunter sehr hohe Korrelationen aufweisen, wie Abbildung 2 verdeutlicht.

Abbildung 2 Korrelation von Ausprägungen ausgewählter Ziel 2 Variablen mit der Anschriftengröße



Hier sind die Anteile der Ausprägungen der Ziel 2 Variablen Geschlecht (SEX), Nationalität (NAT) sowie höchster Bildungsabschluss (EF310) gegen deren jeweilige Korrelationen mit der Anschriftengröße dargestellt.[8]

Es ist leicht ersichtlich, dass mit steigendem Anteilswert tendenziell auch die Korrelation mit der Anschriftengröße zunimmt. Insgesamt bewegen sich die Korrelationen aller Ausprägungen dieser Ziel 2 Variablen jedoch ohnehin auf einem hohen Niveau, so dass ein merklicher Gewinn an Präzision durch Verwendung des GREG-Schätzers zu erwarten ist.

3 Optimale Allokation unter Nebenbedingungen

Auf Basis der oben definierten Anforderungen und Grundlagen empfiehlt sich ein geschichtetes Auswahlverfahren. Aufgrund der Tatsache, dass die Schichtvarianz $S^2_{R,h,Y}$ der Anschriftengröße aus dem Melderegister zur Verfügung steht, liegt es nahe, die optimale Allokation nach Neyman-Tschuprov (Tschuprov 1923; Neyman 1934) zur Aufteilung des Gesamtstichprobenumfangs n auf die H Schichten zu verwenden. Dies ist wünschenswert, da optimale Allokationsverfahren zu Stichproben mit minimaler Varianz führen. Allerdings gilt dies nur insoweit, als ein hoher Zusammenhang zwischen der bei der Allokation verwendeten $S^2_{R,h,Y}$ und den tatsächlichen Schichtvarianzen im Zensus $S^2_{Z,h,Y}$ besteht. Da es sich bei den $S^2_{R,h,Y}$ um Melderegisterdaten handelt, kann hier von einer hohen Übereinstimmung mit $S^2_{Z,h,Y}$ ausgegangen werden und daher $S^2_{R,h,Y} = S^2_{Z,h,Y} = S^2_{h,Y}$ angenommen werden.

Die Neyman-Tschuprov Allokation eines Gesamtstichprobenumfangs n auf H Schichten ist gegeben durch

$$n_h^{\text{opt,NT}} = n \cdot \frac{N_h S_{h,Y}}{\sum_{\ell=1}^{H} N_\ell S_{\ell,Y}} = n \cdot \frac{d_h}{\sum_{\ell=1}^{H} d_\ell} \quad .$$

Bei identischen Schichtvarianzen entspricht die optimale Allokation der proportionalen Aufteilung. Doch anders als bei der proportionalen Aufteilung kann es bei der optimalen Aufteilung vorkommen, dass in einer Schicht mehr Anschriften ausgewählt werden sollen als Elemente in der Population existieren, also $n_h > N_h$

8 Die sehr hohe Korrelation etwa der Ausprägung männlich beziehungsweise weiblich der Ziel 2 Variablen Geschlecht bedeutet inhaltlich, dass tendenziell in einer Anschrift gleich viele Frauen wie Männer leben.

resultiert. Dies kann passieren, wenn die Varianz $S^2_{h,Y}$ in einer Schicht besonders groß ist. Je nach Schichtungsvariante kommt dies in der praktischen Umsetzung durchaus vor, vor allem in den größten Anschriftengrößenklassen. Ein Teilstichprobenumfang mit $n_h > N_h$ kann im Ziehungsmodell ohne Zurücklegen natürlich nicht realisiert werden und sollte von einem geeigneten Allokationsverfahren von vornherein vermieden werden. Darüber hinaus sollte auch kein zu geringer Auswahlsatz innerhalb einer Schicht vorkommen, da hiermit möglicherweise sehr hohe RRMSEs verbunden sind. Verwendet man modellbasierte Verfahren, dann sollte man auch darauf achten, dass die Designgewichte nicht zu unterschiedlich ausfallen. In einem diskutierten Papier stellt Gelman (2007) die Problematik von Designgewichten bei statistischer Modellbildung heraus. Diese Problematik spielt bei der Anwendung von Small Area-Verfahren durchaus eine Rolle und führt zu einer weiteren Einschränkung der Gewichte in Form von oberen und unteren Grenzen, so genannten Box-Constraints.

Diese Restriktionen führen dazu, dass die folgenden Box-Constraints als Anforderung an die zu bestimmenden Stichprobenumfänge n_h definiert werden

$$m_h \leq n_h \leq M_h$$

mit bekannten Constraints m_h und M_h. Darüber hinaus darf ein vorab festgelegter Gesamtumfang an Personen nicht überschritten werden. Aufgrund der Tatsache, dass in einer Schicht Anschriften von verschiedener Größe zusammengefasst sind, kann die Personenzahl in einer Schicht von Stichprobe zu Stichprobe variieren. Daher kann auch nur eine mittlere erwartete Personenzahl bei gegebenem $n_{A,h}$ angegeben werden. Die mittlere erwartete Personenzahl in Schicht h ist gegeben durch $n_{A,h} \cdot \frac{N_{P,h}}{N_{A,h}}$. Soll insgesamt der Anteil θ von Personen in Deutschland ausgewählt werden, erhalten wir als zusätzliche Nebenbedingung

$$\sum_{h=1}^{H} n_{A,h} \cdot \frac{N_{P,h}}{N_{A,h}} = N_P \cdot \theta \ .$$

Diese Anforderungen können von der naiven Neyman-Tschuprov-Allokation im Rahmen geschichteter Zufallsstichproben nicht mehr ohne weiteres erfüllt werden. Das so gestellte Problem einer nicht-linearen Optimierung unter Nebenbedingungen kann jedoch zufriedenstellend gelöst werden, wie Gabler et al. (2010) zeigen. Ein einfacher Algorithmus ermöglicht die optimale Aufteilung eines Gesamtstich-

probenumfangs n auf H Schichten, wobei sowohl untere als auch obere Grenzen für die Stichprobenumfänge in den Schichten eingehalten werden und auch die oben genannte zusätzliche Nebenbedingung erfüllt wird. Dabei macht sich der Algorithmus die Tatsache zunutze, dass die Schichten exakt drei Klassen zugeordnet werden können. In Schichten, die der ersten Klasse U_1 angehören, wird der Stichprobenumfang exakt auf die untere Schranke m_h gesetzt, in Schichten der zweiten Klasse U_2 wird n_h exakt auf die obere Schranke M_h gesetzt und in der dritten Klasse U_3 wird der verbleibende Stichprobenumfang $n - \sum_{h \in U_1} m_h - \sum_{h \in U_2} M_h$ optimal im Sinne von Neyman aufgeteilt. Das Problem besteht somit darin, diejenige Zusammensetzung der Klassen zu finden, für die insgesamt eine bestimmte Zielfunktion minimiert wird. Der Algorithmus löst dieses Problem dadurch, dass zunächst zwei geordnete Reihen gebildet werden, in denen die Schichten entsprechend ihrer Ausprägung auf $N_h \cdot S_{h,Y}$ in aufsteigender, beziehungsweise absteigender Reihenfolge angeordnet sind. Anschließend werden die Kombinationen dieser Ordnungen Schritt für Schritt abgearbeitet. Die erste Lösung, bei der alle Elemente aus U_3 die Nebenbedingungen erfüllen, ist die angestrebte Lösung.

Details des Algorithmus inklusive Beispiel sind in Gabler et al. (2010) aufgeführt. Der Algorithmus wird für die Aufteilung des Gesamtstichprobenumfangs auf die Schichten im Rahmen des Forschungsprojekts verwendet. Im folgenden Abschnitt wird der Einfluss verschiedener Schichtungsvarianten und Box-Constraints auf die zu erwartende Präzision des GREG-Schätzers untersucht.

4 Ergebnisse

Der Grad der zu erwartenden Einhaltung der Präzisionsanforderungen spielt eine zentrale Rolle in der Bewertung der Eignung der potentiellen Allokations- und Auswahlverfahren. Die folgenden Darstellungen beruhen auf den Ergebnissen von Phase 1-Melderegisterdaten. Die in den Abbildungen dargestellten Box-Plots verdeutlichen die zu erwartenden Verteilungen des RRMSE nach SMP-Typen (SMP-Typ 0, SMP-Typ 1, SMP-Typ 2 und SMP-Typ 3). Bei den auf Bundesländern bezogenen Darstellungen kann es vorkommen, dass für verschiedene Typen keine Box-Plots ausgewiesen sind. Dies ist auf den Umstand zurückzuführen, dass in diesen Bundesländern gewisse SMP-Typen nicht vorkommen (siehe hierzu auch Tabelle 2).

Neben den oben beschriebenen Schichtungsvarianten (ADK1, ADK2 und ADK3) werden in den Simulationen darüber hinaus drei Varianten der Kombination von unterer und oberer Grenze der Schichtumfänge sowie unterschiedliche Grade

des Zusammenhangs ρ der Untersuchungsvariable Y mit den Anschriften aus den Melderegisterdaten angenommen. Allen Simulationen liegt ein beispielhafter Wert von $\theta = 9{,}15\,\%$ zugrunde. Schließlich werden noch drei Varianz-Varianten (minimal, mittel, maximal) angenommen, welche die Art und Weise beeinflussen, wie die Untersuchungsvariable auf die Anschriften verteilt ist.[9]

In Abbildung 3 ist die Verteilung des relativen RMSE für eine Ziel 1 Fragestellung (also die amtliche Einwohnerzahl) dargestellt, wobei die mittlere Varianz-Variante und eine Korrelation der Anschriftengröße aus dem Melderegister mit der tatsächlichen Anschriftengröße von $\rho = 0{,}993$ angenommen wird.[10] Innerhalb eines Panels verdeutlichen die vier Box-Plots die Verteilung des RRMSEs nach SMP-Typen. Der Box-Plot des SMP-Typs 1 setzt sich dabei zum Beispiel aus 1.621, der Box-Plot des SMP-Typs 0 dagegen aus 52 Datenpunkten zusammen. Der durchgezogene vertikale Strich markiert den höchstens erlaubten RRMSE, die gestrichelte vertikale Linie den Durchschnitt aller RRMSEs. Darüber hinaus markiert die durchgezogene Linie in einem Boxplot den Mittelwert der RRMSEs innerhalb eines SMP-Typs, der Punkt den Median. Die drei Panels in der unteren Reihe der Abbildung beziehen sich auf die Variante mit 1 % und 50 % als untere bzw. obere Grenze der Anteile $n_{A,h} / N_{A,h}$ in Prozent, die Panels der mittleren Reihe auf die Variante mit 2 % und 40 % und die Panels der oberen Reihe auf die Variante mit 5 % und 20 %. Analog beruhen die Verteilungen, die in den Panels der ersten Spalte dargestellt sind, auf der Schichtungsvariante ADK1, diejenigen der zweiten Spalte auf ADK2 und die der dritten Spalte auf ADK3.

Aus Abbildung 3 wird ersichtlich, dass sich die Verteilungen der RRMSEs zwischen den einzelnen Panels zwar im Niveau unterscheiden, sich gewisse Muster aber in einer Mehrzahl der Panels finden. So ist etwa der Median der RRMSEs von SMPs vom Typ 2 in den Panels der unteren beiden Reihen der Abbildung 3 jeweils am höchsten, der Median der RRMSEs von SMPs vom Typ 1 dagegen am geringsten. Legt man das Kriterium an, dass der mittlere RRMSE über alle SMP-Typen am kleinsten sein soll, so ist von den Schichtungsvarianten ADK1 zu bevorzugen. Mit dem gleichen Kriterium würde innerhalb von ADK1 die Box-Constraints Variante mit 2 %-40 % favorisiert werden. Auch unter der Maßgabe, dass bei Ziel 1 ausschließlich die SMP-Typen 0 und 1 eine Rolle spielen, ist diese Box-Constraints-Variante zu bevorzugen.

9 Alle drei Schichtungsvarianten wurden zu Testzwecken untersucht. Letztendlich fiel die Entscheidung für acht Schichten in Personen-gleicher Allokation (ADK3).

10 Der Wert 0,993 geht auf Abschätzungen des Auftraggebers zurück.

Abbildung 3 RRMSE bei Ziel 1 Fragestellung; mittlere Varianz-Variante; ρ = 0,993

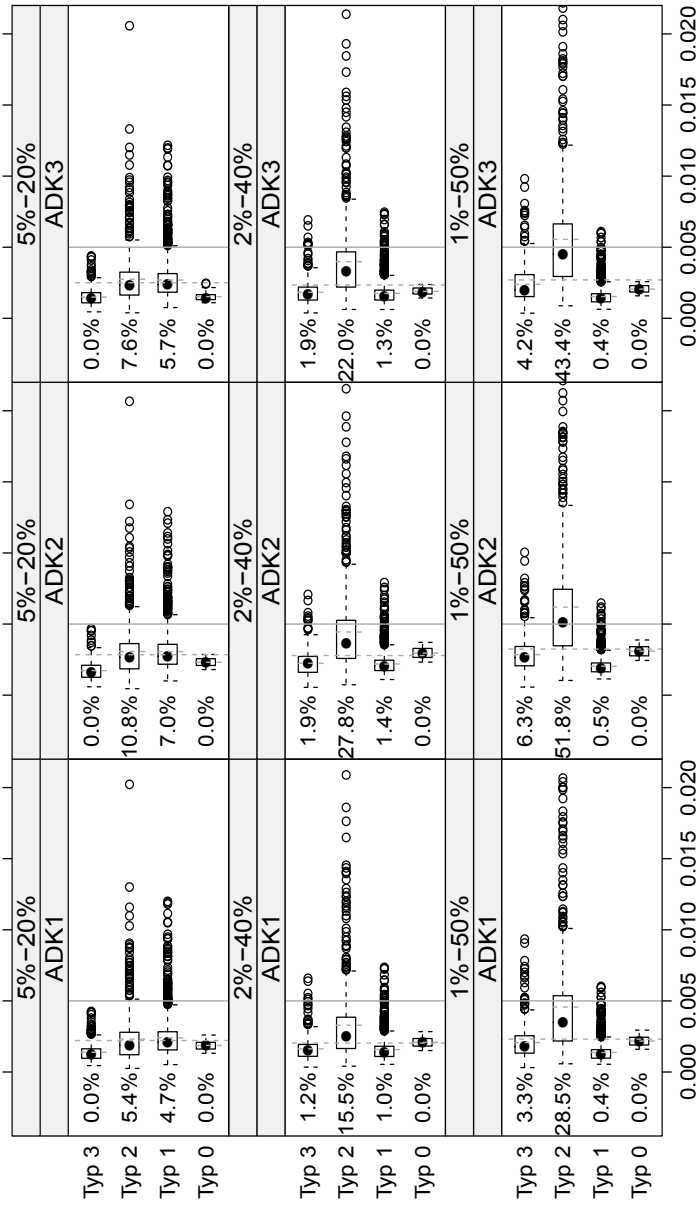


Abbildung 4 zeigt die Verteilung der RRMSEs analog bei einer (synthetischen) Ziel 2 Variable mit 50 % Anteilswert einer interessierenden Ausprägung und einer nur geringen Korrelation mit der Anschriftengröße von $\rho = 0{,}6$ und mittlerer Varianz-Variante.

Nachfolgend werden nun die RRMSEs der Schätzung auf Basis von Phase 2-Daten grafisch verdeutlicht. In Abbildung 5 sind die RRMSEs der Erwerbstätigen-Schätzung für Rheinland-Pfalz und Nordrhein-Westfalen kartografisch aufgearbeitet dargestellt. Die schwarzen Linien stellen dabei die Begrenzungen der SMPs dar. Das Ausmaß der RRMSEs ist farblich gekennzeichnet. Dabei stehen hellgraue Färbungen für niedrige RRMSEs. Je dunkler die Einfärbung, desto höher ist der RRMSE in dem jeweiligen SMP. SMPs mit RRMSEs über 10 % sind in sehr dunklem Grau gekennzeichnet.

Dargestellt werden die Ergebnisse für die in Abschnitt 2.3 beschriebenen vier Schätzer. Unten links der HT Schätzer, unten rechts der Small Area GREG-Schätzer, oben links ein Schätzer beruhend auf einem synthetischen Unit-Level Modell und oben rechts der EBLUPA. Als Hilfsvariablen für die Modelle wurden jeweils die Anschriftengröße, der Anteil an Ausländern sowie die Anschriftengrößenklasse verwendet (unter 3 Personen, 3 bis unter 7 Personen und ab 7 Personen).

Beim Vergleich der beiden designbasierten Verfahren (HT und GREG) schneidet der GREG leicht besser als der HT ab. Insbesondere die beim HT dunkleren SMPs sind beim GREG etwas heller. Auch weist der GREG keinen SMP mit einem RRMSE über 10 % auf. Jedoch ist die Verbesserung des GREG gegenüber dem HT nicht besonders stark ausgeprägt. Dies rührt daher, dass die verwendeten Hilfsvariablen keine hohe Erklärungskraft für die Untersuchungsvariable aufweisen. Da als Hilfsvariablen lediglich Informationen verwendet werden können, die für ganz Deutschland vorhanden sind, können nur Melderegistervariablen als Hilfsinformation herangezogen werden. Dazu gehören Variablen wie Alter, Geschlecht, Nationalität oder Wohnstatus. Für viele Ziel 2 Fragestellungen bieten diese Variablen wenig Erklärungskraft. Weitere Hilfsvariablen aus anderen Registern, etwa erwerbsstatistischen Registern, versprechen hierbei jedoch enorme Verbesserungen der Schätzqualität.

Abbildung 4 RRMSE bei Ziel 2 Fragestellung; mittlere Varianz-Variante; 50 % Anteilswert; ρ = 0,6

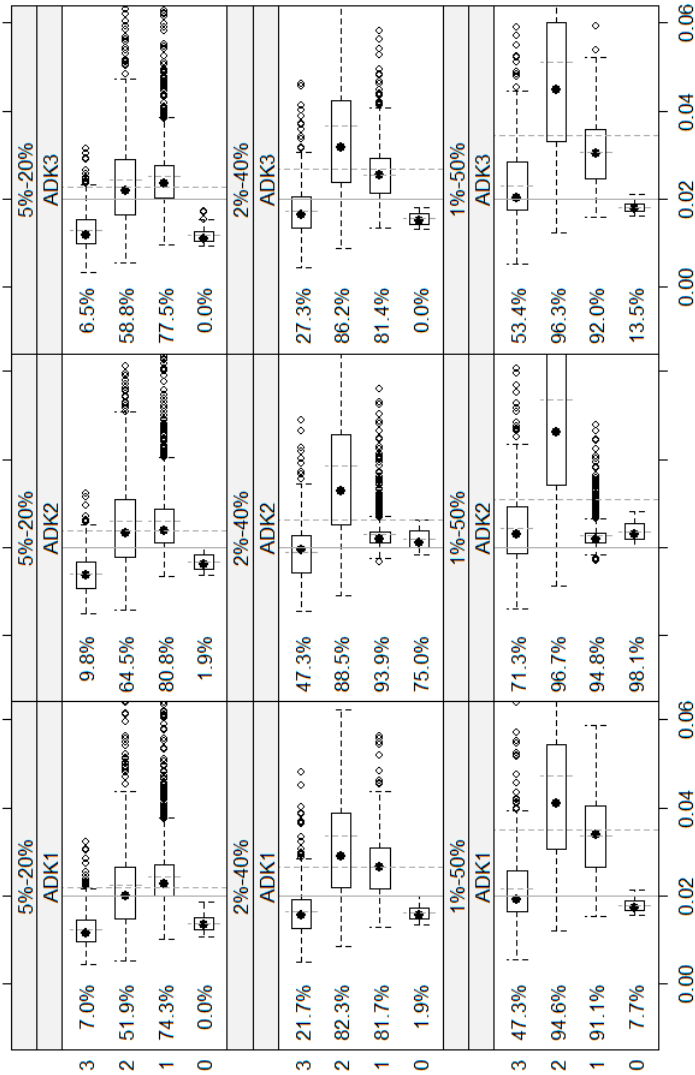
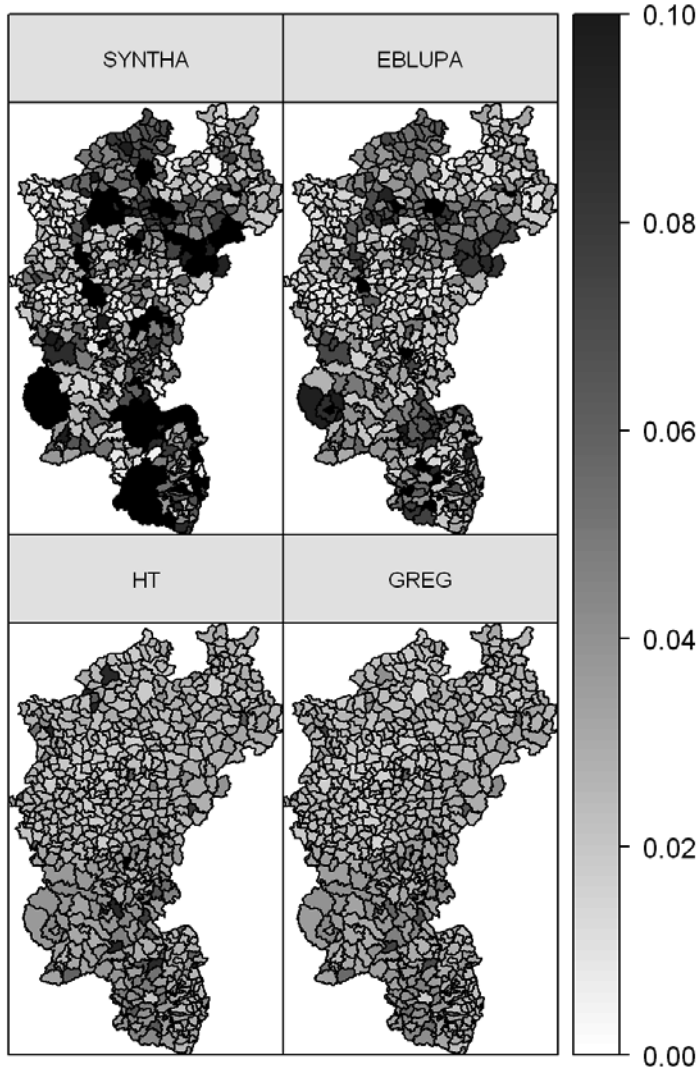


Abbildung 5 Ergebnisse der Small Area-Schätzungen in Rheinland-Pfalz und Nordrhein-Westfalen



Beim GREG werden die Hilfsvariablen lediglich als Unterstützung des designbasierten Schätzers verwendet. Die Erklärungskraft dieser Variablen in Bezug auf die Untersuchungsvariable wirkt sich fast ausschließlich positiv auf die Schätzqualität aus. Dagegen hängt die Qualität der Small Area-Schätzer fast ausschließlich von den Hilfsvariablen ab. Dies bedeutet auch, dass ungünstige Hilfsvariablen einen erheblichen Bias in den Schätzungen hervorrufen können. Daher ist in dieser Situation nicht weiter verwunderlich, dass der synthetische Unit-Level Schätzer bei diesem Vergleich am schlechtesten abschneidet. Die Variablen können nicht die Variabilität zwischen den einzelnen SMPs erklären.

Hierin begründen sich auch die Probleme des EBLUPA, mit welchem versucht wird, die Area-Schätzwerte aufgrund eines synthetischen Schätzers zu ermitteln, dabei aber korrigiert um beobachtete Abweichungen von diesem Modell. Hierbei fällt auf, dass der EBLUPA zwar in einigen Areas erwartungsgemäß sehr niedrige RRMSEs ausweist, jedoch in zahlreichen SMPs aufgrund der nicht passenden Modellierung schlechtere Ergebnisse als der GREG liefert.

Aus diesen einfachen Resultaten lässt sich unmittelbar das Potential der Small Area Modelle erkennen. Allerdings bergen sie auch die Gefahr, erhebliche Effizienzverluste zu verursachen, jedenfalls bei unsachgemäßem Einsatz, etwa wenn keine guten Hilfsinformationen verwendet werden. Gerade bei sehr kleinen (Teil-)Stichprobenumfängen wird man jedoch nicht mehr auf die Small Area Modelle verzichten können.

In zahlreichen Simulationen kristallisiert sich heraus, dass je nach Fragestellung entweder der GREG-Schätzer oder der EBLUPA verwendet werden soll. Mittlerweile steht auch eine erweiterte Version des EBLUPA zur Verfügung, welche bei der Small Area-Schätzung Gewichte verwendet. Diese Methode erweist sich beim derzeit umgesetzten Stichprobendesign bisher als besser geeignet als der EBLUPA.

5 Schlussfolgerungen

Der vorliegende Artikel gibt einen Überblick über die Herausforderungen und erste Ergebnisse innerhalb des Projektes *Zensus 2011 – Projekt zur methodischen Grundlagenforschung* und die gewählte Herangehensweise, um diesen Herausforderungen zu begegnen. Bis Ende des Projektes werden noch einige Erweiterungen der Schätzmethodik ausgebaut und evaluiert. Unter anderem muss die MSE-Schätzung von Small Area Methoden noch verfeinert werden.

Das vorgestellte Stichprobendesign verwendet eine neue Variante einer optimalen Allokation unter Nebenbedingungen. Hierdurch werden verschiedene

Zielsetzungen realisiert. Auf der einen Seite können Mindeststichprobenumfänge formuliert werden, auf der anderen Seite kann die Variabilität von Designgewichten eingeschränkt werden, wodurch der statistische Modellbau erleichtert wird, der auch bei der Verwendung von Small Area Modellen benötigt wird. Dies kommt auch dem inhaltlich interessierten Forscher zugute, der später eigene Analysen mit den Zensusdaten durchführt, sofern Scientific Use Files zur Verfügung gestellt werden. Ganz nebenbei verhindern die Box-Constraints auch eine allzu ungleiche Befragungswahrscheinlichkeit der Bevölkerung.

Nach bisherigen Ergebnissen werden unter Verwendung der Vorgaben vom Auftraggeber die Qualitätsziele für das Ziel 1 erreicht. Eine Verfeinerung der Methoden wird eventuell auch noch einzelne kleinere Verbesserungen ermöglichen. Im Falle von Ziel 2 gestaltet sich die Beurteilung erheblich problematischer, da effiziente Hilfsvariablen kaum zur Verfügung stehen. Ebenso variieren die Genauigkeiten der Schätzergebnisse vielmehr auf Grund der Tatsache, dass in den Sampling Points und Schichten bei unterschiedlichen Variablen die Beobachtungsanteile viel stärker variieren und innerhalb der Schichten heterogene Subgruppen auftreten. Insbesondere bei modellbasierten Schätzverfahren können nur geeignete Modelle diese Strukturen geeignet kompensieren. Der überwiegende Teil der Randschätzungen von Ziel 2-Variablen wird aber voraussichtlich die Präzisionsziele erfüllen, sofern nicht zu geringe Anteile einzelner Ausprägungen auftreten.

Literatur

Burgard, J. P., 2009: Erstellung von Karteileichen und Fehlbestandsmodellen durch Multilevel Modelle. Diplomarbeit an der Universität Trier.

Burgard, P. und R. Münnich, 2010: On the Impact of Over and Undercounts on Small Area Estimates in Register-based Censuses. Computational Statistics and Data Analysis, http://dx.doi.org/10.1016/j.csda.2010.11.002.

Devroye, L., 1986: Non-Uniform Random Variate Generation. Springer.

Gabler, S., M. Ganninger und R. Münnich, 2010: Optimal Allocation of the Sample Size to Strata Under Box Constraints. Metrika, DOI: 10.1007/s0018401003193.

Gelman, A., 2007: Struggles With Survey Weighting and Regression Modeling. Statistical Science 22 (2): 153–164.

Kleber, B., A. Maldonado, D. Scheuregger und K. Ziprik, 2009: Aufbau des Anschriften und Gebäuderegisters für den Zensus 2011. Wirtschaft und Statistik 7: 629-640.

Münnich, R., 2008: Varianzschätzung in komplexen Erhebungen. Austrian Journal of Statistics 37: 319–334.

Münnich, R., P. Burgard und M. Vogt, 2009: Small Area Estimation for Population Counts in the German Census 2011. Section on Survey Research Methods JSM 2009.

Münnich, R., S. Gabler und M. Ganninger, 2007: Some Remarks on the Registerbased Census 2010/2011 in Germany. Proceedings of the Workshop Innovative Methodologies for Censuses in the New Millennium Southampton.

Münnich, R. und J. Schürle, J., 2003: On the Simulation of Complex Universes in the Case of Applying the German Microcensus. Technical report, DACSEIS research paper series 4.

Münnich, R. und M. Vogt, 2011: Small Area Methoden: Modelle, Anwendungen und Praxis. Wirtschafts- und Sozialstatistisches Archiv. In Fertigstellung.
Neyman, J., 1934: On the two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society 97: 558–606.
Rao, J. N. K., 2003: Small Area Estimation. Wiley Series in Survey Methodology. New York: John Wiley and Sons.
Schäfer, J., 2004: Ergänzende Verfahren für einen künftigen registergestützten Zensus. Statistische Analysen und Studien NRW, 17: 20–27.
Tschuprov, A. A., 1923: On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations. Metron, 2: 461–493, 646–683.
Wolter, K. M., 2007: Introduction to Variance Estimation. New York: Springer, 2 edition.

Anschrift der Autoren

Prof. Dr. Ralf Münnich
Fachbereich IV – VWL
Lehrstuhl für Wirtschafts- und Sozialstatistik
Universitätsring 15
54286 Trier
muennich@uni-trier.de

PD Dr. Siegfried Gabler
GESIS – Leibniz-Institut für Sozialwissenschaften
B 2, 1
68159 Mannheim
siegfried.gabler@gesis.org

Dr. Matthias Ganninger
GESIS – Leibniz-Institut für Sozialwissenschaften
B 2, 1
68159 Mannheim
matthias.ganninger@gesis.org

Jan Pablo Burgard
Fachbereich IV – VWL
Lehrstuhl für Wirtschafts- und Sozialstatistik
Universitätsring 15
54286 Trier
jpburgard@uni-trier.de

Jan-Philipp Kolb
Fachbereich IV – VWL
Lehrstuhl für Wirtschafts- und Sozialstatistik
Universitätsring 15
54286 Trier
kolb@uni-trier.de

Konstruktion und Validierung eines allgemeinen Index für die Arbeitsbelastung in beruflichen Tätigkeiten anhand von ISCO-88 und KldB-92

Construction and Validation of a General Index for Job Demands in Occupations Based on ISCO-88 and KldB-92

Lars Eric Kroll

Zusammenfassung

In diesem Beitrag werden zusammenfassende Skalen zur allgemeinen, physischen und psychosozialen Arbeitsbelastung (AB_{ges}, AB_{phy}, AB_{psy}) auf Basis der Erwerbstätigenbefragung 2006 für die Berufsklassifikationen KldB-92 und ISCO-88 entwickelt und validiert. Ziel ist es, eine leicht anzuwendende Kontrollvariable für Studien bereitzustellen, in denen keine umfangreichen Instrumente zur Messung von Arbeitsbelastungen eingesetzt werden können. Einleitend wird ein Überblick über verschiedene Formen von Arbeitsbelastungen sowie deren Messung gegeben. Anschließend werden die drei Skalen anhand hierarchischer Regressionsmodelle in einem dreistufigen Verfahren entwickelt. Der Index wird zuletzt auf Basis der Daten der BIBB/BAuA-Erwerbstätigenbefragung 2006 und des Telefonischen Gesundheitssurveys „Gesundheit in Deutschland Aktuell" (GEDA) 2009 des Robert Koch-Instituts anhand von

Abstract

This paper describes the construction and validation of comprehensive scales for overall, physical and psycho-social job demands (AB_{ges}, AB_{phy}, AB_{psy}) that were constructed using a large-scale representative survey from 2006 conducted by the German Federal Institute for Vocational Education and Training (BIBB). The overall goal is to provide comprehensive scales that can be applied in studies that are not able to measure job demands more thoroughly. The scales are based on standard occupational classifications. They were constructed using multilevel regression models in a three-stage procedure. The resulting index has been validated using seven different health indicators with the data of the German BIBB/BAuA-workforce survey and additional data of the GEDA: German Telephone Health Survey 2008/2009. Results indicated significant associations with health outcomes such as

Gesundheitsindikatoren intern und extern validiert. Insgesamt erweist sich der Index bei der Analyse von gesundheitlichen Beeinträchtigungen, von wahrgenommenen Gesundheitsrisiken am Arbeitsplatz und bei der Analyse von krankheitsbedingten Fehlzeiten auf Basis beider Datensätze als aussagekräftig. Er wird daher für Forschungszwecke bereitgestellt und lässt sich anhand der Schlüsselvariablen KldB-92 oder ISCO-88 beliebigen Datensätzen zuspielen.

self-perceived health, perceived health risks at work or sick absence days. The index is free to use for scientific research and can be matched to any data source with data on occupations classified by KldB-92 or ISCO-88.

1 Hintergrund

Die Teilnahme am Erwerbsleben stellt für einen Großteil der Bevölkerung die Basis zur Sicherung des eigenen Lebensunterhalts dar. Sie vermittelt den Zugriff auf wichtige Ressourcen, wie Einkommen aber auch soziales Kapital oder Prestige. Allerdings bringt die Erwerbstätigkeit auch Einschränkungen mit sich, neben zeitlichen Restriktionen stehen dabei insbesondere Arbeitsbelastungen im Fokus, denen die Erwerbstätigen am Arbeitsplatz ausgesetzt sind und die ihre Gesundheit potentiell schädigen können (Babitsch et al. 2006; Peter 2006; Siegrist 1996, RKI/LGA Brandenburg 2002; Griefahn 1996; Schlick et al. 2010).

Arbeitsbelastungen sind definiert als Bedingungen mit potentiellen physiologischen und/oder psychologischen Auswirkungen auf den menschlichen Organismus, die sich aus den Merkmalen der Tätigkeit selbst oder aus ihren äußeren Bedingungen ergeben (Griefahn 1996; Schlick et al. 2010). Beispiele für körperliche Arbeitsbelastungen sind ergonomische Belastungen des Muskel-Skelett-Systems durch anstrengende oder einseitig belastende Tätigkeiten, Unfallgefahren bei der Arbeit oder der Kontakt mit gesundheitsschädlichen Substanzen bei der Arbeit (Griefahn 1996; Schlick et al. 2010). Neben diesen Gesundheitsrisiken, mit denen häufig direkte physische Schädigungen einhergehen, gibt es am Arbeitsplatz allerdings auch weniger offensichtliche Gesundheitsrisiken, die sich auf die Organisation des Arbeitsprozesses oder die soziale Dynamik am Arbeitsplatz zurückführen lassen und gesundheitsschädlichen Stress erzeugen können (Karasek/Theorell 1990; North et al. 1996; Siegrist 1996; Peter 2006). Beispiele für solche psychosozialen Arbeitsbelastungen sind etwa die Unsicherheit des Arbeitsplatzes aufgrund befristeter Beschäftigungsverhältnisse, starker Termin- und Leistungsdruck bei der Arbeit, soziale Konflikte zwischen den Beschäftigten oder mit den Vorgesetzten und auch unangemessene Belastungs- und Belohnungskonstellationen.

Gesundheitsgefährdenden Arbeitsbelastungen wird sowohl vom Gesetzgeber als auch von den Arbeitnehmervertretern, Betrieben und Sozialversicherungsträgern eine große Bedeutung beigemessen. Dies führt dazu, dass die Arbeitsschutzregelungen in Deutschland – auch im internationalen Vergleich – vergleichsweise gut ausgebaut sind (Kaufmann 2003). Im europäischen Vergleich ist der Anteil von Erwerbstätigen, die ihre Gesundheit durch ihre eigene Arbeit gefährdet sehen, nach Ergebnissen des European Working Conditions Survey in Deutschland so niedrig wie in keinem anderen Land (Parent-Thirion et al. 2007). Dessen ungeachtet muss auch für Deutschland konstatiert werden, dass weiterhin ein bedeutender Teil des gesundheitlichen Versorgungsbedarfs der Erwerbstätigen sowie auch der Fehlzeiten am Arbeitsplatz auf Fehlbelastungen bei der Arbeit zurückzuführen sind, was erhebliche volkswirtschaftliche Folgekosten verursacht (RKI 2007; RKI 2006; Bödeker et al. 2006).

Zur Beschreibung der Verbindung zwischen Arbeitsbelastungen und der Gesundheit bzw. dem Wohlbefinden der Arbeitenden hat sich in den Arbeitswissenschaften, der Arbeitsmedizin und der Arbeitspsychologie in Deutschland das Belastungs-Beanspruchungs-Konzept durchgesetzt (Rohmert 1984; Griefahn 1996; Schlick et al. 2010). Demnach wirken sich Belastungen nicht zwangsläufig negativ auf den Organismus aus, sondern entfalten in Abhängigkeit von der individuellen Leistungsfähigkeit unterschiedliche Wirkungen. Die Leistungsfähigkeit variiert einerseits zwischen den Arbeitnehmerinnen und Arbeitnehmern und andererseits auch mit der Art der Belastung (bspw. ionisierende vs. nichtionisierende Strahlung). Sie kann dabei durch Arbeitsschutzmaßnahmen gesteigert werden (bspw. Schutzkleidung oder Training), um das Risiko von gesundheitsschädigenden Beanspruchungen zu verringern. Je stärker die Belastung im Verhältnis zur individuellen Leistungsfähigkeit ist, desto höher ist die Beanspruchung der Arbeitnehmer und damit auch das Risiko von Gesundheitsstörungen und Krankheiten (Schlick et al. 2010).

Auf der Individualebene gibt es aufbauend auf arbeitswissenschaftlichen, arbeitspsychologischen und arbeitsmedizinischen Erkenntnissen eine Reihe von Instrumenten, anhand derer sich die Arbeitsbelastung von Erwerbstätigen ermitteln lässt. So stehen umfangreiche Kataloge physischer Arbeitsbelastungen zur Verfügung, deren körperliche Folgen erforscht sind (Parent-Thirion et al. 2007; Hall 2009; Schlick et al. 2010). Für psychosoziale Belastungen stehen ebenfalls verschiedene validierte Skalen bereit, die als Indikatoren für das Auftreten von Gesundheitsproblemen dienen können (Fields 2002). Mit dem Anforderungs-Kontroll-Modell und dem Modell beruflicher Gratifikationskrisen gibt es hier zudem zwei besonders häufig verwendete Konzepte mit umfassend validierten Erhebungsinstrumenten (Siegrist 1996; Karasek/Theorell 1990). Forschungspraktisch haben

die vorhandenen Skalen jedoch den Nachteil, dass sie umfangreiche Itembatterien enthalten und so in vielen wissenschaftlichen Studien nicht oder nur unzureichend eingesetzt werden können. Als Alternative zur individuellen Abfrage von Arbeitsbelastungen gibt es die Möglichkeit der Bildung von Tätigkeitsprofilen im Zuge von sog. Job-Exposure Matrizen (JEM), die beruflichen Tätigkeiten zugeordnet werden können (vgl. zu dieser Methode u. a. Goldberg et al. 1993). Dabei werden auf Basis von umfangreichen repräsentativen Befragungen und ggf. durch die Nutzung von verfügbaren Sekundärdaten Belastungsprofile für berufliche Tätigkeiten ermittelt. JEMs, die in den letzten Jahren in Deutschland entwickelt wurden, haben dabei berufliche Tätigkeiten hinsichtlich verschiedener Kriterien, wie beispielsweise psychosoziale Belastungen oder Wissensanforderungen an die Erwerbstätigen, klassifiziert (Bödeker 2002; Pollmann-Schult/Büchel 2002; Friedel 2003; Dragano 2007; Tiemann 2010). Die meisten Studien berechnen dazu Mittelwerte hinsichtlich ihrer jeweiligen Konstrukte (etwa Verhältnis von Verausgabung und Belohnung) für Kategorien beruflicher Tätigkeiten nach der Klassifikation der Berufe 1992 (KldB-92). Die resultierenden JEM können auch auf andere Studien angewendet werden. Bisher stehen allerdings nur für wenige Berufsgruppen umfassende JEM auf Basis von Experteneinschätzungen zur Verfügung, während JEM, die sich auf alle Tätigkeiten anwenden lassen, nur ausgewählte Arten von Belastungen erfassen. Ein allgemeiner Belastungsindex, der berufliche Tätigkeiten in eine Rangfolge hinsichtlich der damit verbundenen körperlichen und psychischen Arbeitsbelastungen bringt und bestehenden Datensätzen als Kontrollvariable zugespielt werden kann, steht bisher nicht zur Verfügung.

In diese Studie wird ein Index zur allgemeinen, physischen und psychosozialen Arbeitsbelastung entwickelt und validiert.[1] Der Index soll über die Klassifikation der Berufe des Statistischen Bundesamtes von 1992 (KldB-92) sowie die Klassifikation der Berufe der Internationalen Arbeitsorganisation (ILO) von 1988 (ISCO-88) bestehenden Datensätzen zugespielt werden können (vgl. zu beiden Klassifikationen Geis/Hoffmeyer-Zlotnik 2001).[2] Ziel ist es, eine einfach anzuwendende Kontrollvariable bereitzustellen, die in Studien, in denen keine umfassende Belastungsmes-

1 Die Skalen sind über die Webseiten der MDA verfügbar (http://www.gesis.org/publikationen/zeitschriften/mda/jg-5-2011-heft-1/). Der Autor bedankt sich bei Dr. Eckard Bergmann und Dr. Thomas Lampert sowie bei zwei anonymen Gutachtern für ihre Hinweise zu einer früheren Fassung des Manuskriptes.

2 Beide Skalen werden derzeit aktualisiert und an neue Entwicklungen auf dem Arbeitsmarkt angepasst (Arbeitsgruppe KldB 2010 2008; ILO 2008). Die überarbeiteten Instrumente waren aber für die Erwerbstätigenbefragung 2006 noch nicht verfügbar. Eine Aktualisierung der Skalen wird vom Autor angestrebt, sobald eine neue Erwerbstätigenbefragung verfügbar ist, welche die beiden neuen Schlüsselvariablen enthält.

sung möglich ist, die Berücksichtigung von Belastungen durch die Erwerbstätigkeit ermöglicht. Auf Basis der Literaturrecherche wurden fünf relevante Dimensionen identifiziert, die gleichbedeutend in den vorgeschlagenen Index eingehen sollen: (1) ergonomische Belastungen bei der Arbeitsausführung (etwa durch das Bewegen schwerer Lasten oder einseitige Bewegungsabläufe), (2) Belastungen durch die Arbeitsumgebung (wie Gifte, Gase, klimatische Belastungen), (3) psychische Belastungen am Arbeitsplatz (wie Überforderung, Unterforderung, geringe Fehlertoleranz bei der Arbeitsausführung), (4) zeitliche Belastungen (Termindruck, Schichtarbeit, übermäßig lange Arbeitszeiten) und (5) soziale Belastungen am Arbeitsplatz (Konflikte mit Kollegen oder Vorgesetzten, fehlende Kontrollmöglichkeiten).

Zur Konstruktion des Index werden die Daten der BIBB/BAuA-Erwerbstätigenbefragung 2005/2006 verwendet. In dieser Studie stehen insgesamt 39 Indikatoren zur Verfügung, die sich den fünf Belastungsdimensionen zuordnen lassen. Der Index wird anschließend auf Basis von Gesundheitsindikatoren aus der Erwerbstätigenbefragung und anhand von ausgewählten Indikatoren aus dem repräsentativen Telefonischen Gesundheitssurvey GEDA 2009 des Robert Koch-Instituts validiert (zu den Datensätzen vgl. Hall 2009; Kurth et al. 2009; RKI 2010). Bei der Validierung des Indexes wird untersucht, ob er mit der gesundheitlichen Lage der Erwerbstätigen assoziiert ist.

2 Daten und Methoden

Die BIBB/BAuA-Erwerbstätigenbefragung 2005/2006 ist eine Repräsentativbefragung von 20.000 Erwerbstätigen in Deutschland, die gemeinsam vom Bundesinstitut für Berufsbildung (BIBB) und der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA) durchgeführt wird (Hartmann 2006a). Die Daten wurden von TNS Infratest Sozialforschung in München durch computerunterstützte, telefonische Interviews zwischen Oktober 2005 bis März 2006 erhoben. Thematisch im Zentrum der Befragung stehen Fragen zum Arbeitsplatz (u. a. Tätigkeitsschwerpunkte, Arbeitsanforderungen, Arbeitsbedingungen und Arbeitsbelastungen) sowie zur Bildung und zur Ausbildung der Erwerbstätigen. Die Grundgesamtheit der Untersuchung bilden erwerbstätige Personen im Alter ab 15 Jahren, die regelmäßig mindestens zehn Stunden pro Woche gegen Bezahlung arbeiten („Kernerwerbstätige").

Gesundheitsrelevante Tätigkeitsmerkmale, die auf die Belastung und Beanspruchung der Erwerbstätigen schließen lassen, werden an verschiedenen Stellen des Fragebogens erhoben. Als Indikatoren für berufliche Belastung wurden insgesamt 39 Items aus fünf verschiedenen Fragekomplexen (F21; F22; F411; F600;

F700) herangezogen. Die Items wurden zumeist mit einer Frequenzskala zur Häufigkeit der jeweiligen Belastung beantwortet (Antwortvorgaben „häufig", „manchmal", „selten" und „nie").[3] Im Anschluss an die Frequenzskala wurden Teilnehmer, die nicht „nie" geantwortet haben gefragt, ob sie sich durch die jeweilige Belastung auch subjektiv belastet fühlen (Antwortvorgaben „Ja", „Nein"). Im Einklang mit dem Belastungs- und Beanspruchungskonzept (Rohmert 1984) wird im Folgenden davon ausgegangen, dass entsprechende Einschätzungen weniger mit arbeitsplatzbezogenen Faktoren, sondern vor allem mit der individuell unterschiedlichen Leistungsfähigkeit der Erwerbstätigen zusammenhängen. Bei der Skalenbildung werden die entsprechenden Merkmale daher nicht berücksichtigt.

Tabelle 1 Stichprobenbeschreibung der Erwerbstätigenbefragung 2006

Variable	Kategorien	Fallzahl	Stichprobe	Grundgesamtheit
Geschlecht	Männer	10209	51,5%	56,1%
	Frauen	9614	48,5%	43,9%
Alter	15-34 Jahre	5432	27,4%	27,8%
	35-49 Jahre	9832	49,6%	46,7%
	50-65 Jahre	4559	23,0%	25,5%
Berufliche Tätigkeit	Arbeitszeit in Std. pro Woche	19823	39,1	39,0
	Beschäftigt in jetziger Tätigkeit in Jahren	17444	8,3	8,7
	KldB-92 liegt vor	19738	99,6%	99,6%
	ISCO 1998 liegt vor	19741	99,6%	99,6%
Bildung	niedrig	4634	23,5%	34,4%
	mittel	9623	48,8%	43,6%
	hoch	5462	27,7%	22,0%
Anzahl Arbeitsbelastungen Mittelwert (SD)	ergonomisch	19812	1,3 (1,3)	1,4 (1,3)
	psychisch	19666	3,6 (2,2)	3,4 (2,2)
	sozial	19274	0,8 (1,1)	0,8 (1,1)
	umgebungsbezogen	19800	1,2 (1,7)	1,4 (1,8)
	zeitlich	19782	2,0 (1,6)	2,0 (1,6)

SD: Standardabweichung vom arithmetischen Mittelwert. Bildung: Die Bildung der Befragten wurde anhand der CASMIN Klassifikation operationalisiert (vgl. Brauns et al. 2003).
Datenbasis: Erwerbstätigenbefragung 2006, Alter 15 bis 65 Jahre.

3 Liste der einbezogenen Variablen laut der Variablenliste der Erwerbstätigenbefragung 2006: *Ergonomische Belastungen* (F600_01; *F600_02; F600_03; F600_07), *Psychische Belastungen* (F411_01; F411_04–F411_09; F411_11–F411_13), *Soziale Belastungen* (*F700_02; *F700_03; *F700_06; *F700_07; F700_08; F700_09; *F700_10- F700_13) *Umgebungsbezogene Belastungen* (F600_04-F600_06;F600_08-F600_12; F600_14) *Zeitliche Belastungen* (F216; F218; F221; F224; F210; az). Legende: * umgekehrte Kodierung des Items (nie vs. häufig, manchmal, selten).

Insgesamt wurden in der Studie 62.253 Personen befragt, 20.000 davon waren Erwerbstätige, die das gesamte Frageprogramm durchlaufen haben, die übrigen Befragten haben lediglich einen Kurzfragebogen beantwortet. Nachfolgend werden nur Erwerbstätige im Alter zwischen 15 und 65 Jahren berücksichtigt, die den ganzen Fragebogen ausgefüllt haben (n=19.823, Tabelle 1). Ein vollständiges Interview benötigte im Durchschnitt 40 Minuten. Die Ausschöpfung der Bruttostichprobe betrug nach Berücksichtigung qualitätsneutraler Ausfälle 44 %. Im Verlauf der Untersuchung wurde eine Quotierung der Befragten nach beruflicher Stellung eingeführt, da Arbeiter in der Nettostichprobe im Verhältnis zu ihrem Bevölkerungsanteil deutlich unterrepräsentiert sind. Dies wurde durch eine Designgewichtung berücksichtigt (Hartmann 2006b). Zusätzlich wurde eine Anpassungsgewichtung nach Alter, Geschlecht, Bildung und beruflicher Stellung auf Basis des Mikrozensus 2005 durchgeführt. Um die Erwerbsstruktur Deutschlands repräsentativ abzubilden, wird in den nachfolgenden Analysen in der Regel dieser Hochrechnungsfaktor verwendet, sofern nichts anderes vermerkt ist.

Die beruflichen Tätigkeiten der Befragten liegen im Datensatz bereits vierstellig kodiert nach KldB-92 und ISCO-88 vor. Die KldB-92 beschreibt die Berufe in Deutschland besonders differenziert, während sich die Klassifikation ISCO-88 besonders für den internationalen Vergleich eignet. Beide Berufsklassifikationen sind hierarchisch aufgebaut, je mehr Stellen berücksichtigt werden, desto enger umrissen sind die jeweiligen Tätigkeiten (Geis/Hoffmeyer-Zlotnik 2001). Die ISCO-88 Klassifikation ermöglicht eine Differenzierung von Berufsbereichen über einstellige Codes. In der Klassifikation der Berufe KldB-92 erfolgt die Gruppierung der Berufsbereiche durch das Zusammenfassen mehrerer zweistelliger Codes. In Tabelle 2 wird die Struktur der beiden Klassifikationen am Beispiel der jeweiligen Zuordnung von „Bankkaufleuten ohne nähere Angabe" veranschaulicht.

In den nachfolgenden Analysen werden zuerst die Skalen zur Arbeitsbelastung nach beruflichen Tätigkeiten bestimmt. Anschließend folgt eine Validierung der Skalen anhand verschiedener Gesundheitsindikatoren. Dazu werden zuerst die Daten der Erwerbstätigenbefragung 2006 verwendet, da in dieser auch eine umfangreiche Liste von gesundheitlichen Beschwerden erhoben wurde, die in Zusammenhang mit der Tätigkeit der Befragten auftreten. Um zu überprüfen, ob sich die vorgefundenen Zusammenhänge auch in anderen Datensätzen zeigen, wird auch eine externe Validierung der Skala anhand von Gesundheitsindikatoren aus dem Telefonischen Gesundheitssurvey GEDA 2009 vorgenommen. Der Index wird den Befragten in der GEDA-Studie dazu auf Basis des Berufsschlüssels KldB-92 zugespielt. Dieses Vorgehen bei der Anwendung von JEM wurde auch schon von anderen Autoren angewendet (vgl. etwa Dragano 2007; Bödeker 2002). Für die Datenanalysen wird das Statistikprogramm Stata SE in der Version 11.0 verwendet.

Tabelle 2 Struktur von KldB-92 und ISCO-88 am Beispiel der Zuordnung von Bankkaufleuten ohne nähere Angabe

Ebene	Stelle	Bezeichnung	Einordnung
Klassifikation der Berufe von 1992 (KldB-92)			
1	1-2 (kat.)	Berufsbereiche (n=6)	*V Dienstleistungsberufe* (**66-93**)
2	1-2 (kat.)	Berufsabschnitte (n=33)	*Vb Dienstleistungskaufleute* (**69-70**)
3	1-2	Berufsgruppen (n=88)	**69** *Bank-, Bausparkassen- und Versicherungsfachleute*
4	1-3	Berufsordnungen (n=369)	**691** *Bankfachleute*
5	1-4	Berufsklassen (n=2287)	**6910** *Bank-, Sparkassenfachleute, allgemein*
International Classification of Occupations von 1988 (ISCO-88)			
1	1	Hauptgruppe (n=9)	**4** *Bürokräfte, kaufmännische Angestellte*
2	1-2	Untergruppe (n=28)	**41** *Büroangestellte ohne Kundenkontakt*
3	1-3	Gattung (n=116)	**412** *Angestellte im Rechnungs- Statistik- und Finanzwesen*
4	1-4	Unit Groups (n=390)	**4121** *Statistik- und Finanzangestellte*

Quelle: Geis/Hoffmeyer-Zlotnik (2001), Zuordnung des Beispiels Bankkaufleute laut Erwerbstätigenbefragung 2006.

3 Vorgehen bei der Berechnung der Skalen zur allgemeinen und bereichsspezifischen Arbeitsbelastung

Die Indizes zur allgemeinen und zur physischen sowie zur psychosozialen Arbeitsbelastung werden in einem vierstufigen Verfahren berechnet:

1. Berechnung der individuellen Summenscores für die fünf Teildimensionen beruflicher Belastungen.
2. Zusammenfassung der Summenscores in einem Gesamtscore (allgemeine Arbeitsbelastung) und zwei Teilscores (physische und psychosoziale Belastung).
3. Berechnung der tätigkeitsbezogenen JEM für ISCO-88 und KldB-92 anhand von Mehrebenenmodellen.
4. Standardisierung der Scores auf den Variationsbereich 1 bis 10 zur Abbildung von Dezilen von Tätigkeiten nach beruflicher Belastung.

In den ersten beiden Schritten werden allgemeine, körperliche und psychosoziale Arbeitsbelastungen bei den Befragten der Erwerbstätigenbefragung ermittelt und über drei Scores abgebildet. Dazu werden die 39 Items mit Bezug zu Arbeitsbelastungen dichotomisiert und fünf Teildimensionen („Ergonomische Belastungen" (EB), „Umgebungsbelastungen" (UB), „Psychische Belastungen" (PB), „Soziale Belastungen" (SB) und „Zeitliche Belastungen" (ZB)) zugeordnet. Anschließend wird für jeden Befragten die Anzahl von Belastungen in den Teildimensionen berechnet. Die resultierenden Punktsummenscores werden z-standardisiert und zu einem

Gesamtscore (allgemeine Belastung AB_{ges} = EB+UB+PB+SB+ZB) und den zwei bereichsspezifischen Scores (körperliche Belastung AB_{phy} = EB+UB; psychosoziale Belastung AB_{psy} = PB+SB+ZB) aufsummiert. Durch die vorangegangene Standardisierung gehen die Teildimensionen dabei gleichbedeutend in den jeweiligen Index ein.

Im dritten und vierten Schritt werden die drei Scores den beruflichen Tätigkeiten der Befragten zugeordnet. Die Zuweisung von Arbeitsbelastungen zu Berufen (Job-Exposure Matrizen) erfolgt in dieser Studie – im Unterschied zu früheren Arbeiten (Tiemann 2010; Dragano 2007; Bödeker 2002) – nicht durch die Berechnung von tätigkeitsspezifischen Mittelwerten, sondern durch sog. hierarchische Regressionsmodelle bzw. Mehrebenenanalysen (vgl. Langer 2008; de Leeuw/Meijer 2008). Bei der Berechnung von berufsspezifischen Mittelwerten wird implizit vorausgesetzt, dass die beobachteten Werte allein auf die Eigenschaften der Tätigkeiten zurückzuführen sind und keine weiteren systematischen und für das abhängige Merkmal bedeutsamen Unterschiede zwischen den Personen bestehen, die diese Tätigkeiten ausführen. Diese Annahme ist fraglich, da die Anteile von männlichen und weiblichen Beschäftigten, Teilzeitbeschäftigungsquoten oder die mittlere Dauer in der aktuellen Tätigkeit deutlich zwischen den Tätigkeiten variieren. Diese Merkmale haben zwar einen Einfluss auf die Häufigkeit von Arbeitsbelastungen, dieser Einfluss liegt aber nicht in der Tätigkeit begründet. Das statistische Verfahren der Mehrebenenanalyse erlaubt es, den Einfluss intervenierender Merkmale zu kontrollieren (Formel 1). Es liefert dadurch für kleine Stichproben, wie etwa seltene berufliche Tätigkeiten, robustere Schätzwerte für Parameter als die einfache Berechnung von Mittelwerten.

$$Y_{ik} = b_{0k} + b_{1k} X_{1k} + e_i \qquad (1)$$

$Y_{i,k}$: Abhängiges Merkmal, das zwischen Individuen (i) und Kontexten (k) variiert

$b_{o,k}$: Konstante, die zwischen Kontexten variiert, sie lässt sich in einen fixen und einen variablen Anteil zerlegen ($b_{ok} = b_o + u_{o,k}$)

$b_{1,k}$: Effekt der erklärenden Variable X_1, die zwischen den Kontexten variiert, er lässt sich in einen fixen und einen variablen Anteil zerlegen ($b_{1k} = b_1 + u_{1,k}$)

e_i: Residuum des Individuums (i)

In der Mehrebenenanalyse werden zwei Formen von Modellen unterschieden, die sog. Random-Intercept- und die sog. Random-Effect-Modelle (Langer 2008; de Leeuw/Meijer 2008). Während in Random-Intercept-Modellen nur die Variation der Konstante über die Kontexte betrachtet wird, wird in Random-Effect-Modellen auch die Variation der Effekte der erklärenden Variablen als kontextabhängig modelliert. In dieser Studie werden nachfolgend nur Random-Intercept-Modelle

verwendet, da die Variation der Arbeitsbelastungen über die Kontexte der Berufe untersucht werden soll. Nicht betrachtet wird die Variation der Kontrollvariablen über die Kontexte, um die Modelle möglichst einfach zu halten. Im Vergleich zum Grundmodell wird allerdings eine Erweiterung vorgenommen, da nicht nur eine, sondern drei Ebenen sozialer Kontexte – die zwei-, drei- und vierstelligen Berufsklassifikationen – betrachtet werden (Formel 2).

$$\mathrm{Y}_{\mathrm{i},k_1,k_2,k_3} = \mathrm{b}_0 + u_{k_1} + u_{k_2} + u_{k_3} + \mathbf{bX} + \mathrm{e}_\mathrm{i} \tag{2}$$

u_{k1}: Abweichung der Tätigkeit laut 1. und 2. Stelle der Berufklassifikation von b_0

u_{k2}: Abweichung der Tätigkeit laut 3. Stelle der Berufklassifikation von b_0 und u_{k1}

u_{k3}: Abweichung der Tätigkeit laut 4. Stelle der Berufklassifikation von b_0, u_{k1}, u_{k2}

bX: Vektor von Kontrollvariablen (X) und zugehöriger Effektkoeffizienten (b) auf Individualebene (sog. ‚fixed-part' des Modells)

e_i: Residuum des Individuums (i)

Die individuelle Arbeitsbelastung ergibt sich demnach als Summe der geschätzten Parameter für die allgemeine Arbeitsbelastung in Deutschland (b_0), der berufsgruppenspezifischen Belastung auf Ebene der Zweisteller der Berufsklassifikation (u_{k1}), der bereichsspezifischen Belastung auf Ebene der Dreisteller (u_{k2}), der tätigkeitsspezifischen Belastung auf Ebene der Viersteller (u_{k3}) sowie des individuellen Fehlerterms (e_i).

Der Intraklassen-Korrelationskoeffizient ρ liefert Hinweise darüber, welcher Anteil der Variation des abhängigen Merkmals (berufliche Belastungen) auf welche Differenzierungsebene der Tätigkeitsschlüssel zurückzuführen ist (Formel 3).

$$\begin{aligned}
\rho_{k_1} &= \frac{{\sigma_{u_{k_1}}}^2}{{\sigma_{u_{k_1}}}^2 + {\sigma_{u_{k_2}}}^2 + {\sigma_{u_{k3}}}^2 + {\sigma_{e_i}}^2} \\
\rho_{k_2} &= \frac{{\sigma_{u_{k_2}}}^2}{{\sigma_{u_{k_1}}}^2 + {\sigma_{u_{k_2}}}^2 + {\sigma_{u_{k3}}}^2 + {\sigma_{e_k}}^2} \\
\rho_{k_2} &= \frac{{\sigma_{u_{k_3}}}^2}{{\sigma_{u_{k_1}}}^2 + {\sigma_{u_{k_2}}}^2 + {\sigma_{u_{k3}}}^2 + {\sigma_{e_i}}^2}
\end{aligned} \tag{3}$$

Je größer die Intraklassen-Korrelation auf einer Ebene der Berufsklassifikation ist, desto höher ist der Anteil der Gesamtvariation der Arbeitsbelastung zwischen den Individuen, der auf diese Ebene zurückzuführen ist. Aus der Summe der Intraklassen-Korrelationen der drei Ebenen ergibt sich folglich der im Modell durch die beruflichen Tätigkeiten erklärte Anteil der Variation der Arbeitsbelastung.

Die Job-Exposure Matrizen für die Verknüpfung von Arbeitsbelastung und beruflichen Tätigkeiten werden aus den vorhergesagten Werten der Regressionsmodelle berechnet. Bei der Vorhersage wird die Variation der Belastung zwischen den beruflichen Tätigkeiten (Formel 2: u_1, u_2, u_3), nicht aber der Einfluss der individuellen Merkmale der Befragten (Formel 2: $\mathbf{bX}+\mathbf{e}_i$), berücksichtigt (sog. Intercept-as-Outcome-Modell). Die resultieren JEM's abstrahieren dadurch von den individuellen Merkmalen der Befragten. Die vorhergesagten Werte aus dem Regressionsmodell werden zum Abschluss in Dezile eingeteilt. Dadurch resultiert für die drei Teildimensionen ein Score, der – je nach allgemeiner, körperlicher oder psychosozialer Belastung – Werte zwischen 1 (= die 10 % der Berufe mit der niedrigsten Arbeitsbelastung) und 10 (= die 10 % der Berufe mit der höchsten Arbeitsbelastung) annimmt.

4 Ergebnisse

4.1 Berechnung der Indizes

In Tabelle 3 ist die Zuordnung der 39 Items mit Bezug zu Arbeitsbelastungen aus der Erwerbstätigenbefragung 2006 zu den Teildimensionen „Ergonomische Belastungen" (EB), „Umgebungsbelastungen" (UB), „Psychische Belastungen" (PB), „Soziale Belastungen" (SB) und „Zeitliche Belastungen" (ZB) dargestellt. Die Antworten der Befragten lagen als Frequenzskalen vor und wurden dichotomisiert. Belastungen wurden mit „0" kodiert, wenn sie nicht „häufig" und mit „1" kodiert, wenn sie häufig vorkamen. Bei gesundheitsförderlichen Aspekten der Tätigkeit wurde diese Zuweisung invertiert (0: „häufig", „manchmal" oder „selten"; 1: „nie").[4] Durch dieses starke Kriterium sollten nur Belastungen berücksichtigt werden, die kennzeichnend für den Arbeitsplatz sind. Die Abgrenzung orientiert sich dabei an den Berichten zur Sicherheit und Gesundheit bei der Arbeit in Deutschland (BAuA 2010).

4 Vier Items, die nur eine sehr geringe Korrelation mit den übrigen Items in ihren jeweiligen Bereichen aufwiesen, wurden bei der Berechnung der Scores nicht berücksichtigt, um die interne Konsistenz der Skalen zu erhöhen. Ausgeschlossene Belastungen: *Soziale Belastungen* (F700_04: Emotionale Beanspruchung bei der Arbeit); *Umgebungsbelastungen* (F600_13: Umgang mit mikrobiologischen Stoffen), *Zeitliche Belastungen* (F201: Häufigkeit von Überstunden/Mehrarbeit; F208: Unzureichende Berücksichtigung familiärer/privater Interessen).

Tabelle 3 Einbezogene Items beruflicher Belastung und Entlastung nach Bereichen

Bereich	Verwendete Items
Ergonomische Belastungen (4 Items)	Häufig: Stehen
	Nicht häufig: Sitzen
	Häufig: Heben und tragen schwerer Lasten (Männer >=20 kg, Frauen >=10 kg)
	Häufig: Arbeiten in Zwangshaltungen (in gebückter, hockender, kniender oder liegender Stellung arbeiten oder Arbeiten über Kopf)
Psychische Belastungen (10 Items)	Häufig: Termin-Leistungsdruck
	Häufig: Vor neue Aufgaben gestellt werden
	Häufig: Verbessern von Verfahren
	Häufig: Bei der Arbeit gestört/unterbrochen werden
	Häufig: Mindestleistung erfüllen müssen
	Häufig: Dinge tun, die nicht gelernt
	Häufig: Verschiedenartige Arbeiten gleichzeitig ausführen
	Häufig: Kleine Fehler große Folgen
	Häufig: Bis an Grenze der Leistungsfähigkeit gehen müssen
	Häufig: Sehr schnell arbeiten müssen
Soziale Belastungen (10 Items)	Nie: Möglichkeit, Arbeit selbst zu organisieren
	Nie: Einfluss auf Arbeitsmenge
	Nie: Entscheidungsfreiheit bei Pauseneinteilung
	Nie: Gefühl, dass Arbeit wichtig
	Häufig: Nicht rechtzeitig über Entwicklungen im Betrieb informiert
	Häufig: Nicht rechtzeitig notwendige Informationen zur eigenen Tätigkeit erhalten
	Nie: Gemeinschaftsgefühl
	Nie: Gute Zusammenarbeit mit Kollegen
	Nie: Unterstützung durch Kollegen
	Nie: Unterstützung durch direkten Vorgesetzten
Umgebungs-belastungen (9 Items)	Häufig: Arbeiten bei Rauch, Staub oder unter Gasen, Dämpfen
	Häufig: Arbeiten bei Kälte, Hitze, Nässe, Feuchtigkeit, Zugluft
	Häufig: Öl, Fett, Schmutz, Dreck ausgesetzt
	Häufig: Starken Erschütterungen, Stößen, Schwingungen ausgesetzt
	Häufig: Grellem Licht oder schlechter Beleuchtung ausgesetzt
	Häufig: Arbeiten mit gefährlichen Stoffen, unter Einwirkung von Strahlung
	Häufig: Schutzkleidung oder Schutzausrüstung bei der Arbeit tragen
	Häufig: Arbeiten bei Lärm
	Häufig: An einem Platz, an dem geraucht wird arbeiten
Zeitliche Belastungen (6 Items)	Tatsächliche Arbeitszeit >=48 Stunden pro Woche
	Häufig: Bereitschaftsdienst/Rufbereitschaft
	Häufig: Samstagsarbeit
	Häufig: Sonntags-/Feiertagsarbeit
	Häufig: zwischen 23 und 5 Uhr
	Häufig: Schichtarbeit

Quelle: Erwerbstätigenbefragung 2006.

In Tabelle 4 ist für die resultierenden Subskalen ihre interne Konsistenz (gemessen mit Cronbachs α) sowie der Anteil fehlender Werte ausgewiesen. Insgesamt variiert Cronbachs α dabei zwischen 0,50 bis 0,70. Besonders hoch ist die interne Konsistenz der Skalen zur ergonomischen und umgebungsbezogenen Belastung, besonders gering ist die Konsistenz der Skala für soziale Belastungen. In der Subskala soziale Belastungen sind Items enthalten, die auf gruppendynamische Aspekte (Konflikte bzw. Zusammenarbeit mit Kollegen und den Vorgesetzten) aber auch auf organisatorische Prozesse (Informationsfluss im Betrieb, Einfluss auf die Arbeitsmenge) abzielen, die Heterogenität dieser Faktoren wird durch das geringe α der Skala abgebildet. Bei der Bildung des Belastungs-Scores soll trotzdem nicht auf diese Skala verzichtet werden, um keinen Informationsverlust zu erleiden. Die interne Validität des Gesamtscores (AB_{ges}) von 0,54 ist angesichts der großen Zahl von einbezogenen Items (39) akzeptabel, zudem ist auch der Anteil von Befragten mit fehlenden Werten (3,7 %) relativ gering.

Tabelle 4 Cronbach's α und gültige Werte der Subskalen und der Gesamtskala

Bereich	Subskalen	Cronbach's α	Gültige Werte %
Physische Belastungen	Ergonomische Belastungen (EB)	0,718	99,9
	Umgebungsbelastungen (UB)	0,730	99,9
	AB_{phy} = EB- und UB-Score	0,668	99,8
Psychosoziale Belastungen	Psychische Belastungen (PB)	0,658	99,2
	Soziale Belastungen (SB)	0,526	97,2
	Zeitliche Belastungen (ZB)	0,646	99,8
	AB_{psy} = PB-, SB- und ZB-Score	0,242	96,4
Gesamt	AB_{ges} = EB-, UB-, PB-, SB-, ZB-Score	0,540	96,3

Datenbasis: Erwerbstätigenbefragung 2006.

In einem zweiten Schritt wurde nun für die drei zusammenfassenden Scores ein lineares hierarchisches Regressionsmodell mit Random-Intercept (Mehrebenenmodell) berechnet, um die mittlere Arbeitsbelastung in beruflichen Tätigkeiten zu ermitteln (Tabelle 5). Die Ebenen des Modells ergeben sich dabei aus den Kategorien und Hierarchien der beiden Berufsklassifikationen. Bei Anwendung dieses Modells ist der Anteil fehlender Werte vergleichsweise hoch, da alle Personen ausgeschlossen wurden, von denen nicht bekannt ist, wie lange sie die derzeitige berufliche Tätigkeit bereits ausüben (12 %). Berücksichtigt man diese Einschränkung, dann liegt der Anteil fehlender Werte bei den übrigen Merkmalen unter 4 %. Die Verwendung der

Mehrebenenmethodik soll die möglichen Verzerrungen der Stichprobe hinsichtlich Alter, Geschlecht und Arbeitszeit der Befragten ausgleichen.

Tabelle 5 Lineares Mehrebenenmodell der Variation der Arbeitsbelastung über berufliche Tätigkeiten

	KldB-92			ISCO-88		
	AB_{ges}	AB_{phy}	AB_{psy}	AB_{ges}	AB_{phy}	AB_{psy}
Fixed-Effects						
Frauen	-0,003	-0,004	0,007	-0,011*	-0,026*	0,000
Männer	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Alter (z-standardisiert)	-0,024***	-0,046***	-0,036***	-0,024***	-0,050***	-0,036***
ln(Arbeitszeit in Std.)	0,161***	0,139***	0,381***	0,164***	0,136***	0,392***
ln(Jahre in Tätigkeit)	0,011***	0,030***	0,008	0,012***	0,037***	0,006
Konstante	-0,496***	-0,496***	-0,496***	-0,496***	-0,496***	-0,496***
Intra-Klassenkorrelation (Random Intercept Modell)						
ICC: Berufscode 1. und 2. Stelle	0,295	0,454	0,062	0,203	0,349	0,043
ICC: Berufscode 3. Stelle	0,075	0,081	0,047	0,079	0,084	0,045
ICC: Berufscode 4. Stelle	0,069	0,072	0,050	0,086	0,097	0,052
chi² Random-Effects	8306	12449	2055	7759	11630	1897
N	16773 (85%)	17409 (88%)	16793 (85%)	16704 (84%)	17338 (87%)	16724 (84%)
LL_0	-5995	-22262	-16317	-5971	-22183	-16244
LL_1	-1349	-15695	-14746	-1603	-16021	-14767

*AB_{ges}: Allgemeine Arbeitsbelastung; AB_{phy}: Körperliche Arbeitsbelastung; AB_{psy}: Psychosoziale Arbeitsbelastung; Signifikanz: * p<0,05; ** p<0,01; *** p<0,001; ICC: Intraklassen-Korrelation für die jeweilige Differenzierungsebene der Berufsklassifikation (2.-,3- oder 4-Steller); chi²: Chi²-Test Modell mit vs. Modell ohne Effekte auf Ebene der beruflichen Tätigkeiten; p: p-Wert des chi²-Tests; N: Gültige Fälle (Anteil an der Stichprobe); LL_0: Log Likelihood des Null-Modells nur mit Konstante ohne random-part; LL_1: Log Likelihood des Modells mit Konstante, erklärenden Variablen im fixed-part und random-part. Die dargestellten Modelle wurden ungewichtet berechnet, da alle Beobachtungen mit einer beruflichen Tätigkeit als Gleichbedeutung für das Ergebnis angesehen wurden. Datenbasis: Erwerbstätigenbefragung 2006.*

Insgesamt zeigt sich, dass die meisten Kontrollvariablen einen signifikanten Einfluss auf die Wahrnehmung der allgemeinen und bereichsspezifischen Arbeitsbelastungen haben. Die Ergebnisse deuten dabei auf eine Abnahme der physischen und psychosozialen Arbeitsbelastung mit dem Alter hin, unabhängig von der jeweiligen Tätigkeit. Eine höhere Wochenarbeitszeit ist dagegen auch mit einer höheren Arbeitsbelastung verbunden, dies gilt besonders für psychosoziale Belastungen. Signifikante Geschlechtsdifferenzen hinsichtlich der Belastungen zeigen sich dagegen innerhalb der Tätigkeiten nicht mehr. Die Arbeitsbelastungen variieren aber stark zwischen den beruflichen Tätigkeiten, wie die hohen Werte der Intraklassen-Korrelationen für beide Berufklassifikationen ausweisen, und zwar von 0,43 (KldB-92, allgemeine

Arbeitsbelastung, Summe der ICC über alle drei Ebenen) bzw. 0,37 (ISCO-88). Die Erklärungskraft der Klassifikation des Statistischen Bundesamtes ist dabei etwas größer als die Erklärungskraft der internationalen Klassifikation. Besonders hoch ist der Zusammenhang zwischen Arbeitsbelastung und Tätigkeit auf der obersten Differenzierungsebene der Berufsgruppen, die unteren beiden Ebenen erklären demgegenüber einen deutlich geringeren Teil der Variation der Arbeitsbelastung zwischen den Befragten. Im Vergleich von allgemeiner, physischer und psychosozialer Arbeitsbelastung zeigt sich, dass der Zusammenhang zwischen beruflichen Tätigkeiten und physischen Arbeitsbelastungen besonders ausgeprägt ist, während sich der Zusammenhang hinsichtlich der psychosozialen Arbeitsbelastungen als deutlich schwächer erweist. Sie eignen sich demnach weniger für das Verfahren der Job-Exposure Matrizen. Die Intraklassen-Korrelation des allgemeinen Scores kann insgesamt als zufriedenstellend erachtet werden. Aufgrund der besseren Anpassung der KldB-92 werden nachfolgend nur noch die Job-Exposure Matrizen auf Basis dieser Klassifikation berücksichtigt.

Tabelle 6 Vergebene Punktwerte für die Kategorien von ISCO-88 und KldB-92 nach Stellen

Klassifikation	Anzahl Kategorien	Anzahl in der EWT2006	Abdeckung durch den Belastungsindex %	Mittlere Anzahl von Beobachtungen pro Kategorie
ISCO-88 (2-Steller)	28	28	100,0	708,0
ISCO-88 (3-Steller)	116	110	94,8	180,2
ISCO-88 (4-Steller)	390	306	78,5	64,8
KldB-92 (2-Steller)	88	88	100,0	222,7
KldB-92 (3-Steller)	369	355	96,2	55,7
KldB-92 (4-Steller)	2287	1310	57,3	15,1

Datenbasis: Erwerbstätigenbefragung 2006.

In Tabelle 6 ist dargestellt, welchem Anteil der beruflichen Tätigkeiten, die in der KldB-92 und der ISCO-88 erfasst sind, ein Wert zur allgemeinen und bereichsspezifischen Arbeitsbelastung zugeordnet werden konnte und auf wie vielen Befragten die Zuordnung dabei im Mittel beruhte. Auf Basis der Ergebnisse wird deutlich, dass sich die Datenbasis in der Erwerbstätigenbefragung mit zunehmendem Grad der Differenzierung der Tätigkeiten verschlechtert. So können auf Basis der 2-Steller von ISCO-88 und KldB-92 noch alle Bereiche abgedeckt werden, wobei die Punktwerte jeweils auf einer mittleren Anzahl von Beobachtungen 708 (ISCO-88) bzw. 223 (KldB-92) basieren. Während auch auf der Ebene der dreistelligen Berufscodes noch etwa 95 % der möglichen Berufe erfasst werden, beträgt die Abdeckung bei

den vierstelligen Codes lediglich noch 79 % (ISCO-88) bzw. 57 % (KldB-92). Sollten in der Anwendung der Skalen Berufscodes vorkommen, denen auf der Ebene der Viersteller kein Wert zur Belastung zugeordnet ist, lässt sich – durch die hierarchische Konstruktion der Skalen – aber eine Zuordnung anhand der Drei- oder sogar Zweisteller vornehmen, da spätestens allen Zweistellern ein Skalenwert zugeordnet werden konnte.

Tabelle 7 Auszug aus der JEM: Berufliche Tätigkeiten nach KldB-92 (2-Steller) mit der niedrigsten und höchsten allgemeinen Arbeitsbelastung (AB_{ges})

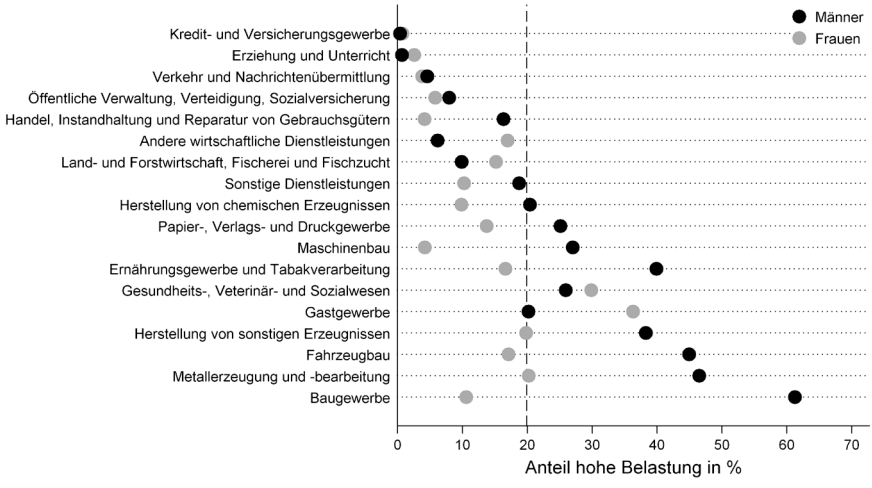
Code	Berufsgruppen nach KldB-92	Anteil	Arbeitsbelastung
78	Büroberufe, Kaufmännische Angestellte	9,2	1. Dezil
69	Bank-, Bausparkassen-, Versicherungsfachleute	2,8	1. Dezil
64	Technische Zeichner und verwandte Berufe	0,2	1. Dezil
	[...]	78,3	2.-9. Dezil
93	Reinigungs- und Entsorgungsberufe	2,2	10. Dezil
27	Maschinenbau- und -wartungsberufe	1,6	10. Dezil
25	Metall- und Anlagenbauberufe	1,1	10. Dezil
48	Ausbauberufe	0,9	10. Dezil
22	Berufe in der spanenden Metallverformung	0,9	10. Dezil
50	Berufe in der Holz- und Kunststoffverarbeitung	0,9	10. Dezil
51	Maler, Lackierer und verwandte Berufe	0,8	10. Dezil
46	Tiefbauberufe	0,5	10. Dezil
19	Berufe in der Hütten- und Halbzeugindustrie	0,2	10. Dezil
23	Berufe in der Metalloberflächenveredlung und Metallvergütung	0,1	10. Dezil
20	Gießereiberufe	0,1	10. Dezil
42	Berufe in der Getränke-, Genußmittelherstellung	0,1	10. Dezil
33	Spinnberufe	0,1	10. Dezil
7	Bergleute	0,1	10. Dezil
11	Baustoffhersteller	<0,1	10. Dezil

Tätigkeiten, die sich nicht im obersten oder untersten Dezil der allgemeinen Arbeitsbelastung befinden, werden hier – unabhängig von ihrer Lage hinsichtlich der bereichsspezifischen Arbeitsbelastungen – nicht aufgeführt. Datenbasis: Erwerbstätigenbefragung 2006.

In Tabelle 7 ist ein Auszug aus der Liste der Berufsgruppen laut KldB-92 (2-Steller) für das oberste und unterste Zehntel des Belastungsindexes abgebildet. Demnach finden sich die ca. 10 % am geringsten belasteten Beschäftigten in den Berufsgruppen „Technische Zeichner und verwandte Berufe", „Bank-, Bausparkassen-, Versicherungsfachleute", sowie „Büroberufe, Kaufmännische Angestellte". Besonders stark belastet sind dagegen etwa Bergleute und Reinigungs- und Entsor-

gungsberufe. Die vollständigen Listen und Zuordnungsschlüssel können vom Autor bezogen werden (vgl. Anmerkung 1).[5]

Abbildung 1 Anteil von hoch belastenden Tätigkeiten (AB_{ges}) nach Branche und Geschlecht



Für deskriptive Darstellungen wurde die Ordinalskala aus den zehn Indexwerten (1-10), in denen sich jeweils etwa 10 % der Erwerbstätigen des Jahres 2006 befinden, in drei Gruppen unterteilt: Als „niedrig" belastet wird das untere Fünftel der Berufe (Indexwerte: 1-2), als „mittel" belastet die mittleren drei Fünftel (Indexwerte 3-8) und als „hoch" belastet das obere Fünftel der Tätigkeiten bezeichnet. Legt man diese Kategorien zugrunde, liegt der Anteil hoher Belastungen definitionsgemäß bei ca. 20 % der Berufe. Im Branchenvergleich variiert der Anteil von „hoch" belastenden Tätigkeiten allerdings deutlich. Bei Männern im Baugewerbe (61 %) und bei Frauen im Gastgewerbe (36 %) sowie im Gesundheits-, Veterinär- und Sozialwesen ist er besonders hoch (vgl. Abbildung 1).

5 Während sich bei der Besetzung der Dezile auf den tieferen Ebenen von KldB-92 und ISCO-88 nur geringe Abweichungen vom Zielwert 10 % zeigen (Spannweite der Abweichungen 1 %), ist die Abweichung auf der höheren Ebene der 2-Steller durch die ungleiche Besetzung der Gruppen deutlich größer (Spannweite der Abweichungen 4 %).

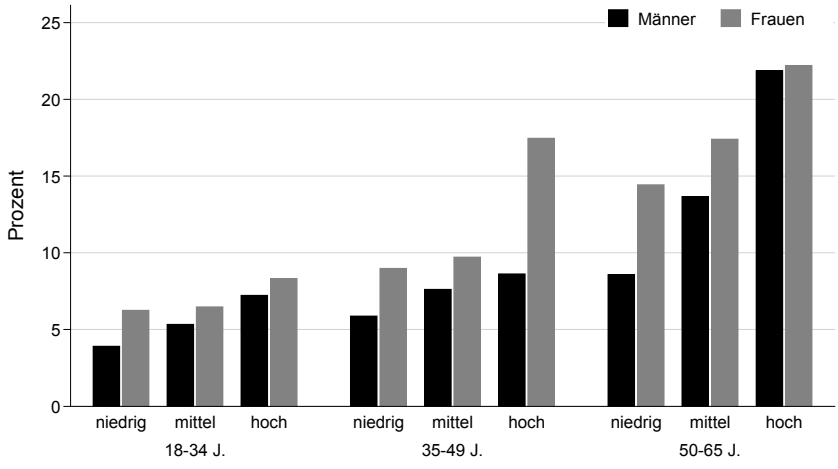
4.2 Validierung der JEM anhand von Gesundheitsindikatoren in der Erwerbstätigenbefragung 2006

Die ermittelten Werte zur Arbeitsbelastung in beruflichen Tätigkeiten auf Ebene der dreistelligen Berufscodes nach KldB-92 werden nachfolgend anhand von Gesundheitsindikatoren aus der Erwerbstätigenbefragung 2006 einer internen Validierung unterzogen. Die Scores auf Basis der KldB-92 sind weitgehend mit den Scores auf Basis von ISCO-88 vergleichbar, beide Skalen korrelieren zwischen 0,80 und 0,95. Aus Platzgründen wird nur eine Validierung der KldB-92 dargestellt.

Zur Validierung werden der allgemeine Gesundheitszustand, die Anzahl von Krankheitssymptomen während der oder nach der Ausübung der Tätigkeit sowie die Anzahl krankheitsbedingter Fehltage verwendet. Der subjektive Gesundheitszustand hat sich in vielen Studien – auch unabhängig von medizinisch objektivierbaren Befunden – als guter Prädiktor für die Lebenserwartung und die Inanspruchnahme des medizinischen Versorgungssystems erwiesen (Idler/Benyamini 1997). In der Erwerbstätigenbefragung 2006 wird der Indikator über die Frage *„Wie ist Ihr allgemeiner Gesundheitszustand?"* erhoben. Die Antwortmöglichkeiten reichen auf einer fünfstufigen Skala von *„ausgezeichnet"* bis *„schlecht"*. Untersucht wird der Anteil von Befragten, die ihre Gesundheit als *„weniger gut"* oder *„schlecht"* bewerten. Zudem stehen im Datensatz Informationen zu gesundheitlichen Beschwerden zur Verfügung, die häufig während oder unmittelbar nach der Arbeit auftreten. Insgesamt werden 23 verschiedene gesundheitliche Beschwerden abgefragt. Es handelt sich dabei einerseits um allgemeine Symptome für körperliche und psychische Befindlichkeitsstörungen (bspw. Allgemeine Müdigkeit, Niedergeschlagenheit, Burnout), aber auch um spezifische Symptome für Einschränkungen im Bewegungsapparat (bspw. Nacken-, Knie- oder Hüftschmerzen), für Reizungen der Haut und der Schleimhäute oder Störungen im Herz-Kreislaufsystem (Herzstiche, Engegefühl in der Brust). Nachfolgend wird die Anzahl der berichteten Beschwerden als Gesundheitsindikator verwendet. Als dritter Indikator wird die Anzahl von Tagen herangezogen, an denen die Beschäftigten krankheitsbedingt ihrem Arbeitsplatz ferngeblieben sind. Die Angaben der Befragten beziehen sich dabei auf das letzte Jahr vor der Befragung.

In Abbildung 2 ist der Anteil von Befragten, die ihren eigenen Gesundheitszustand als „weniger gut" oder „schlecht" einschätzen, differenziert nach der allgemeinen Arbeitsbelastung (AB_{ges}) in ihren Tätigkeiten dargestellt. Insgesamt geben 8,9 % bzw. 11,1 % der erwerbstätigen Männer und Frauen im Alter zwischen 18 und 65 Jahren an, dass ihre Gesundheit weniger gut oder schlecht ist. Es zeigen sich dabei allerdings deutliche Differenzen je nach allgemeiner Arbeitsbelastung.

Abbildung 2 Anteil Gesundheitszustand weniger gut/schlecht nach beruflicher Belastung (AB_{ges}), Alter und Geschlecht



Unter niedrig belasteten Erwerbstätigen (unteres Fünftel der nach Belastung geordneten Tätigkeiten) betragen die entsprechenden Anteile 6,2 % bzw. 9,7 %, während sie unter hoch belasteten Erwerbstätigen (oberes Fünftel) bei 11,0 % bzw. 16,1 % liegen. Vergleichbare Differenzen zwischen niedrig und hoch belasteten Erwerbstätigen zeigen sich auch hinsichtlich des Anteils mit oder der Anzahl von Gesundheitsproblemen, die in Zusammenhang mit der Tätigkeit auftreten. Während in den Tätigkeiten mit niedriger Arbeitsbelastung nur 29,0 % der Männer und 19,1 % der Frauen beschwerdefrei sind, beträgt dieser Anteil bei den Erwerbstätigen in den Tätigkeiten mit hoher Arbeitsbelastung lediglich 14,5 % bzw. 12,2 %. Betrachtet man die Anzahl der berichteten Beschwerden, so nennen gering belastete Männer und Frauen im Durchschnitt 2,7 bzw. 3,6 Symptome, während hoch belastete Erwerbstätige 4,6 bzw. 5,3 Beschwerden nennen. Diese gesundheitlichen Differenzen spiegeln sich auch in der Anzahl der Fehltage im letzten Jahr wider: Männer und Frauen in gering belastenden Tätigkeiten geben 5,9 bzw. 7,4 Fehltage an, für die Vergleichsgruppe in hoch belastenden Tätigkeiten betragen die entsprechenden Werte dagegen 12,2 und 11,5 Tage.

In Tabelle 8 sind die Ergebnisse von multivariaten Regressionsmodellen der Determinanten des Risikos eines weniger guten oder schlechten Gesundheitszustandes (Logit Modell) der Anzahl von Gesundheitsbeschwerden (Poisson Modell) sowie der Anzahl von krankheits- oder unfallbedingten Fehltagen im Beruf in Abhängigkeit von der allgemeinen (Modell 1) und bereichsspezifischen (Modell 2)

Arbeitsbelastung in beruflichen Tätigkeiten dargestellt. Die Effekte der Arbeitsbelastungen auf Gesundheitsindikatoren werden dabei für das Alter, die Wochenarbeitszeit, die Länge der Beschäftigung in der aktuellen Tätigkeit und die schulische und berufliche Qualifikation (nach CASMIN Klassifikation, vgl. Brauns et al. 2003) kontrolliert. Die schulisch-/berufliche Bildung der Beschäftigten wurde berücksichtigt, um zu überprüfen, ob der Effekt der Arbeitsbelastung auch unabhängig von der Qualifikation der Beschäftigten zum Tragen kommt.

Tabelle 8 Gesundheitszustand nicht gut/sehr gut, Anzahl von berichteten Beschwerden während oder nach der Arbeit und krankheitsbedingte Fehltage nach Belastung und Geschlecht

	Gesundheitszustand nicht gut/sehr gut		Anzahl Symptome bei der Tätigkeit		Anzahl krankheitsbedingte Fehltage	
	Männer	Frauen	Männer	Frauen	Männer	Frauen
Modell 1:	OR	OR	IRR	IRR	IRR	IRR
Allgemein						
niedrig	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
mittel	1,23	1,13	1,24***	1,23***	1,18***	1,25***
Hoch	1,33*	1,72***	1,49***	1,48***	1,39***	1,51***
Modell 2:	OR	OR	IRR	IRR	IRR	IRR
Physisch						
niedrig	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
mittel	1,21	1,27*	1,21***	1,13***	1,55***	1,31***
hoch	1,32	2,35***	1,46***	1,31***	1,81***	1,05**
Psychosozial						
niedrig	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
mittel	1,50**	0,99	1,11***	1,13***	1,12***	1,11***
hoch	1,47**	1,00	1,16***	1,27***	1,03*	1,41***
N	8608 (84%)	8723 (91%)	8574 (84%)	8687 (90%)	8587 (84%)	8688 (90%)

*Kontrolliert für Alter, Wochenarbeitszeit, Dauer der Betriebszugehörigkeit und Bildungsabschluss (CASMIN); OR: Odds Ratio (Logit Modell); IRR: Incidence Rate Ratio (Poisson Modell für Zähldaten). Ref.: Referenzkategorie; Signifikanz: * p<0,05; ** p<0,01; *** p<0,001; N: Gültige Fälle (Anteil an der Stichprobe).*
Datenbasis: Erwerbstätigenbefragung 2006, KldB-92.

Insgesamt zeigen die Ergebnisse deutliche Zusammenhänge zwischen Arbeitsbelastungen und den betrachteten Gesundheitsindikatoren auf. Nach Berücksichtigung von Bildungsunterschieden ist das Risiko eines weniger guten Gesundheitszustandes bei Männern mit einer hohen Arbeitsbelastung signifikant 1,3-fach gegenüber Männern mit einer geringen Arbeitsbelastung erhöht. Bei Frauen in Tätigkeiten mit hoher Arbeitsbelastung ist das Risiko um das 1,7-Fache erhöht. Bei Männern ist nur die Subskala für physische Belastungen, bei Frauen nur die Subskala für psychi-

sche Belastungen signifikant mit dem Gesundheitszustand assoziiert. Die Anzahl von Symptomen, die während oder nach der Arbeit auftreten, variiert bei Männern und Frauen mit der allgemeinen und den bereichspezifischen Arbeitsbelastungen in ihren Berufen. Bei Männern und Frauen mit einer hohen allgemeinen Arbeitsbelastung ist die Anzahl von Symptomen auch nach Kontrolle für Alter, Bildung und weitere berufsbezogene Merkmale 1,5-fach höher als in der Vergleichsgruppe mit geringer Belastung. Die Anzahl von Symptomen ist zudem signifikant mit physischen und mit psychosozialen Arbeitsbelastungen assoziiert, wobei der statistische Effekt der physischen Arbeitsbelastungen überwiegt. In der Länge krankheits- oder unfallbedingter Fehlzeiten bei der Arbeit zeigen sich ebenfalls deutliche Differenzen zwischen den Gruppen. Männer in hoch belasteten Tätigkeiten geben 1,4-mal längere Fehlzeiten im letzten Jahr an als solche in gering belasteten Tätigkeiten. Bei Frauen ist die entsprechende Anzahl 1,5-fach erhöht. Dieser Zusammenhang geht bei Männern vor allem auf körperliche und bei Frauen vor allem auf psychosoziale Belastungen zurück. Die dargestellten Analysen wurden mit den Indexwerten, die auf Basis der Klassifikation ISCO-88 gewonnen wurden, repliziert (nicht dargestellt). Dabei haben sich insgesamt – bei leicht verringerten Effektstärken – vergleichbare Zusammenhänge mit den Gesundheitsindikatoren gezeigt.

4.3 Validierung der JEM anhand von Gesundheitsindikatoren aus der GEDA-Studie 2009

Der allgemeine und die beiden bereichsspezifischen Scores für das Ausmaß von Arbeitsbelastungen in beruflichen Tätigkeiten haben sich bei den Analysen auf Basis der Erwerbstätigenbefragung als signifikant assoziiert mit der allgemeinen gesundheitlichen Lage sowie mit krankheitsbedingten Fehlzeiten erwiesen. Allerdings besteht die Möglichkeit, dass dieses Ergebnis nicht auf die Bedeutung der Tätigkeiten für die Arbeitsbelastungen, sondern auf die Selektivität der Beschäftigten in den Tätigkeiten in der Erwerbstätigenbefragung zurückzuführen ist. Damit wäre zwar eine gewisse interne Validität der verwendeten Scores als Einflussfaktoren für Gesundheitsrisiken und -chancen auf der Individualebene gegeben, sie wären allerdings nicht geeignet, um auf der aggregierten Ebene der beruflichen Tätigkeiten verwendet zu werden. Nachfolgend soll daher eine externe Validierung der Scores zur allgemeinen und bereichsspezifischen Arbeitsbelastung anhand eines anderen Datensatzes durchgeführt werden. Dazu werden die Daten der Studie Gesundheit in Deutschland Aktuell 2009 verwendet (RKI 2010). Es handelt sich dabei um einen repräsentativen telefonischen Survey, der in den Jahren 2008 und 2009 am Robert Koch-Institut in Berlin bei 21.262 Erwachsenen durchgeführt wurde.

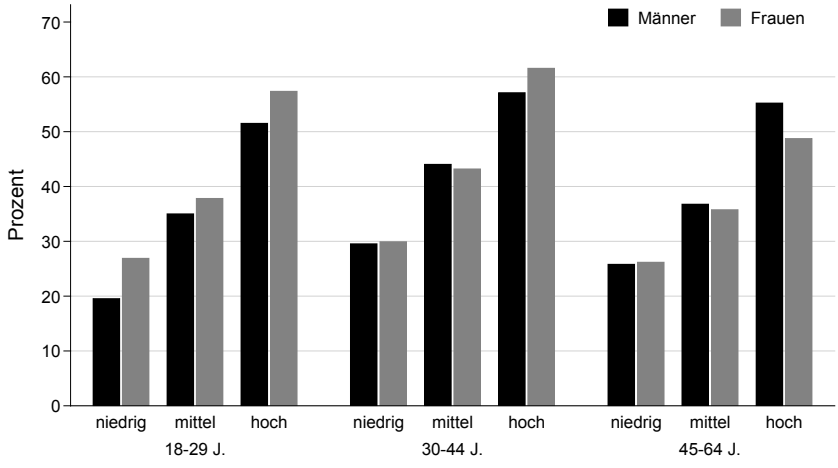
Bei den verwendeten Daten der GEDA Studie 2009 handelt es sich um einen Auszug aus dem Datensatz, in dem nur die erwerbstätigen Befragten im Alter zwischen 18 und 64 Jahren berücksichtigt wurden (n=13.044). Im Datensatz stehen Informationen zur aktuellen beruflichen Tätigkeit auf Basis der Klassifikation KldB-92 und ISCO-88 zur Verfügung, die von der Gesellschaft sozialwissenschaftlicher Infrastruktureinrichtungen e. V. (GESIS) für das Robert Koch-Institut auf Basis von Angaben zu Stellung im Beruf, Beruflicher Tätigkeit (Freitext) sowie Branche und Bildung der Befragten kodiert wurden. Für die nachfolgenden Analysen wurde der Schlüssel KldB-92 verwendet, um den erwerbstätigen Befragten die berechneten Scores zur allgemeinen, physischen, und psychosozialen Arbeitsbelastung zuzuspielen. Fehlende Werte bei den Angaben zur beruflichen Tätigkeit sind im Datensatz sehr selten, insgesamt konnte 99,2 % der erwerbstätigen Befragten ein Berufscode zugeordnet werden. Die Zuspielung der Belastungsindizes erfolgte sowohl über die KldB-92 als auch über die ISCO-88, nachfolgend werden aber nur die Ergebnisse auf Basis der KldB-92 dargestellt. Die Zuspielung der Werte der drei Scores auf Basis der KldB-92 verlief schrittweise:

1. Zuspielung der drei Belastungsscores auf Basis der vierstelligen Berufscodes. Ergebnis 95,7 % der erwerbstätigen Befragten mit gültigem Berufscode konnten die Scores zugespielt werden.
2. Zuspielung der Scores auf Basis der dreistelligen Berufscodes, sofern noch fehlende Werte aus (1). Ergebnis: 98,6 % der erwerbstätigen Befragten mit gültigem Berufscode hatten einen Score.
3. Zuspielung der Scores auf Basis der zweistelligen Berufscodes, sofern noch fehlende Werte aus (1) oder (2). Ergebnis: Alle erwerbstätigen Befragten mit gültigem Berufscode hatten einen Score.

Der Schwerpunkt der GEDA Studie 2009 liegt auf Indikatoren zur Gesundheit und zum Gesundheitsverhalten. Zur Validierung des Scores für berufliche Belastungen werden vier Indikatoren aus der Studie herangezogen. Der erste Indikator beschreibt die Selbsteinschätzung der Befragten hinsichtlich ihrer gesundheitlichen Beanspruchung durch die eigene berufliche Tätigkeit. Die Formulierung „Glauben Sie, dass Ihre Gesundheit durch Ihre Arbeit gefährdet ist?" orientiert sich dabei an einer entsprechenden Fragestellung im European Working Conditions Survey (Parent-Thirion et al. 2007). Die anderen verwendeten Indikatoren stammen aus dem 4-Item-Healthy-Days-Core-Module der CDC (CDC 2009). Bei den CDC Indikatoren werden die Befragten gefragt, wie viele Tage es ihnen in den letzten vier Wochen aufgrund von körperlichen bzw. emotionalen Problemen nicht gut ging und an welchen Tagen sie ihre normalen Alltagstätigkeiten durch diese gesundheitlichen Probleme nicht mehr

ausführen konnten.[6] Insgesamt gibt es bei nur 3,7 % der erwerbstätigen Befragten fehlende Werte bei einem der vier Gesundheitsindikatoren in der GEDA-Studie 2009.

Abbildung 3 Wahrnehmung von gesundheitsgefährdenden Arbeitsbedingungen nach beruflicher Belastung, Alter und Geschlecht



Analog zur internen Validierung auf Basis der Erwerbstätigenbefragung 2006 wurden die zehnstufigen Skalen zur allgemeinen und bereichsspezifischen Arbeitsbelastung für die Analysen in drei Bereiche (AB_{ges} 1-2: niedrig; 3-8 mittel; 9-10 hoch) unterteilt. In Abbildung 3 wird die Wahrnehmung von gesundheitsgefährdenden Arbeitsbedingungen in Abhängigkeit von der allgemeinen Arbeitsbelastung dargestellt. Insgesamt beträgt der Anteil von Erwerbstätigen, die sich als belastet ansehen, in der GEDA-Studie 2009 39,6 % bei Männern und 30,7 % bei Frauen. Die wahrgenommene gesundheitliche Belastung variiert bei Männern und Frauen in allen Altersgruppen deutlich mit der auf Basis der Erwerbstätigenbefragung 2006 ermittelten Belastung. Der Anteil von Befragten, die ihre Gesundheit durch ihre Arbeit als gefährdet ansehen, ist bei Männern und Frauen in hoch belasteten Tätigkeiten zumeist etwa doppelt so hoch wie in der Vergleichsgruppe in gering belasteten Tätigkeiten.

6 Die Formulierungen der entsprechenden Fragen lauten (1) „Wenn Sie an ihre körperliche Gesundheit denken - dazu zählen körperliche Krankheiten und Verletzungen - an wie vielen Tagen in den letzten vier Wochen ging es Ihnen dann wegen Ihrer körperlichen Gesundheit nicht gut?“ und (2) „Wenn Sie an Ihr seelisches Befinden denken - dazu zählen auch Stress, Depressionen oder Ihre Stimmung - ganz allgemein, an wie vielen Tagen in den letzten vier Wochen ging es Ihnen dann wegen Ihres seelischen Befindens nicht gut?“.

Tabelle 9 Logistisches Regressionsmodell des Risikos von krankheits- oder unfallbedingten Fehlzeiten in den letzten vier Wochen und des aktuellen Rauchens

	Zustimmung Arbeit gefährdet eigene Gesundheit		Anzahl Tage mit körperlichen Einschränkungen		Anzahl Tage mit emotionalen Einschränkungen		Anzahl Tage mit funktionalen Einschränkungen	
	Männer	Frauen	Männer	Frauen	Männer	Frauen	Männer	Frauen
Modell 1:	OR	OR	IRR	IRR	IRR	IRR	IRR	IRR
Allgemein								
niedrig	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
mittel	1,70***	2,00***	1,11***	1,00	0,99	1,03*	1,09**	1,00
hoch	3,28***	4,38***	1,33***	1,06**	1,04	1,15***	1,25***	1,16***
Modell 2:	OR	OR	IRR	IRR	IRR	IRR	IRR	IRR
Physisch								
niedrig	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
mittel	1,61***	1,55***	1,18***	0,91***	1,03	0,99	1,09**	0,91***
hoch	2,90***	2,38***	1,37***	0,98	0,99	0,99	1,11**	1,09*
Psychosozial								
niedrig	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
mittel	1,23*	1,26**	0,99	1,07***	1,07**	1,02	0,99	1,06*
hoch	1,52***	2,15***	1,00	1,17***	1,07*	1,13***	0,99	1,17***
N	5823 (95%)	6530 (94%)	5861 (96%)	6556 (95%)	5852 (96%)	6525 (94%)	5870 (96%)	6596 (95%)

*Kontrolliert für Alter, Wochenarbeitszeit, Dauer der Betriebszugehörigkeit und Bildungsabschluss (CASMIN); OR: Odds Ratio (Logit Modell); IRR: Incidence Rate Ratio (Poisson Modell für Zähldaten). Ref.: Referenzkategorie; Signifikanz: * p<0,05; ** p<0,01; *** p<0,001; N: Gültige Fälle (Anteil an der Stichprobe).*
Datenbasis: GEDA 2009, KldB-92.

In Tabelle 9 werden die deskriptiven Ergebnisse durch logistische Regressionsmodelle und durch Poisson-Modelle für Zähldaten multivariat abgesichert. In Modell 1 wird jeweils nur der Score für die allgemeine Arbeitsbelastung in den Tätigkeiten betrachtet, in Modell 2 werden die Scores für physische und psychosoziale Belastungen gleichzeitig einbezogen. Abhängige Variablen sind das Risiko einer wahrgenommenen Gesundheitsgefährdung durch die eigene Arbeit sowie die Anzahl von Tagen mit körperlichen und emotionalen Beschwerden oder funktionellen Beeinträchtigungen. Die Ergebnisse machen deutlich, dass Arbeitsbelastungen auch unabhängig von Alter, Arbeitszeit, Dauer der Betriebszugehörigkeit und Bildung der Beschäftigten signifikant mit den betrachteten Gesundheitsindikatoren assoziiert sind. Dieses Ergebnis spricht für die externe Validität der verwendeten Scores. Männer und Frauen in Tätigkeiten mit hoher allgemeiner Arbeitsbelastung haben nach Kontrolle der genannten Faktoren ein 3,3-fach bzw. 4,4-fach erhöhtes Risiko, eine Gesundheitsgefährdung durch die eigene Tätigkeit wahrzunehmen. Sie berichten zudem von signifikant längeren Dauern körperlicher, emotionaler (nur Frauen)

und funktioneller Einschränkungen im letzten Monat vor der Befragung. Signifikant mehr Tage mit körperlichen Problemen werden von Männern – allerdings nicht von Frauen – in physisch hoch belastenden Tätigkeiten genannt. Signifikant mehr Tage mit emotionalen Problemen werden dagegen von Männern und Frauen in psychosozial hoch belastenden Tätigkeiten genannt. Die Länge von Tagen mit funktionellen Einschränkungen variiert bei Männern nur mit physischen und bei Frauen sowohl mit physischen als auch mit psychosozialen Belastungen.

5 Schlussfolgerungen

Arbeitsbelastungen sind wichtige Determinanten von Gesundheitschancen (vgl. u. a. Griefahn 1996). In dieser Studie wurden JEM für allgemeine, physische und psychosoziale Arbeitsbelastungen entwickelt und anhand von Gesundheitsindikatoren aus der Erwerbstätigenbefragung und GEDA-Studie 2009 validiert. Insgesamt haben sich die Skalen dabei als signifikante Einflussfaktoren für das Auftreten von Beschwerden während der Arbeit, Krankheiten und Unfallverletzungen und den selbstberichteten Gesundheitszustand von Erwerbstätigen erwiesen. Zudem ist die Skala zur allgemeinen Belastung eng mit der Selbstwahrnehmung der Betroffenen zu Gesundheitsgefährdungen durch die Arbeit assoziiert. Die Skala bietet sich damit für Studien, in denen Arbeitsbelastungen nicht direkt erhoben werden können an, um als Proxy-Indikator für Arbeitsbelastungen verwendet zu werden. Im Vergleich der drei entwickelten Skalen hat sich die JEM für allgemeine Arbeitsbelastungen am aussagekräftigsten erwiesen, während die JEM für physische und psychosoziale Arbeitsbelastungen geringere und geschlechtsspezifische Effekte auf die Gesundheit von Erwerbstätigen zeitigten. Hierbei muss methodisch berücksichtigt werden, dass die psychosozialen Arbeitsbelastungen auch – gemessen an der der Intraklassen-Korrelation – deutlich weniger eng mit den Tätigkeiten assoziiert sind als die physischen Arbeitsbelastungen.

Die verwendete mehrstufige Methode (lineares hierarchisches Regressionsmodell mit drei Dimensionen) schätzt zuerst die Variation der Belastungen auf der obersten Ebene der Berufskodierung (2-stellig kodierte „Berufsgruppen" in ISCO-88 und KldB-92) vom allgemeinen Durchschnitt der Belastung unter den Erwerbstätigen, um anschließend auf den unteren Ebenen die Abweichungen der Belastungen in den spezifischen Tätigkeiten von ihrer jeweiligen Berufsgruppe zu ermitteln. Sie hat gegenüber der einfachen Bildung von berufsspezifischen Mittelwerten mehrere Vorzüge: Durch die Aufnahme von Drittvariablen in den fixed-part des Modells kann für Aspekte der Heterogenität von Berufsgruppen und dadurch bedingte Ver-

zerrungen in den Ausgangsdaten kontrolliert werden. Zudem ist die verteilungsbasierte Schätzung berufsspezifischer Belastungen stabiler als eine tätigkeitsdifferenzierte Mittelwertberechnung. Die Varianzzerlegung im Zuge der hierarchischen Regressionsmodelle liefert außerdem Informationen darüber, in welchem Maße die jeweiligen Belastungen im Vergleich der betrachteten Tätigkeiten variieren.

Limitationen der vorliegenden Analysen ergeben sich insbesondere aus der verwendeten Datenbasis und aus der offenen Herangehensweise beim Einbezug der arbeitsplatzbezogenen Merkmale. Die Erwerbstätigenbefragung 2006 ist eine telefonische Befragung der Erwerbstätigen und als solche anfällig für Verzerrungen und die Selektivität der Stichprobe. So kann ein möglicher Response Bias, etwa in Richtung einer geringeren Antwortbereitschaft unter stark belasteten Beschäftigten, nicht ausgeschlossen werden. Für die Verwendung der Erwerbstätigenbefragung sprach, dass sie mit ihrer umfangreichen Abbildung von arbeitsplatzbezogenen Merkmalen für 20.000 Beschäftigte in Deutschland einzigartig ist. Das statistisch anspruchsvolle Verfahren der Mehrebenenanalyse wurde verwendet, um den inferenzstatistischen Problemen der Erhebung zu begegnen. Das Verfahren ermöglicht es, die mögliche Heterogenität der Befragten im Modell zu berücksichtigen und die Ergebnisse dafür zu bereinigen. Die den Skalen zugrunde liegenden Scores wurden außerdem in Dezile transformiert, um statistische Unsicherheiten aufgrund kleiner Fallzahlen in den einzelnen Berufen zu verringern. Dadurch musste ein Informationsverlust in Kauf genommen werden, der aber angesichts der beabsichtigten Verwendung der Skalen als Proxy-Indikatoren für Arbeitsbelastungen vertretbar erschien. Weiterhin muss bei der Interpretation der Ergebnisse auf Basis der drei Skalen berücksichtigt werden, dass den Skalen kein Krankheitswert zugesprochen werden kann. Empirische Zusammenhänge zwischen den drei konstruierten Skalen und spezifischen Erkrankungen sollten aus diesem Grund immer auch durch vertiefende ätiologische Analysen erklärt werden.

Zusammengenommen bieten die in dieser Studie konstruierten und validierten Skalen für die Surveyforschung und auch für epidemiologische Studien eine einfache Möglichkeit, erste Erkenntnisse zum Einfluss der Arbeitsbelastung von Erwerbstätigen auf abhängige Merkmale zu gewinnen, ohne umfangreiche Instrumente zur Erfassung spezifischer Arbeitsbelastungen in eine Studie integrieren zu müssen. Die generierten Skalen zur allgemeinen, physischen und psychosozialen Arbeitsbelastung können mit geringem Aufwand allen Datensätzen zugespielt werden, die Informationen zur beruflichen Tätigkeit auf Basis der Klassifikationen ISCO-88 oder KldB-92 enthalten.

Literatur

Arbeitsgruppe KldB, 2010 (2008): Exposé Klassifikation der Berufe 2010 15. April 2008. Nürnberg und Bundesagentur für Arbeit und Institut für Arbeitsmarkt- und Berufsforschung.

Babitsch, B., T. Lampert, S. Müters und M. Morfeld, 2006: Ungleiche Gesundheitschancen bei Erwachsenen: Zusammenhänge und mögliche Erklärungsansätze. S. 221-240 in: M. Richter und K. Hurrelmann (Hg): Gesundheitliche Ungleichheit. Grundlagen, Probleme, Konzepte. Wiesbaden: VS-Verlag.

BauA, 2010: Sicherheit und Gesundheit bei der Arbeit 2008. Berlin: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.

Bödeker, W., H. Friedel, M. Friedrichs und C. Röttger, 2006: Kosten arbeitsbedingter Erkrankungen. Schriftenreihe der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin FB 946. Bremerhaven; Wirtschaftsverlag NW.

Bödeker, W., 2002: Die Job-Exposure-Matrix als Instrument für eine arbeitsweltbezogene Auswertung von Morbiditätsdaten der Krankenkassen. Zeitschrift für Arbeitswissenschaft 56 (5): 330-339.

Brauns, H., S. Scherer und S. Steinmann, 2003: The CASMIN Educational Classification in International Comparative Research. S. 221-224 in: J. H.P. Hoffmeyer-Zlotnik und C. Wolf (Hg): Advances in Cross-National Comparison. New York: Kluwer.

CDC, 2009: Measuring Healthy Days. Population Assessment of Health-Related Quality of Life. Atlanta: Centers for Disease Control and Prevention.

De Leeuw, J. und E. Meijer, 2008: Handbook of Multilevel Analysis. New York: Springer.

Dragano, N., 2007: Arbeit, Stress und krankheitsbedingte Frührenten: Zusammenhänge aus theoretischer und empirischer Sicht. Wiesbaden: VS Verlag für Sozialwissenschaften.

Fields, D. L., 2002: Taking the Measure of Work. Thousand Oaks: Sage.

Friedel, H., 2003: Differenzielle Assoziationen zwischen hohen psychischen Arbeitsanforderungen und dem Arbeitsunfähigkeitsgeschehen. Das Gesundheitswesen 65: 181-186.

Geis, A. J. und J. H.P. Hoffmeyer-Zlotnik, 2001: Kompatibilität von ISCO-68, ISCO-88 und KldB-92. ZUMA-Nachrichten 48: 117-138. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_48.pdf (1.12.2010).

Goldberg, M., H. Kromhout, P. Guenel, A. C. Fletcher, M. Gerin, D. C. Glass, D. Heederik, T. Kauppinen und A. Ponti, 1993: Job Exposure Matrices in Industry. Int J Epidemiol 22 Suppl 2: S.10-15. http://ije.oxfordjournals.org/content/22/Supplement_2/S10.full.pdf (1.12.2010).

Griefahn, B., 1996: Arbeitsmedizin. Stuttgart: Thieme.

Hall, A., 2009: Die BIBB /BAuA-Erwerbstätigenbefragung 2006 – Methodik und Frageprogramm im Vergleich zur BIBB/IAB-Erhebung 1998. Wissenschaftliche Diskussionspapiere BiBB 106: 1-54. http://www.bibb.de/dokumente/pdf/wd_107_bibb_baua_erwerbstaetigenbefragung_2006.pdf (1.12.2010).

Hartmann, J., 2006a: BIBB/BAuA-Erwerbstätigenbefragung 2005/2006 Feldbericht. München: TNS Infratest Sozialforschung.

Hartmann, J., 2006b: BIBB/BAuA-Erwerbstätigenbefragung 2005/2006 - Strukturkontrolle, Steuerung und Gewichtung der Stichprobe. München: TNS Infratest Sozialforschung.

Idler E. L. und Y. Benyamini, 1997: Self-rated Health and Mortality: A Review of Twenty-Seven Community Studies. J Health Soc Beh, 38 (1): 21-37.

ILO, 2008: Resolution Concerning Updating the International Standard Classification of Occupations. Geneva: International Labour Office.

Karasek, R. und T. Theorell, 1990: Healthy Work: Stress, Productivity, and the Reconstruction of Working Life. New York: Basic Books.

Kaufmann, F.-X., 2003: Varianten des Wohlfahrtsstaats. Der deutsche Wohlfahrtsstaat im internationalen Vergleich. Edition Suhrkamp 2301. Frankfurt a. M: Suhrkamp.

Kurth, B. M., C. Lange, P. Kamtsiuris und H. Hölling, 2009: Gesundheitsmonitoring am Robert Koch-Institut. Bundesgesundheitsblatt 52 (5): 557-570.

Langer, W., 2008: Mehrebenenanalyse: Eine Einführung für Forschung und Praxis. Wiesbaden: VS Verlag für Sozialwissenschaften.

Mackenbach J. P. und A. E. Kunst, 1997: Measuring the Magnitude of Socio-Economic Inequalities in Health: An Overview of Available Measures Illustrated with Two Examples from Europe. Soc Sci Med 44 (6): 757-771.

North, F. M., S. L. Syme, A. Feeney, M. Shipley und M. Marmot, 1996: Psychosocial Work Environment and Sickness Absence Among British Civil Servants: The Whitehall II Study. Am J Public Health 86 (3): 332-40.

Parent-Thirion, A., E. F. Macias, J. Hurley und G. Vermeylen, 2007: Fourth European Working Conditions Survey. Luxembourg, Office for Official Publications of the European Communities.

Peter, R., 2006: Psychsoziale Belastungen im Erwachsenenalter. S. 109-123 in: M. Richter und K. Hurrelmann (Hg): Gesundheitliche Ungleichheit. Grundlagen, Probleme, Konzepte. Wiesbaden: VS-Verlag.

Pollmann-Schult, M. und F. Büchel, 2002: Generierung eines Proxys zum Job-Anforderungsniveau aus den Informationen zu ausgeübtem Beruf und beruflicher Stellung. Ein neues Tool für die deutsche Überqualifikations-Forschung. ZUMA-Nachrichten (51): 78-93. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_51.pdf (1.12.2010).

RKI und LGA Brandenburg, 2002: Arbeitsweltbezogene Gesundheitsberichterstattung in Deutschland. Berlin: Robert Koch-Institut.

RKI, 2010: Daten und Fakten: Ergebnisse der Studie „Gesundheit in Deutschland Aktuell 2009". Beiträge zur Gesundheitsberichterstattung des Bundes. Berlin: Robert Koch-Institut.

RKI, 2007: Arbeitsunfälle und Berufskrankheiten. Gesundheitsberichterstattung des Bundes Heft 38. Berlin: Robert Koch-Institut.

RKI, 2006: Gesundheitsbedingte Frühberentung. Gesundheitsberichterstattung des Bundes Heft 30. Berlin: Robert Koch-Institut.

Rohmert, W., 1984: Das Belastungs-Beanspruchungs-Konzept. Zeitschrift für Arbeitswissenschaft 38 (10): 196-200.

Schlick, C., R. Bruder und H. Luczak, 2010: Arbeitswissenschaft. Heidelberg: Springer.

Siegrist, J., 1996: Soziale Krisen und Gesundheit. Göttingen, Hogrefe.

StataCorp, 2009: Stata Longitudinaldata/Paneldata XT – Reference Manual Release 11. College Station, TX, Stata Press.

Tiemann, M., 2010: Wissensintensive Berufe – Empirische Forschungsarbeit (Vorabdruck). Wissenschaftliche Diskussionspapiere BiBB 114. http://www.bibb.de/veroeffentlichungen/de/publication/download/id/6176 (1.12.2010).

Anschrift des Autors

Dr. Lars Eric Kroll
Robert Koch-Institut
- Fachgebiet 24 -
Postfach 65 02 61
13302 Berlin
Kroll@rki.de

Standardisierte Erfassung von Verweigerungsgründen in Face-to-Face-Umfragen

Suggestions for the Standardized Collection of Reasons for Refusals in Face-to-Face Surveys

Natalja Menold und Cornelia Züll

Zusammenfassung
Verweigerungen stellen einen beträchtlichen Anteil systematischer Ausfälle in Umfragen dar. Einige Umfragen erheben Verweigerungsgründe während des Datenerhebungsprozesses. Neben einer Zusatzinformation zum Ablauf der Datenerhebung liefern diese Daten auch Informationen, die zur Reduzierung der Verweigerungsraten genutzt werden können. Allerdings fehlen derzeit standardisierte Instrumente zur Erfassung von Verweigerungsgründen, die eine zuverlässige Datenerhebung sichern würden. In diesem Artikel stellen wir ein standardisiertes Instrument – ein Kategorienschema – zur Klassifizierung der Verweigerungsgründe vor, das als Teil einer Inhaltsanalyse entwickelt wurde. Als Datenbasis nutzten wir offen erhobene Angaben der Interviewer in Kontaktprotokollen im ALLBUS 2008. Die Intercoder-Reliabilität der Codierung beträgt 0,84 und ist zufriedenstellend. Abschließend geben wir Hinweise zur Nutzung des Kategorienschemas für die Codierung der offen erhobenen Daten sowie zur Datenerhebung direkt durch den Interviewer im Feld.

Abstract
Refusals are a considerable source of nonresponse in surveys. During the field period, some surveys collect reasons for refusals as survey para-data. In addition to providing information for research these data can also provide information about nonrespondents that can be used to reduce refusal rates. However, there is a lack of standardized instruments of data collection, the fact of which declines reliability, validity, and objectivity of the data collected. This article presents a standardized instrument – a categorization scheme – for classifying reasons for refusals. The instrument was developed using content analysis of open-ended comments by interviewers in the German General Social Survey (ALLBUS 2008). The interrater-reliability of the developed categorization scheme is .84 and thus satisfying. We give some suggestions on how the categorization scheme can be used by surveyors as well as by interviewers when collecting reasons for refusals.

1 Einleitung

Die Teilnahmebereitschaft an Umfragen nimmt immer mehr ab (De Leeuw/De Heer 2002), und in der Bevölkerung entwickelt sich ein zunehmender Widerwillen, an Befragungen teilzunehmen (Groves/Heeringa 2006). In der allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS)[1] betrug die Teilnahmeverweigerung im Jahr 2008 ca. 50 %. Auch in anderen Umfragen ist der Anteil der Ausfälle durch Teilnahmeverweigerung hoch (im ESS 2008[2] betrugen die Verweigerungsraten bis zu 40 % in den einzelnen Ländern).

Forschung in Bezug auf Verweigerer und Reduzierung der Verweigerungsraten ist wichtig: Wenn die Verweigerer sich von anderen Gruppen, z. B. den Befragten, unterscheiden, können hohe Verweigerungsraten einen Surveybias verursachen (Groves et al. 2004; Esser 1973; Reuband 1975; Zeh 1976).

Bei einigen Umfragen – in diesem Artikel beschränken wir uns auf face-to-face-Interviews – werden deshalb Daten zu Gründen für Verweigerungen als ein Teil der Paradaten erfasst. Paradaten sind Daten, die Prozesse beschreiben (Faulbaum/Prüfer/Rexroth 2009). Zu Paradaten in der Surveyforschung zählen beispielsweise Daten des Case-Managements (Anzahl und Zeit der Kontakte pro Fall oder Interviewdauer) und Daten über die Interaktion des Interviewers mit der Zielperson. Paradaten zu Verweigerungsgründen können auf zwei verschiedene Arten verwendet werden:

a) Zum einen, um Forschung über Verweigerer durchzuführen, z. B. um zukünftige Antwortbereitschaft bei einer neuen Kontaktierung vorauszusagen (Bates et al. 2008; Kreuter/Kohler 2009).

b) Zum anderen können sie in der Feldphase einer Umfrage zur Reduzierung der Ausfälle durch Verweigerung verwendet werden. Hier gibt es folgende Möglichkeiten:

   - Konvertierung von Verweigerern: Die genannten Verweigerungsgründe können darauf hinweisen, wie leicht ein Verweigerer konvertiert werden könnte. Die höchsten Konvertierungsraten wurden bei den Verweigerungsgründen „keine Zeit", „kein Interesse" oder „generelle Verweigerung" festgestellt (Fuse/Xie 2007; Neller 2005; Reuband/Blasius 2000; Schnauber/Daschmann 2008).
   - Supervision der Interviewer: Bei der Nutzung der Daten während der Survey-Durchführung können die Daten zur Supervision der Interviewer genutzt werden, beispielsweise zwecks flexibler Anpassung ihres Verhaltens bei der Kontaktierung von Zielpersonen (vgl. Stoop 2004; Neller 2005; Durrant/Steele 2009).

1 Siehe http://www.gesis.org/dienstleistungen/daten/umfragedaten/allbus/.
2 ESS (European Social Survey), http://ess.nsd.uib.no/ess/round4/.

- Follow-up Kontakte/Nacherfassung: Die Information über den Grund der Verweigerung kann zudem dazu genutzt werden, Befragte in einem neuen Anschreiben doch noch zur Teilnahme zu bewegen (Groves/Couper 1998; Stoop 2004). Neller (2005) beschreibt eine flexible Verwendung von unterschiedlichen Überzeugungsstrategien während der Kontaktierung – abhängig vom genannten Verweigerungsgrund. Weiterhin können die Verweigerungsgründe als eine Basis zur Verwendung flexibler (variabler) Incentives dienen. Eine derartige Flexibilisierung der Anreize schlagen Költringer (1992) und Stoop (2004) vor, um Ausschöpfungsquoten zu erhöhen.

In Hinblick auf diese skizzierte Nutzung der Daten über Verweigerungsgründe für die Forschung sowie im Rahmen der Datenerhebung stellt sich die Frage der Zuverlässigkeit dieser Daten, und es ergibt sich die Notwendigkeit reliabler Erhebungsinstrumente. Im Gegensatz zu Fragebögen, bei denen die psychometrische Güte der Erhebung (Reliabilität, Validität und Objektivität) oftmals untersucht wird, wird die Güte bei den Instrumenten zur Erhebung von Paradaten bisher kaum thematisiert. Dabei stellen Protokolle zur Dokumentation der Kontaktversuche per se ein Erhebungsinstrument dar. Insbesondere dann, wenn die gesammelten Daten zu Forschungszwecken genutzt werden, sollte die Datenqualität gesichert werden.

AAPOR (2009) empfiehlt in seinen „Standard Definitions" bei der Erhebung von Paradaten auch eine Erhebung von Verweigerungsgründen (siehe dazu auch Durrant/Steele 2009). Jedoch fehlen bisher diesbezüglich Erfahrungen und Standards. Die Verweigerungsgründe werden des Öfteren „nebenbei" erfasst und eingeschränkt oder gar nicht genutzt. Manche Surveys (wie ALLBUS) erheben Verweigerungsgründe im offenen Antwortformat. Die Nutzung solcher Daten erfordert aufwändige Inhaltsanalysen und wird dadurch erschwert. In anderen Surveys (Beispiel: ESS) werden Kategorien zur Erhebung genutzt. Es fehlen jedoch Angaben zur Güte entsprechender Erhebungsinstrumente.

In diesem Artikel stellen wir ein Kategoriensystem als standarisiertes Instrument zur Erfassung der Verweigerungsgründe vor. Wir entwickelten dieses Instrument mit Hilfe einer Inhaltsanalyse der Kommentare der Interviewer zu Verweigerungen in Kontaktprotokollen im ALLBUS 2008. Im Folgenden werden die einzelnen Entwicklungsschritte beschrieben und das Kategorienschema mit Angaben zur Reliabilität der Codierungen vorgestellt. Abschließend wird die Anwendung des Schemas in Umfragen zur Reduktion der Verweigerungsraten exemplarisch für den Nutzer aufgezeigt.

2 Ansätze für ein Instrument zur Erfassung von Verweigerungsgründen

In einem ersten Schritt zur Entwicklung eines Kategoriensystems zur Erfassung von Verweigerungsgründen haben wir Kategorien identifiziert, die in anderen Studien verwendet werden. Eine Studie, die Verweigerungsgründe in einem geschlossenen Antwortformat abfragt und diese Paradaten im Gegensatz zu anderen Studien für Sekundäranalysen zur Verfügung stellt, ist der ESS. In den ersten vier Runden des ESS werden bis zu 14 Kategorien zur Codierung von Verweigerungsgründen vorgegeben. Dazu gehören die Kategorien: „keine Zeit", „kein Interesse", „Geldverschwendung", „Zeitverschwendung", etc. (siehe Tabelle 1). Eine Analyse über alle Länder hinweg zeigte, dass der Anteil der Angaben in der Restkategorie relativ hoch ist (ESS 1: 12,3 %; ESS 2: 10 %; ESS 3: 11,7 %; ESS 4: 11,5 %).

Tabelle 1 Verweigerungsgründe im ESS 4 (exemplarisch ausgewählte Länder)

Verweigerungsgrund	Dänemark		Schweiz		Deutschland	
	N	%	N	%	N	%
Bad timing, otherwise engaged	122	21,8	158	10,3	358	12,8
Not interested	-*		434	28,3	826	29,5
Do not know subject / too difficult	17	3,0	14	0,9	-*	-*
Waste of time	51	9,1	34	2,2	-*	-*
Waste of money	4	0,7	11	0,7	-*	-*
Interferes with my privacy	6	1,1	113	7,4	78	2,8
Never do surveys	96	17,2	155	10,1	685	24,5
Co-operated too often	28	5,0	8	0,5	53	1,9
Do not trust surveys	10	1,8	69	4,5	107	3,8
Previous bad experience	3	0,5	14	0,9	-*	-*
Do not like subject	8	1,4	11	0,7	96	3,4
No approval to cooperate	4	0,7	21	1,4	128	4,6
Afraid to let strangers in	15	2,7	5	0,3	-*	-*
Other	195	34,9	485	31,6	465	16,6

** In Dänemark fehlte die Kategorie „not interested". In Deutschland standen die Kategorien „do not know", „waste of time", „waste of money", „previous bad experience" und „afraid to let strangers in" nicht zur Verfügung. Daten: http://ess.nsd.uib.no/.*

Tabelle 1 präsentiert die Häufigkeiten der Verweigerungsgründe exemplarisch in den drei Ländern mit dem höchsten Anteil an „other" im ESS 4. Es ist zu sehen, dass in diesen Beispielen der Anteil an „other" hoch ist (16,6 % - 34,9 %), wobei einige andere Kategorien nur marginal besetzt sind (teilweise unter einem Prozent). Zugleich finden sich in den Instruktionen für die Interviewer keine näheren Erläu-

terungen, wie die Kategorien zur Erhebung der Verweigerungsgründe zu verstehen und zu nutzen sind (vgl. Fieldwork Instructions ESS 1-4[3]). Dies zeigt, dass die ESS Kategorien nicht optimal sind, hier wäre eine Verbesserung der Kategorien sinnvoll.

Im Gegensatz zum ESS werden im ALLBUS keine Kategorien bzgl. der Verweigerungsgründe vorgelegt, sondern die Verweigerungsgründe in Form von Erläuterungen zum Ergebnis der Kontaktaufnahme offen erhoben. Ein standardisiertes Instrument würde hier nicht nur die Güte der Daten (Reliabilität und Objektivität) erhöhen, sondern auch deren Nutzung – für Forschungszwecke sowie zur Reduzierung der Verweigerungsraten während der Feldphase – vereinfachen.

3 Beschreibung der Datenbasis

Zur Entwicklung eines solchen standardisierten Instruments anhand offen erhobener Daten eignet sich die Methode der Inhaltsanalyse, deren Kern die Entwicklung eines Kategoriensystems darstellt. Das im Folgenden beschriebene Kategorienschema wurde an Hand der (unbereinigten) Kontaktprotokolle bis zur Woche 34, der letzten Woche der Datenerhebung (CAPI-Protokolle) des ALLBUS 2008, entwickelt. Bei diesen Protokollen handelt es sich um Berichte der Interviewer.

Die ALLBUS-Interviewer mussten, wenn ein Interview nicht realisiert wurde, zunächst einen Ausfallcode angeben. Es konnte dabei aus den folgenden Kategorien auswählt werden:

- Adresse falsch, existiert nicht mehr
- Zielperson (ZP) verstorben
- ZP nicht (mehr) unter der angegebenen Adresse
- ZP lebt in Anstalt und nicht in Privathaushalt
- Im Haushalt niemand angetroffen
- ZP aus Zeitgründen nicht zum Interview bereit
- ZP nicht zum Interview bereit (hier sollten dann die Gründe im Feld für offene Angaben eingetragen werden)
- ZP spricht nicht hinreichend gut Deutsch
- ZP dauerhaft krank und nicht in der Lage, dem Interview zu folgen.

Zusätzlich stand ein Feld für Bemerkungen zur Adresse und zur Befragung zu Verfügung. Im Falle der Verweigerung sollte hier eine Erläuterung gegeben werden.

3 Siehe http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=119&Itemid=367.

Dieses Feld stand somit nicht nur für Kommentare bei Verweigerungen, sondern auch für andere Anmerkungen der Interviewer zur Verfügung. Wir haben für die Entwicklung des oben beschrieben Kategorienschemas und die anschließende Codierung Anmerkungen nur ausgewählt, wenn „ZP nicht zum Interview bereit" angekreuzt war.

Die im ALLBUS erhobenen Daten der Kontaktprotokolle stellen das Basismaterial für eine Inhaltsanalyse zur Entwicklung eines standardisierten Erhebungsinstrumentes dar. Dieses sollte in Form einer – im Vergleich zu den Kategorien im ESS - optimierten Kategorienliste mit Definitionen der Kategorien und Ankerbeispielen vorliegen, die eine zuverlässige Zuordnung zu Kategorien ermöglichen.

4 Kategorienschema

Bei der Entwicklung unseres Kategorienschemas verwendeten wir in einem ersten Schritt die ESS-Kategorien (siehe Tabelle 1). Die ersten Ergebnisse zeigten, dass die ESS-Kategorien für eine Codierung der Verweigerungsgründe im ALLBUS nicht ausreichend waren. Wir hatten dieselben Probleme wie beim ESS: hohe Häufigkeit der Kategorie „Sonstiges" und nur geringe Besetzung anderer Kategorien, wie z. B. „Geldverschwendung". Wir ergänzten daher das Schema um neue Kategorien und fassten andere Kategorien zusammen, um diese Codierergebnisse zu verbessern.

Tabelle 2 Verweigerungsgründe: Kategorien

Hauptgruppe	Kategorie
Verweigerung allgemein	Generelle Verweigerung Kein Interesse Keine Zeit Verweigerung durch Dritte/Teilnahmeverbot
Alter und Gesundheitszustand	Alter der ZP Gesundheitszustand
Politische Situation	Unzufrieden mit der politischen Situation Teilnahmeverweigerung, weil Ausländer
Negative Einstellung gegenüber Umfragen	Umfragen bringen nichts Zu viele Umfragen Schlechte Erfahrungen mit Interviews
Survey Prozess	Datenschutz und Verletzung der Privatsphäre Freiwilligkeit der Teilnahme Methodik von Surveys
Sonstiges	Andere Gründe Nicht zuordenbar

Das Kategorienschema zur Codierung der Verweigerungsgründe umfasst in seiner Endfassung 15 Kategorien in fünf Hauptgruppen (vgl. Tabelle 2) und vier weitere Kategorien, die keinen Verweigerungsgrund beschreiben (vgl. Tabelle 3). Diese vier Kategorien werden vorgestellt, um das vorliegende Datenmaterial lückenlos zu charakterisieren.

Tabelle 3 Restkategorien, die keine Verweigerung beschreiben

Hauptgruppe	Kategorie
Nicht erreicht/Spätere Bereitschaft	Keine Verweigerung, sondern kein Kontakt, Bereitschaft oder Erreichbarkeit später
Probleme der Umfrageorganisation/ Durchführung	Adressfehler
	Kontaktaufnahme trotz bisher geäußerter Verweigerung
	Probleme bei der Surveydurchführung

In der folgenden Beschreibung sind alle Kategorien im Kategorienschema verkürzt dargestellt. Dabei verwendeten wir die für eine Kategorie besonders typischen Beispiele. Das ausführliche Kategorienschema mit den vollständigen Definitionen und der umfassenden Auflistung der Beispiele kann in Menold/Züll (2010) nachgelesen werden.

Verweigerung allgemein

Die ersten vier Kategorien beziehen sich auf allgemeine Verweigerungen, ohne die Angabe eines Grundes oder aus fehlendem Interesse oder Zeitmangel.

Generelle Verweigerung (Code 110): Dieser Code wird immer vergeben, wenn die Zielperson ein Interview oder Interviews im Allgemeinen verweigert/ablehnt. Dabei lassen die Aussagen keine Schlussfolgerungen auf den Grund der Verweigerung zu. Codiert werden hier neben Angaben wie „verweigert", „will nicht" oder „hat keine Lust" auch Aussagen, dass dem Interviewer der Zutritt verweigert wurde.

Kein Interesse (Code 120): Die Zielperson sagt explizit, dass kein Interesse an Umfragen im Allgemeinen, an dieser Umfrage oder am Thema der Umfrage besteht (z. B. „kein Interesse an Umfragen, danke") oder das Thema abgelehnt wird.

Keine Zeit (Code 130): Die Zielperson gibt an, keine Zeit zu haben, oder es werden Zeitprobleme signalisiert (schwierige Terminfindung, Prüfungen, Pflegefall im Haushalt, etc.).

Verweigerung durch Dritte/Teilnahmeverbot (Code 140): Der Zielperson wird die Teilnahme durch Dritte, z. B. durch die Eltern oder den Ehepartner, verboten oder die Teilnahme wird durch Dritte *abgelehnt* oder verweigert (z. B. „meine Frau nimmt an so etwas nicht teil" oder „Vater strikt dagegen").

Alter und Gesundheit

In diese Kategoriengruppe fallen die beiden Dimensionen Alter und Gesundheitszustand, die der Verweigerung zugrunde gelegt werden.

Alter der Zielperson (Code 210): Die Zielperson ist zu alt (Einschätzung des Interviewers), fühlt sich selbst zu alt oder ist nach Aussage Dritter zu alt.

Gesundheitszustand (Code 220): Hier werden Verweigerungen aus Gesundheitsgründen codiert. Codiert wird, wenn die Zielperson krank ist, z. B. an Demenz erkrankt, aber auch, wenn die Zielperson sich in einer Reha-Maßnahme oder in Kur befindet oder eine Operation unmittelbar bevorsteht.

Politische Situation

Eine dritte Codegruppe deckt die politische Situation als Verweigerungsgrund ab. Dazu gehört sowohl die Unzufriedenheit mit der aktuellen politischen Situation, aber auch die Situation der Zielperson, die sich als Ausländer(in) nicht in die Umfragesituation begeben möchte.

Unzufrieden mit der politischen Situation (Code 310): Die Zielperson gibt explizit ihre Unzufriedenheit mit der Politik, mit Politikern oder dem Staat als Grund der Verweigerung an, z. B. „Politiker machen eh was sie wollen" oder „an uns Hartz IV Empfänger ist niemand interessiert".

Teilnahmeverweigerung, weil Ausländer (Code 320): Die Teilnahme wird mit einem Hinweis verweigert, dass die Zielperson Ausländer ist und dass sie die Situation in Deutschland nichts angeht[4].

4 Verständigungsprobleme aufgrund mangelnder Deutschkenntnisse wurden im ALLBUS 2008 gesondert (außerhalb der Verweigerungsgründe) als Ausfälle aufgenommen (Wasmer/Scholz/Blohm 2010). Die Kategorie „Verständigungsprobleme aufgrund mangelnder Sprachkenntnisse" wäre für andere Umfragen wichtig und im Schema zu ergänzen.

Negative Einstellung gegenüber Umfragen

Diese Gruppe von Verweigerungsgründen betrifft (a) den Nutzen und Sinn von Umfragen, (b) die Menge an verschiedenen Umfragen, und (c) schlechte Erfahrungen mit Interviews.

Umfragen bringen nichts (Code 410): Hier werden alle Aussagen zur Wirkung und zum Sinn dieser Umfrage oder von Umfragen allgemein codiert (z. B. „Umfragen bringen nichts", „habe nichts davon"). Codiert werden auch Aussagen, die auf Zeit- oder Geldverschwendung hinweisen.

Zu viele Umfragen (Code 420): Die Zielperson verweigert, weil sie zu oft wegen Umfragen angesprochen wird/wurde.

Schlechte Erfahrung mit Interviews (Code 430): Codiert werden Aussagen der Zielperson, dass sie selbst schlechte Erfahrungen mit Interviews gemacht hat oder von anderen negativen Erfahrungen gehört hat und deshalb eine Teilnahme ablehnt, z. B. „Mein Nachbar hat nach einer Umfrage jede Menge Werbung bekommen".

Survey Prozess

Die fünfte Kategoriengruppe betrifft den Datenschutz, die Privatsphäre, die Freiwilligkeit und die Methodik von Surveys.

Datenschutz und Verletzung der Privatsphäre (Code 510): Die Verweigerung erfolgt aus Zweifel am Datenschutz oder aufgrund der Angst vor einer Verletzung der Privatsphäre. Auch generelles Misstrauen und Angst fallen unter diese Kategorie. Im Gegensatz Kategorie 430, in der schlechte Erfahrungen des Befragten oder seines Umfeldes codiert werden, werden hier Bedenken der Zielperson ohne konkrete Erfahrung codiert.

Freiwilligkeit der Teilnahme (Code 520): Eine Zielperson beruft sich darauf, dass die Teilnahme an der Umfrage freiwillig ist und sie deshalb nicht teilnimmt (z. B. „Wenn ich verpflichtet wäre, ja").

Methodik von Surveys (Code 530): Die Methodik der Befragung kann sich auf verschiedene Aspekte beziehen:

a) Der Erhebungsmodus: Eine Zielperson würde z. B. bei einer schriftlichen Befragung teilnehmen, verweigert aber einem Interviewer den Zutritt.
b) Incentives: Eine Zielperson würde bei (höherer) Belohnung/Entschädigung teilnehmen.
c) Länge des Interviews: Die Zielperson schreckt vor der Dauer des Interviews zurück.

Nicht Erreicht/Spätere Bereitschaft

Keine Verweigerung, sondern die Zielperson wurde nicht erreicht oder (mögliche) Teilnahmebereitschaft (Code 610): Hier werden alle Interviewerangaben codiert, die, obwohl vom Interviewer „ZP nicht zum Interview bereit" ausgewählt wurde, keine Verweigerung beschreiben. Dazu gehören Aussagen des Interviewers, dass ein Interview durchgeführt wurde (auch teilweise) oder dass ein Interview zu einem späteren Zeitpunkt möglich ist (wäre). Hier wird auch codiert, wenn niemand angetroffen wurde oder der bereits vereinbarte Termin seitens der Zielperson nicht eingehalten wurde.

Probleme der Umfrageorganisation/Durchführung

Hier werden organisatorische Probleme codiert. Es sind in der Regel keine Verweigerungsgründe, sondern Fehler/Probleme des Feldinstituts und der Stichprobenziehung. Diese organisatorischen Probleme werden in drei Kategorien definiert:

a) *Adressfehler (Code 710)*
b) *Kontaktaufnahme trotz einer bisher geäußerten Verweigerung (Code 720)*
c) *Andere Probleme bei der Durchführung/Organisation (Code 730).*

Sonstige Angaben

Andere Gründe (Code 810): Der Code wird immer vergeben, wenn keiner der im Kategorienschema definierten Gründe als Verweigerungsgrund zutrifft (z. B. „Kenne die Firma (Infratest) nicht und habe noch nichts davon gehört").

Nicht codierbar (Code 910): Aussagen, die nicht codierbar sind, werden unter der Kategorie 910 subsumiert (z. B. unlesbare oder unverständliche Aussagen).

Keine Angabe (Code 999): Das entsprechende Feld wurde vom Interviewer nicht ausgefüllt oder enthält die explizite Aussage „Keine Angabe".

5 Textstatistiken und Reliabilität des Kategorienschemas

Unter Verwendung des entwickelten Kategorienschemas wurde die Codierung der Verweigerungsgründe im ALLBUS 2008 durch eine Codiererin von Hand vorgenommen.

Das Textmaterial war umfangreich: 210 Interviewer schrieben insgesamt 6.363 Kontaktprotokolle, die bis zu drei verschiedene Verweigerungsgründe enthalten konnten. Insgesamt wurden 6.868 Verweigerungsgründe protokolliert (siehe Tabelle 5). Die Angaben variierten in der Ausführlichkeit: Im Durchschnitt wurden in Protokollen 9 Wörter mit 45 Zeichen geschrieben (siehe Tabelle 4). Die ermittelte Anzahl an Verweigerungsgründen ist im Durchschnitt hoch: jeder Interviewer protokollierte durchschnittlich 33 Gründe in seinen Protokollen.

Tabelle 4 Textstatistik

	M	SD	min	max
Durchschnittliche Zahl der Zeichen pro Protokoll	44,80	25,43	0	137
Durchschnittliche Zahl der Wörter	9,47	5,98	0	31
Durchschnittliche Zahl der pro Interviewer angegebenen Verweigerungsgründe	33,05	27,08	0	131

Die erarbeiteten Codierungen wurden anhand einer zweiten Kontrollcodierung überprüft. Dazu wurde eine 10 % Stichprobe aus allen Antworten gezogen. Es wurden nur Antworten ausgewählt, zu denen auch tatsächlich eine offene Angabe vorlag. Die Angaben dieser Textstichprobe wurden durch eine zweite Codiererin noch einmal codiert. Die Intercoder-Reliabilität wurde nach dem von Früh (2007) vorgeschlagenen einfachen Reliabilitätsmaß berechnet:

$$CR = \frac{2 * Anzahl\ der\ Übereinstimmungen}{NCodes1 + NCodes2}$$

dabei sind *NCodes1* die Zahl der von der Codiererin vergebenen Codes und *NCodes2* die Zahl der Codes in der Kontrollcodierung.

Der Intercoder-Reliabilitätskoeffizient lag für unsere Codierungen bei 0,84, d. h. 84 % aller Angaben wurden in der gleichen Art und Weise durch beide Codiererinnen codiert. Dies ist für das vorliegende Kategorienschema mit den oben beschriebenen Kategorien und der Qualität der Interviewerangaben zufriedenstellend.

6 Deskriptive Ergebnisse zur Nennung der Verweigerungsgründe im ALLBUS 2008

Die Codierungen der ALLBUS-Kontaktprotokolle ergeben die in Tabellen 5 und 6 gezeigten Häufigkeiten. Tabelle 5 stellt Kategorien zur Beschreibung der Verweigerungsgründe dar, in der Tabelle 6 werden Häufigkeiten zu den weiteren Kategorien (keine Verweigerung, sondern nicht erreicht oder spätere Bereitschaft) gezeigt.

Tabelle 5 Kategorienhäufigkeit der Verweigerungsgründe

Hauptgruppe	Kategorie	Häufigkeit	%
Verweigerung allgemein	Generelle Verweigerung	2295	33,4
	Kein Interesse	1807	26,3
	Keine Zeit	1285	18,7
	Verweigerung durch Dritte/Teilnahmeverbot	316	4,6
Alter und Gesundheitszustand	Alter der ZP	200	2,9
	Gesundheitszustand	185	2,7
Politische Situation	Unzufrieden mit der politischen Situation	81	1,2
	Teilnahmeverweigerung, weil Ausländer	25	0,4
Negative Einstellung gegenüber Umfragen	Umfragen bringen nichts	191	2,8
	Zu viele Umfragen	66	1,0
	Schlechte Erfahrungen mit Interviews	63	0,9
Survey Prozess	Datenschutz und Verletzung der Privatsphäre	136	2,0
	Freiwilligkeit der Teilnahme	88	1,3
	Methodik von Surveys	90	1,3
Sonstiges	Andere Gründe	17	0,2
	Nicht zuordenbar	23	0,3
Total		6868	100

Bei der Berechnung der Häufigkeiten wurden bis zu drei Nennungen einer Zielperson berücksichtigt. Zudem wurden bei der Kategorie „keine Zeit" auch die Nennungen berücksichtigt, bei denen der Interviewer explizit „ZP aus Zeitgründen nicht zum Interview bereit" angekreuzt, aber keine Anmerkung hierzu notiert hatte (n=829). Die Prozente beziehen sich auf die Zahl der Nennungen.

Tabelle 6 Kategorienhäufigkeiten weiterer Kategorien

Hauptgruppe	Kategorie	Häufigkeit
Nicht erreicht/Spätere Bereitschaft	Keine Verweigerung, sondern kein Kontakt, Bereitschaft oder Erreichbarkeit später	282
Probleme der Umfrageorganisation/ Durchführung	Adressfehler	63
	Kontaktaufnahme trotz einer bisher geäußerten Verweigerung	194
	Probleme bei der Surveydurchführung	97
Total		636

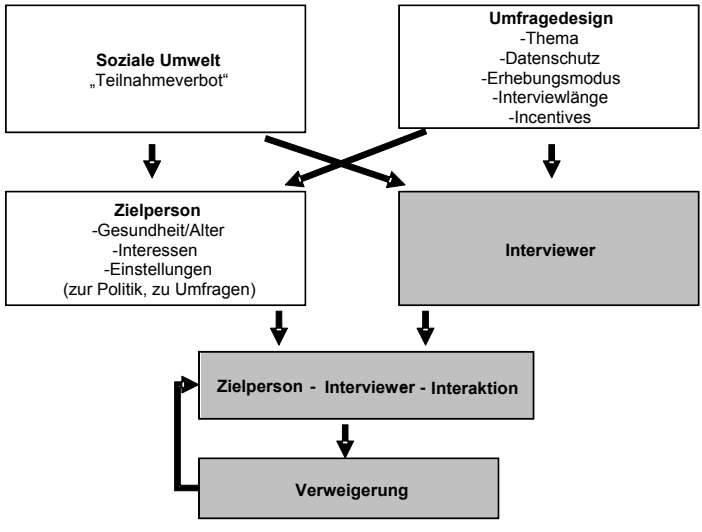
7 Diskussion der Vorgehensweise und Ergebnisse

Mit Hilfe einer Inhaltsanalyse offener Angaben der Interviewer im ALLBUS 2008 entwickelten wir ein Kategorienschema zur Kategorisierung von Verweigerungsgründen. Als Ausgangsbasis nutzten wir Kategorien im ESS. Wir fassten die im ESS selten genutzten Kategorien zusammen und definierten diese sowie andere Kategorien neu. Darüber hinaus fanden wir neue informative Kategorien, beispielsweise Angaben bezüglich der Umfragemethodik. Der prozentuale Anteil der Kategorie „Sonstiges" konnte hierdurch auf 0,2 % reduziert werden. Die entwickelte Kategorienliste zur Beschreibung der Verweigerungsgründe ist nicht wesentlich länger als die Kategorienliste im ESS (15 Kategorien im Vergleich zu 14 im ESS genutzten Kategorien). Zusätzlich sind Definitionen der Kategorien, Zuordnungsregeln und Ankerbeispiele vorhanden, die eine zuverlässigere Codierung ermöglichen (siehe Menold/Züll 2010). Das entwickelte Instrument zeigt dementsprechend eine zufriedenstellende Intercoder-Reliabilität (0,84) und kann zur Kategorisierung offener Angaben der Interviewer in Umfragen eingesetzt werden.

Die Analyse der codierten Daten zeigt – wie frühere Studien auch (DeMaio 1980; Erbslöh/Koch 1988; Groves/Cooper 1998; Költringer 1992; Neller 2005) – dass als Verweigerungsgrund sehr häufig „keine Zeit" und „kein Interesse" notiert wird. Die am häufigsten verwendete Kategorie ist allerdings „generelle Verweigerung" (30 %). Diese Kategorie existierte im vom ESS vorgegeben Schema nicht, und das Fehlen könnte eine Erklärung für die Häufigkeit der Kategorie „other" im ESS sein. Speziell in Deutschland ist diese Kategorie aber wichtig, denn im Falle einer generellen Verweigerung der Teilnahme darf die Zielperson nicht wieder kontaktiert werden (vgl. ADM 2005). Eine weitere Kategorie, die in unserem Schema neu aufgenommen wurde, ist „Survey Prozess" (4 % aller Verweigerungsgründe). Die Kommen-

tare der Interviewer im ALLBUS liefern wichtige Hinweise auf die Faktoren, die die Teilnahmebereitschaft beeinflussen. Sie geben Informationen über die Zielperson und ihre soziale Umwelt sowie auch über die Merkmale des Umfragedesigns in der Wahrnehmung der Zielpersonen wieder (Groves/Cooper 1998, vgl. Abbildung 1).

Abbildung 1 Zuordnung der Verweigerungsgründe zu den Einflussgrößen der Teilnahmebereitschaft an Surveys nach Groves und Couper (1998)



8 Anwendung des Kategorienschemas in Umfragen

Unser Kategorienschema kann zur nachträglichen Codierung aller offenen Interviewerangaben von Verweigerungsgründen eingesetzt werden. Dies ist eine Möglichkeit seiner Verwendung, die für die Umfragen wie ALLBUS oder auch ESS (zur Kategorisierung der Angaben in der Restkategorie) von Bedeutung wäre. Darüber hinaus kann das Kategorienschema als Basis für eine standardisierte Erfassung von Verweigerungsgründen durch Interviewer direkt im Feld in face-to-face-Umfragen dienen. Im Fall der Verwendung der Kategorien im Feld zur Erhebung der Verweigerungsgründe erwarten wir einen Rückgang in der Kategorie „generelle Verweigerung“ und einen Anstieg der Verwendung von speziellen Kategorien (z. B. „Methodik von Surveys“): Werden Kategorien mit differenzierten Verweigerungsgründen vorgege-

ben, wird dem Interviewer deutlich, dass man an dieser Information interessiert ist, sowie dass solche Äußerungen in Protokollen festzuhalten sind.

Die Verwendung eines standardisierten Instruments mit vorgegeben Kategorien würde auch die Fehler durch den Interviewer bei der Dokumentation der Ausfälle reduzieren, z. B. falsche Zuordnung von „später erreichbar" oder „Adresse falsch" als Verweigerung. Solche Fehler haben Einfluss auf die Berechnung von Teilnahme- und Ausfallraten. Speziell im ALLBUS, bei dem die Verweigerungsrate hoch ist, sind solche Korrekturen wichtig und wären auch anhand der Analysen der zur Zeit im offenen Format erhobenen Kommentare ex-post zu korrigieren.

Eine Kategorisierung der Verweigerungsgründe mit Hilfe eines standardisierten Instruments macht die Verwendung der Angaben während der Feldphase zwecks Reduzierung der Verweigerungsraten einfacher. Möglichkeiten, hierzu die Verweigerungsgründe während der Datenerhebung zu nutzen, sind z. B.

1. Die Angaben können für weitere Kontaktversuche genutzt werden. Eine einfache Konvertierung ist bei den Verweigerungsgründen „kein Interesse" und „keine Zeit" möglich (Fuse/Xie 2007; Schnauber/Daschmann 2008; Neller 2005; Reuband/Blasius 2000). Diese wurden von einer großen Gruppe der Zielpersonen im ALLBUS 2008 genannt (ca. 45 % aller Gründe).
2. Abhängig vom Verweigerungsgrund kann eine neue Kontaktierung erfolgreich sein, wenn die Strategie der Kontaktierung geändert wird (siehe Dillman 2007; Dillman/Smyth/Christian 2008). Verweigerungsgründe geben Auskunft darüber, was Befragte als Belastung bei einer möglichen Befragung empfinden (im Einklang mit "leverage-salience theory", Groves/Singer/Corning 2000). Abhängig vom Verweigerungsgrund können beim Eröffnen des Interviews flexibel spezifische Informationen vermittelt werden, die es den Zielpersonen ermöglichen, die Kosten- und Nutzen-Aspekte der Teilnahme weniger oberflächlich zu bewerten und sie hierdurch zu einer Teilnahme zu motivieren. Führt der Befragte an, kein Interesse an einer Befragung zu haben, kann ihm z. B. das Umfragethema selbst und die Wichtigkeit dieses Themas erläutert werden. Im Falle einer Antwort „zu alt" kann der Zielperson erklärt werden, dass gerade auch die Meinung älterer Personen für die Umfrage von Bedeutung sind. Wird als Verweigerungsgrund die politische Situation genannt, kann auf die Nützlichkeit der Ergebnisse der Umfrage für die Gesellschaft (nicht nur für Politiker[5]) verwiesen werden. Das Interview kann gegebenenfalls auch mit Informationen über die Survey-Methode oder den Datenschutz eröffnet werden. Weitere Strategien könnten ein verkürzter Fragebogen mit nur einigen grundsätzlichen Fragen oder die Zusage von (höheren/anderen) Incentives sein, wenn die Länge der Befragung oder ein nicht attraktives Incentive Grund für eine Teilnahmeverweigerung ist. Sollte die Zielperson nicht bereit sein, an

5 Der ALLBUS 2008 war eine politikwissenschaftlich ausgerichtete Studie.

einer persönlichen Befragung teilzunehmen, könnte sie um ein (ggfs. verkürztes) telefonisches Interview oder schriftliche Teilnahme gebeten werden.

3. Des Weiteren können die Verweigerungsgründe zur Supervision der Interviewer genutzt werden, beispielsweise zwecks flexibler Anpassung ihres Verhaltens bei der Kontaktierung von Zielpersonen (vgl. Neller 2005).
4. Zudem können Verweigerungsgründe zur Verringerung des nonresponse bias oder zu einer Verbesserung der Gewichte verwendet werden, z. B. die Kategorie „Verweigerung, weil Ausländer" (je nach Nationalität), falls Randsummen bekannt sind. Man könnte sich auch vorstellen, dass Verweigerungsgründe als Proxy für substantielle Variablen verwendet werden könnten, etwa „an uns Hartz IV Empfänger ist niemand interessiert" zur Einschätzung von Einkommen, „Alter/dement" zur Schätzung des Gesundheitszustands, „kein Interesse an politischen Themen" zur Abschätzung von politischem Interesse. Diesbezüglich wäre es auch vorstellbar, Informationen über die Situation der Verweigerer (Berufstätigkeit, Bildung) zukünftig durch die gezielten Nachfragen der Interviewer zu sammeln und das vorliegende Kategorienschema entsprechend weiterzuentwickeln.

Das entwickelte Kategorienschema enthält übergreifende Kategorien (Hauptkategorien) mit Subkategorien. Je nach dem Zweck der Anwendung kann dieses Kategorienschema flexibel gehandhabt werden. Wenn beispielsweise die Verweigerung aus politischen Gründen im Fokus der Erhebung steht (entsprechend der gestellten Forschungsfrage), können alle Subkategorien der Hauptgruppe „politische Situation" zur Kategorisierung verwendet werden. Ist die politische Situation nicht primär das Interesse der Erhebung, kann nur die Hauptkategorie beim Codieren und in den Analysen verwendet werden. Dabei können die Kategorien „unzufrieden mit der politischen Situation" und „Teilnahmeverweigerung, weil Ausländer" als Ankerbeispiele verwendet werden.

Um die Verwendung des Kategorienschemas zur Erhebung der Verweigerungsgründe durch die Interviewer zu ermöglichen, sind die folgenden weiteren Entwicklungsschritte geplant:

- Überarbeitung der Instruktionen, wobei anstelle der Hinweise zum Codieren Hinweise zur Zuordnung zu den einzelnen Kategorien formuliert werden müssen.
- Besonders aussagekräftige Ankerbeispiele sollten gefunden werden.
- Beispiele sollten identifiziert werden, die bei Schwierigkeiten eine Zuordnung ermöglichen (seltene, schwer zuordenbare Aussagen).

Im elektronischen Erhebungsmodus (CAPI, CATI) ist eine adaptive Handhabung der Kategorien möglich. So können die Definitionen der Kategorien ausgeblendet werden, wenn die Interviewer diese nicht mehr benötigen. Bei Zuordnungsschwierig-

keiten könnten weiterhin weitere Ankerbeispiele abgerufen werden, die eine sichere Zuordnung ermöglichen.

Nach dem Abschluss dieser Entwicklungsschritte soll das Kategorienschema in Hinblick auf die erforderliche Flexibilität, Benutzerfreundlichkeit und psychometrische Güte der erhobenen Daten u. a. in experimentellen Studien evaluiert werden.

Literatur

AAPOR, 2009: Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions1&Template=/CM/ContentDisplay.cfm&ContentID=1814 (3.3.2011).

ADM, 2005: Richtlinie zum Umgang mit Adressen in der Markt- und Sozialforschung. http://www.adm-ev.de/fileadmin/user_upload/PDFS/R07_D.pdf (3.3.2011).

Bates, N., J. Dahlhamer und E. Singer, 2008: Privacy Concerns, Too Busy, or Just Not Interested: Using Doorstep Concerns to Predict Survey Nonresponse. Journal of Official Statistics 24: 591-612.

De Leeuw, E. und W. De Heer, 2002: Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. S. 41-54 in: R. Groves, D. A. Dillman, J. L. Eltinge, and R. J. Little (Hg.): Survey Nonresponse. New York: Wiley.

DeMaio, T. J., 1980: Refusals: Who, Where and Why. Public Opinion Quarterly 44: 223-233.

Dillman, D. A., 2007 (2nd edition): Mail and Internet Surveys. The Tailored Design Method. New York: Wiley.

Dillman, D. A., J. D. Smyth und L. M.Christian, 2008: Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method. New York: Wiley.

Durrant, G. B. und F. Steele, 2009: Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys. Journal of the Royal Statistical Society Series A 172: 361-381.

Erbslöh, B. und A. Koch, 1988: Die Non-Response-Studie zum ALLBUS 1986: Problemstellung, Design, erste Ergebnisse. ZUMA Nachrichten, 22: 29-44. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_22.pdf (3.3.2011).

Esser, H., 1973: Kooperation und Verweigerung im Interview. S. 69-141 in: E. Erbslöh, H. Esser, W. Reschka und D. Schöne (Hg.): Studien zum Interview. Meisenheim am Glan: Hain.

Faulbaum, F., P. Prüfer und M. Rexroth, 2009: Was ist eine gute Frage? Die systematische Evaluation der Fragenqualität. Wiesbaden: VS Verlag.

Früh, W., 2007: Inhaltsanalyse: Theorie und Praxis. Konstanz: UKV.

Fuse, K. und D. Xie, 2007: A Successful Conversion or Double Refusal: A Study of the Process of Refusal Conversion in Telephone Survey Research. Social Science Journal 44: 434-446.

Groves, R. M. und M. P. Couper, 1998: Nonresponse in Household Interview Surveys. New York: Wiley.

Groves, R. M. and S. G. Heeringa, 2006: Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. Journal of the Royal Statistical Society, Series A, 169 (3), 439-459.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer und R. Tourangeau, 2004: Survey Methodology. New York: Wiley.

Groves, R. M.; E. Singer und A. Corning, 2000: Leverage-Saliency Theory of Survey Participation: Description and an Illustration. Public Opinion Quarterly, 64: 299-308.

Költringer, R., 1992: Die Interviewer in der Markt- und Meinungsforschung. Wien: Service Fachverlag.

Kreuter, F. und U. Kohler, 2009: Analyzing Contact Sequences in Call Record Data. Potential and Limitations of Sequence Indicators for Nonresponse Adjustments in the European Social Survey. Journal of Official Statistics 25 (2): 203-226.

Menold, N. und C. Züll, 2010: Codierung von Gründen der Verweigerung der Teilnahme an Interviews: ein Kategorienschema. GESIS-Technical Reports 2010/11. Bonn: GESIS. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2010/ (3.3.2011).

Neller, K., 2005: Kooperation und Verweigerung: Eine Non-Response-Studie. ZUMA Nachrichten 57: 9-36. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/zuma_nachrichten/zn_57.pdf (3.3.2011).

Reuband, K.-H., 1975: Ausfälle in einer mündlichen Befragung. Unveröffentlichtes Manuskript. Köln.

Reuband, K.-H. und J. Blasius, 2000: Situative Bedingungen des Interviews, Kooperationsverhalten und Sozialprofil konvertierter Verweigerer. Ein Vergleich von telefonischen und face-to-face-Befragungen. S. 139-167 in: V. Hüfken (Hg.): Methoden in Telefonumfragen. Opladen: Westdeutscher Verlag.

Schnauber, A. und G. Daschmann, 2008: States oder Traits? Was beeinflusst die Teilnahmebereitschaft an telefonischen Interviews? Methoden, Daten, Analysen 2 (2): 97-123. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.2_Heft_2/03_Schnauber.pdf (3.3.2011).

Stoop, I. A. L., 2004: Surveying Nonrespondents. Field Methods 16: 23-54.

Wasmer, M., Scholz, E. und Blohm, M., 2010: Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" ALLBUS (2008). GESIS-Technical Reports 2010/04. Bonn: GESIS. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2010/ (3.3.2011).

Zeh, J., 1976: Der Verzerrungsfehler durch Ausfälle bei Meinungsbefragungen. Dissertation, Bonn.

Anschrift der Autorinnen

Dr. Natalja Menold
GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Natalja.Menold@gesis.org

Cornelia Züll
GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Cornelia.Zuell@gesis.org

Persönliche Codes bei Längsschnitt-untersuchungen III

Personal Codes in Longitudinal Studies III

Fehlertolerante Zuordnung unverschlüsselter und verschlüsselter selbst-generierter Codes im empirischen Test

The Empirical Test of Fault-Tolerant Linkage of Unencrypted and Encrypted Self-Generated Codes

Andreas Pöge

Zusammenfassung

In Längsschnittuntersuchungen werden oftmals selbstgenerierte persönliche Codes verwendet, um Daten aus mehreren Erhebungszeitpunkten miteinander zu verknüpfen. Aus Gründen der Ausschöpfung und Datenqualität muss die Zuordnungsmethode dabei fehlertolerant sein. In diesem Artikel wird die Qualität der fehlertoleranten Zuordnung von *verschlüsselten* Codes mit der von *unverschlüsselten* Codes verglichen. Um geeignete Daten zu erheben, wurde ein Feldexperiment mit Studierenden durchgeführt. Für die Zuordnung der verschlüsselten Codes wurden die von Schnell, Bachteler und Reiher (2009a) entwickelten und in dieser Zeitschrift vorgestellten Programme „Merge Toolbox", „BloomEncoder" und „Bloom-Comparator" eingesetzt. Die Ergebnisse der Analysen zeigen, dass die fehlertolerante Zuordnungsmethode mit verschlüsselten Codes herkömmlichen Methoden mit unverschlüsselten sogar überlegen ist.

Abstract

In this article the quality of a fault-tolerant linkage of unencrypted and encrypted self-generated codes is compared. To obtain suitable data a field experiment with students was carried out. For the linkage of the encrypted codes the programs "Merge Toolbox", "BloomEncoder" and "BloomComparator" developed by Schnell, Bachteler and Reiher (2009a) were used. The results show that the fault-tolerant linkage with encrypted codes works even better than using unencrypted codes under special conditions.

1 Einleitung

In Panelerhebungen mit sensiblem Befragungsthema besteht oftmals die Notwendigkeit, die Befragten, die zu mehreren Zeitpunkten teilnahmen, einander mit Hilfe von selbstgenerierten persönlichen Codes zuzuordnen. Zum Teil ergibt sich dies aus Gründen der Ausschöpfungsoptimierung, indem durch Zusicherung der Anonymität positive Effekte auf die Teilnahmebereitschaft erzielt werden sollen, zum Teil auch aus datenschutzrechtlichen Vorgaben. Im Bereich der Sozialwissenschaften werden zu diesem Zweck häufig Codes eingesetzt, die durch Fragen zu zeitstabilen persönlichen Merkmalen gebildet werden (vgl. Grube/Morgan/Kearney 1989: 159). Die Daten aus unterschiedlichen Erhebungszeitpunkten werden dann über die gebildeten Codes zugeordnet. Dieses Verfahren ist aus unterschiedlichen Gründen nicht unproblematisch: Der Code muss über eine ausreichende Anzahl geeigneter Fragen gebildet werden, so dass mit ausreichender Sicherheit gewährleistet ist, dass unterschiedliche Befragte auch unterschiedliche Codes aufweisen. Entscheidend ist hier das Zusammenspiel aus der Länge des Codes und der Varianz der einzelnen Stellen bzw. der möglichen Antworten zu den diesbezüglichen Fragen (vgl. Pöge 2005b, 2008).

Darüber hinaus zeigt die Erfahrung, dass die konsistente Beantwortung der persönlichen Fragen zu mehreren Zeitpunkten häufig in nicht unbeträchtlichem Ausmaß scheitert. Dies führt zu „fehlerhaften" Codes, wodurch eine fehlertolerante Zuordnung nötig wird, soll nicht auf einen erheblichen Teil der Befragungspersonen – mit möglicherweise negativen Auswirkungen auf die Datenqualität[1] – verzichtet werden (vgl. detailliert Pöge 2005b: 66f., 2008: 67; Galanti/Siliquini/Cuomo u. a. 2007; Yurek/Vasey/Havens 2008). Im Zusammenhang mit der geforderten Codelänge ist hierbei zu bedenken, dass auch die fehlerhaften Codes eindeutig sein müssen, so dass nicht mehrere Personen dieselben Codes aufweisen. Konkret heißt das, mehrere Codes sollten nicht durch zufällige Fehler an ein oder mehreren Stellen identisch werden, obwohl sie unterschiedlich sein müssten. Wendet man eine einfache fehlertolerante Zuordnung an, bei der Fehler zugelassen werden, indem einzelne Codestellen bei der Zuordnung ignoriert werden, müssen die Codes auch bei Reduktion um die fehlerhafte Stelle eindeutig bleiben.

Reicht diese Form der Anonymisierung über selbstgenerierte persönliche Codes nicht aus, wenn etwa bei besonders sensiblem Untersuchungsgegenstand

1 Es sind unter anderem Verzerrungen im Hinblick auf Geschlecht und (Schul-) Bildung zu erwarten. Dies kann sich beispielsweise verzerrend auf Kriminalitätsraten auswirken (siehe Pöge 2005b: 66f., 2008: 67).

(Kriminalität, Sexualität, Gesundheit etc.) und/oder besonders schützenswerter Population (zum Beispiel minderjährige Befragte) befürchtet wird, aus dem gebildeten Codewort ließen sich Rückschlüsse auf persönliche Merkmale ziehen und somit eine Identifizierung der Probanden ermöglichen, kann eine zusätzliche Verschlüsselung der Codes gewünscht oder sogar gefordert sein.[2] Gleiches gilt für Zuordnungsverfahren, welche auf „Codes" basieren, die aus persönlichen Merkmalen wie beispielweise Namen, Adressen, Geburtsorten etc. in Reinform bestehen und die trotzdem über eine Verschlüsselung ein sehr hohes Maß an Anonymität sichern möchten (siehe Schnell/Bachteler/Reiher 2009a: 204). Auch mit diesen Merkmalen kann es bei mehrfacher Erhebung zu fehlerbehafteten Codes kommen. Sollen diese Codes verschlüsselten einander zugeordnet werden, ist auch hier ein fehlertolerantes Verfahren notwendig.

Bislang war das Wissen über Zuordnungsmethoden mit verschlüsselten Codes limitiert. Von Schnell, Bachteler und Reiher (2009a, b) wurde nun jedoch unlängst ein Verfahren vorgestellt, welches solch ein Vorgehen ermöglicht. Die Autoren stellen darüber hinaus im Rahmen ihres SAFELINK-Projektes mit den frei verfügbaren, plattformübergreifenden Java-Programmen „BloomEncoder" und „BloomComparator"[3] Software zur Verfügung, mit der sowohl die Verschlüsselung als auch die Zuordnung von Codes für die Anwenderin bzw. den Anwender komfortabel handhabbar wird.

Dieses im nächsten Abschnitt näher erläuterte Verfahren soll hier getestet werden. Dabei wird so vorgegangen, dass Codes, die mit Hilfe eines Feldexperimentes gewonnen wurden, zum einen unverschlüsselt und zum anderen über das SAFELINK-Verfahren verschlüsselt zugeordnet werden. Beide Herangehensweisen sollen im Hinblick auf die Zuordnungsperformanz verglichen werden. Es wird herausgearbeitet, dass der Einsatz von verschlüsselten Codes mit dem SAFELINK-Verfahren praktikabel ist und deutliche Vorteile für die Zuordnung im Vergleich zu einem Vorgehen auf Basis von unverschlüsselten Codes bietet. Da der in dem genannten Feldexperiment eingesetzte Code in ähnlicher Weise seit dem Jahr 2002 in dem DFG-Projekt „Kriminalität in der modernen Stadt" eingesetzt wird (siehe Pöge 2005b, 2008), können aufschlussreiche Vergleiche mit der Praxis angestellt werden. Des Weiteren sollen in dieser Arbeit Anwendungsempfehlungen für den Einsatz des SAFELINK-Verfahrens und die bei der eingesetzten Software einzustellenden Parameter entwickelt werden.

2 Diese Variante war in dem DFG-geförderten Forschungsprojekt „Kriminalität in der modernen Stadt" diskutiert worden (vgl. Pöge 2005b, 2008), ließ sich jedoch wegen der damals noch nicht zur Verfügung stehenden Möglichkeiten nicht realisieren.

3 Die genannten Programme können über http://www.uni-due.de/soziologie/schnell_forschung_safelink.php bezogen werden.

1.1 Das SAFELINK-Verfahren

Das SAFELINK-Verfahren basiert darauf, dass Codewörter zunächst in sogenannte *N-Gramme*, das heißt eine Folge aus *N* Zeichen, zerlegt werden. Die Anzahl der Zeichen kann dabei frei bestimmt werden. Daneben kann ausgewählt werden, ob dem ersten und letzten Teilstring ein Leerzeichen voran- bzw. hintangestellt wird. Setzt man die Zahl der Zeichen beispielsweise auf 2, verwendet man also sogenannte *Bigramme*, und verwendet die Leerzeichen zu Beginn und Ende, so würde ein hypothetisches Codewort „ABC1DEF" zerlegt in: „_A AB BC C1 1D DE EF F_". Diese Teilstrings werden dann mit einer wählbaren Anzahl an *kryptografischen Hashfunktionen*[4] jeweils in Form von Einsen an bestimmten Positionen in einem sogenannten *Bloomfilter* mit ebenfalls wählbarer Länge gespeichert (vgl. Schnell/Bachteler/Reiher 2009a: 207ff.). Im Ergebnis wird somit jedes Codewort in einen Bitvektor, also einen Vektor, bestehend nur aus Nullen und Einsen, mit vorzugebender Länge umgewandelt und kann bei einer genügend großen Zahl an eingesetzten Hashfunktionen nicht mehr rekonstruiert werden. Dadurch, dass die Codewörter zunächst in N-Gramme aufgespalten und dann erst im Bitvektor gespeichert werden, ergibt sich trotzdem die Möglichkeit der fehlertoleranten Zuordnung. Wenn nämlich beispielsweise nur an einer Stelle ein „Fehler" zwischen zwei zuzuordnenden Codes besteht, werden nur die N-Gramme, die diese fehlerhafte Codestelle enthalten, an unterschiedlichen Positionen auf dem Bitvektor gespeichert. Die fehlerfreien Stellen werden nach wie vor an denselben Positionen in den Bitvektor als Einsen geschrieben. Ähnliche Codes ähneln sich daher auch in der Form ihrer zugehörigen Bitvektoren (vgl. Schnell/Bachteler/Reiher 2009a: 208).

In Bezug auf die Sicherheit der Verschlüsselung ist zu bemerken, dass eine Entschlüsselung, beispielsweise durch einen illegalen sogenannten „Wörterbuchangriff"[5], immer schwieriger wird, je mehr Hashfunktionen verwendet werden (vgl. Schnell/Bachterler/Reiher 2009a: 211). Insofern ist eine hohe Zahl an Hashfunktionen aus Sicherheitsaspekten wünschenswert. Allerdings steigt bei zunehmender Zahl an Hashfunktionen auch die Wahrscheinlichkeit falsch positiver Treffer, das heißt die Identifikation von zusammengehörigen Codepärchen, die in Wirklichkeit nicht zusammengehören. Insofern muss hier eine Balance gefunden werden. Schnell, Bachteler und Reiher (2009a: 211) halten eine Hashfunktionszahl zwischen

4 Bei diesen Einwegfunktionen kann vom Ergebnis nicht mehr auf den Ausgangswert geschlossen werden (vgl. Schnell/Bachteler/Reiher 2009a: 207).

5 Die Autoren verwenden selbst den Begriff des „Wörterbuchangriffs", der allerdings nur dann kritisch wäre, wenn die zu entschlüsselnde Zeichenkette aus einer sinnvollen Zeichenkette bestünde, was hier nicht der Fall ist. In diesem Zusammenhang wäre ein „Häufigkeitsangriff" eher problematisch, der ebenfalls immer schwieriger wird, je mehr Hashfunktionen verwendet werden.

10 und 30 für vernünftig, eine Anzahl von 15 für „akzeptabel".[6] Ein Ziel der hier durchgeführten Analysen ist die Überprüfung dieser Schwellenwerte auf realer Datenbasis.

1.2 Distanzen und Ähnlichkeiten bei unverschlüsselten und verschlüsselten Codes

Um die Fälle zweier Datensätze mit Hilfe von Codes einander zuzuordnen, müssen diese Codes verglichen und deren Ähnlichkeit analysiert werden. Sind die Codes unverschlüsselt, kann im einfachsten Fall Codestelle für -stelle auf Übereinstimmung oder Fehler überprüft und die Anzahl der Fehler aufsummiert werden. Dieses Vorgehen entspricht der Ermittlung der Hamming-Distanz (HD; Hamming 1950; Leitgöb 2010: 479f.). Dieses Vorgehen kann auch als Konzeption einer Überführung des ersten in den zweiten Code betrachtet werden: Bei der Hamming-Distanz ist hierbei nur eine Operation möglich, nämlich die Substitution einer Stelle. Für jede Substitutionsoperation, die nötig ist, um den ersten in den zweiten Code zu überführen, werden Kosten von 1 berechnet. Die Hamming-Distanz kann dann für zwei Codes *a* und *b* der Länge *T* (Anzahl der Stellen) über diese Kosten bestimmt werden:

$$\mathrm{HD}(a,b) = \sum_{t=1}^{T} c_{st} \ ,$$

mit der Substitutionskostenfunktion:

$$c_{st} = \begin{cases} 0, a_t = b_t \\ 1, a_t \neq b_t \end{cases} .$$

Die maximale Distanz zwischen zwei Codes entspricht damit der Anzahl der Stellen *T* und damit der insgesamt möglichen Fehler. Bei der minimalen Distanz von 0 sind beide Codes identisch. Vergleicht man beispielsweise die fiktiven Codes „ABC1DEF" und „CAB1EFD", ergibt sich eine Distanz von 6, da 6 Stellen des Codes nicht übereinstimmen und bei einer Überführung substituiert werden müssen. Mit Hilfe der Hamming-Distanz kann auch eine auf das Intervall [0,1] normierte Hamming-Ähnlichkeit (HÄ) bestimmt werden:

$$\mathrm{HÄ} = 1 - \frac{\mathrm{HD}}{T} \ .$$

6 Diese Zahlen wurden für Trigramme entwickelt.

Im obigen Beispiel beträgt sie 1 - 6/7 = 0,143.

Eine Erweiterung der Hamming-Distanz stellt die Levenshtein-Distanz (LD; Levenshtein 1966; Leitgöb 2010: 481ff.) dar. Sie ermöglicht zusätzlich zu der Substitutionsoperation noch zwei weitere Operationen (sogenannte Indel-Operationen), das „Einfügen" und das „Löschen", für die ebenfalls (Indel-)Kosten veranschlagt werden. Mit diesen zusätzlichen Operationen wird auch der Vergleich von Strings mit ungleicher Länge möglich. Die Levenshtein-Distanz wird für zwei Codes *a* und *b* der Längen *T* und *T** rekursiv berechnet:

$$\begin{aligned}\mathrm{LD}(a_{t,}b_{t*}) &= \mathrm{d}(a_t, b_{t*}) \\ &= \min\begin{cases} d(a_{t-1}, b_{t*}) + c_i(a_t, \varphi) \rightarrow \text{Löschen von } a_t \\ d(a_{t-1}, b_{t*-1}) + c_s(a_t, b_{t*}) \rightarrow \text{Ersetzen von } a_t \text{ durch } b_t \\ d(a_t, b_{t*-1}) + c_i(\varphi, b_{t*}) \rightarrow \text{Einfügen von } b_t \end{cases}\end{aligned}$$

mit φ als Platzhalter.

Es ist bei der Levenshtein-Distanz prinzipiell möglich, unterschiedliche Kosten für die unterschiedlichen Operationen zu vergeben. Die Software „Merge Toolbox" (Schnell/Bachteler/Reiher 2005)[7], mit der die hier aufgeführten Analysen durchgeführt werden, vergibt allerdings gleiche Kosten für alle Operationen. Wird auf diese Weise die Distanz für die Codes „ABC1DEF" und „CAB1EFD" berechnet, ergibt sich eine Distanz von 4. Der erste Code kann nämlich in den zweiten Code überführt werden, indem zunächst das „C" und das „D" gelöscht (Kosten: 2) und dann an den korrekten Stellen wieder eingefügt (Kosten: 2) werden.[8] Auch mit der Levenshtein-Distanz kann analog eine auf [0,1] normierte Levenshtein-Ähnlichkeit (LÄ) bestimmt werden:

$$\mathrm{LÄ} = 1 - \frac{\mathrm{LD}}{(T + T^*)/2}.$$

Im Beispiel ergibt sich – deutlich abweichend zu der Hamming-Ähnlichkeit – eine Levenshtein-Ähnlichkeit von 1 - 4/7 = 0,429. In den Analysen von Schnell, Bachteler und Reiher (2010) zeigt die Levenshtein-Distanz bzw. -Ähnlichkeit bei der Zuordnungsermittlung mit Hilfe von selbstgenerierten Codes sehr gute Ergebnisse.

7 Das Programm kann über http://www.uni-due.de/soziologie/schnell_forschung_safelink_mtb.php bezogen werden.

8 Zur Berechnung siehe auch http://odur.let.rug.nl/~kleiweg/lev/.

Werden die Codes, die für die Zuordnung verwendet werden sollen, auf die oben geschilderte Art und Weise mit Hashfunktionen und Bloomfiltern verschlüsselt, müssen die gebildeten Filter verglichen werden. Da sie Bitvektoren, bestehend aus Nullen und Einsen, einer bestimmter Länge sind, bietet sich als Ähnlichkeitsmaß der Dice-Koeffizient (DK; Dice 1945) an. Für zwei Bloomfilter F_1 und F_2 wird er folgendermaßen bestimmt:

$$\mathrm{DK}(F_1, F_2) = \frac{2 \cdot c}{a \cdot b} \ ,$$

mit *c* als Anzahl der übereinstimmend auf eins gesetzten Bits in beiden Filtern und *a* sowie *b* als Anzahl der insgesamt auf eins gesetzten Bits in Filter 1 bzw. Filter 2. Der Dice-Koeffizient hat ebenfalls einen Wertebereich von 0 bis 1.

1.3 Zuordnungsverfahren mit unverschlüsselten und verschlüsselten Codes

In der empirischen Praxis steht eine Zuordnung über unverschlüsselte und verschlüsselte Codes vor einer Reihe von Problemen. Zunächst ist häufig nicht bekannt, von wie vielen Personen überhaupt Daten aus *beiden* Erhebungszeitpunkten (t_1 *und* t_2) vorliegen und damit auch, wie viele echt positive Treffer günstigstenfalls gefunden werden können. Darüber hinaus ist unsicher, wie fehlerhaft diese echt positiven Treffer sind, das heißt, wie viele Fehler bei einer fehlertoleranten Zuordnung zugelassen werden müssen.[9] Daneben existiert das Problem der falsch positiven Treffer: Es kann in der Praxis in durchaus beträchtlichem Ausmaß vorkommen, dass zu einem Code ein Code aus dem jeweils anderen Erhebungszeitpunkt weniger Fehler aufweist, als der korrekt dazugehörige Code, und daher möglicherweise falsch zugeordnet würde.[10] Ein falsch positiver Treffer liegt ebenfalls vor, wenn zu einem Code, der überhaupt kein passendes Pendant aus dem zweiten Erhebungszeitpunkt hat, ein Code zugeordnet würde, der zufällig auf einem relativ niedrigem Fehlerniveau vermeintlich „passt".

Ohne eine Validierung der Codezuordnungen sind die aufgezeigten Probleme kaum zu lösen. Eine Möglichkeit der Validierung besteht darin, Handschriftenvergleiche entweder der Codeblätter – wenn sie handschriftlich ausgefüllt wurden – oder der zugehörigen Fragebögen durchzuführen (vgl. Pöge 2005b, 2008). Dies

9 Nach unseren Erfahrungen sind dies mindesten zwei bzw. drei Fehler, bezogen auf einen sechs- bzw. siebenstelligen Code (vgl. Pöge 2005b, 2008 sowie die weiteren Ausführungen in diesem Artikel).

10 Mit zunehmender Länge der Codes nimmt dieses Problem allerdings deutlich ab.

ist allerdings nur dann möglich, wenn handschriftliche Angaben vorliegen und zu einem Vergleich geeignet sind (zum Beispiel offene Fragen).[11] Bei Daten, die keine handschriftlichen Angaben enthalten, kann versucht werden, eine Zuordnungsvalidierung durch die im Fragebogen gegebenen Antworten zu erreichen. Dies verspricht nur Aussicht auf Erfolg, wenn Fragen zu (relativ) zeitstabilen Merkmalen verwendet werden (beispielsweise Geschlecht, Kinderzahl, Körpergröße, Nationalität etc.). Häufig sind diese Angaben allerdings ebenfalls fehlerbehaftet bzw. werden in nicht unerheblichem Ausmaß zu zwei Erhebungszeitpunkten inkonsistent beantwortet oder sind (gerade im Heranwachsendenalter) dann letztlich doch zeitlich instabil. Aus diesen sowie aus Datenschutzgründen muss diese letztere Vorgehensweise als recht problematisch angesehen werden.

Steht eine hinreichend verlässliche Validierungsmethode zur Verfügung, kann bei unverschlüsselten Codes ein mehrstufiges, hierarchisches Zuordnungsverfahren angewendet werden, das auf der Anzahl der Fehler bei Beantwortung der Codefragen basiert (Hamming-Ähnlichkeit) und in Pöge (2005b, 2008) beschrieben wird: In einem ersten Schritt werden alle Codepärchen herausgesucht, die eine Übereinstimmung in dem kompletten Code aufweisen (0 Fehler). Alle Pärchen werden nun einer Validierung unterzogen und die sich als passend erwiesenen aus den Daten herausgenommen. Mit den verbleibenden Codes werden weitere Zuordnungs- und Validierungsschritte unter Zulassung von immer mehr Fehlern durchgeführt. Nach unseren Erfahrungen steigt der Kontrollaufwand mit der Anzahl der zugelassenen Fehler deutlich an. Bei einem siebenstelligen Code und der Zulassung von bis zu drei Fehlern, waren bei einer Zuordnung von Duisburger Jugendlichen zwischen den Jahren 2003 und 2004 beispielsweise rund 3.850 Handschriftenkontrollen nötig, um 2.600 echt positive Treffer zu erhalten (Verhältnis rund 1,5 zu 1; Pöge 2008: 65). Neben der Voraussetzung, überhaupt ein Validierunginstrument zur Verfügung zu haben, ist diese Vorgehensweise vor allem aufgrund des hohen Arbeitsaufwandes problematisch. Sowohl die Durchführung der Handschriftenvergleiche als auch die datenverarbeitungstechnische Aufbereitung (Erstellung von Kontrolllisten und Herausnahme der Codes bei validierter Zuordnung bzw. Führen von Listen mit als nicht passend identifizierten Pärchen) ist sehr arbeits- und damit kostenintensiv.

Soll die sich als sehr geeignet herausgestellte Levensthein-Ähnlichkeit (vgl. Schnell/Bachteler/Reiher 2010) angewendet werden oder sollen verschlüsselte Codes zum Einsatz kommen, ist eine hierarchische fehlertolerante Vorgehensweise

11 Dies Verfahren ist nicht unproblematisch, da nach unseren Erkenntnissen viel Erfahrung nötig ist, um zusammengehörige Handschriften zu erkennen. Hier spielt sicherlich das Alter der Befragten und damit zusammenhängend deren veränderliche Handschriften eine Rolle. Aber auch der Einfluss zum Beispiel der Stiftfarbe darf bei den Kontrollen nicht unterschätzt werden.

auf Basis von Fehlern in der Beantwortung der einzelnen Codefragen nicht mehr möglich.[12] Soll dennoch hierarchisch vorgegangen werden, kann man zunächst über die vorgestellten Distanzmaße die Ähnlichkeiten zwischen *allen* Codes zweier Erhebungszeitpunkte bestimmen. Dann kann aus der entstehenden Ähnlichkeitsmatrix pro Code des einen Erhebungszeitpunktes der Code aus dem jeweils anderen Erhebungszeitpunkt ermittelt werden, zu dem die größte Ähnlichkeit besteht.[13] Nun können die Codezuordnungen hierarchisch validiert werden, das heißt, die Codepärchen sollten mit absteigenden Ähnlichkeiten, beispielsweise mit Hilfe von Handschriftenvergleichen der Fragebogen, überprüft werden. Passende Pärchen können herausgeschrieben und die Codes, so sie mit niedrigerer Ähnlichkeit nochmals vorkommen, aus den Daten gelöscht werden. Wird so vorgegangen, stellt sich die Frage, bis zu welchem Ähnlichkeitsniveau der Codepärchen Zuordnungsvalidierungen sinnvollerweise durchgeführt sollten. Im optimalen Fall sollten die Validierungen nur bis zu dem Ähnlichkeitsniveau des unähnlichsten echt positiven Treffers vorgenommen werden. Leider ist dies Niveau in der Praxis unbekannt und man ist auf die Anwendung mehr oder weniger plausibler Schwellenwerte angewiesen. Diese Schwellenwerte entsprechen dann der minimalen Ähnlichkeit, die zwei Codes mindestens aufweisen müssen, um als zusammengehörig identifiziert zu werden.[14]

Der Vorteil dieser Vorgehensweise liegt darin, dass ein beliebiges Ähnlichkeitsmaß verwendet werden kann, welches nicht zwingend eine Entsprechung in der Fehleranzahl der zugrunde liegenden Codes hat, dafür aber möglicherweise besser geeignet ist[15] und/oder auch bei verschlüsselten Codes verwendet werden kann.

Um die Qualität der fehlertoleranten Zuordnung verschlüsselter Codes im Vergleich zu unverschlüsselten zu testen, wird in den nachfolgenden Ausführungen die zuletzt geschilderte Vorgehensweise gewählt. Die Datengrundlage bildet ein selbstgenerierter persönlicher Code, der seit dem Jahr 2002 in der DFG-geförderten Panel-

12 Bei der Levensthein-Ähnlichkeit können Codepärchen mit gleicher Fehlerzahl in den zugrunde liegenden Codefragen unterschiedliche Ähnlichkeiten aufweisen, und bei der Verwendung von verschlüsselten Codes gehen die Informationen über die Fehler in den Codefragen gänzlich verloren. Als Ausnahme gilt hier allerdings die Zuordnung exakt gleicher Codes. Diese Pärchen weisen mit Levenshtein-Ähnlichkeit bzw. in verschlüsselter Form ebenso wie mit der Hamming-Ähnlichkeit immer die maximale Ähnlichkeit auf.

13 Die Praxis zeigt, dass dann allerdings noch eine nicht unerhebliche Anzahl an falsch positiven Treffern identifiziert wird. Existieren nämlich Codes, die zu keinem Code aus dem jeweils anderen Datensatz gehören, werden diese Codes dennoch herausgeschrieben, da sie zu irgendeinem Code eine größtmögliche Ähnlichkeit aufweisen (und sei sie auch Null).

14 Im Falle der Hamming-Ähnlichkeit und unverschlüsselten Codes entspricht dies dem oben vorgestellten Verfahren.

15 Die unten vorgestellten Analysen zeigen beispielsweise, dass die Levenshtein-Ähnlichkeit der Hamming-Ähnlichkeit bei unverschlüsselten Codes deutlich überlegen ist. Schnell, Bachteler und Reiher (2010) schlagen als Schwellenwert für die Mindestähnlichkeit zweier Codes bei der Levenshtein-Distanz einen Wert von 0,34 vor, was einer Levenshtein-Ähnlichkeit von 0,66 entspricht.

studie „Kriminalität in der modernen Stadt“[16] eingesetzt wird (Pöge 2005a, 2007; Pollich 2010) und stetig weiterentwickelt und verbessert wurde (Pöge 2005b, 2008). Für die hier durchgeführten Analysen wurde der Code in seiner letzten, siebenstelligen Version verwendet und in einem Experiment mit Bielefelder Studierenden erhoben. Um die tatsächlich zusammengehörigen Codes identifizieren zu können, wurde zusätzlich die Matrikelnummer der Probanden erfragt. Die Beschreibung dieses Feldexperiments und der resultierenden Daten ist Gegenstand des nächsten Abschnittes.

2 Experiment

Am 21. Oktober 2009 wurde das in Abbildung 1 dargestellte Codeblatt den Teilnehmerinnen und Teilnehmern der Pflichtvorlesung „Einführung in die Methoden der quantitativen empirischen Sozialforschung“ zum Ausfüllen vorgelegt. Der Besuch dieser Veranstaltung ist laut Studienablaufempfehlung für das erste Semester vorgesehen und – neben anderen – für die BA-Studiengänge „Soziologie“, „Sozialwissenschaften“ und „Politikwissenschaft“ obligatorisch. Es gaben 354 Studierende ein ausgefülltes Codeblatt ab, bei 3 Blättern wurde die Angabe der Matrikelnummer verweigert (351 gültige Fälle zu t_1).

Ein halbes Jahr später, am 21. April 2010, wurde das leicht modifizierte Codeblatt[17] den Teilnehmerinnen und Teilnehmern der Pflichtvorlesung „Statistik I“ zur Beantwortung vorgelegt. Der Besuch dieser Veranstaltung wird laut Studienablaufempfehlung für das zweite Semester empfohlen und ist ebenfalls – neben anderen – für die BA-Studiengänge „Soziologie“, „Sozialwissenschaften“ und „Politikwissenschaft“ verpflichtend. Bei einem regulären Studienablauf sollte der Teilnehmerkreis der beiden Veranstaltungen demnach in weiten Teilen deckungsgleich sein. Zum zweiten Zeitpunkt füllten 245 Personen ein Codeblatt aus, bei 2 Blättern wurde die Angabe der Matrikelnummer verweigert (243 gültige Fälle zu t_2).

Eine erste Analyse der erhobenen Daten aus beiden Zeitpunkten zeigt zunächst, dass alle aus der Beantwortung der Fragen gebildeten (Komplett-)Codes in t_1 und in t_2 nur je einmal vorkommen. Sie sind also eindeutig und gut geeignet, die Fälle zu identifizieren bzw. zu differenzieren. Ein Problem mit der Identifikation der Personen eines Zeitpunktes gibt es mit dem siebenstelligen Code bei der Anzahl der Teilnehmerinnen und Teilnehmer dementsprechend nicht. Dies ist aufgrund der relativ geringen Teilnehmerzahl und der Länge auch nicht verwunderlich (vgl. Pöge 2005b: 57ff., 2008: 61ff.).

16 Siehe http://www.krimstadt.de/.

17 Zusätzlich zu den vorhandenen Fragen sollte nun auch das Geschlecht und das Geburtsjahr angegeben werden.

Abbildung 1 Das eingesetzte Codeblatt (Zeitpunkt t_1)

Matrikelnummer: _____

Erstellung des persönlichen Codes

Liebe Teilnehmerin, lieber Teilnehmer,

da wir Ihren Fragebogen dem des letzten Jahres ohne Ihren Namen zuordnen wollen, ist es wichtig, dass Sie sich an Ihren persönlichen Code vom letzten Jahr erinnern. Denn nur so können Ihre Fragebögen einander zugeordnet werden, ohne dass jemand herausfinden kann, wer diese Fragebögen ausgefüllt hat. Wichtig ist also, dass Sie denselben Code noch wissen. Aus diesem Grund haben wir die nachfolgenden Fragen formuliert, die Ihnen helfen sollen, sich an Ihre persönliche Kombination zu erinnern.

Bitte kreuzen Sie bei jeder der sieben Fragen immer nur ein Feld an!
Wenn Sie eine der Fragen überhaupt nicht beantworten können, kreuzen Sie bitte kein Feld an!

Hier nun die sieben Fragen zur Erstellung Ihres persönlichen Codes:

1	Bitte kreuzen Sie den **ersten** Buchstaben des Vornamens Ihres **Vaters** (oder einer Person, die für Sie einem Vater am nächsten kommt) an. (z. B. Anton, Bernd, Hans-Peter usw.). a b c d e f g h i j k l m n o p q r s t u v w x y z ä ö ü ß
2	Bitte kreuzen Sie den **ersten** Buchstaben des Vornamens Ihrer **Mutter** (oder einer Person, die für Sie einer Mutter am nächsten kommt) an. (z. B. Anna, Beate, Jutta, Maria, usw.). a b c d e f g h i j k l m n o p q r s t u v w x y z ä ö ü ß
3	Bitte kreuzen Sie den **ersten** Buchstaben Ihres **Vornamens** an (z. B. Michael, Thomas, Ute usw.). a b c d e f g h i j k l m n o p q r s t u v w x y z ä ö ü ß
4	Bitte kreuzen Sie den **Tag** Ihres **Geburtsdatums** an (z. B. Geburtstag am 7. Januar = 7, am 12. Mai = 12, am 31. Oktober = 31). 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
5	Bitte kreuzen Sie den **letzten** Buchstaben Ihrer natürlichen **Haarfarbe** an. (z. B. braun, Glatze, schwarz, usw.). a b c d e f g h i j k l m n o p q r s t u v w x y z ä ö ü ß
6	Bitte kreuzen Sie den **letzten** Buchstaben Ihrer **Augenfarbe** an. (z. B. braun, grün, grau, usw.). a b c d e f g h i j k l m n o p q r s t u v w x y z ä ö ü ß
7	Bitte kreuzen Sie den **letzten** Buchstaben Ihres **Nachnamens** an (Sollten Sie Ihren Namen gewechselt haben, nehmen Sie Ihren Geburtsnamen!) a b c d e f g h i j k l m n o p q r s t u v w x y z ä ö ü ß

Um zu überprüfen, wie viele Fehler die Befragten im Vergleich der beiden Erhebungszeitpunkte beim Ausfüllen der Codeblätter machten bzw. wie viele Inkonsistenzen in den einzelnen Fragen auftreten, wurden die Fälle aus beiden Zeitpunkten mit gleichen Matrikelnummern einander zugeordnet. Wie Tabelle 1 entnommen werden kann, beläuft sich die Gesamtzahl der Personen auf n=187. Auch wenn

nicht völlig auszuschließen ist, dass hier Fälle zugeordnet wurden, die durch einen Fehler gleiche Matrikelnummern aufweisen, aber nicht zusammengehören, erscheint dies doch äußerst unwahrscheinlich. Im weiteren Verlauf der Analysen werden die genannten 187 Pärchen als tatsächlich zusammengehörend behandelt.

Vergleicht man die Codes, die diese Pärchen zwischen t_1 und t_2 aufweisen, kann zunächst die Anzahl der Beantwortungsfehler analysiert werden. Das Ergebnis dieser Analyse findet sich in Tabelle 1 – es fällt relativ ernüchternd aus. Von den 187 Personen waren lediglich drei Viertel (75,4 %) in der Lage, in beiden Erhebungszeitpunkten die sieben Codefragen gleich zu beantworten. Vergleicht man diese Ergebnisse mit denen aus der oben genannten Schüleruntersuchung, in der ein sehr ähnlicher Code verwendet wurde, ist die Quote jedoch deutlich besser: Bei der Zuordnung zwischen den Jahren 2002/2003 waren nur 45,3 % der 16-jährigen Münsteraner Jugendlichen bei einem ähnlichen (fünfstelligen jedoch schwierigeren) Code zu fehlerfreier Beantwortung in der Lage (vgl. Pöge 2005b: 54). In Duisburg 2003/2004 lag die Quote mit 58,3 % im Vergleich zu Münster deutlich höher, wobei hier durchschnittlich 15-jährige Schülerinnen und Schüler mit einem sechsstelligen Code befragt wurden (vgl. Pöge 2008: 65). Wenn man allerdings berücksichtigt, dass an dem hier durchgeführten Experiment nur Personen mit deutlich höherem Bildungsgrad (Studierende mit zumeist Abitur) und deutlich höherem Alter (durchschnittlich 23 Jahre zu t_2) teilnahmen und sich den bewusst einfachen Charakter der Fragen verdeutlicht, ist das Ergebnis absolut gesehen überraschend schlecht. Es verdeutlicht noch einmal drastisch das generelle Problem der Zuordnung über persönliche Codes und die Notwendigkeit einer fehlertoleranten Zuordnungsmethode.

Tabelle 1 Anzahl der Fehler bei der Beantwortung der Codefragen zu beiden Entstehungszeitpunkten (n=187)

Fehlerzahl	Anzahl	Prozent
0	141	75,4
1	39	20,9
2	5	2,7
3	2	1,1
Gesamt	187	100,0

Eine differenziertere Betrachtung der Fragen, bei denen Beantwortungsfehler gemacht wurden, offenbart, dass deren Anzahl durchaus abhängig vom „Schwierigkeitsgrad" der jeweiligen Fragen ist (siehe Tabelle 2). Die Fragen nach dem *ersten Buchstaben* der Elternvornamen, des eigenen Vornamens und des Tages des Geburtsdatums funktionieren augenscheinlich recht zuverlässig. Die Fehlerquote

liegt hier bei maximal 1,6 %. Demgegenüber waren die Fragen nach den jeweils *letzten Buchstaben* der eigenen Haar- und Augenfarbe aber auch des eigenen Nachnamens deutlich weniger erfolgreich (Fehlerquote zwischen 6,4 und 10,7 %, siehe auch Yurek/Vasey/Havens 2008). Obwohl man annehmen sollte, dass alle Fragen von der hier befragten Klientel korrekt zu beantworten sein müssten, ist indes schlicht zu konstatieren, *dass dies nicht der Fall ist.* Hauptgrund mögen Konzentrations- oder Motivationsprobleme sein, die aller Wahrscheinlichkeit nach jedoch nicht spezifische Probleme des hier geschilderten Experimentes sind. Vielmehr lässt sich vermuten, dass diese Probleme in dem überwiegenden Teil aller realen Befragungssituationen zutage treten. Jedenfalls ist nicht verwunderlich, dass bei den genannten Schülerbefragungen mit deutlich jüngeren Jugendlichen aus allen Schulformen die Ergebnisse in Bezug auf die Reproduktion des Codes noch deutlich schlechter als die der Studierenden sind.[18]

Tabelle 2 Fehler bei der Beantwortung der einzelnen Codefragen (*n*=187)

Frage		Anzahl	Prozent
1	Vorname Vater (erster Buchstabe)	3	1,6
2	Vorname Mutter (erster Buchstabe)	2	1,1
3	eigener Vorname (erster Buchstabe)	1	0,5
4	Tag Geburtsdatum	2	1,1
5	eigene Haarfarbe (letzter Buchstabe)	20	10,7
6	eigene Augenfarbe (letzter Buchstabe)	15	8,0
7	eigener Nachname (letzter Buchstabe)	12	6,4

Es stellt sich mit den vorherigen Resultaten die Frage, welche Auswirkungen bzw. Verzerrungen die oben genannten Sachverhalte mit sich bringen. Um den Befragungsaufwand möglichst gering zu halten, wurde in dem hier durchgeführten Experiment allerdings – neben der Frage nach dem Geburtsjahr – einzig die Frage nach dem Geschlecht der Befragten mit erhoben.[19] Eine entsprechende Auswertung zeigt die schon bekannten Befunde (vgl. Pöge 2005b: 66): Frauen gelingt es deutlich besser, die Codefragen zu zwei Zeitpunkten gleich zu beantworten (siehe Tabelle 3).

18 Am Rande sei erwähnt, dass sich hier die Fragen aufdrängen, wie korrekt inhaltliche Fragen mit meistenteils deutlich höheren Schwierigkeitsgraden generell in schriftlichen Umfragen beantwortet werden und welche Auswirkungen die sicherlich auftretenden Fehler in Bezug auf die Datenqualität haben.

19 Unsere Erkenntnisse aus den Schülerbefragungen lassen eine Verzerrung im Hinblick auf den Bildungsgrad (Schulform) der Befragten vermuten (vgl. Pöge 2005b: 66), die hier aufgrund der diesbezüglichen Homogenität der Stichprobe ohnehin nicht messbar wäre.

Tabelle 3 Fehler bei der Beantwortung der Codefragen nach Geschlecht

	Fehler											
	0		1		2		3		Gesamt		Gesamt t_2	
Geschlecht	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
weiblich	80	57,1	17	43,6	1	20,0	1	50,0	99	53,2	120	49,6
männlich	60	42,9	22	56,4	4	80,0	1	50,0	87	46,8	122	50,4
Gesamt	140	100,0	39	100,0	5	100,0	2	100,0	186	100,0	242	100,0

Während wir in der Gesamtdatei des Zeitpunktes t_2 ein nahezu paritätisches Geschlechterverhältnis vorfinden, ist dies bei den Personen, die keinen Fehler bei der Beantwortung machten, deutlich in Richtung der Frauen verschoben (Frauen: 57,1 %; Männer: 42,9 %). Nur durch eine fehlertolerante Zuordnungsmethode ist es möglich, das diesbezügliche Missverhältnis zu verbessern. Das Geschlechterverhältnis aller 186 zusammengehörenden Fälle (mit gültigen Angaben bei der Frage nach dem Geschlecht), die nur durch das Zulassen von bis zu 3 Fehlern bei den Codefragen ermittelt werden können, beträgt 53,2 % (Frauen) zu 46,8 % (Männer). Die Differenz zu dem Verhältnis des Gesamtdatensatzes (t_2) lässt sich vermutlich durch ein verzerrtes generelles Verweigerungsverhalten erklären. Anscheinend haben mehr Männer die Teilnahme komplett verweigert oder das Studium abgebrochen, so dass sie an der zweiten Erhebung nicht mehr teilnahmen.

3 Analyse

Um einen Test der Zuordnungsmethoden (unverschlüsselt vs. verschlüsselt) durchzuführen, wird eine Vorgehensweise gewählt, die derjenigen im tatsächlichen Anwendungsfall entspricht oder entsprechen könnte (siehe Abschnitt 1.3). Das bedeutet, die Anwenderin bzw. der Anwender hat vorab keine Kenntnis darüber, wie viele Personen, die zu t_1 befragt wurden, zu t_2 ebenfalls an der Befragung teilnahmen. Die mögliche Vorgehensweise wurde in Abschnitt 1.3 vorgestellt und wird in allen durchgeführten Versuchen gleichermaßen angewandt. Sie besteht aus vier Schritten:

Schritt 1: Berechnung der Ähnlichkeit zwischen allen Fällen aus t_1 und allen Fällen aus t_2 auf Grundlage des gewählten Distanzmaßes.[20]

Schritt 2: Zu jedem Fall aus t_1 wird der Fall aus t_2 ermittelt, zu dem auf Grundlage des jeweiligen Distanzmaßes die größte Ähnlichkeit besteht. Gibt es mehrere Pärchen mit maximaler Ähnlichkeit, werden sie alle verwendet. Die Datei wird abgespeichert und ausgewertet.

Schritt 3: Analog wird zu jedem Fall aus t_2 der Fall aus t_1 bestimmt, zu dem auf Grundlage des jeweiligen Distanzmaßes die größte Ähnlichkeit besteht. Gibt es mehrere Pärchen mit maximaler Ähnlichkeit, werden sie alle verwendet. Die Datei wird abgespeichert und ausgewertet.

Schritt 4: Beide Dateien werden zusammengeführt, Duplikate werden entfernt; danach wird die Gesamtdatei ausgewertet.

Im realen Anwendungsfall würde sich, wie in Abschnitt 1.3 geschildert, eine (hierarchische) Zuordnungsvalidierung der herausgeschriebenen Codepärchen, beispielsweise über Handschriftenvergleiche, anschließen. Auf diesen Schritt wird hier aus Zeit- und Kostengründen verzichtet. Stattdessen soll die Anzahl der ermittelten Codepärchen für einen Vergleich der Leistungsfähigkeit der jeweiligen Zuordnungsmethoden verwendet werden. Je geringer die Anzahl der als zusammengehörig herausgeschriebenen Codes ist, je geringer also die Anzahl der in der Praxis zu validierenden Codepärchen ist, desto effizienter ist das Verfahren, sofern auch alle echt positiven Treffer identifiziert werden.

Um zu verdeutlichen, warum die Schritte 2 *und* 3 drei nötig sind, das heißt die Zuordnungsrichtung von t_1 nach t_2 ($t_1 \rightarrow t_2$) und von t_2 nach t_1 ($t_1 \leftarrow t_2$) bzw. die Zusammenführung der Kombinationen ($t_1 \leftrightarrow t_2$) sei hier ein fiktives Beispiel vorgestellt: Zum Zeitpunkt t_1 liegen die Fälle a und b, zum Zeitpunkt t_2 die Fälle c und d vor. Die Ähnlichkeiten zwischen diesen vier Fällen seien die in Tabelle 4 dargestellten.

20 Auf ein technisches Problem bei der Durchführung sei hingewiesen: Es ist zu berücksichtigen, dass in Schritt 1 zunächst große Datenmengen anfallen, da zu jedem Fall aus t_1 die Ähnlichkeit zu jedem Fall aus t_2 bestimmt werden muss. In dem hier geschilderten Experiment entstehen Dateien mit 351*243 = 85.293 Fällen. Bei zwei Datensätzen mit je 2.000 Fällen ergäbe sich eine Fallzahl von 4 Millionen. Hier stellt sich die Frage, ob das Statistikprogramm im konkreten Anwendungsfall mit einer solch hohen Fallzahl umgehen kann. Die „Merge Toolbox" ermöglicht eine Beschränkung der auszugebenden Fälle und ist auch daher sehr empfehlenswert.

Tabelle 4 Ähnlichkeiten für ein fiktives Beispiel mit 4 Fällen

	c	d
a	1	3
b	2	4

Geht man nach Schritt 1 vor, sucht also zu jedem Fall aus t_1 den Fall mit der höchsten Ähnlichkeit aus t_2, ergeben sich die Kombinationen a-d und b-d. Geht man umgekehrt vor und sucht zu jedem Fall aus t_2 den Fall mit der höchsten Ähnlichkeit aus t_1, ergeben sich die Kombinationen c-b und d-b, wovon letztere schon vorhanden ist. Mit Schritt 4 würden somit 3 Kombinationen verbleiben (a-d, b-d und c-b), die untersucht werden müssten. Im weiteren Verlauf der hier vorgestellten Analysen wird deutlich, dass dieses Vorgehen tatsächlich nötig ist, um immer die komplette Anzahl der tatsächlich zusammengehörigen Pärchen zu ermitteln.

3.1 Zuordnung der unverschlüsselten Codes

Zunächst wird das oben genannte Vorgehen auf Grundlage des nicht verschlüsselten kompletten Codes und der Hamming- sowie der Levenshtein-Ähnlichkeit durchgeführt. Die Bestimmung der Ähnlichkeiten erfolgt mit Hilfe des oben erwähnten Programmes „Merge Toolbox" (MTB). Die Ergebnisse sind in Tabelle 5 dargestellt. Es findet sich in der ersten Spalte das gewählte Ähnlichkeitsmaß, in der zweiten Spalte die Zuordnungsrichtung, so wie im vorangegangenen Abschnitt beschrieben. In der Spalte „Total" wird die Anzahl der Kombinationen bzw. Pärchen ausgewiesen, die sich für die jeweiligen Zuordnungsrichtungen ergeben. In der ersten Zeile bedeutet der ausgewiesene Wert, dass sich für die 351 Fälle aus t_1 786 Kombinationen ergeben, wenn zu jedem einzelnen Fall aus t_1 die Fälle mit maximaler Ähnlichkeit aus t_2 zugeordnet werden. Es sind deshalb mehr als 351 Fälle, da bei mehrfachem Vorkommen der maximalen Ähnlichkeit alle Kombinationen in den Daten belassen werden.

In der Spalte „Treffer" wird aufgelistet, wie viele korrekte Pärchen bzw. echt positive Treffer in diesen Kombinationen enthalten sind. In der Spalte „Treffer in %" wird der prozentuale Absolutwert dargestellt (187 korrekte Treffer sind es insgesamt, daher entspricht 187 100 %). In der Spalte „Trefferquote in %" findet sich das prozentuale Verhältnis der echt positiven Treffer zu den insgesamt ermittelten Kombinationen („Total"). In der Spalte „kein Treffer" steht die Anzahl der falsch positiven Treffer, was der Differenz zwischen „Total" und „Treffer" entspricht.

Tabelle 5 Ergebnisse für den unverschlüsselten kompletten Code (Ähnlichkeitsmaße: Hamming, Levenshtein)

	Richtung	kein Treffer	Treffer	Treffer in %	Total	Trefferquote in %	minimale Ähnlichkeit Treffer ($ä_{min}$)	Fälle mit $ä \geq ä_{min}$
Hamming	$t_1 \rightarrow t_2$	599	187	100,00	786	23,79	0,571	513
	$t_1 \leftarrow t_2$	252	187	100,00	439	42,60	0,571	330
	$t_1 \leftrightarrow t_2$	789	187	100,00	976	19,16	0,571	600
Levenshtein	$t_1 \rightarrow t_2$	449	187	100,00	636	29,40	0,571	358
	$t_1 \leftarrow t_2$	173	187	100,00	360	51,94	0,571	275
	$t_1 \leftrightarrow t_2$	578	187	100,00	765	24,44	0,571	413

Grau hinterlegt: Versuche, in denen alle Treffer identifiziert werden.

Die Spalte „minimale Ähnlichkeit Treffer ($ä_{min}$)" weist die Ähnlichkeit des unähnlichsten echt positiven Treffers (minimale Ähnlichkeit) aus. Dieser Wert lässt sich für die Hamming-Distanz anschaulich interpretieren: 0,571 ergibt sich nach der oben dargestellten Formel bei 3 Fehlern im siebenstelligen Code (1-3/7) und wie in Tabelle 1 aufgezeigt, beträgt die Maximalzahl der Fehler bei den korrekt zusammengehörigen Pärchen 3.

Für die Zuordnung im praktischen Anwendungsfall bzw. die Zuordnungsvalidierung ist eine weitere Kennzahl bedeutsam: Wie viele Pärchen werden identifiziert, deren Codes eine gleiche bzw. größere Ähnlichkeit zeigen als der unähnlichste echt positive Treffer? Geht man nämlich bei der Zuordnungsvalidierung, wie oben beschrieben, hierarchisch vor, wäre dies die Anzahl der Codezuordnungen, die kontrolliert werden müssten, um alle echt positiven Treffer zu identifizieren. In der Spalte „Fälle mit $ä \geq ä_{min}$" wird daher die Anzahl der Pärchen angegeben, die gleiche oder größere Ähnlichkeit als der unähnlichste echt positive Treffer haben. Beispielsweise existieren 600 Codekombinationen mit einem Ähnlichkeitsniveau von größer/gleich 0,571 (also mit drei oder weniger Fehlern), wenn man die Hamming-Ähnlichkeit und die Zuordnungsrichtung $t_1 \leftrightarrow t_2$ betrachtet. Diese müssten in der Praxis, wie in Abschnitt 1.3 dargestellt, hierarchisch validiert werden. Es gilt dabei prinzipiell: Je niedriger diese Zahl ist, desto besser werden die Treffer von den Nicht-Treffern separiert und desto weniger Kontrollen müssten im praktischen Anwendungsfall durchgeführt werden. Bei der effizienteren Levenshtein-Ähnlichkeit sind es zum Beipiel nicht mehr 600 sondern nur 413 Codekombinationen.

Diese Anzahl reduziert sich in der Praxis allerdings noch dadurch, dass bei der erfolgreichen Validierung eines Codepärchens auf höherem Ähnlichkeitsniveau alle weiteren Zuordnungsvorschläge mit den beiden entsprechenden Codes auf niedrigerem Ähnlichkeitsniveau eliminiert werden können. Der Umfang dieser

Reduzierung ist schwer abzuschätzen. Unserer Erfahrung nach ist es möglich, bei unverschlüsselten Codes und Benutzung der Hamming-Distanz auf einen Wert von 1,5 zu 1 (Handschriftenkontrollen : Treffer) zu kommen (siehe oben). Diese Reduzierung bei hierarchischer Vorgehensweise ergibt sich indes bei allen Distanzmaßen und auch bei der Verwendung von verschlüsselten Codes. Um den zu erwartenden Aufwand für Validierungen zu vergleichen, sollen die in der Spalte „Fälle mit ä ≥ $ä_{min}$" ausgewiesenen Werte als Indikatoren betrachtet und verglichen werden, obwohl sich deren absolute Werte im realen Anwendungsfall noch deutlich reduzieren würden.

Neben der Möglichkeit zur vergleichenden Analyse sind die ausgewiesenen Werte auch im Hinblick auf die Entwicklung von Schwellenwerten bedeutsam, bis zu welchem Ähnlichkeitsniveau eine kosten- und zeitintensive Validierung sinnvoll ist. Das Ziel ist ja, möglichst alle echt positiven Treffer zu identifizieren. Unsere Ergebnisse legen hier mit 0,571 eine untere Grenze für einen Schwellenwert nahe, der die Identifikation falsch negativer Treffer verhindert. Das heißt, es wird mit diesem Wert im vorgestellten Fall ausgeschlossen, dass korrekte Treffer nicht gefunden werden. Dieser Schwellenwert kann in der Praxis allerdings zu vielen falsch positiven Treffern führen. Hier muss im Anwendungsfall entschieden werden, ob der durchzuführende Validierungsaufwand im Verhältnis zu dem zu erwartenden Ertrag an echt positiven Treffern steht und der Schwellenwert unter Umständen nachjustiert werden. Schnell, Bachteler und Reiher (2010) empfehlen für die Levenshtein-Ähnlichkeit einen etwas höheren Wert (0,66), der sich im Hinblick auf einen möglichst guten Ausgleich zwischen den Anzahlen an falsch negativen und falsch positiven Pärchen als günstig erwiesen hat. Dieser Wert schließt jedoch nicht aus, dass korrekte Treffer nicht identifiziert werden.

Vergleicht man die Ergebnisse der Zuordnungen über die Hamming- und die Levenshtein-Ähnlichkeit, kann man festhalten, dass in beiden Fällen alle 187 passenden Pärchen ermittelt werden. Dies gilt für die Zuordnungsrichtungen von t_1 nach t_2, von t_2 nach t_1 und logischerweise auch für die zusammengesetzte Datei aus beiden Zuordnungsrichtungen. Die Zuordnungen über die beiden Ähnlichkeitsmaße unterscheiden sich allerdings nicht unerheblich in der Anzahl der falsch positiven Pärchen (Spalten „kein Treffer") und damit auch in der Gesamtzahl der Pärchen (Spalten „Total" und „Trefferquote in %"). Hier schneidet die Levenshtein-Ähnlichkeit deutlich besser ab. Die Trefferquote, also der prozentuale Anteil der Treffer an der Gesamtzahl, liegt mit 19 % (Hamming) und 24 % (Levenshtein) auf einem recht niedrigen Niveau. Nach den hier dargestellten Ergebnissen könnte es empfehlenswert sein, in einem ersten Schritt nur die Zuordnungsrichtung von t_2 nach t_1 zu betrachten (Zeilen $t_1 \leftarrow t_2$). Es werden alle echt positiven Pärchen identifiziert und die Trefferquote liegt mit 43 % (Hamming) und 54 % (Levenshtein) deutlich höher.

Zu beachten ist allerdings, dass diese Quoten nur im Hinblick auf die theoretisch bestmöglichen Zuordnungsquoten beurteilt werden dürfen. Mit der gewählten Vorgehensweise ist die Minimalzahl der herausgesuchten Pärchen mit maximaler Ähnlichkeit und damit die bestmögliche Quote nämlich vorgegeben (Spalte „Total"): Da zu t_1 351 Fälle vorliegen, stellt diese Zahl das Minimum für $t_1 \rightarrow t_2$ dar (wenn alle Fälle aus t_1 jeweils nur ein Pendant aus t_2 zugewiesen bekommen). Für diese Zuordnungsrichtung kann also auch keine bessere Trefferquote als 53,3 % erreicht werden. Für die Richtung $t_2 \rightarrow t_1$ ist das Minimum 243, da in t_2 243 Fälle enthalten sind. Daher kann hier eine Trefferquote von 77,0 % nicht überschritten werden. Für $t_1 \leftrightarrow t_2$ hängt das Minimum von einer weiteren Bedingung ab. Sind alle 243 Pärchen ($t_1 \leftarrow t_2$) in den 351 ($t_1 \rightarrow t_2$) enthalten, liegt es bei 351. Sind alle 243 nicht in den 351 enthalten, liegt es bei 594, ansonsten im Bereich dazwischen. Das entspricht maximalen Trefferquoten von 53,3 % bis 31,5 %. Die in der Tabelle dargestellten Trefferquoten müssen also immer in Relation zu diesen theoretisch überhaupt nur möglichen Quoten interpretiert werden.[21]

Auch die Indikatoren für die zu erwartende Anzahl der im konkreten Anwendungsfall durchzuführenden Validierungen („Fälle mit ä ≥ $ä_{min}$") zeigen für alle Zuordnungsrichtungen eine deutliche Überlegenheit der Levenshtein-Ähnlichkeit.

Insgesamt stellt sich in den hier dargestellten Ergebnissen eine deutlich bessere Performanz der Levenshtein- gegenüber der Hamming-Ähnlichkeit bei der Zuordnung von unverschlüsselten Codes dar. Bei der Zuordnung von unverschlüsselten Codes sollte daher der Levenshtein-Ähnlichkeit Vorzug vor der Hamming-Ähnlichkeit gegeben werden.

3.2 Zuordnung der verschlüsselten Codes

Um die Leistungsfähigkeit des Zuordnungsverfahrens mit verschlüsselten Codes zu überprüfen und sie mit derjenigen eines Verfahrens über unverschlüsselte Codes zu vergleichen, wurden die in Tabelle 6 aufgelisteten 28 Versuche durchgeführt (84 Teilversuche). Das Programm „BloomEncoder" ermöglicht, wie oben bereits ausgeführt, unter anderem die Variation der Länge der Bloomfilter (Bitsize) und der Anzahl der Hashfunktionen, die zur Verschlüsselung verwendet werden. Daneben kann eingestellt werden, welche Länge die Teilstrings haben sollen, in die der komplette Code aufgesplittet wird („Ngramme") und ob dem ersten und letzten N-Gramm ein Leerzeichen voran- bzw. hintangestellt wird („Padded"). Bei allen hier

21 Das Maximum liegt immer bei 243 * 351 = 85.293, was eine Untergrenze der Trefferquote von 0,002 % ergibt.

durchgeführten Versuchen wird die Länge der Bloomfilter konstant bei 1.000 Bit belassen (Voreinstellung) und die Option „Padded" gesetzt, das heißt, es werden immer Leerzeichen am Anfang und Ende verwendet. Systematisch variiert wird die Anzahl der Hashfunktionen, die beginnend mit 1 stetig erhöht wird. Für jede Hashfunktionsanzahl wird je ein Versuch mit Monogrammen und Bigrammen durchgeführt (Ngramme = 1 bzw. Ngramme = 2). Bei zwei Versuchen (Versuch t3 und v3, Tabelle 6) werden exemplarisch Trigramme verwendet (Ngramme = 3). Ausgewertet werden dann jeweils die Zuordnungsrichtungen $t_1 \rightarrow t_2$, $t_1 \leftarrow t_2$ und $t_1 \leftrightarrow t_2$.

Tabelle 6 Durchgeführte Versuche mit verschlüsselten Codes

Versuch	Bitsize	Hash-Fuktionen	Ngramme	Padded
a	1000	1	1	1
b	1000	1	2	1
c	1000	2	1	1
d	1000	2	2	1
e	1000	5	1	1
f	1000	5	2	1
g	1000	10	1	1
h	1000	10	2	1
i	1000	15	1	1
j	1000	15	2	1
k	1000	20	1	1
l	1000	20	2	1
m	1000	25	1	1
n	1000	25	2	1
o	1000	50	1	1
p	1000	50	2	1
q	1000	100	1	1
r	1000	100	2	1
s	1000	200	1	1
t	1000	200	2	1
t3	1000	200	3	1
u	1000	300	1	1
v	1000	300	2	1
v3	1000	300	3	1
w	1000	400	1	1
x	1000	400	2	1
y	1000	500	1	1
z	1000	500	2	1

Im Ergebnis zeigt sich zunächst, dass bis zu einer Hashfunktionsanzahl von einschließlich 200 in den hier durchgeführten Versuchen (a bis t3) in zumindest einer Variante (N-Gramme und Zuordnungsrichtung) alle zusammengehörigen Fälle auch identifiziert werden (siehe Tabellen 7 und 8, grau hinterlegte Zeilen). Auffällig ist, dass dies jedoch nur bei der Verschlüsselung über Bigramme (Ngramme = 2) und nicht über Monogramme (Ngramme = 1) gelingt. Auch der mit 200 Hashfunktionen durchgeführte Versuch t3 mit Trigrammen (Ngramme = 3) identifiziert nicht alle Treffer korrekt. Mit einer Hashfunktionsanzahl größer als 200 (Versuche u bis z) funktioniert die Methode dann jedoch nicht mehr zuverlässig: In keinem der Versuche mit mehr als 200 Hashfunktionen werden alle zusammengehörigen Pärchen gefunden.

Tabelle 7 Ergebnisse der Versuche a bis p

	Richtung	kein Treffer	Treffer	Treffer in %	Total	Trefferquote in %	Minimale Ähnlichkeit Treffer ($ä_{min}$)	Fälle mit $ä \geq ä_{min}$
a	$t_1 \rightarrow t_2$	291	186	99,47	477	38,99	0,476	–
	$t_1 \leftarrow t_2$	93	185	98,93	278	66,55	0,476	–
	$t_1 \leftrightarrow t_2$	358	186	99,47	544	34,19	0,476	–
b	$t_1 \rightarrow t_2$	282	186	99,47	468	39,74	0,444	–
	$t_1 \leftarrow t_2$	84	187	100,00	271	69,00	0,444	246
	$t_1 \leftrightarrow t_2$	339	187	100,00	526	35,55	0,444	352
c	$t_1 \rightarrow t_2$	242	186	99,47	428	43,46	0,741	–
	$t_1 \leftarrow t_2$	80	186	99,47	266	69,92	0,741	–
	$t_1 \leftrightarrow t_2$	300	186	99,47	486	38,27	0,741	–
d	$t_1 \rightarrow t_2$	215	186	99,47	401	46,38	0,500	–
	$t_1 \leftarrow t_2$	70	187	100,00	257	72,76	0,500	221
	$t_1 \leftrightarrow t_2$	261	187	100,00	448	41,74	0,500	265
e	$t_1 \rightarrow t_2$	188	186	99,47	374	49,73	0,727	–
	$t_1 \leftarrow t_2$	67	184	98,40	251	73,31	0,727	–
	$t_1 \leftrightarrow t_2$	236	186	99,47	422	44,08	0,727	–
f	$t_1 \rightarrow t_2$	176	186	99,47	362	51,38	0,487	–
	$t_1 \leftarrow t_2$	57	187	100,00	244	76,64	0,487	220
	$t_1 \leftrightarrow t_2$	214	187	100,00	401	46,63	0,487	267
g	$t_1 \rightarrow t_2$	170	185	98,93	355	52,11	0,738	–
	$t_1 \leftarrow t_2$	62	183	97,86	245	74,69	0,727	–
	$t_1 \leftrightarrow t_2$	219	186	99,47	405	45,93	0,727	–
h	$t_1 \rightarrow t_2$	170	185	98,93	355	52,11	0,509	–
	$t_1 \leftarrow t_2$	59	186	99,47	245	75,92	0,509	–
	$t_1 \leftrightarrow t_2$	213	187	100,00	400	46,75	0,509	270
i	$t_1 \rightarrow t_2$	171	185	98,93	356	51,97	0,742	–
	$t_1 \leftarrow t_2$	62	182	97,33	244	74,59	0,729	–
	$t_1 \leftrightarrow t_2$	218	186	99,47	404	46,04	0,729	–
j	$t_1 \rightarrow t_2$	168	185	98,93	353	52,41	0,540	–
	$t_1 \leftarrow t_2$	56	187	100,00	243	76,95	0,540	216
	$t_1 \leftrightarrow t_2$	209	187	100,00	396	47,22	0,540	254
k	$t_1 \rightarrow t_2$	167	186	99,47	353	52,69	0,743	–
	$t_1 \leftarrow t_2$	62	183	97,86	245	74,69	0,743	–
	$t_1 \leftrightarrow t_2$	214	186	99,47	400	46,50	0,743	–
l	$t_1 \rightarrow t_2$	165	186	99,47	351	52,99	0,552	–
	$t_1 \leftarrow t_2$	56	187	100,00	243	76,95	0,552	216
	$t_1 \leftrightarrow t_2$	206	187	100,00	393	47,58	0,552	250
m	$t_1 \rightarrow t_2$	167	185	98,93	352	52,56	0,749	–
	$t_1 \leftarrow t_2$	62	183	97,86	245	74,69	0,739	–
	$t_1 \leftrightarrow t_2$	213	186	99,47	399	46,62	0,739	–
n	$t_1 \rightarrow t_2$	166	185	98,93	351	52,71	0,564	–
	$t_1 \leftarrow t_2$	56	187	100,00	243	76,95	0,564	216
	$t_1 \leftrightarrow t_2$	206	187	100,00	393	47,58	0,564	256
o	$t_1 \rightarrow t_2$	166	185	98,93	351	52,71	0,785	–
	$t_1 \leftarrow t_2$	62	181	96,79	243	74,49	0,785	–
	$t_1 \leftrightarrow t_2$	212	185	98,93	397	46,60	0,785	–
p	$t_1 \rightarrow t_2$	165	186	99,47	351	52,99	0,642	–
	$t_1 \leftarrow t_2$	56	187	100,00	243	76,95	0,642	214
	$t_1 \leftrightarrow t_2$	205	187	100,00	392	47,70	0,642	245

–: Nicht berechnet, da nicht alle Treffer identifiziert werden.
Grau hinterlegt: Versuche, in denen alle Treffer identifiziert werden.

Tabelle 8 Ergebnisse der Versuche q bis z

	Richtung	kein Treffer	Treffer	Treffer in %	Total	Treffer-quote in %	Minimale Ähnlichkeit Treffer ($ä_{min}$)	Fälle mit $ä \geq ä_{min}$
q	$t_1 \rightarrow t_2$	169	182	97,33	351	51,85	0,841	–
	$t_1 \leftarrow t_2$	60	183	97,86	243	75,31	0,826	–
	$t_1 \leftrightarrow t_2$	213	184	98,40	397	46,35	0,826	–
r	$t_1 \rightarrow t_2$	165	186	99,47	351	52,99	0,744	–
	$t_1 \leftarrow t_2$	56	187	100,00	243	76,95	0,744	215
	$t_1 \leftrightarrow t_2$	207	187	100,00	394	47,46	0,744	264
s	$t_1 \rightarrow t_2$	172	182	97,33	354	51,41	0,879	–
	$t_1 \leftarrow t_2$	65	178	95,19	243	73,25	0,894	–
	$t_1 \leftrightarrow t_2$	214	182	97,33	396	45,96	0,879	–
t	$t_1 \rightarrow t_2$	165	186	99,47	351	52,99	0,866	–
	$t_1 \leftarrow t_2$	60	183	97,86	243	75,31	0,866	–
	$t_1 \leftrightarrow t_2$	217	187	100,00	404	46,29	0,866	360
t3	$t_1 \rightarrow t_2$	172	179	95,72	351	51,00	0,861	–
	$t_1 \leftarrow t_2$	69	174	93,05	243	71,60	0,875	–
	$t_1 \leftrightarrow t_2$	233	182	97,33	415	43,86	0,861	–
u	$t_1 \rightarrow t_2$	176	177	94,65	353	50,14	0,895	–
	$t_1 \leftarrow t_2$	69	174	93,05	243	71,60	0,937	–
	$t_1 \leftrightarrow t_2$	224	177	94,65	401	44,14	0,895	–
v	$t_1 \rightarrow t_2$	178	173	92,51	351	49,29	0,912	–
	$t_1 \leftarrow t_2$	147	163	87,17	310	52,58	0,933	–
	$t_1 \leftrightarrow t_2$	322	178	95,19	500	35,60	0,912	–
v3	$t_1 \rightarrow t_2$	203	148	79,14	351	42,17	0,967	–
	$t_1 \leftarrow t_2$	91	153	81,82	244	62,70	0,944	–
	$t_1 \leftrightarrow t_2$	293	156	83,42	449	34,74	0,944	–
w	$t_1 \rightarrow t_2$	184	173	92,51	357	48,46	0,950	–
	$t_1 \leftarrow t_2$	73	172	91,98	245	70,20	0,946	–
	$t_1 \leftrightarrow t_2$	232	174	93,05	406	42,86	0,946	–
x	$t_1 \rightarrow t_2$	330	165	88,24	495	33,33	0,950	–
	$t_1 \leftarrow t_2$	199	156	83,42	355	43,94	0,955	–
	$t_1 \leftrightarrow t_2$	525	172	91,98	697	24,68	0,950	–
y	$t_1 \rightarrow t_2$	13380	172	91,98	13552	1,27	0,961	–
	$t_1 \leftarrow t_2$	11839	173	92,51	12012	1,44	0,959	–
	$t_1 \leftrightarrow t_2$	13671	175	93,58	13846	1,26	0,959	–
z	$t_1 \rightarrow t_2$	19443	164	87,70	19607	0,84	0,973	–
	$t_1 \leftarrow t_2$	16713	163	87,17	16876	0,97	0,965	–
	$t_1 \leftrightarrow t_2$	25045	172	91,98	25217	0,68	0,965	–

–: Nicht berechnet, da nicht alle Treffer identifiziert werden.
Grau hinterlegt: Versuche, in denen alle Treffer identifiziert werden.

In Bezug auf die Zuordnungsrichtung ist festzustellen, dass bei der Richtung $t_1 \rightarrow t_2$ in keinem Versuch alle 187 tatsächlich zusammengehörigen Fälle ermittelt werden. In der Gegenrichtung $t_1 \leftarrow t_2$ werden in deutlich mehr Fällen die korrekten Treffer identifiziert. Bei den Versuchen mit Bigrammen im Bereich von 1 bis 200

Hashfunktionen gelingt dies in 8 von 10 Versuchen (b, d, f, j, l, n, p und r). In allen 10 Versuchen mit Bigrammen im Bereich von 1 bis 200 Hashfunktionen werden alle 187 korrekten Treffer nur in der zusammengesetzten Datei (Zuordnungsrichtung $t_1 \leftrightarrow t_2$) identifiziert (b, d, f, h, j, l, n, p, r und t).

Betrachtet man die Gesamtzahl der ermittelten Pärchen mit maximaler Ähnlichkeit (Spalte „Total"), dann zeigt sich schon im diesbezüglich schlechtesten Versuch mit allen korrekt ermittelten Treffern (Versuch b) eine deutliche Verbesserung zur Zuordnung mit unverschlüsselten Codes (siehe Tabelle 5). So treten bei der Zuordnungsrichtung $t_1 \leftrightarrow t_2$ im verschlüsselten Fall (Bigramme, 1 Hashfunktion) 526 Pärchen auf, in denen die 187 korrekten Treffer enthalten sind (Quote: 35,55 %). Im unverschlüsselten Fall waren dies bei der Hamming-Ähnlichkeit 976 Pärchen (Quote: 19,16 %) und bei der Levenshtein-Ähnlichkeit 765 Pärchen (Quote: 24,44 %). Dieses Verhältnis wird bei zunehmender Zahl von Hashfunktionen (im Bereich bis 200 Funktionen) immer besser zu Gunsten der Zuordnung über verschlüsselte Codes. Das beste Ergebnis wird hier bei Versuch p erreicht (Bigramme, Bitsize = 1.000, 50 Hashfunktionen): Bei Zuordnungsrichtung $t_1 \leftrightarrow t_2$ werden lediglich 392 Pärchen ermittelt, in denen die 187 korrekten Treffer enthalten sind (Quote: 47,70 %). Offensichtlich wirkt die Verschlüsselung hier positiv differenzierend, ohne dass die Fähigkeit, die korrekten Treffer zu identifizieren, verloren geht. Damit zusammenhängend ist die Anzahl der Pärchen mit gleicher oder größerer Ähnlichkeit als der Treffer mit der minimalsten Ähnlichkeit im besagten Versuch p am niedrigsten.[22] Hier liegen 245 Pärchen vor, die die 187 korrekten Treffer enthalten. Im konkreten Anwendungsfall müssten also maximal nur 245 Pärchen mit einem Ausschuss von 58 falsch positiven Treffern kontrolliert werden, um alle korrekten Treffer zu erhalten.[23] Bei der Zuordnung über unverschlüsselte Codes liegt die Anzahl der maximal zu kontrollierenden Pärchen bei 600 (Hamming) bzw. 413 (Levensthein) und damit in beiden Fällen erheblich höher.

In den hier analysierten Versuchen mit einer Hashfunktionsanzahl größer 200 funktioniert die Zuordnung nicht mehr befriedigend, denn es werden nicht mehr alle korrekten Treffer identifiziert. Dies liegt höchstwahrscheinlich an dem Verhältnis von Hashfunktionsanzahl und Länge der Bloomfilter.[24] Es werden zu viele Einsen in die Filter geschrieben, so dass die Ähnlichkeit in allen Fällen zu sehr an-

22 Diese Anzahl wurde nur bestimmt, wenn auch alle Treffer korrekt ermittelt wurden.

23 Auch hier gilt, dass sich diese Anzahl bei hierarchischer Vorgehensweise noch weiter reduzieren kann.

24 Zwar wird bei der Ermittlung des Dice-Koeffizienten die Länge der Filter nicht explizit berücksichtigt, diese gibt jedoch die Anzahl der möglichen Stellen vor, an denen überhaupt Einsen gespeichert werden können. Es kommt daher bei zu kurzen Filtern in Bezug auf die Hashfunktionszahl vermutlich häufig zu Kollisionen.

steigt, um eine genügende Differenzierung zwischen fehlerhaften und korrekten Codepärchen zu ermöglichen. Ein deutliches Indiz dafür ist, dass ab 300 Hashfunktionen die Versuche mit Monogrammen eine höhere Trefferquote haben als die mit Bigrammen. Bei der Aufsplittung der Codes in Bigramme und nachfolgender Verschlüsselung über die Hashfunktionen werden nämlich mehr Einsen in die Filter geschrieben als bei Monogrammen.

4 Fazit und Ausblick

Die hier vorgestellten Ergebnisse zeigen das zunächst erstaunliche Resultat, dass die Zuordnung von verschlüsselten Codes der Zuordnung von unverschlüsselten deutlich überlegen ist. Werden Bigramme und eine nicht zu hohe Anzahl an Hashfunktionen bei der Verschlüsselung in Bloomfilter der Länge 1.000 Bit verwendet, so ist die Rate der falsch positiven Treffer deutlich kleiner, ohne dass die Treffergenauigkeit leidet. Auch im Hinblick auf die im Anwendungsfall durchzuführenden Validierungen (beispielsweise mit Handschriftenvergleichen) ist eine deutlich niedrigere Anzahl zu erwarten.

In den durchgeführten Versuchen zeigte sich die beste Performanz bei einer Hashfunktionsanzahl von 50 mit dem verwendeten siebenstelligen[25] Code. Diese Anzahl reicht im Zusammenspiel mit den anderen Parametern (Bigramme, Bitsize = 1.000) völlig aus, um eine zureichende Verschlüsselungssicherheit zu realisieren. Die Empfehlung von 15 Funktionen (Schnell/Bachteler/Reiher 2009a: 211) kann auf Basis der hier vorgestellten Resultate deutlich nach oben korrigiert werden.

Falls im konkreten Anwendungsfall die Anzahl der falsch positiven Treffer weiter gesenkt werden muss, beispielsweise um einen eventuellen Kontrollaufwand über Handschriftenvergleiche etc. zu minimieren, kann auf Grundlage der vorliegenden Erkenntnisse eine Einschränkung der Zuordnungsrichtung in Betracht gezogen werden. Hierbei ist es sinnvoll, die kleinere von zwei vorliegenden Dateien als Grundlage zu verwenden. Hier war dies die Datei aus dem zweiten Erhebungszeitpunkt t_2. Allerdings muss einschränkend hinzugefügt werden, dass in den durchgeführten Versuchen bei einer solchen Einschränkung nicht immer alle korrekten Treffer identifiziert wurden. Will man diesbezüglich sichergehen, müssen beide Zuordnungsrichtungen beachtet werden.

25 Zu beachten ist, dass bei der Aufsplittung in N-Gramme der eigentlich siebenstellige Code in manchen Fällen zu einem achtstelligen wird. Dies ist der Fall, wenn der Tag des Geburtstages zweistellig ist. In den Analysen zeigte sich, dass dies Verhalten einen positiven und erwünschten Effekt hat, daher wurde auf eine komplette Umstellung auf sieben Stellen, durch Voranstellen einer Null bei einstelligen Geburtstagen, verzichtet.

Als untere Schranke des Ähnlichkeitsniveaus, bis zu dem eine Validierung sinnvollerweise durchgeführt werden sollte, legen unsere Ergebnisse bei der unverschlüsselten Zuordnung (Hamming- und Levenshtein-Ähnlichkeit) einen Schwellenwert von 0,571 nahe. Bei der verschlüsselten Zuordnung ist ein solcher Schwellenwert schwierig anzugeben, da er augenscheinlich von den gewählten Verschlüsselungsparametern abhängt. Zumindest weist in allen durchgeführten Versuchen kein korrekter Treffer eine Ähnlichkeit von unter 0,444 auf. Insofern ist dieser Wert möglicherweise eine realistische untere Schranke. Die angegebenen Schwellenwerte stehen dabei unter der Prämisse, dass auf keinen echt positiven Treffer verzichtet werden soll. Kann dies hingegen akzeptiert werden, bietet sich an, die Werte nach oben zu verschieben, um den Validierungsaufwand zu verringern.

Alles in allem scheint der Einsatz des vorgestellten Verfahrens bei der Zuordnung des hier verwendeten verschlüsselten Codes uneingeschränkt empfehlenswert. Die Ergebnisse fordern geradezu, solch einer Zuordnung den Vorzug gegenüber einer Methode mit unverschlüsselten Codes zu geben. Inwieweit die Befunde auf den Einsatz und die Zuordnung anderer selbstgenerierter Codes übertragbar sind, ist mit den vorliegenden Analysen allerdings nicht zu beantworten. Es kann vermutet werden, dass die Ergebnisse auf ähnliche, selbstgenerierte persönliche Codes übertragbar sind. Um dies zu überprüfen, sind weitere Untersuchungen hilfreich und sinnvoll. Die von Schnell, Bachteler und Reiher unentgeltlich zur Verfügung gestellten Programme „Merge Toolbox", „BloomEncoder" und „BloomComparator" ermöglichen in jedem Fall die komfortable und zuverlässige Durchführung eines solchen Vorhabens.

Literatur

Dice, L. R., 1945: Measures of the Amount of Ecologic Association Between Species. Ecology 26 (3): 297-302.

Galanti M. R., R. Siliquini, L. Cuomo, J. C. Melero, M. Panella, F. Faggiano, and the EU-DAP study group, 2007: Testing Anonymous Link Procedures for Follow-Up of Adolescents In A School-Based Trial: The EU-DAP pilot study. Preventive Medicine 44 (2): 174-177.

Grube, Joel 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 (2): 159-171.

Hamming, R. W., 1950: Error Detecting and Error Correcting Codes. The Bell System Technical Journal 26 (2): 147-160.

Leitgöb, H., 2010: Klassifikation von Verläufen mittels Optimal Matching. S. 475-492 in: J. Bacher, A. Pöge und K. Wenzig (Hg.): Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren. 3. Aufl. München: Oldenbourg.

Levenshtein, V. I., 1966: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Cybernetics and Control Theory 10 (8): 707-710.

Pöge, A., 2005a: Methodendokumentation der kriminologischen Schülerbefragung in Münster 2000-2003 (Vier-Wellen-Panel). Bd. 9. Schriftenreihe „Jugendkriminalität in der modernen Stadt – Methoden". Münster, Trier.

Pöge, A., 2005b: Persönliche Codes bei Längsschnittstudien: Ein Erfahrungsbericht. ZA-Information 56: 50-69. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/za_information/ZA-Info-56.pdf (15.4.2011).

Pöge, A., 2007: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2005 (Vier-Wellen-Panel). Bd. 13. Schriftenreihe „Jugendkriminalität in der modernen Stadt – Methoden". Münster, Bielefeld.

Pöge, A., 2008: Persönliche Codes ›reloaded‹. Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung 2 (1): 59-70. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.2_Heft_1/2008_MDA1_Poege.pdf (15.4.2011).

Pollich, Daniela, 2010: Methodendokumentation der kriminologischen Schülerbefragung in Duisburg 2002 bis 2007 (Sechs-Wellen-Panel). Bd. 16. Schriftenreihe „Jugendkriminalität in der modernen Stadt – Methoden". Münster, Bielefeld.

Schnell, R., T. Bachteler und J. Reiher, 2005: MTB: Ein Record-Linkage-Programm für die empirische Sozialforschung. ZA-Information 56: 93-103. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/za_information/ZA-Info-56.pdf (15.4.2011).

Schnell, Rainer, T. Bachteler und J. Reiher, 2009a: Entwicklung einer neuen fehlertoleranten Methode bei der Verknüpfung von personenbezogenen Datenbanken unter Gewährleistung des Datenschutzes. Methoden – Daten – Analysen. Zeitschrift für Empirische Sozialforschung 3 (2): 203-217. http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/mda/Vol.3_Heft_2/06_Schnell_et_al.pdf (15.4.2011).

Schnell, R., T. Bachteler und J. Reiher, 2009b: Privacy-Preserving Record Linkage Using Bloomfilters. BMC Medical Informatics and Decision Making 41 (9).

Schnell, R., T. Bachteler und J. Reiher, 2010: Improving the Use of Self-Generated Identification Codes. Evaluation Review 34 (5): 391-418.

Yurek, L. A., J. Vasey und D. S. Havens, 2008: The Use of Self-Generated Identification Codes in Longitudinal Research. Evaluation Review 32 (5): 435-452.

Anschrift des Autors

Akad. Rat Dr. Andreas Pöge
Universität Bielefeld
Fakultät für Soziologie
Postfach 10 01 31
33501 Bielefeld
andreas.poege@uni-bielefeld.de

Rezensionen

Christof Wolf & Henning Best (Hg.), 2010: Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften. ISBN: 978-3-531-16339-0, 1.098 Seiten, 79,95 EUR.

Christof Wolf (GESIS und Universität Mannheim) und Henning Best (GESIS, Mannheim) haben sich mit dem Handbuch der sozialwissenschaftlichen Datenanalyse das Ziel gesetzt, der quantitativen empirischen Sozialforschung ein deutschsprachiges Einführungs-, Überblicks- und Nachschlagewerk zu elaborierten Analyseverfahren zur Verfügung zu stellen. Primäre Adressaten des Herausgeberbandes sind alle Anwender fortgeschrittener quantitativ-empirischer Methoden und damit auch Doktoranden und Studierende der Sozialwissenschaften in höheren Fachsemestern (S. 4). Auf insgesamt rund 1.100 Seiten (sic!) wird – unter Mitzählung der einleitenden Bemerkungen von Henning Best und Christof Wolf – in 40 Einzelbeiträgen ein methodologischer und methodischer Bogen gespannt, der von A wie *Analyse kategorialer Daten* (Kapitel 18 von Hans-Jürgen Andreß) bis Z wie *Zeitreihenanalyse* (Kapitel 40 von Rainer Metz) reicht. Sofern die spezifische Zielsetzung eines Kapitels nicht ein abweichendes Vorgehen erfordert, folgen alle Beiträge dabei einem formal gleichen Aufbau. Zunächst wird das behandelte Analyseverfahren in allgemeinen Worten vorgestellt, bevor seine mathematischen Grundlagen dargestellt und diskutiert werden. Im Anschluss werden exemplarische Anwendungen aufgezeigt, wobei in den meisten Beiträgen möglichst einheitlich ALLBUS- oder SOEP-Daten als Datenbasis herangezogen werden. Den Abschluss eines jeden Kapitels bilden Hinweise zu häufig vorkommenden Anwendungsfehlern, die insbesondere für diejenigen von Interesse sind, die sich in eine bestimmte Analysemethode neu einarbeiten wollen, und kommentierte Hinweise auf weiterführende Literatur zum behandelten Verfahren. Jedem Kapitel ist zudem ein viertel- bis einseitiger Abstract vorangestellt, der einen kurzen Überblick über die behandelten Inhalte, den konkreten Aufbau und die Kernaussagen des Beitrages gibt.

Thematisch ist das Handbuch der sozialwissenschaftlichen Datenanalyse in sechs Sinnabschnitte gegliedert. Den ersten Teil *Einführung* bilden die einleitenden Worte von Best und Wolf sowie ein Beitrag von Karl-Dieter Opp zur *Kausalität als Gegenstand der Sozialwissenschaften und der multivariaten Statistik* (Kapitel 2). In der für ihn charakteristischen, prägnanten und präzisen Art behandelt Opp das Kausalitätsproblem, das auf Grund der Struktur der nutzbaren Daten vielen Analysen der empirischen Sozialforschung immanent ist. Obwohl es sich bei diesem Kapitel im engeren Sinne nicht um einen Beitrag zu Datenanalyseverfahren handelt, schärfen Opps Ausführungen noch einmal den Blick dafür, was statistische Analyseverfahren in den Sozialwissenschaften zu leisten vermögen und was nicht. Insofern ist der Beitrag wichtig und mit seiner Einordnung am Anfang des Handbuchs richtig platziert.

Der zweite Teil des Handbuchs *Grundlagen der Datenanalyse* versammelt – wie der Titel bereits nahelegt – Arbeiten, die sich mit den Basics der sozialwissenschaftlichen Datenanalyse befassen. Manuela Pötschke widmet sich in ihrem Beitrag *Datengewinnung und Datenaufbereitung* (Kapitel 3)

den Problemen, die durch die Auswahl des Erhebungsverfahrens und im Prozess der Datengewinnung und -aufbereitung entstehen können. Im Zentrum der Betrachtung stehen dabei die verschiedenen Formen der Befragung. Cornelia Weins gibt in ihrem Beitrag *Uni- und bivariate deskriptive Statistiken* (Kapitel 4) einen Überblick über die gängigen Lage- und Streuungsmaße sowie grundlegende Verfahren der bivariaten Datenauswertung. Horst Degen widmet sich der *Graphische(n) Datenexploration* (Kapitel 5), wobei er den Schwerpunkt auf die adäquate Visualisierung univariater Verteilungen legt. *Der Umgang mit fehlenden Werten* (Kapitel 6) wird anschließend von Martin Spieß thematisiert. Er beleuchtet dabei vor allem Möglichkeiten und Grenzen von Imputationsverfahren, greift aber auch Aspekte der Datengewichtung auf. Explizit um *Gewichtung* (Kapitel 7) geht es in dem Beitrag von Siegfried Gabler und Matthias Ganninger. Die Autoren erklären Prozesse und Ziele von Gewichtungsverfahren anschaulich und illustrativ am Beispiel der dritten Erhebungswelle des ESS. Der Beitrag *Grundlagen des Statistischen Schließens* (Kapitel 8) von Steffen M. Kühnel und Dagmar Krebs schafft dann einen Übergang zur Inferenzstatistik. Den Schwerpunkt setzen Kühnel und Krebs bei den gängigen Schätz- und Testverfahren. Einen entgegengesetzten Weg beschreitet Susumu Shikano in seiner *Einführung in die Inferenz durch den nichtparametrischen Bootstrap* (Kapitel 9) in dem er das weniger genutzte (und wohl auch weniger bekannte) Bootstrapping in seiner Logik und seiner Anwendung anschaulich erläutert. Den Abschluss des zweiten Teils des Handbuches bilden Thomas Gautschis Ausführungen zur *Maximum-Likelihood Schätztheorie* (Kapitel 10). Sein Beitrag setzt zwar ein mathematisches Verständnis auf etwas höherem Niveau voraus, bietet aber gleichzeitig eine wirklich gute Gelegenheit, die Logik von ML-Verfahren kompakt nachzuvollziehen.

Im dritten Teil des Handbuchs geht es um *Messen und Skalieren*. Beatrice Rammstedt widmet sich hier zunächst ganz grundsätzlich den Themen *Reliabilität, Validität, Objektivität* (Kapitel 11), bevor sich Joachim Gerich mit der *Thurstone- und Likertskalierung* (Kapitel 12) sowie der *Guttman- und Mokkenskalierung* (Kapitel 13) befasst. Zusammengenommen bieten beide Kapitel nicht nur einen guten Einblick in die Logik der jeweiligen Skalierungsverfahren, sondern lassen auch die Unterschiede zwischen ihnen pointiert hervortreten. Im 14. Kapitel thematisieren Christian Geiser und Michael Eid dann die *Item-Response-Theorie*. Im Zentrum stehen dabei Variationen des Rasch-Modells, an dem sie exemplarisch und anschaulich die Logik und Anwendung von IRT-Modellen verdeutlichen. *Hauptkomponentenanalyse und explorative Faktorenanalyse* (Kapitel 15) sind das Thema des Beitrages von Hans-Georg Wolff und Johann Bacher. Die beiden Autoren bieten einen kompakten und zugleich doch umfassenden ersten Einblick in diese Form der Datenanalyse bzw. Datenreduktion. Den Abschluss des dritten Teils bilden Beiträge von Jörg Blasius zur *Korrespondenzanalyse* (Kapitel 16) und von Ingwer Borg zur *Multidimensionalen Skalierung* (Kapitel 17). Beide Autoren gehören auf ihrem jeweiligen Gebiet ohne Zweifel zu den führenden Experten, was den Beiträgen unschwer anzumerken ist.

Der vierte Teil des Handbuchs *Analyse von Häufigkeiten, Gruppen und Beziehungen* wird mit einem Beitrag von Hans-Jürgen Andreß zur *Analyse kategorialer Daten* (Kapitel 18) eröffnet, der sich darauf konzentriert, Besonderheiten und Unterschiede von Logit-Modellen, log-linearen Modellen und der gewichteten Regressionsanalyse nach Grizzle, Starmer und Koch herauszuarbeiten. Im Anschluss geben Manuel C. Völkle und Edgar Erdfelder einen gelungenen Einblick in die *Varianz- und Kovarianzanalyse* (Kapitel 19), bevor Reinhold Decker, Silvia Rašković und Kathrin Brunsiek kompetent

über das Verfahren der *Diskriminanzanalyse* (Kapitel 20) informieren. Die *Clusteranalyse* (Kapitel 21) ist Thema des anschließenden Beitrages von Michael Wiedenbeck und Cornelia Züll. Sie konzentrieren sich auf die Behandlung der Clusterzentrenanalyse und agglomerativer Verfahren der Clusterbildung. Johann Bacher und Jeroen K. Vermunt geben einen instruktiven ersten Einblick in die *Analyse latenter Klassen* (Kapitel 22), wobei sie u.a. die Probleme der Validitätsprüfung ins Zentrum ihrer Ausführungen rücken. Im letzten Kapitel des vierten Teils behandeln Hans J. Hummell und Wolfgang Sodeur – zwei Wegbereiter der sozialwissenschaftlichen Netzwerkforschung in Deutschland – mit großer Expertise Logik und Verfahren der *Netzwerkanalyse* (Kapitel 23).

Mit dem fünften Teil *Regressionsverfahren für Querschnittsdaten* wird ein Einstieg in die nach wie vor wachsende Familie der regressionsanalytischen Methoden gegeben. Nachdem Christof Wolf und Henning Best hier zunächst kompakt und kompetent die *Lineare Regressionsanalyse* (Kapitel 24) vorstellen, erörtert Dieter Ohr vertiefend und versiert *Modellannahmen und Regressionsdiagnostik (der linearen Regression)* (Kapitel 25), bevor Henning Lohmann spezifisch und instruktiv *Nicht-Linearität und Nicht-Additivität in der multiplen Regression* (Kapitel 26) diskutiert. Zusammengenommen bieten die drei Kapitel einen wirklich guten Einstieg für die fortgeschrittene Handhabung linearer Regressionsmodelle. Ben Jann widmet sich in seinem anschließenden Beitrag zur *Robuste(n) Regression* (Kapitel 27) den spezifischen Problemen, die die so genannten Ausreißer in OLS-Regressionen hervorrufen können, und gibt einen Einblick in regressionsdiagnostische Verfahren jenseits der üblichen Residuenanalyse. Wolfgang Langer offeriert im Anschluss in gewohnter Kompetenz einen Überblick über die *Mehrebenenanalyse mit Querschnittsdaten* (Kapitel 28), bevor Jost Reinecke und Andreas Pöge sich versiert der Logik und der Handhabung von *Strukturgleichungsmodelle(n)* (Kapitel 29) annehmen und Petra Stein die besonderen Probleme der *Regression mit unbekannten Subpopulationen* (Kapitel 30) diskutiert. Das 31. Kapitel *Logistische Regression* von Henning Best und Christof Wolf stellt – analog dem 24. Kapitel – eine kompetent geschriebene Einführung in den Umgang mit (binären) logistischen Regressionsanalysen dar. Ergänzt und vertieft wird die Darstellung durch die nachfolgende versierte Abhandlung von Steffen M. Kühnel und Dagmar Krebs zur *Multinomiale(n) und ordinale(n) Regression* (Kapitel 32), bevor Gerhard Tutz instruktiv die Besonderheiten der *Regression für Zählvariablen* (Kapitel 33) darlegt und Gerrit Bauer zum Abschluss des fünften Teils Hinweise für die *Graphische Darstellung regressionsanalytischer Ergebnisse* (Kapitel 34) gibt.

Der sechste und letzte Teil des Handbuchs schlägt schließlich den Bogen zur *Analyse von zeitbezogenen Daten*. Markus Gangl thematisiert hier zunächst die *Nichtparametrische Schätzung kausaler Effekte mittels Matchingverfahren* (Kapitel 35). Josef Brüderl geht im Anschluss auf die *Kausalanalyse mit Paneldaten* (Kapitel 36) ein, bevor Hans-Peter Blossfeld die *Survival- und Ereignisanalyse* (Kapitel 37) vorstellt. Das Thema des Beitrages von Florian Schmiedek und Julia K. Wolff sind *Latente Wachstumskurvenmodelle* (Kapitel 38) und Stefani Scherer gibt gemeinsam mit Josef Brüderl einen Einblick in die *Sequenzdatenanalyse* (Kapitel 39), bevor Rainer Metz last but not least die *Zeitreihenanalyse* (Kapitel 40) eingehend erörtert. Summarisch lässt sich festhalten, dass alle sechs Beiträge dieses Teils die jeweiligen Autorinnen und Autoren als versierte Kenner der behandelten Materie ausweisen und in der Regel zugleich auch ihr didaktisches Können dokumentieren. Gleichwohl bringt die longitudinale Perspektive der Analyseverfahren eine gewisse Komplexitätssteigerung mit sich, die die Auseinandersetzung – möglicherweise – erschwert. Interessierten Neueinsteigern sei daher empfohlen, sich auf jeden Fall

zunächst noch einmal ihrer Grundkenntnisse der sozialwissenschaftlichen Datenanalyse zu versichern, bevor sie sich mit dem spezifischen Feld der Analyse zeitbezogener Daten auseinandersetzen.

Ein Mammutwerk wie das Handbuch der sozialwissenschaftlichen Datenanalyse bietet zwangsläufig auch Anhaltspunkte für Detailkritik und es wäre billig, hier exemplarisch einzelne Aspekte herauszugreifen. Dies würde jedoch in keiner Weise der Leistung des Herausgeberbandes gerecht und daher soll im Folgenden eher eine Gesamtwürdigung versucht werden. An allererster Stelle ist dabei herauszustellen, dass die Umsetzung des Konzepts insgesamt als sehr gelungen bezeichnet werden kann und dass das Handbuch der sozialwissenschaftlichen Datenanalyse in keiner einschlägigen Fachbibliothek fehlen sollte. Ungeachtet aller qualitativen Schwankungen, die in einem solchen Werk zwischen einzelnen Beiträgen stets bestehen, ist der Herausgeberband insgesamt sowohl mit Blick auf die versammelte fachliche Expertise wie auch hinsichtlich der didaktischen Aufbereitung der Inhalte auf einem sehr hohen Niveau anzusiedeln. Dazu dürfte nicht zuletzt beigetragen haben, dass alle Einzelabhandlungen einem eingehenden Review-Verfahren unterzogen wurden (S. 5). Als weiterer wichtiger Pluspunkt ist in diesem Zusammenhang die Begleithomepage zum Handbuch zu nennen (http://www.handbuch-datenanalyse.de/). Auf ihr können nicht nur Zusammenfassung und Gliederung der einzelnen Kapitel sowie die Kurzportraits der Autorinnen und Autoren eingesehen werden. Zudem sind dort die – wohl unausweichlichen – Errata dokumentiert und es besteht die Möglichkeit, die Analysesyntax zu den in den einzelnen Kapiteln behandelten Auswertungen abzurufen oder über ein dafür eingerichtetes Kontaktformular Anfragen und Kommentare an die Herausgeber zu senden.

Grundsätzlich kritisch anzumerken sind lediglich vier Punkte: Zunächst muss festgehalten werden, dass das Handbuch über kein Stichwortregister verfügt, was für ein Werk mit dem oben skizzierten Anspruch einen deutlichen Mangel darstellt. Davon abgesehen fällt in der Gesamtschau auf, dass die Aufgabe, Hinweise zu häufigen Fehlern zu geben, von den Autorinnen und Autoren mit sehr unterschiedlichem Einatz gelöst wurde. Während viele Kapitel sich dadurch auszeichnen, dass die Autoren mit viel Engagement bemüht sind, Novizen des diskutierten Verfahrens vor den typischen Anfängerfehlern zu bewahren, vermitteln einzelne Kapitel hier eher den Eindruck einer unliebsamen formalen Anforderung nachzukommen. Dabei dürfte gerade die Diskussion häufig auftretender Fehler für diejenigen, die sich an einem für sie neuen Analyseverfahren erproben wollen, eine Schlüsselpassage des jeweiligen Kapitels sein – ganz zu schweigen davon, dass sie auch für fortgeschrittene Anwender eine Art Vergewisserung darstellen kann, methodisch-analytisch sauber zu arbeiten. Kurzum: Hier hätten die Herausgeber strenger auf die durchgängige Einhaltung der aufgestellten Maßstäbe pochen sollen.

Des Weiteren ist anzumerken, dass insbesondere einige der im zweiten und dritten Teil des Handbuchs versammelten Beiträge nach Dafürhalten des Rezensenten einen anderen Personenkreis, als die von den Herausgebern anvisierten Gruppen, adressieren. Ohne dadurch die inhaltliche und didaktische Qualität der Beiträge schmälern zu wollen, behandeln sie dennoch Gegenstände, die eher der grundständigen universitären Ausbildung in empirischer Sozialforschung zuzurechnen sind und die spätestens bei fortgeschrittenen Studierenden als selbstverständliches Basiswissen voraussetzbar sein sollten. (Wobei natürlich auch denkbar ist, dass der Rezensent sich an dieser Stelle irrt.) Selbstverständlich spricht nichts dagegen, in einem Handbuch der sozialwissenschaftlichen Datenanalyse auch grundlegende Inhalte zu behandeln und den Adressatenkreis des Werkes entsprechend auf Studierende in der Anfangsphase ihrer

Ausbildung auszuweiten. Nur wäre es dann vielleicht empfehlenswert gewesen, den basalen Themenbereichen mehr Raum zu gewähren und sie in kleineren Sinneinheiten detaillierter zu behandeln.

Schließlich ist festzuhalten, dass für ein Handbuch der sozialwissenschaftlichen Datenanalyse noch einige weitere Kapitel zu einer Reihe anderer Themenbereiche denkbar gewesen wären. Die Herausgeber weisen zwar in ihren einleitenden Bemerkungen zu Recht darauf hin, dass ein vollständiger Überblick über den Stand des Faches in *einem* Buch schlechterdings nicht möglich ist (S. 5), gleichwohl hat man bereits bei der Durchsicht des Inhaltsverzeichnisses das Gefühl, dass bestimmte Themenkomplexe noch hätten Berücksichtigung finden können. Beispielsweise hätte sich der Rezensent Kapitel zu Analyse*strategien*, zur Conjoint-Analyse und zu vielversprechenden Innovationen in der sozialwissenschaftlichen Datenanalyse gewünscht. Anderen Leserinnen und Lesern des Handbuchs wird es mit anderen Themen wahrscheinlich ähnlich ergehen. So paradox es daher für ein Buch von über eintausend Seiten klingen mag: Mehr wäre mehr gewesen. Allerdings muss es ja bei Neuauflagen – zu denen es aufgrund des Mehrwerts des Sammelbandes und der daraus resultierenden Nachfrage ohne Zweifel kommen wird – nicht zwangsläufig bei einem einbändigen Werk bleiben.

Ulrich Rosar, Universität Düsseldorf

* * * * *

Sabine Fromm 2010: Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten. Wiesbaden: VS Verlag für Sozialwissenschaften. ISBN: 978-3-531-14792-5, 257 Seiten, 24,95 EUR.

Der Band „Datenanalyse mit SPSS für Fortgeschrittene 2: Multivariate Verfahren für Querschnittsdaten" von Sabine Fromm ist der zweite Band einer Reihe. Teil 1 dieser SPSS Lehrbuchreihe beschäftigt sich mit den grundlegenden Problemen der Vorbereitung quantitativer Auswertungen (z. B. die Schritte vom ausgefüllten Fragebogen zum analysefähigen Datensatz, wie werden Daten bereinigt und wie können neue Variablen berechnet werden, etc.), während Teil 2 weiterführende Verfahren für die Analyse von Querschnittsdaten beinhaltet.

Das erste Kapitel führt in Mittelwertvergleiche ein. Das einführende Beispiel zur Varianzanalyse beschreibt einen Anwendungsfall mit drei Faktoren (16 Untergruppen), der relativ gut erläutert wird. Besonders gelungen ist die Veranschaulichung kleiner und großer Varianzen in Teilgruppen (S. 32 und 33). Ob ein Beispiel mit drei Faktoren unnötig komplex und damit als Einstieg in die Thematik weniger geeignet ist, kann bei der Darstellung der Aufteilung der Gesamtstreuung gefragt werden. Die Vielzahl an Effekten scheint eine handhabbare Menge leider deutlich zu überschreiten. Ein weiteres Beispiel ist zum Glück einfacher gehalten und illustriert in durchaus gelungener Weise die Funktionsweise der Varianzanalyse.

Anschließend wird die Faktoren- und Reliabilitätsanalyse vorgestellt (Kapitel 2).

Die Faktorenanalyse wird als „Verallgemeinerung der Dimensionsanalyse nach dem Modell der Likert-Skalierung" dargestellt (S. 59). Die rotierte Matrix stellt das Anwendungsbeispiel anschaulich dar. Anschließend wird Cronbachs Alpha zur Reliabilitätsanalyse erläutert. Ob man Variablen, die nach einer Faktorenanalyse zu einem Faktor zusammengestellt wurden, zusätzlich einer Reliabilitätsanalyse unterziehen muss, kann zwar hinterfragt werden; prinzipiell sind die Erläuterungen zur Reliabilitätsanalyse jedoch gut gelungen.

Kapitel drei stellt die lineare Regression vor. Die Einführung in die statistischen Grundlagen der linearen Regression erfolgt hierbei routiniert. Der Schwerpunkt der Diskussion der Modellvoraussetzungen liegt auf der Multikollinearität. Bei der Vorstellung des Variance Inflation Factor (bzw. Toleranz) wäre es allerdings hilfreich gewesen zu erfahren, wo sich die Grenzwerte befinden, die bedenklich hohe Multikollinearität anzeigen. In der Durchführung der Regressionsanalyse wird den LeserInnen zunächst das schrittweise Prüfen auf Signifikanz einzelner Variablen nahegelegt. Hier wäre ein Hinweis darauf, dass diese Option im Zweifelsfall zum „theorielosen Auswerten" verleitet, durchaus sinnvoll gewesen. Anschließend wird die gleichzeitige Aufnahme von Variablen in die Berechnung besprochen (Method = Enter).

Kapitel vier behandelt die logistische Regression. Auch hier erfolgt die Einführung in das Thema fachkundig. Gut gelungen ist auch der Abschnitt zur Transformation kategorialer unabhängiger Variablen. Die Erläuterungen zu den einzelnen Optionen der Syntax logistischer Regressionen sind vielseitig, fallen aber recht knapp aus. Leider ist das Beispiel der Wahlbeteiligung nicht glücklich gewählt, da NichtwählerInnen durch das Modell fast nicht vorhergesagt werden können (siehe S. 130). So gesehen „nützt" es auch nichts, wenn der Hosmer-Lemeshow Test akzeptable Werte ausweist. Zum HL-Test hätte man sich zudem eine etwas detailliertere Ausführung gewünscht, was genau hier auf „Goodness of Fit" geprüft wird. Im folgenden Schritt wird zur Erläuterung der multinominalen logistischen Regression die Parteipräferenz mit fünf Ausprägungen untersucht. Anschließend folgt die Einführung in die Diskriminanzanalyse, wofür dieselben beiden abhängigen Variablen verwendet werden wie im vorausgegangenen Abschnitt zur logistischen Regression. Dies macht durchaus Sinn, da beide Verfahrenstypen in der vorgeschlagenen inhaltlichen Anwendung ähnlich sind.

Das nächste Kapitel ist der Clusteranalyse gewidmet. Hier wird zwischen Verfahren zur vorgegebenen Anzahl von Clustern (Austauschverfahren) und hierarchischen Verfahren, die die geeigneten Clusteranzahlen ermitteln, unterschieden. Im Anwendungsbeispiel wird die Klassifikation von Ländern vorgenommen. Anschließend wird der interpretative Schritt der Klassendiagnose durchgeführt und die Cluster werden über einen Mittelwertvergleich inhaltlich näher beschrieben.

Im letzten Kapitel wird die Korrespondenzanalyse behandelt. Es handelt sich hierbei um „ein exploratives Verfahren zur Visualisierung der Datenstruktur einer Kontingenztabelle" (S. 223). Die Beispieldaten sind hier allerdings etwas schwierig in der Handhabung, ebenso die Syntax, die nicht klar auf die Datensätze zugreift. In der graphischen Darstellung werden Länder und Einstellungen auf zwei Dimensionen dargestellt. Je geringer die Distanz von Land und Einstellung, desto stärker die Assoziation.

Allgemein ist anzumerken, dass Studierende es eventuell verwirren könnte, wenn Effekte auf dem Niveau von z.B. $\alpha = 0{,}053$ oder 0,099 an der einen Stelle als signifikant bezeichnet werden und an einer anderen Stelle der T-Wert von 1,96 als Grenzwert kommuniziert wird, der bekanntlich dem 5%-Niveau entspricht. Natürlich ist das nicht falsch, da man prinzipiell Signifikanzniveaus

frei wählen kann. Dennoch hätte man zu sonst üblichen Konventionen klar Stellung beziehen können.

Generell ist das Layout des Buches zu beanstanden. Tabellen sind abgeschnitten über verschiedene Seiten verteilt, Zahlenwerte in Tabellen sind nicht am Komma ausgerichtet oder sehr stark verkleinert und einige Spaltenbeschriftungen sind unleserlich. Einerseits liegt das natürlich an den SPSS Outputs, andererseits hätte man diese Tabellen durchaus nachbearbeiten können, um die Lesbarkeit zu verbessern. Leider sind auch Formeln oder formelhafte Erläuterungen optisch überwiegend wenig überzeugend gesetzt. Von Lesbarkeit und Ästhetik abgesehen sollten zumindest Abbildungen den laufenden Text oder andere Abbildungen nicht verdecken (S. 175). Durch die Verwendung alternativer Software, wie beispielsweise LaTeX, hätte man das Layout des Buches sicherlich verbessern können.

Die Beispieldatensätze, die von http://www.vs-verlag.de/ herunter geladen werden können, stellen eine gelungene Ergänzung zum Buch dar und regen an, nicht nur die besprochenen Analysen zu replizieren, sondern diese auch zu variieren und eigene Analysen zu erstellen. Zu Kapitel 2 bis 7 sind auch die Syntaxdateien bereitgestellt.

Alles in Allem kann das Buch unter Anleitung in der Lehre eingesetzt werden. Für das Selbststudium ist das Lehrbuch meiner Meinung nach weniger gut geeignet, da sich zum einen einige Bereiche zu unübersichtlich gestalten und oft in einer Vielzahl knapp beschriebener Details verlieren, und zum anderen wurden bestimmte Aspekte zu oberflächlich dargestellt. Trotz des relativ geringen Umfangs von ca. 250 Seiten wird dennoch ein angemessener Überblick zu multivariaten statistischen Verfahren sozialwissenschaftlicher Analyse für Querschnittsdaten geboten.

Peter Kriwy, Universität Erlangen-Nürnberg

Frank Faulbaum & Christof Wolf (Hg.), 2010: Gesellschaftliche Entwicklungen im Spiegel der empirischen Sozialforschung. Wiesbaden: VS Verlag für Sozialwissenschaften. ISBN: 978-3-531-17525-6, 254 Seiten, 29,95 EUR.

Das in der neu gegründeten Schriftenreihe der ASI (Arbeitsgemeinschaft Sozialwissenschaftlicher Institute) erschienene Buch „*Gesellschaftliche Entwicklungen im Spiegel der empirischen Sozialforschung*", herausgegeben von Frank Faulbaum und Christof Wolf, zeichnet die Entwicklung der empirischen Sozialforschung in der Bundesrepublik Deutschland in den letzten 60 Jahren nach. In vier Teilen und neun Kapiteln wird ein umfassendes Bild der Entwicklung der empirischen Sozialforschung gegeben. Teil I beschäftigt sich mit der *sozialen und demographischen Entwicklung*, Teil II mit dem *Wandel von Einstellungen und Werten*, Teil III widmet sich den *Entwicklungen in der politischen Sozialforschung* und Teil IV bietet eine *Bestandsaufnahme der methodisch-statistischen Forschung*. Wie Frank Faulbaum und Christof Wolf in ihrer Einleitung schreiben, soll der vorliegende Sammelband „eine Orientierungshilfe für Lehrende und Studierende im Bereich der empirischen Sozialforschung zur Verfügung stellen" (S. 7).

Im ersten Kapitel stellt Martin Diewald anhand ausgewählter Indikatoren die Entwicklung sozialer Ungleichheit in den letzen 60 Jahren in Deutschland dar. Dabei kommt er zu dem Schluss, dass sich die relativ stabile wirtschaftliche Entwicklung positiv auf den allgemeinen Wohlstand ausgewirkt hat,

wobei sich aber vor allem die zunehmende Massenarbeitslosigkeit negativ auf die Chancen „gering Qualifizierter" (S. 27) niedergeschlagen hat. Auch die Chancenstruktur hat sich positiv entwickelt, es lassen sich jedoch noch immer Nachteile für Migranten und Ostdeutsche finden. Diese positive Entwicklung ist vor allem seit Mitte der neunziger Jahre rückläufig. Für Deutschland charakteristisch ist, dass die Ungleichheit schneller und stärker gestiegen ist als in anderen OECD Ländern.

Rosemarie Nave-Herz beschäftigt sich in ihrem Beitrag mit dem Wandel der Familie. In diesem Rahmen versucht die Autorin „einige derzeit gängige theoretische Thesen über den abgelaufenen familialen Wandel von 1949 bis heute mit den Ergebnissen der empirischen Sozialforschung zu konfrontieren" (S. 40). Nave-Herz zeigt in ihrem Beitrag, dass die Familie nach wie vor wichtig ist. Trotz einer Zunahme an unterschiedlichen Lebens- und Haushaltsformen kam es nicht zu einer Verdrängung der Lebensform Familie, sie ist nach wie vor zentraler Bestandteil der Gesellschaft. Da „im Alltag des Familienlebens moderne und traditionelle Trends nebeneinander und sogar miteinander verzahnt verlaufen" (S. 53) wird jedoch eine Darstellung der Entwicklung von Familie erschwert.

Heiner Meulemann analysiert in seinem Beitrag Wertewandel und Kulturumbruch. Eine der Fragen, die der Aufsatz stellt, ist ob sich die Werte in Ost- und West-Deutschland seit der Wiedervereinigung angenähert haben. Der Autor findet sowohl Unterschiede als auch Gemeinsamkeiten zwischen beiden Landesteilen. Die Unterschiede lassen sich zum Teil als „beabsichtigte" Folge des „Experimentes DDR" erklären (bspw. der Einfluss des sozialistischen Regimes auf Moralität und Religiosität) und zum Teil als unbeabsichtigte Folge (bspw. die Entwicklung einer spezifisch ostdeutschen Mentalität). Die Unterschiede zwischen Ost- und Westdeutschland lassen sich nicht auf regionale Differenzen zurückführen, wie sie sich bspw. zwischen Nord- und Süddeutschland zeigen. Aus der Tatsache, dass sich auch im Vergleich zwischen West- und Osteuropa keine vergleichbaren Unterschiede zeigen, zieht der Autor den Schluss, dass es sich hierbei um ein deutsches Phänomen handelt.

Anhand von Allensbach-Studien zeichnet Renate Köcher Einstellungen und Befindlichkeiten in der BRD in den letzten 60 Jahren nach. Dabei behandelt sie verschiedene Themen, wie bspw. Krieg, wirtschaftliche Verhältnisse und Inflation, sowie die Wahrnehmung von sozialen Netzwerken und generalisiertem Vertrauen. Auch die Einstellung zur Politik wird näher betrachtet. Generell haben sich die unterschiedlichen Einstellungen positiv entwickelt. Das Vertrauen in andere Menschen ist größer geworden, ebenso wie das Interesse an Politik in den letzten 60 Jahren zugenommen hat. Dies spiegelt sich auch in einem gesteigerten Selbstbewusstsein wider. So erfährt nicht nur die Vorstellung von der Entwicklung Deutschlands als Erfolgsmodell breite Unterstützung, auch das Ansehen von Deutschland in der Welt wird von den Befragten im Zeitverlauf immer positiver eingeschätzt.

Matthias Kepplinger untersucht in seinem Beitrag die Entwicklung von Medien und Politik sowie ihr Verhältnis zueinander. Die Medien stellen Politik vor allem im Zusammenhang mit Problemen dar. Den Leistungen der Politik wird dagegen weitaus weniger Aufmerksamkeit entgegengebracht. Die Darstellung von Politikern durch Äußerungen in den öffentlichen Medien hat deutlich abgenommen, d.h. Zitate von Politikern nehmen einen immer geringeren Platz in den Nachrichten ein. Daneben hat auch die Emotionalisierung der Politik zugenommen. Der Machtanspruch zwischen beiden ist unausgeglichen. Während die Politik sich einen geringeren Einfluss der Medien wünscht, präferieren die Medien mehr Einfluss. Dieses Ungleichgewicht führt der Autor auf den Schutz der Medien durch das Grundgesetz und auf die historisch begründete Defensivität der Politik gegenüber den Medien zurück.

Rüdiger Schmitt-Beck, Hans Rattinger, Sigrid Roßteutscher und Bernhard Weßels stellen in ihrem Beitrag die deutsche Wahlforschung und insbesondere die neu ins Leben gerufene „German Longitudinal Election Study" (GLES) vor. Dieses Projekt soll eine nationale Wahlstudie installieren, die der Forschung zur Verfügung steht und sicherstellt, dass die Bundestagswahlen wissenschaftlich begleitet werden. GLES ist eine komplexe Studie, die verschiedene Komponenten miteinander vereint, zum Beispiel indem Quer- und Längsschnittkomponenten miteinander kombiniert werden um eine fundierte Analyse des Wahlverhaltens zu ermöglichen. Nach einer ausführlichen Darstellung der einzelnen Studienteile zeigen die Autoren erste Analysen mit den Daten der GLES.

Christian Fleck untersucht in seinem Beitrag die Entwicklung der empirischen Sozialforschung in den letzen 60 Jahren in einer vergleichenden Perspektive. Hierzu vergleicht der Autor die Anzahl an empirischen Artikeln in der „American Sociological Review" und der „Kölner Zeitschrift für Soziologie und Sozialpsychologie". Hierbei offenbart sich eine deutlich stärkere Theorieorientierung der deutschen Nachkriegssoziologie im Vergleich mit der US-amerikanischen Sozialforschung, die sich allerdings in jüngerer Zeit nicht mehr finden lässt. In ihren Beiträgen zur empirischen Sozialforschung hat die deutsche Forschung international „Resonanz" gefunden. Insgesamt hält Fleck fest, dass die amerikanische Sozialforschung sich schneller weiterentwickelt hat als die deutsche.

Hans-Jürgen Andreß betrachtet die Entwicklung der sozialwissenschaftlichen Datenanalyse. Die Entwicklung und die Anwendung von Analysemethoden sind auch mit der Entwicklung von geeigneter Hard- und Software verbunden. So benötigt man genügend Rechenleistung und bedienbare Software um ohne großen Aufwand komplexe Analysen durchzuführen. Des Weiteren, so der Autor, bedarf es auch Daten (Umfragen, Prozessdaten, usw.) um statistische Analysen durchführen zu können. Daneben sind selbstverständlich auch Experten äußerst wichtig, die neue Analyseverfahren nicht nur entwickeln, sondern auch einen Beitrag zu ihrer Verbreitung liefern. Zuletzt sind auch Institutionen wie beispielsweise die Arbeitsgemeinschaft sozialwissenschaftlicher Institute (ASI) aber auch Fachvereine wie die Deutsche Gesellschaft für Soziologie (DGS) eine wichtige Instanz für die Weiterentwicklung der Datenanalyse.

Marek Fuchs widmet sich in seinem Beitrag einem wichtigen Thema der Umfrageforschung: der Datenqualität. Entgegen der üblichen Praxis, die Qualität von Datensätzen einzig an Indikatoren wie Fallzahl und Ausschöpfungsquote festzumachen, plädiert Fuchs dafür, den Total Survey Error (TSE) zu berücksichtigen, der ein umfassenderes Maß für die Datenqualität darstellt. Der TSE nämlich „integriert und systematisiert die verschiedenen Komponenten, die einen potenziellen negativen Einfluss auf die Datenqualität einer Umfrage bzw. eines einzelnen Schätzers haben können" (S. 228). Anschließend werden verschiedenen Umfragemethoden wie bspw. Onlinebefragungen und die damit verbundenen Herausforderungen an die Umfrageforschung diskutiert.

Der vorliegende Sammelband enthält ein breites Themenspektrum, das von sozialer Ungleichheit und Familie über Werte und Einstellungen und Politik bis zu Entwicklungen in der Methodenforschung reicht. Daneben ist auch der Ansatzpunkt der einzelnen Autoren sehr unterschiedlich; so reicht der Stil der Artikel von klassischer Literaturanalyse über rein deskriptive Auswertungen bis zur Inhaltsanalyse.

Die einzelnen Artikel sind ansprechend geschrieben. Die Überblicke über soziale Ungleichheit und die Entwicklung der Familie bringen die Entwicklungen in diesem Fachbereich auf den Punkt und sind als Einstiegstexte für die universitäre Lehre

hervorragend geeignet. Die Darstellung der Werteentwicklung ist weitgehend deskriptiv-empirisch und bietet eine gute Vorstellung der Entwicklung in Deutschland, wobei man sich aber durchaus mehr Interpretation und Hintergründe zu manchem Ergebnis gewünscht hätte. Die Darstellung des Verhältnisses von Politik und Medien und die Vorstellung der GLES im dritten Teil sind spannend zu lesen. Ein zusätzlicher Fokus auf die historische Entwicklung der Wahlforschung wäre wünschenswert gewesen und hätte zusätzlich auch die Vorteile der GLES verdeutlicht. Im vierten und letzten Teil, der sich mit der Entwicklung der methodischen Forschung beschäftigt, finden sich nicht nur interessante Befunde zur Verbreitung der quantitativen Forschung in der deutschen Forschungslandschaft, sondern auch Erkenntnisse über die Entwicklung der Datenanalyse und die Herausforderungen, der sich die Sozialforschung in Zukunft stellen muss. Insgesamt lässt sich sagen, dass der Sammelband durch sein breites Themenspektrum einen guten Überblick über die Entwicklung der empirischen Sozialforschung in Deutschland vermittelt.

Christian Deindl, Köln

Ankündigungen

Workshop

An Introduction to the EU-SILC & EU-LFS Data

Organised by
ESDS Government (University of Manchester) and
GESIS (Leibniz Institute for the Social Sciences, Mannheim)

University of Manchester
August 4 - 5, 2011

EU-SILC & EU-LFS Data

The workshop will give an introduction to the cross-national comparative EU-SILC and EU-LFS data and is intended for researchers, postgrads, and others who have not used the data before. The workshop includes presentations by Eurostat, informing about the background and content of the data and how to access it; presentation from researchers who have used either EU-SILC or EU-LFS, highlighting methodological considerations. The workshop offers also a chance to use some of the Microdata in a hands-on practical computing session.

Cost: £110 (£55 for student concessions);
Price includes lunch and refreshments.

For more information (programme, book a place) go to:
www.ccsr.ac.uk/esds/events/2011-08-04/

Contact at GESIS:
heike.wirth@gesis.org

* * * * *

Nutzerkonferenz zu den amtlichen Haushaltsstatistiken: Forschen mit dem Mikrozensus und der Einkommens- und Verbrauchsstichprobe

German Microdata Lab, GESIS &
Statistisches Bundesamt
29. - 30. September 2011

Konferenzort: Rheingoldhalle
Rheingoldstraße 215, 68199 Mannheim

Sozialstruktur, Einkommen und Verbrauch

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

Eine Anmeldung zu der Konferenz ist ab sofort unter folgender Adresse möglich: workshop-mannheim@gesis.org
Der Konferenzbeitrag beträgt € 120 (Studierende € 90).
Weitere Informationen finden Sie unter: www.gesis.org/gml/
Bei Fragen wenden Sie sich bitte an die Organisatoren bei GESIS: georgios.papastefanou@gesis.org und bernhard.schimpl-neimanns@gesis.org

Programm

Donnerstag, 29. September 2011

10:00 – 10:50 Begrüßung und Einführung

Begrüßung
Christof Wolf (GESIS, Mannheim)

Die Einkommens- und Verbrauchsstichprobe im europäischen Kontext
Anette Stuckemeier (Destatis, Bonn)

Der Mikrozensus im nationalen und europäischen Kontext
Hermann Seewald (Destatis, Bonn)

10:50 – 12:50 Soziale Ungleichheit I

Ungleichheit und Armut im Alter. Vergleichende Analysen auf der Basis von Einkommen und Konsumausgaben
Heinz-Herbert Noll und Stefan Weick (GESIS, Mannheim)

Ein alternativer Vorschlag zur Messung von Armut: Der Zerlegungsansatz – Empirische Illustration auf Basis der Einkommens- und Verbrauchsstichprobe 2003
Jürgen Faik (Neue Frankfurter Sozialforschung, Frankfurt, und Universität Lüneburg)

Wechselwirkungen zwischen den Leistungen zur Grundsicherung für Arbeitssuchende (ALG II) und Wohngeld – Eine Bilanzierung auf Haushaltsebene
Tim Clamor und Nicole Horschel (Institut der deutschen Wirtschaft, Köln)

12:50 – 13:50 Mittagspause

13:50 – 15:50 Soziale Ungleichheit II

Do Time Poor Individuals Pay More?
Tim Rathjen (Universität Lüneburg)

Armuts- und Familiendynamik mit dem Mikrozensus-Panel 2006-2009
Torsten Lietzmann und Helmut Rudolph (Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg)

Auswirkungen sozialer Ungleichheit auf das Gesundheitsverhalten. Auswertungen auf Basis des Mikrozensus 2005
Sophie Meyer (Universität Wuppertal)

15:50 – 16:20 Kaffeepause

16:20 – 18:20 Migration und Integration, Soziale Lage

Fortschreitende Integration oder dauerhafter Ausschluss? Eine Mikrozensusanalyse des Wandels der Arbeitsmarktchancen von Migranten zwischen 1976 und 2005
Andreas Herwig und Dirk Konietzka (Universität Braunschweig)

Aufstieg aus dem Migrationsmilieu in hochqualifizierte Berufe
August Gächter und Stefanie Smoliner (Zentrum für Soziale Innovation, Wien)

Wie leben und arbeiten Hamburgs Eltern? Auftrag und Chance für Hamburger Unternehmen
Christina Boll und Nora Reich (Hamburgisches WeltWirtschaftsInstitut)

19:00 Gemeinsames Abendessen

Freitag, 30. September 2011

09:00 – 11:00 Arbeitsmarkt

Identifizierung von Existenzgründungen und deren Erfolg auf Basis des Mikrozensus-Panel
Marc Langhauser und René Leicht (Universität Mannheim)

Verbleibsanalysen mit Querschnittsdaten? Die Veränderung der Alterserwerbsbeteiligung in Deutschland im Spiegel des Mikrozensus 1991 bis 2007
Martin Brussig (Universität Duisburg-Essen)

Maternal employment transitions across Bundesländer: a latent curve model approach
Pierre Walthery (University of Manchester)

11:00 – 11:20 Kaffeepause

11:20 – 12:40 Datenqualität und Methoden

Rekonstruktion bildungsspezifischer Fertilitätsraten mit Daten des Mikrozensus 1991 - 2003: Ein Schätzkonzept
Marc Hannappel und Damian Macura (Universität Koblenz)

Zur Datenqualität der Angaben zum Schulbesuch im Mikrozensus 2008
Bernhard Schimpl-Neimanns (GESIS, Mannheim)

12:40 – 13:40 Mittagspause

13:40 – 15:00 Bildung und Arbeitsmarkt

Relative Humankapitalausstattung und Erwerbsbeteiligung. Ergebnisse auf Basis der Mikrozensen 1976 bis 2005
Peter Kriwy (Universität Erlangen-Nürnberg)

Systematisierung der Lehrerforschung und Verbesserung ihrer Datenbasis. Möglichkeiten des Mikrozensus zur Analyse der sozialen Situation der pädagogischen Berufe unter besonderer Berücksichtigung der Lehrerschaft
Radoslaw Huth (Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt)

15:00 – 15:30 Abschlussdiskussion

Workshop

Interviewers' Deviant Behaviour – Reasons, Detection, Prevention

Organised by
GESIS (Leibniz Institute for the Social Sciences, Mannheim) and
Justus-Liebig University of Giessen

Castle Rauischholzhausen – Conference Centre of the Justus-Liebig University of Giessen
October 27 – 28, 2011

Interviewer Behaviour

This workshop is organised as part of the project "Identification of Falsifications in Survey Data" within the DFG Priority Programme "Survey Methodology" (DFG SPP 1292). It will provide a platform to present and discuss research methods and results related to deviant interviewer behaviour, in particular interviewers' deviances from standardised procedures while contacting target persons or conducting interviews. It will also include presentations of first project results regarding falsifications and statistical methods for detecting "at risk interviewers". The organisers invite contributions dealing with topics including prevention methods, e. g. incentives for interviewers and payment of interviewers' work, kinds of deviant interviewers' behaviour in surveys, effects on data quality and other related topics.

There is no conference fee. Furthermore, cost of local accommodation (in the castle) will be covered by the organisers for up to 10 active contributors (presenters). Some contribution to travel cost is also possible. The maximum number of participants is limited to 25.

If you are interested to participate in or contribute to this workshop, please contact Peter Winker (Peter.Winker@wirtschaft.uni-giessen.de).

The deadline for submissions of abstracts for presentations is **July, 31st 2011**; deadline for registration is **August, 31st 2011**.

Organisers: Christoph J. Kemper (GESIS), Natalja Menold (GESIS), Nina Storfinger (JLU), Peter Winker (JLU)

For more information go to:
http://www.uni-giessen.de/cms/ueber-uns/rhh/

Hinweise für unsere Autorinnen und Autoren

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Das Literaturverzeichnis ist wie folgt zu gestalten:

Buchveröffentlichungen:

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

Zeitschriftenbeiträge:

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

Beiträge in Büchern:

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

Internetquellen:

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

Datenfile:

Forschungsgruppe Wahlen, Mannheim: Zur politischen Lage in Niedersachsen im Januar 2008. GESIS Köln, Deutschland ZA Studie Nr. 4863; doi: 10.4232/1.4863.

ISSN 1864-6956

5. Jahrgang 2011